计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(5):107−115 [doi: 10.15888/j.cnki.csa.009881] [CSTR: 32024.14.csa.009881]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于图像对比增强的大型视觉语言模型物体幻觉缓解①

卜立平[1,2], 常贵勇[1,2], 于碧辉[1,2], 刘大伟[1,2], 魏靖烜[1,2], 孙林壮[1,2], 刘龙翼[1,2]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学, 北京 100049)
通信作者: 常贵勇, E-mail: changguiyong22@mails.ucas.ac.cn

摘　要: 大型视觉语言模型 (LVLM) 在理解视觉信息和运用语言表达方面展现出了非凡的能力. 然而, 在 LVLM 进行问答的过程中, 它通常受到物体幻觉问题的困扰, 具体表现为生成的文本内容看似合理, 但实际上却与图片中的信息不相符, 造成了文本与图片之间的不匹配现象. 为解决这一问题, 本文通过实验发现, 物体注意力的缺失是导致物体幻觉的关键因素. 为缓解此问题, 本文引入了图像对比增强方法 (ICE). ICE 是一种无需训练、操作简便的方法, 通过对比原始视觉输入与增强视觉输入所产生的输出分布, 有效提升模型对图片的感知能力, 确保生成的内容与视觉输入紧密契合, 从而生成上下文一致且准确的输出. 实验结果显示, ICE 方法在无需额外训练或外部工具的情况下, 便能显著减轻不同 LVLM 的物体幻觉问题, 并在大型视觉语言模型基准 MME 测试中同样表现出色, 验证了其广泛的适用性和有效性. 本文代码链接: ChangGuiyong/ICE.

关键词: 大型视觉语言模型; 物体幻觉; 图像对比增强; 人工智能

引用格式: 卜立平,常贵勇,于碧辉,刘大伟,魏靖烜,孙林壮,刘龙翼.基于图像对比增强的大型视觉语言模型物体幻觉缓解.计算机系统应用, 2025,34(5):107–115. http://www.c-s-a.org.cn/1003-3254/9881.html

Mitigating Object Hallucination in Large Visual Language Model Through Image Contrast Enhancement

BU Li-Ping[1,2], CHANG Gui-Yong[1,2], YU Bi-Hui[1,2], LIU Da-Wei[1,2], WEI Jing-Xuan[1,2], SUN Lin-Zhuang[1,2], LIU Long-Yi[1,2]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Large visual language model (LVLM) demonstrate remarkable capabilities in understanding visual information and generating verbal expressions. However, LVLM are often affected by the phenomenon of object hallucinations, where the outputs appear plausible but do not align with the visual information in the images. This discrepancy between the generated text and the images presents a significant challenge in achieving accurate image-to-text alignment. To address this issue, this study identifies the lack of object attention as a key factor contributing to object hallucinations. To mitigate this, the proposed image contrast enhancement (ICE) method is introduced. ICE is a simple, user-friendly approach that compares the output distributions from both the original and the augmented visual inputs. This method enhances the model's ability to perceive images more accurately, ensuring that the generated content aligns closely with the visual input and produces contextually consistent outputs. Experimental results demonstrate that the ICE method effectively mitigates object hallucinations across various LVLM without requiring additional training or external tools. Furthermore, the

① 基金项目: 2023 年度沈阳市科学技术计划 (23407329)
收稿时间: 2024-10-16; 修改时间: 2024-11-29; 采用时间: 2025-01-21; csa 在线出版时间: 2025-03-31
CNKI 网络首发时间: 2025-04-01

method performs well on the MME benchmark test for large-scale visual language models, indicating its broad applicability and effectiveness. The code will be released at ChangGuiyong/ICE.

Key words: large visual language model (LVLM); object hallucination; image contrast enhancement (ICE); artificial intelligence

大语言模型 (large language model, LLM)[1–4]的发展使通用人工智能 (artificial general intelligence, AGI) 领域发生了翻天覆地的变革. ChatGPT[5]与 LLaMA[6]等模型在文本理解和生成方面展现出了令人惊叹的能力. 然而, 它们在处理多模态数据, 尤其是图像等非文本数据时能力有限. 为解决这一难题, 多种大型视觉语言模型 (large visual language model, LVLM)[7–13]应运而生, 其中包括 GPT-4V[7]和 Gemini[14], 这些模型已成功地实现了接近人类水平的多模态性能. 这一新兴领域的探索不仅推进了人工智能技术的应用, 也预示着未来的技术前景.

然而, LVLM 虽然在视觉语言任务上能力出色, 但仍会出现模型响应与视觉中物体不匹配现象, 称之为物体幻觉[15]. 在金融、医疗、智能安防等领域对数据精准性和可信性要求严格, 幻觉引发的错误数据、解读及判断可能导致经济损失、规划失误和安全隐患等严重后果, 因此尽快解决幻觉问题对提升模型在这些领域的适用性极为关键.

近年来, 已有众多研究致力于解决 LVLM 的物体幻觉问题. 例如, WoodPecker[16]和 LURE[17]通过对生成的响应进行后处理以减少幻觉; LRV-Instruction[18]和 RLHF-V[19]等方法则提议通过对 LVLM 进行指令调整来减轻幻觉. 然而, 这些研究通常依赖于外部工具 (如 GPT-3.5-turbo[20]), 这需要昂贵的人工反馈收集, 并且还需要对 LVLM 进行额外的训练. 虽然这些干预措施已被证明可以有效减少物体幻觉, 但它们需要大量的人力参与, 并且会因额外的培训或补充模块的集成而产生大量的计算成本.

本项工作对 LVLM 的幻觉现象进行了深入分析与验证, 结果表明, LVLM 在处理物体数量多的图像时容易受到干扰, 对单个物体的注意力分散, 导致关联偏差引发的物体注意力减弱, 最终产生物体幻觉. 基于此分析, 本文提出了图像对比增强方法 (image contrast enhancement, ICE). 该方法通过对比原始视觉输入与增强视觉输入而生成模型输出, 以增强模型对物体的注意力, 进而增强模型对图片的感知能力. 此方法的独特创新之处在于无需进行训练, 且与底层 LVLM 无关. 它通过原始图像和增强图像来比较概率分布, 从而增强视觉因素对编码的影响, 有效减轻 LVLM 的物体幻觉. 在对 POPE (polling-based object probing evaluation)[15]和 MME (MLLM evaluation)[21]等辨别幻觉基准进行的综合实验中, ICE 对包括 Qwen-VL[9]和 LLaVA-1.5[11]在内的最先进 LVLM 进行了评估, 结果显示这些方法在减轻物体幻觉方面展现出了显著的成效. 此外, 本文提出的方法还提升了 LVLM 在感知与识别任务中的性能. 本文的主要贡献如下.

(1) 分析了 LVLM 的物体幻觉现象并进行了验证, 确认图像中物体注意力的缺失是导致该现象的原因. 这一发现为深入理解并减轻物体幻觉开辟了新的途径.

(2) 设计了 ICE 方法, 该方法通过对比原始视觉和增强视觉的输出分布来校准模型输出, 能够有效缓解 LVLM 中的幻觉现象, 且无需额外训练.

(3) 通过综合实验, 充分证明了所提出的图像对比增强方法在减轻物体幻觉和增强一般感知能力方面的卓越功效. ICE 方法无需额外培训或借助外部工具即可实现显著改进.

1 相关工作

1.1 大型视觉语言模型 (LVLM)

LVLM 旨在赋予大语言模型视觉感知能力, 使其能够结合视觉和语言处理多模态数据, 推动图像字幕生成、视觉问答等任务的进展. LVLM 还在人工智能助手、机器人等领域有广泛应用.

整合视觉与语言两种模式的方式主要有两类方法: 第 1 类方法基于预训练的单模态模型, 通过连接器将视觉编码器与 LLM 相结合. 具体来说, 连接器包括两种: 1) 基于可学习查询的方法, 以 Q-Former[22]为例, 用于 Qwen-VL 和 Instruct-BLIP[8], 利用一组可学习查询标记通过交叉注意力捕获视觉信号; 2) 基于投影层的方法 (如 LLaVA 和 Shikra[23]) 则通过线性投影层或多

层感知器 (multi-layer perceptron, MLP) 转换视觉特征. 第 2 类方法, 如 Fuyu-8B[24]和 Gemini 等模型, 采用了端到端的训练策略, 不依赖于预先训练的视觉编码器, 而是直接将图像块作为输入, 并将其映射 (或投影) 为嵌入向量进行后续处理. 这些方法体现了大型视觉语言模型从预训练到端到端训练的探索过程.

1.2 LVLM 中的物体幻觉

在 LVLM 中存在物体幻觉的现象, 其表现为模型产生详细、流畅且连贯的响应, 然而这些响应却未能准确反映视觉上下文的真实情况, 常包含错误的对象、属性以及它们之间的关系. 近期, 已有数项研究致力于提出解决方案, 以解决 LVLM 中存在的物体幻觉问题. 如 LURE 和 Woodpecker 采用后处理方法对生成的响应进行修改, 通过训练修订器或使用 GPT-3.5-turbo 微调方法. LRV-Instruction 和 RLHF-V 通过利用额外数据进行指令调整以减轻幻觉, 然而该方法需要大量的数据收集和培训资源. 鉴于 LVLM 中参数众多, 这种方法在计算上效率较低. 因此, 那些用于改进解码过程且无需额外训练的方法近来备受关注. 具体而言, OPERA[25]探索了导致幻觉的聚集模式, 并利用这一洞察来抑制呈现出这些模式的标记的生成. VCD (visual contrastive decoding)[26]利用模型在响应扭曲图像时倾向于优先考虑先验知识而非视觉信息的特点. 因此, 在处理扭曲图像与原始图像时, 模型所产生的反应在幻觉标记上展现出显著差异. VCD 正是通过对比这些差异, 来有效地减轻物体幻觉现象的影响. HALC[27]观察到, 当将具有不同视场的图像输入大型语言模型时, 真实标记的概率变化远大于幻觉标记的概率变化. 这一观察结果有助于识别清晰描绘物体的视觉上下文候选者, 通过对比这些候选者, HALC 可以减少幻觉. 与现有技术不同, 本文提出一种利用 SAM (segment anything model)[28]模型的新解码方法. 具体来说, 本文方法使用 SAM 模型对原始图片进行增强, 然后将增强图像生成的响应与原始图像生成的响应进行对比. 通过这一对比过程, 能够识别并对比幻觉令牌的分布情况, 从而有效地减轻物体幻觉现象.

2 方法介绍

本节主要介绍 LVLM 的生成原理, 说明文本的生成主要依赖每一个时间步的自回归. 在本节中, 首先从直观观察入手, 然后查明物体幻觉的原因, 最后提出图像对比增强算法对存在的幻觉进行缓解.

2.1 大型视觉语言模型生成原理

LVLM 的生成原理涉及处理文本查询 t 和视觉输入 v. 该模型旨在生成文本查询的相关响应, y_i为在给定视觉上下文的情况下保持与输入信息的一致性. 假设有一个由参数 θ 参数化的模型. 生成过程表示如下:

$$y_i \sim p_\theta(y_i \mid v,t,y<i) \propto \exp(\log it_\theta(y_i \mid v,t,y<i)) \quad (1)$$

其中, $p_\theta(y_i \mid v,t,y<i)$表示在给定视觉上下文 v, 文本查询 t 和先前生成的标记序列$y<i$的情况下, 生成标记 y_i 的概率. $\log it_\theta(y_i \mid v,t,y<i)$是 y_i 对应的$\log it(\log -odds)$值, 代表模型对每个 token 生成的“置信度”.

生成过程是自回归的, 每个时间步生成一个标记, 直到生成完整的响应序列. 在解码阶段, 模型利用生成的标记序列来生成下一个标记, 直到产生完整的响应. 这种生成原理使得模型在处理多模态输入时能够通过考虑文本和视觉信息之间的相关性来生成相关响应.

2.2 物体幻觉产生的机制分析

物体幻觉是 LVLM 在视觉语言任务中出现的模型响应与视觉中物体不匹配的现象, 即模型在图像理解过程中错误地预测图像中存在某些物体, 导致生成与实际内容不符的描述. 尤其在包含多个物体或存在遮挡的复杂场景中, 物体幻觉现象尤为明显, 其根本原因是模型对物体注意力的缺失. 本文探讨了物体注意力缺失对物体幻觉的影响.

在 LVLM 的预训练过程中, 模型往往使用大规模的多模态数据集, 这些数据集中存在着显著的统计偏差[29–31]. 例如, 常见物体的高频共现、类不平衡以及某些物体之间的强先验关联性, 使得模型在学习过程中对特定的物体形成了固有的认知习惯. 在推理过程中, 模型可能会更多依赖这些频繁共现的物体, 而忽略对视觉内容的细致理解, 从而导致物体幻觉的出现. 这种现象的根本原因在于模型缺乏对所有物体的足够注意力, 导致对部分物体的关注度不足.

为验证共现偏差对物体注意力的影响, 设计并进行了相关实验. 图 1 展示了共现偏差对物体注意力的影响, 结果表明, LVLM 更容易对高频共现的物体产生幻觉. 这揭示了 LVLM 在推理过程中, 倾向于依据物体间的共现关联性进行判断, 而不是专注于图像中的实际内容. 例如, 模型在看到“餐桌”时, 可能会错误地产生“椅子”的幻觉, 这种幻觉很大程度上归因于预训

练阶段统计偏差和共现偏差的影响. 这说明模型对这些物体的注意力集中度较高, 对其他物体的注意力不足.

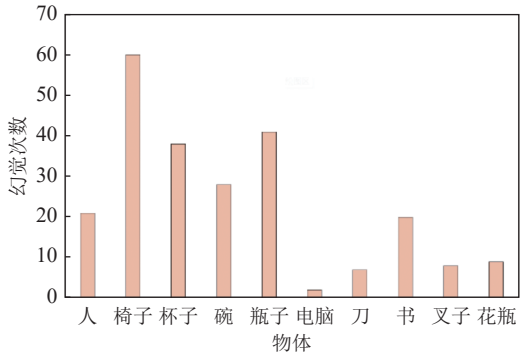

图 1 POPE 数据集中与“餐桌”同时出现的前 10 位物体的幻觉次数

综上所述, 物体幻觉的产生主要归因于 LVLM 在面对复杂场景时的注意力缺失问题. 预训练阶段的数据偏差导致模型倾向于过度关注高频共现的物体, 而忽略其他物体, 从而在多物体场景中产生幻觉现象. 通过实验结果验证可以看出, 物体共现偏差加剧了模型对部分物体的注意力缺失. 因此, 改进注意力机制以实现对所有物体的均衡关注是减少物体幻觉的重要途径.

2.3 图像对比增强方法

第 2.2 节的实验表明, 由于 LVLM 在处理物体数量多的图像时易受干扰, 所以对单个物体的注意力分散, 导致关联偏差引发的物体注意力减弱, 最终产生物体幻觉. 针对这一问题, 本文提出了一种图像对比增强方法. 该方法通过对比原始视觉输入与经过增强处理的视觉输入生成模型输出, 旨在提升模型对物体的注意力, 从而进一步增强模型对图像的整体感知能力.

鉴于 SAM 模型展现出卓越的分割精度与高度的灵活性, 本文依托 SAM 模型创新地提出了 ICE 方法, 旨在实现图像的显著增强. SAM 模型能够达成精确的 mask 分割, 得益于其采用的 ViT (vision Transformer)[32] 架构、大规模数据的训练基础、自适应的 mask 生成机制以及灵活的提示输入设计. 具体而言, ViT 架构首先将输入图像 $I \in \mathbb{R}^{H\times W\times 3}$ 分割为若干小块, 随后进行全局特征的深度提取, 并通过自注意力机制精细计算这些特征, 公式如下:

$$Z_i = Softmax\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d_k}}\right)V_i \tag{2}$$

其中, Q_i、K_i、V_i 分别位查询、键、值向量, d_k 是维度特征.

SAM 能有效捕捉全局和局部信息, 结合点、框等多种提示方式, 将特征映射到掩码生成空间. 通过自适应生成机制, SAM 逐步调整 mask 的边界, 动态生成高精度对象边界, 如图 2 所示. 最终 mask 的生成公式如下:

$$L_{\text{mask}} = -\sum_{i,j}(y_{ij}\log(M_{ij}+(1-y_{ij})\log(1-M_{ij})) \tag{3}$$

大规模训练数据增强了模型的泛化能力, 使其能适应不同场景和对象. SAM 的多尺度处理和边界感知能力进一步提高了分割的准确性, 同时其自回归优化机制确保了 mask 的不断精炼和改进. 这些技巧的结合, 使得 SAM 在复杂分割任务中表现出色.

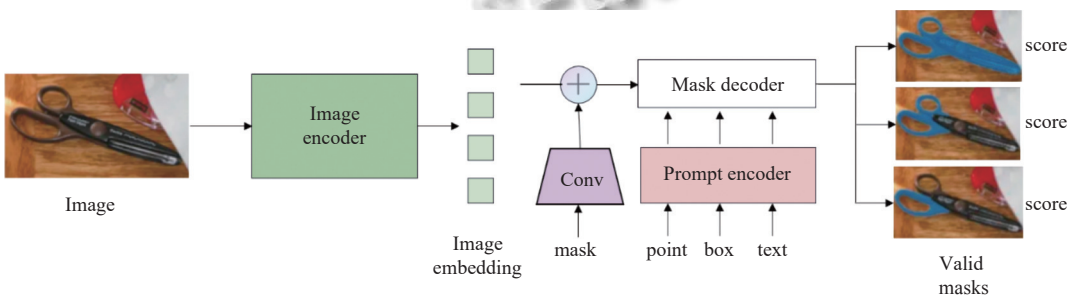

图 2 SAM 模型架构图

ICE 方法是一种基于图像对比增强的创新技术, 使用 SAM 对原图 v 进行有效的处理和增强. 具体实现为对原图进行分割, 从中提取出关键的视觉元素, 确保在增强过程中不损失重要信息. SAM 模块通过精确的分割算法, 识别出图像中的物体、边缘和细节等主要区域. 获得分割结果后, ICE 方法针对提取出的主要区域进行增强处理, 生成增强图 v'. 增强策略会根据不同的视觉特征进行调整. 这种局部增强的方式, 使得关键区域在整体图像中更加突出, 同时保持了与原图的语义一致性, 避免了过度增强可能带来的失真.

整体模型如图 3 所示, 给定文本查询和视觉输入, 模型基于原始及特定变换后的视觉输入生成两种输出分布, 再利用其差异计算新的对比概率分布, 其公式为:

$$v' = SAM(v) \tag{4}$$

$$p(y \mid v, v', I) = Softmax[\alpha \cdot \log it_\theta(y \mid v, t) + (1-\alpha) \cdot \log it(y \mid v', t)] \tag{5}$$

其中, α是权重参数, 表示增强图像和原始图像对模型输出的相对影响.

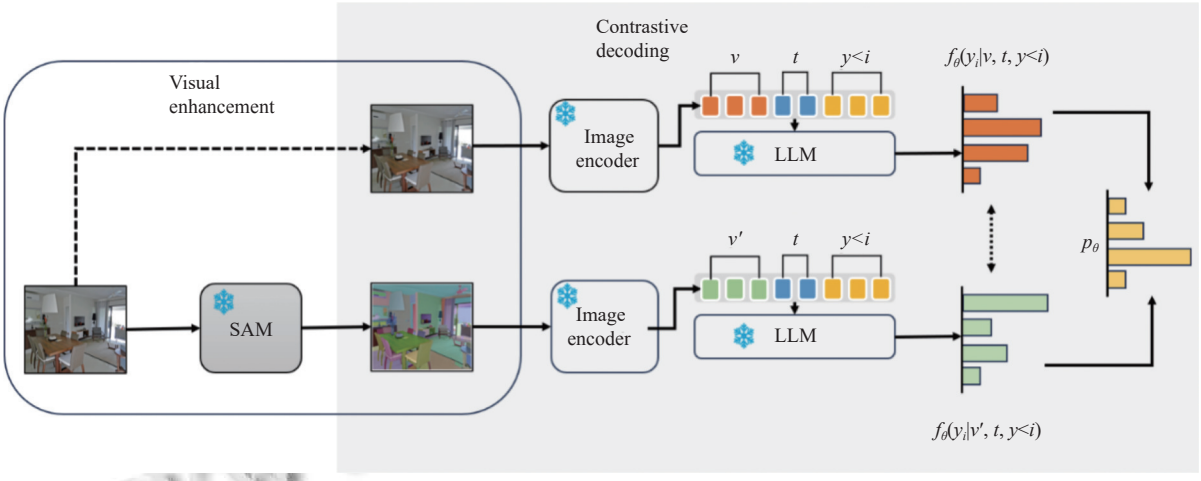

图 3 图像对比增强 (ICE) 方法架构图

2.4 合理性约束

ICE 目标的核心在于促进大型视觉语言模型输出首选的令牌, 同时对受指令干扰影响的令牌实施惩罚措施. 然而, 这一方法可能无意中削弱了那些在标准与干扰指令条件下均可被准确且自信识别的令牌预测. 这些令牌通常因其在视觉上下文 (涵盖对象、动词、属性及关系等) 中的基础性和高可能性而具有显著优势. 相反, 该方法可能错误地赋予那些代表不可信概念的令牌以不当的奖励. 为解决这一问题, 本文借鉴了开放式文本生成领域中的自适应合理性约束理念[33]. 据此, ICE 方法对目标进行了优化, 融入了自适应合理性约束, 以期实现更为精准的预测与控制:

$$\begin{cases} V_{\text{head}}(y<i) = \{y_i \in V : p_\theta(y_i \mid v, t, y<i) \\ \qquad\qquad \geqslant \beta \max p_\theta(\omega \mid v, t, y<i)\} \\ p(y_i \mid v, v', t) = 0, \quad \text{if } y_i \in V_{\text{head}}(y<i) \end{cases} \tag{6}$$

其中, V 是 LVLM 的输出词汇表, β是[0, 1]中的超参数, 用于控制下一个令牌分布的截断. β越大表示截断越激进, 仅保留高概率的标记. 这对于减轻不可信标记的影响尤其重要, 特别是当 LVLM 表现出高置信度并准确锚定在视觉语义中时.

ICE 可以使 LVLM 增强对视觉的感知, 从而缓解 LVLM 中的幻觉. 此外, 自适应合理性约束的集成通过考虑 LVLM 的置信水平进一步磨炼了对比分布, 从而将决策过程缩小到更可靠的候选池. 该方法不仅减少了 LVLM 存在的物体幻觉, 而且还减少了不可信标记的生成, 展示了 ICE 方法在增强模型可靠性和输出有效性方面的功效.

3 实验结果

本节将探讨用于减轻物体幻觉的图像对比增强方法的评估工作, 具体利用 POPE 基准来评估该方法在缓解物体级别幻觉症状方面的效果. 此外, 还进一步扩展了对 MME 基准的分析范围, 以涵盖对象和属性级别上的幻觉症状. 实验使用的环境是 Ubuntu 20.4 操作系统, GPU 为 NVIDIA GeFore RTX4090, 深度学习框架为 PyTorch.

3.1 数据集介绍

POPE 提出了一种用于评估物体幻觉的简化方法. 在该基准测试中, 对 LVLM 进行询问, 以确定给定图像中是否存在特定对象. 其中, 探测存在对象和不存在对象的查询比例保持平衡 (即各占 50%).

它包含 3 种采样设置, 即随机、流行和对抗性, 每种设置在构建负样本方面各具特色. 在随机设置中, 随机选取图像中不存在的对象; 流行设置从高频池中选择缺失的对象; 而在对抗性设置中, 优先考虑图像中不存在但同时出现概率较高的对象. POPE 基准整合来自 3 个不同数据源的数据, 分别是 MSCOCO[34]、A-OKVQA[33]和 GQA[35]. 它涉及每个采样设置下每个数

据集的 500 张图像, 并为每张图像制定 6 个问题, 最终从这些数据集的开发集中产生总计 27 000 个查询答案对. 评估以准确率 (*Accuracy*)、精确率 (*Precision*)、召回率 (*Recall*) 和 *F*1 分数这 4 个关键指标为核心. 具体公式如式 (7)–式 (10):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

其中, *TP* 代表真正例的数量, *FP* 代表假正例的数量, *FN* 代表假负例的数量, *TN* 代表真负例的数量.

MME 数据集是一个全面的基准测试工具, 旨在多维度评估 LVLM 的感知与认知能力. 该基准覆盖了从感知类别的存在、计数、位置、颜色识别到认知类别的常识推理、数值计算等共计 14 项任务, 其中特别设计了针对存在、计数、位置和颜色的任务作为幻觉辨别基准, 以考察模型在对象级与属性级的幻觉问题识别能力. 所有任务均以问答形式呈现, 并通过准确度作为指标量化模型性能.

3.2 模型基线

本文对所提出的 ICE 方法在最先进的 LVLM 上的有效性进行了全面评估. 具体而言, ICE 方法被应用于两个模型: 一是采用 Vicuna 7B[4]作为语言解码器, 并配备线性投影层作为融合模块的 LLaVA-1.5 模型; 二是基于 Qwen 7B[3]主干构建, 且其融合模块采用复杂 Q-Former 架构的 Qwen-VL 模型. 文中详细报告了这两个模型在 POPE 和 MME 基准测试中的具体测试结果.

3.3 实验结果

表 1 所总结的 POPE 实验结果证实了 ICE 方法在 POPE 基准测试的 3 个不同子集 (MSCOCO、A-OKVQA 和 GQA) 中的有效性. 值得注意的是, ICE 方法在这些实验中始终优于基础 LVLM 模型 LLaVA-1.5 和 Qwen-VL. 具体来说, 采用 ICE 方法的 LVLM 相较于基线结果实现了显著的性能提升, 其中准确度提高了 7.46%, *F*1 值提升了 8.25%.

表 1 模型在 POPE 数据集中的实验结果 (%)

数据集	采样类型	方法	LLaVA-1.5				Qwen-VL			
			Accuracy	*Precision*	*Recall*	*F*1	*Accuracy*	*Precision*	*Recall*	*F*1
MSCOCO	Random	Default	83.29	92.13	72.80	81.33	84.73	95.61	72.81	82.67
		+VCD	87.73	91.42	83.28	87.16	88.63	94.64	81.91	87.81
		+Ours (ICE)	**88.12**	**92.34**	**84.27**	**88.12**	**90.28**	**96.31**	**83.54**	**89.47**
	Popular	Default	81.88	88.93	72.80	80.06	84.13	94.31	72.64	82.06
		+VCD	85.38	86.92	83.28	85.06	87.12	91.49	81.85	86.40
		+Ours (ICE)	**86.92**	**88.12**	**84.44**	**86.24**	**88.24**	**93.38**	**83.53**	**88.18**
	Adversarial	Default	78.96	83.06	72.75	77.57	82.26	89.97	72.61	80.37
		+VCD	**80.88**	79.45	83.29	81.33	84.26	85.84	82.05	83.90
		+Ours (ICE)	80.34	**81.11**	**85.73**	**83.36**	**86.10**	**87.21**	**83.83**	**85.47**
A-OKVQA	Random	Default	83.45	87.24	78.36	82.56	86.67	93.16	79.16	85.59
		+VCD	86.15	85.18	87.53	86.34	89.22	90.77	87.32	89.01
		+Ours (ICE)	**89.72**	**87.40**	**90.84**	**89.08**	**91.72**	**93.27**	**89.22**	**91.20**
	Popular	Default	79.90	80.85	78.36	79.59	85.56	90.44	79.53	84.63
		+VCD	81.85	78.60	87.53	82.82	87.85	88.10	87.53	87.81
		+Ours (ICE)	**82.74**	**80.50**	**89.66**	**84.83**	**89.41**	**90.26**	**89.13**	**89.69**
	Adversarial	Default	74.04	72.08	78.49	75.15	79.57	79.77	79.23	79.50
		+VCD	74.97	70.01	87.36	77.73	81.27	77.79	87.53	82.38
		+Ours (ICE)	**76.33**	**71.49**	**88.91**	**79.70**	**82.37**	**79.62**	**90.01**	**84.50**
GQA	Random	Default	83.73	87.16	79.12	82.95	80.97	88.07	71.64	79.01
		+VCD	86.65	84.85	89.24	86.99	85.59	86.88	83.84	85.33
		+Ours (ICE)	**88.10**	**86.35**	**91.61**	**88.47**	**86.41**	**88.64**	**85.23**	**86.90**
	Popular	Default	78.17	77.64	79.12	78.37	75.99	78.62	71.40	74.84
		+VCD	80.73	76.26	89.24	82.24	81.83	80.45	84.09	82.23
		+Ours (ICE)	**82.24**	**78.25**	**91.63**	**84.41**	**83.14**	**82.27**	**86.52**	**84.34**
	Adversarial	Default	75.08	73.19	79.16	76.06	75.46	77.92	71.07	74.33
		+VCD	**76.09**	70.83	88.75	78.78	80.01	77.86	83.85	80.75
		+Ours (ICE)	75.69	**72.15**	**90.47**	**80.28**	**81.09**	**79.13**	**85.22**	**82.06**

POPE 的 4 个关键指标均呈现显著增强, 这有力证明了所提方法的有效性. 此外, 从随机模式到流行模式, 再到对抗性模式的逐步测试, 揭示了性能显著下降的趋势, 这凸显了统计偏差和语言先验在 LVLM 中引发幻觉的影响逐渐增强. 即便面对这些挑战, ICE 方法在所有测试设置中均展现出持续的改进趋势. 该方法通过对比原始视觉输入与增强视觉输入所产生的模型输出, 有效增强模型对物体的注意力, 增强了模型对图片的感知能力, 从而有效缓解了上述问题, 并减轻了对象级幻觉现象. 与 VCD 方法相比, ICE 方法总体性能提升了 2.64%, 这进一步验证了 ICE 方法的有效性.

表 2 展示了 MME 幻觉子集的评估结果. 需要注意的是, 这里所指的“幻觉子集”即表 2 中的存在、计数 (对象级任务) 以及位置、颜色 (属性级任务). 由于幻觉现象不仅存在于对象层面, 也可能出现在属性层面, 因此, 将研究拓展至属性相关任务显得尤为重要. 为此, 本文利用 MME 幻觉子集对 ICE 方法进行了全面评估. 结果显示, ICE 方法在 MME 幻觉子集 4 项任务中均显著优于基线方法 LVLM 和 VCD, 充分展示了其在抑制物体和属性级幻觉方面的卓越效能及显著优势. 值得注意的是, 尽管 VCD 方法在位置幻觉任务上的表现有所下滑, 但 ICE 方法依然保持了稳定的性能. 这一差异彰显了 ICE 方法在处理更广泛幻觉问题上的适应性和有效性, 使其成为 LVLM 中一个更为通用的解决方案.

表 2 所有 MME 感知相关任务的结果

模型	方法	存在	计数	位置	颜色	海报	名人	场景	地标	艺术品	OCR	总分
LLaVA-1.5	Default	170.00	113.33	111.67	150.00	123.13	102.94	138.00	128.39	108.50	87.50	1233.46
	+VCD	**175.00**	116.42	110.33	153.67	124.15	108.00	145.50	**130.90**	110.05	87.50	1261.52
	+Ours (ICE)	**175.00**	**118.33**	**116.67**	**156.67**	**130.78**	**115.29**	**147.25**	112.50	**114.75**	**92.00**	**1279.24**
Qwen-VL	Default	158.33	150.00	98.33	**173.33**	158.16	89.41	152.25	101.76	107.50	57.50	1246.58
	+VCD	170.00	150.00	95.33	164.50	160.54	112.50	**154.25**	**110.30**	114.50	55.00	1286.92
	+Ours (ICE)	**175.00**	**150.00**	**113.33**	170.00	**165.99**	**117.24**	153.00	107.79	**119.75**	**57.50**	**1329.60**

本文提出的 ICE 方法旨在缓解 LVLM 在推理过程中产生的幻觉现象. 为了深入探究该方法是否能在保留 LVLM 基本识别和推理能力的基础上进一步提升其性能, 进行了相关分析. 具体而言, 对整个综合 MME 基准进行了评估, 该基准包含 14 个子任务, 旨在全面检验模型的感知和识别能力. 表 2 和表 3 的数据表明, 在两个骨干模型上应用该方法后, 任务分数显著提升, 超越了基础 LVLM 和已有的 VCD 方法的性能. 这一结果表明, 该方法不仅能够有效控制推理过程中的幻觉, 还能提高 LVLM 在基本任务上的准确性.

表 3 所有 MME 推理相关任务的结果

模型	方法	常识推理	数值计算	文本翻译	代码推理	总分
LLaVA-1.5	Default	97.14	45.00	55.00	50.00	247.14
	+VCD	97.62	45.00	55.00	50.00	247.62
	+Ours (ICE)	**99.29**	**45.00**	**55.00**	**50.00**	**249.27**
Qwen-VL	Default	109.29	42.50	62.50	52.50	266.79
	+VCD	115.71	40.00	65.00	52.50	273.21
	+Ours (ICE)	**118.57**	**52.50**	**77.50**	**62.50**	**311.07**

在更为详尽的模型特定分析中, 观察到 ICE 方法在全部 14 个子任务中, 相较于以 LLaVA-1.5 为主干以及采用相同主干的 VCD 方法, 展现出了极大的性能提升. 相反地, 与基线 LVLM 相比, VCD 方法在推理能力的提升上显得颇为有限, 并且在处理位置信息时甚至出现了性能下滑. 此外, 当 Qwen-VL 作为主干模型时, VCD 在涉及位置识别、颜色判断以及数值计算的任务中遭遇了性能下降. 这引发了一种推测: 尽管利用视觉不确定性有助于在视觉输入中更稳固地锚定预测结果, 但这种方法可能因过度依赖视觉线索而忽视文本指令的基础作用, 从而导致了性能上的缺陷.

同时, 值得注意的是, 当以 Qwen-VL 为模型基础时, 该方法在颜色识别、地标定位以及场景理解等 MME 基准的特定子任务上并未达到最优水平. 这提示研究人员, 尽管图像增强技术可以显著提升模型在某些方面的性能, 但在某些特定任务上, 它也可能产生一定的副作用, 需要谨慎权衡和优化. 这一发现为进一步的研究和优化提供了重要的方向和启示.

4 结论与展望

在本文中, 针对大型视觉语言模型 (LVLM) 存在的物体幻觉问题进行了深入研究. 分析了物体幻觉产生的根本原因, 并通过实验验证了图像中物体注意力的缺失是导致物体幻觉的主要因素. 为此, 提出了图像对比增强方法 (ICE). 该方法通过对比原始视觉输入与

增强视觉输入所生成的模型输出, 强化模型对物体的注意力分配, 进而提升其对图片的全面感知能力. 在多个基准测试和 LVLM 系列中进行的广泛实验, 有力地证实了 ICE 在减少幻觉方面的显著效果, 同时也展现了其在提升 LVLM 整体感知能力方面的巨大潜力.

目前有关 LVLM 物体幻觉问题的研究工作较少, 该领域面临诸多挑战, 还有很多问题值得深入探究. 未来研究可考虑更精确地识别和量化幻觉程度, 以便采取更有针对性的抑制措施. 此外, 针对不同视觉任务和语言场景, 优化图像对比增强方法以适应特定需求, 以及探索新的模型架构和训练策略以降低 LVLM 产生幻觉的可能性, 都是未来研究的重点方向, 有望为该领域带来新突破.

参考文献

1 Zhao WX, Zhou K, Li JY, *et al*. A survey of large language models. arXiv:2303.18223, 2023.

2 Wei J, Wei J, Tay Y, *et al*. Larger language models do in-context learning differently. arXiv:2303.03846, 2023.

3 Bai JZ, Bai S, Chu YF, *et al*. Qwen technical report. arXiv:2309.16609, 2023.

4 Chiang WL, Li Z, Lin Z, *et al*. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. https://vicuna.lmsys.org. (2023-04-14) [2024-10-16].

5 Lund BD, Wang T. Chatting about ChatGPT: How may AI and GPT impact academia and libraries? Library Hi Tech News, 2023, 40(3): 26–29.

6 Touvron H, Martin L, Stone K, *et al*. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.

7 Yang Z, Li L, Lin K, *et al*. The dawn of LMMs: Preliminary explorations with GPT-4V (ision). arXiv:2309.17421, 2023.

8 Dai WL, Li JN, Li DX, *et al*. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 2142.

9 Bai JZ, Bai S, Yang SS, *et al*. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview.net, 2024. 1–24.

10 Liu HT, Li CY, Wu QY, *et al*. Visual instruction tuning. Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2024. 1516.

11 Liu HT, Li CY, Li YH, *et al*. Improved baselines with visual instruction tuning. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 26286–26296.

12 Ye QH, Xu HY, Xu GH, *et al*. mPLUG-Owl: Modularization empowers large language models with multimodality. arXiv:2304.14178, 2023.

13 Zhu DY, Chen J, Shen XQ, *et al*. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview.net, 2024. 1–17.

14 Gemini Team Google, Anil R, Borgeaud S, *et al*. Gemini: A family of highly capable multimodal models. arXiv:2312.11805, 2023.

15 Li YF, Du YF, Zhou K, *et al*. Evaluating object hallucination in large vision-language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 292–305.

16 Yin SK, Fu CY, Zhao SR, *et al*. Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences, 2024, 67(12): 220105. [doi: 10.1007/s11432-024-4251-x]

17 Zhou YY, Cui CH, Yoon J, *et al*. Analyzing and mitigating object hallucination in large vision-language models. Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview.net, 2024. 1–30.

18 Liu FX, Lin K, Li LJ, *et al*. Mitigating hallucination in large multi-modal models via robust instruction tuning. Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview.net, 2024. 1–45.

19 Yu TY, Yao Y, Zhang HY, *et al*. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 13807–13816.

20 Brown TB, Mann B, Ryder N, *et al*. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 159.

21 Fu CY, Chen PX, Shen YH, *et al*. MME: A comprehensive evaluation benchmark for multimodal large language models. arXiv:2306.13394, 2023

22 Li JN, Li DX, Savarese S, *et al*. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. 19730–19742.

23 Chen KQ, Zhang Z, Zeng WL, *et al*. Shikra: Unleashing multimodal LLM's referential dialogue magic. arXiv:2306.15195, 2023.

24 Bavishi R, Elsen E, Hawthorne C, *et al*. Fuyu-8B: A multimodal architecture for AI agents. https://www.adept.ai/blog/fuyu-8b. (2023-10-17)[2024-10-16] .

25 Huang QD, Dong XY, Zhang P, *et al*. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 13418–13427.

26 Leng SC, Zhang H, Chen GZ, *et al*. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 13872–13882.

27 Chen ZR, Zhao ZK, Luo HY, *et al*. HALC: Object hallucination reduction via adaptive focal-contrast decoding. Proceedings of the 41st International Conference on Machine Learning. Vienna: JMLR.org, 2024. 307.

28 Kirillov A, Mintun E, Ravi N, *et al*. Segment anything. Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023. 3992–4003.

29 Agarwal V, Shetty R, Fritz M. Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9687–9695.

30 Agrawal A, Batra D, Parikh D. Analyzing the behavior of visual question answering models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016. 1955–1960.

31 Goyal Y, Khot T, Summers-Stay D, *et al*. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6325–6334.

32 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020.

33 Schwenk D, Khandelwal A, Clark C, *et al*. A-OKVQA: A benchmark for visual question answering using world knowledge. Proceedings of the 17th European Conference on Computer Vision. Tel Aviv: Springer, 2022. 146–162.

34 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision (ECCV 2014). Zurich: Springer, 2014. 740–755.

35 Hudson DA, Manning CD. GQA: A new dataset for real-world visual reasoning and compositional question answering. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6693–6702.

(校对责编: 王欣欣)