计算机研究与发展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(12): 2929−2950, 2025

视觉语言大模型的幻觉综述:成因、评估与治理

李 煦[1] 朱 睿[1] 陈小磊[1] 伍瑾轩[2] 郑 毅[2] 赖承杭[1] 梁宇轩[2] 李 斌[2] 薛向阳[1,2]
[1](复旦大学大数据研究院 上海 200433)
[2](复旦大学计算机科学技术学院 上海 200438)
(jerryryrylx@gmail.com)

Survey of Hallucinations in Large Vision-Language Models: Causes, Evaluations and Mitigations

Li Xu[1], Zhu Rui[1], Chen Xiaolei[1], Wu Jinxuan[2], Zheng Yi[2], Lai Chenghang[1], Liang Yuxuan[2], Li Bin[2], and Xue Xiangyang[1,2]
[1](*Institute of Big Data*, *Fudan University*, *Shanghai* 200433)
[2](*School of Computer Science*, *Fudan University*, *Shanghai* 200438)

Abstract LVLMs (large vision-language models) represent a significant advancement in the intersection of natural language processing and computer vision. By integrating pre-trained visual encoders, vision-language adapters, and large language models, LVLMs can understand both visual and textual information, and generate responses in natural language, making them suitable for a range of downstream vision-language tasks such as image captioning and visual question answering. However, these models commonly exhibit hallucinations — generating inaccurate perceptions of image contents. Such hallucinations significantly limit the application of LVLMs in high-stakes domains like medical image diagnosis and autonomous driving. This survey aims to systematically organize and analyze the causes, evaluations, and mitigation strategies of hallucinations to guide research in the field and enhance the safety and reliability of LVLMs in practical applications. It begins with an introduction to the basic concepts of LVLMs and the definition and classification of hallucinations within them. It then explores the causes of hallucinations from four perspectives: training data, training task, visual encoding, and text generation, while also discussing the interactions among these factors. Following this, it discusses mainstream benchmarks for assessing LVLM hallucinations in terms of task setting, data construction, and assessment metrics. Additionally, it examines hallucination mitigating techniques across five aspects: training data, visual perception, training strategy, model inference, and post-hoc corrections. Finally, the review provides directions for future research in the areas of cause analysis, evaluation, and mitigation of hallucinations in LVLMs.

Key words natural language processing; computer vision; large vision-language models; multimodal large language models; hallucinations

摘 要 视觉语言大模型(large vision-language models, LVLMs)代表了自然语言处理与计算机视觉交叉领域的一项重要进展.通过结合预训练的视觉编码器、视觉语言适配器和大语言模型,LVLMs能够同时理解图像与文本信息,并通过自然语言进行响应,适用于图像描述、视觉问答等多种视觉语言下游任务.然而,这类模型普遍存在幻觉现象,即模型对于图像内容进行了错误感知,制约了其在医学图像诊断、自动驾驶等高风险领域的赋能应用.旨在系统梳理并深入分析幻觉成因、评估方法及治理策略,为LVLMs

收稿日期:2024-06-05;**修回日期**:2025-04-09
基金项目:国家自然科学基金(62176061)
This work was supported by the National Natural Science Foundation of China (62176061).
通信作者:薛向阳(xyxue@fudan.edu.cn)

的可靠性研究提供指导. 首先, 介绍 LVLMs 的基础概念及其幻觉现象的定义与分类; 随后, 从训练数据、训练任务、视觉编码、文本生成 4 方面分析 LVLMs 的幻觉成因, 并讨论这些成因间的交互关系; 接着, 从任务形式、数据构建和评估指标 3 方面介绍 LVLMs 的幻觉评估策略; 此外, 从训练数据、视觉感知、训练策略、模型推理、事后修正 5 方面讨论 LVLMs 的幻觉治理技术; 最后, 为这类幻觉的成因分析、评估和治理 3 方面提供未来的研究方向.

关键词 自然语言处理; 计算机视觉; 视觉语言大模型; 多模态大语言模型; 幻觉

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440444 **CSTR:** 32373.14.issn1000-1239.202440444

自然语言处理与计算机视觉代表了人工智能领域的 2 块关键拼图, 而将两者融合是实现通用人工智能的关键路径[1]. 近年来, 随着大语言模型(large language models, LLMs)[2-8] 和预训练视觉模型[9-14] 的飞速发展, 构建视觉语言大模型(large vision-language models, LVLMs)成为了新的研究热点[15-36]. LVLMs 能够同时处理图像和文本信息, 并通过自然语言与用户交互, 展现出强大的多模态信息理解与文本生成能力. 这些能力来源于 LVLMs 的 3 个主要组件: 预训练视觉编码器、视觉语言适配器和预训练 LLM. 视觉编码器负责从图像中提取视觉特征, 视觉语言适配器负责将视觉特征对齐到 LLM 的嵌入空间, 而 LLM 则负责同时理解图像和文本信息并生成相应的文本回复. 通过多模态指令微调, LVLMs 能够以自然语言交流的方式解决多种视觉语言下游任务, 包括图像描述、视觉问答和光学字符识别等.

LVLMs 凭借其卓越的泛化性和友好的交互性, 在多模态领域展现出强大的应用潜力. 然而, 这类模型的幻觉却成为了阻碍其在现实世界中广泛应用的主要因素[37]. LVLMs 的幻觉是指模型在生成文本中显示出了其对图像内容的错误感知. 例如, 模型错误感知了图中不存在的物体、错误识别了图中的物体关系或错误描述了图中的物体属性等. 这种现象严重削弱了 LVLMs 生成内容的可靠性, 限制了其在关键领域的应用落地. 如图 1 所示, 在医疗影像分析中, 由幻觉造成的错误诊断可能导致不必要的治疗或错误用药, 对患者的身心健康构成潜在风险, 甚至引发医疗事故[38-39]; 在自动驾驶技术中, 对交通标志或路

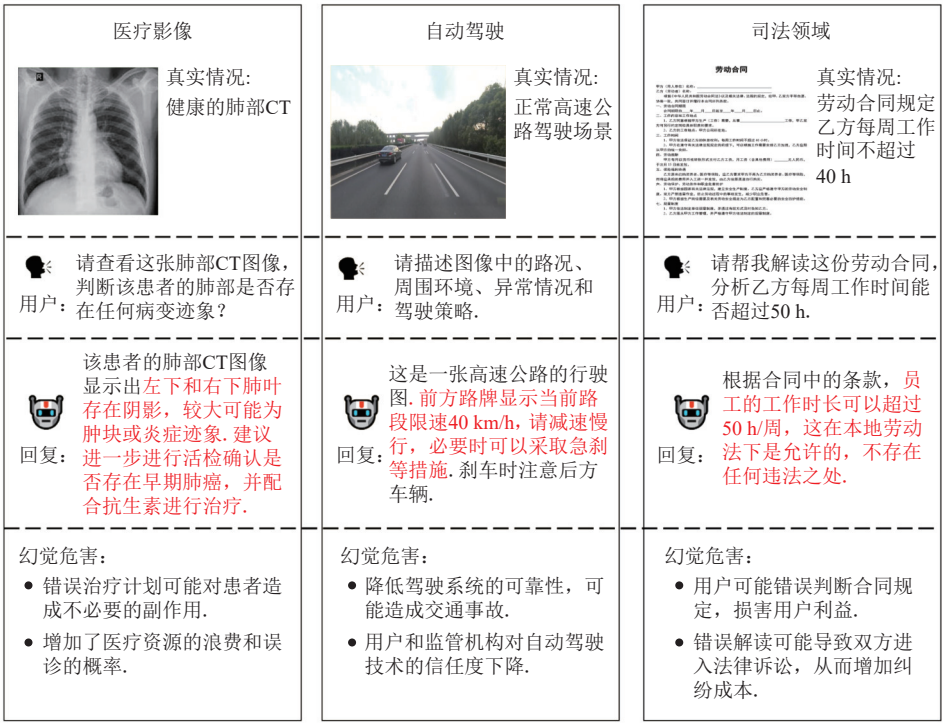


Fig. 1 The hallucinatory scenes in LVLMs

图 1 LVLMs 中的幻觉场景

况的误解可能触发系统性判断失误，进而引发交通事故，威胁道路安全[40-41]；在司法领域中，模型在证据分析或案件审查时出现幻觉，可能导致对案件事实的误读或误判，损害判决的公正性[42-43]. 此外，幻觉现象还会引发法律责任不明和信任危机等伦理问题. 无论是医疗事故、交通事故还是司法误判，都难以明确应由开发者、使用者还是相关机构承担责任. 其次，这些领域高度依赖公众对 AI(artificial intelligence)系统的信任，而频繁的幻觉错误会破坏信任基础，阻碍技术的发展与推广. 因此，确保 LVLMs 输出的准确性和可靠性至关重要.

分析并解决 LVLMs 的幻觉面临着多重挑战. 首先，幻觉的成因复杂多样，涉及训练数据、训练方法、架构设计等多个因素的交互作用，使得幻觉的成因分析变得格外困难[37]. 其次，由于开放式生成任务缺乏统一的评估标准和客观的测量工具，评估幻觉的性质和程度本身就构成了一大挑战，不同研究可能会采用不同的评估方法[17,44-62]. 此外，当前的幻觉治理方法[16,46,48,61,63-67]往往需要在增加计算成本和牺牲模型性能之间做出权衡. 在这一背景下，系统性综述 LVLMs 的幻觉研究显得尤为重要，不仅有助于梳理和分析幻觉成因及其对策，还能为未来研究提供理论基础和实践指导.

本综述旨在全面介绍造成 LVLMs 幻觉的原因、评估方法及治理手段，并探讨潜在的改进方向，以促进该领域的科学发展与实际应用.

1 预备知识

1.1 LVLMs

1.1.1 模型架构

LVLMs 是一种集成了视觉感知能力的 LLMs，能够同时接收图像和文本指令作为输入，并生成对应的文本输出. 如图 2 所示，LVLMs 通常由 3 个部件组成：预训练视觉编码器、视觉语言适配器以及预训练 LLMs. 视觉编码器首先从输入图像中提取视觉特征，随后视觉特征通过视觉语言适配器对齐到 LLMs 的嵌入空间，最终 LLMs 负责联合理解视觉与指令信息，并利用其强大的世界知识和推理能力生成相关的文本响应.

在视觉编码器的选型上，现有研究通常采用经过图像文本对比预训练的 ViT(vision Transformer)模型[10]，如 CLIP-ViT[9] 和 EVA-CLIP[14] 等. 这类编码器提取的视觉特征本身是序列化的，自然满足了 LLMs 对

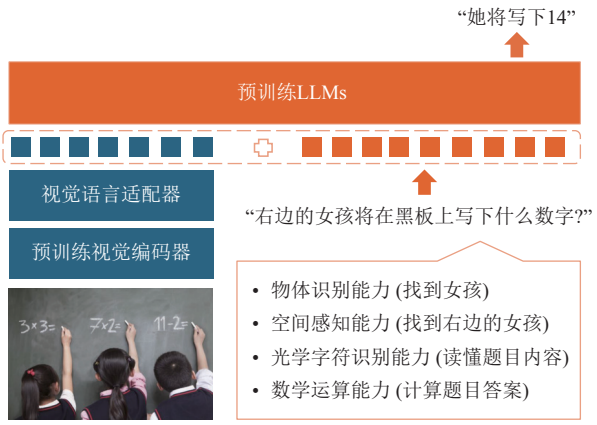


Fig. 2 The typical architecture of LVLMs

图 2 LVLMs 的典型架构

于输入格式的需求. 此外，图像与文本的对比训练使得其所提取的视觉特征含有丰富的语义信息，减小了视觉特征与 LLM 模块的对齐难度.

视觉语言适配器按架构通常分为 3 类. 第 1 类适配器采用单个线性层[17,22,24,68]. 这种设计结构简单、参数量小，但难以实现视觉特征与 LLMs 嵌入空间的充分对齐. 第 2 类适配器是带有非线性激活函数的双层全链接神经网络[21,33-34,36]. 非线性的加入增强了适配器的对齐能力. 第 3 类适配器采用 Q-former 架构[20,25,27,29]，即一个融合了交叉注意力模块与可学习查询向量的 Transformer 编码器. 这类适配器参数量大、训练难度高，但具有 2 点优势：1)通过设置可学习查询向量的个数可以控制视觉特征的序列长度，从而减轻视觉特征对于 LLMs 上下文窗口的占用；2)通过将文本指令与可学习查询向量一同输入适配器，有助于萃取出与文本指令更相关的视觉特征. 亦有早期研究尝试过在 LLMs 内部植入交叉注意力模块来接收视觉特征[31,69-71]，但这种做法需要调整 LLMs 的原生架构，有悖于模块化设计原则，遂在后续研究中被逐渐放弃.

LVLMs 通常采用 2 类 LLM 模块. 一类是只经过预训练的基础模型，如 LLaMA[2]，LLaMA-2[3] 等. 另一类是经过指令微调的模型，如 Vicuna[72]，LLaMA-2-Chat[3] 等. 研究表明，经过指令微调的 LLMs 能够增强 LVLMs 在多模态场景下的指令跟随能力[31]，因此这种语言模型已成为主流选择.

1.1.2 训练方法

LVLMs 的训练通常由跨模态对齐预训练和视觉指令微调[73]2 阶段组成. 跨模态对齐预训练旨在让视觉语言适配器学会将视觉特征有效转换为 LLM 模块能够理解的语义嵌入. 该阶段的训练数据由图像与文本描述的配对组成，训练策略是最大化给定图像

时文本描述在模型输出端的对数似然. 在这一阶段, 视觉编码器和 LLM 模块通常保持参数冻结, 仅调整视觉语言适配器的参数. 视觉指令微调阶段旨在增强模型的多模态指令跟随能力. 该阶段的训练数据由图像、文本指令及文本响应组成的三元组构成, 训练策略是最大化给定图像和文本指令时文本响应在模型输出端的对数似然. 由于 LVLMs 的指令跟随能力主要依赖于其中的 LLM 模块[17], 因此, 这一阶段通常会解冻 LLM 组件(或其中的高效参数微调模块), 使其与视觉语言适配器共同训练, 而视觉编码器则保持全程冻结以减轻计算负担.

1.1.3 下游任务

得益于多样化的训练数据和强大的 LLM 组件, LVLMs 能够以零样本方式解决传统视觉领域或视觉语言领域的诸多下游任务, 包括物体识别、图像描述、图像问答、光学字符识别、图表理解、仇恨检测、目标检测、指代表达理解与生成等. 更为关键的是, 这些任务通过自然语言交互来执行, 极大提升了模型的灵活性和用户友好性.

1.2 LVLMs 中的幻觉

1.2.1 幻觉的定义与分类

LVLMs 的幻觉是指模型生成的文本响应表现出了对图像内容的错误感知. 这种幻觉现象存在一个重要补集: 模型在正确感知视觉内容的基础上, 由于推理错误或知识误用而生成了错误响应. 例如, 当模型在面对图 2 中的图像与指令时, 回复道“她已经写下了 7 乘 2, 将写下 42”. 虽然该文本响应表现出了与视觉上下文的推理不一致, 但并不涉及视觉模态的感知失误, 遂不属于本文所探讨的幻觉范畴. 需要强调的是, 尽管 LVLMs 的幻觉被定义在视觉感知层面, 但其与传统视觉模型(如图像分类模型)所产生的感知错误仍有所不同. 这种差异主要源于 LVLMs 使用自然语言作为任务理解和生成输出的媒介. 而自然语言在表达相同语义时可能具有多种形式, 使得 LVLMs 的感知错误比传统视觉模型更复杂多样. 例如, 图像分类模型的输出通常是固定的类别标签, 其错误主要表现为不正确的类别预测. 然而, 在 LVLMs 中, 由于自然语言的多样性, 针对相同的感知任务, 不同的文本指令会导致不同的幻觉界定. 仍以图 2 为例, 若任务指令是“请描述图中的女孩数量”, 此时, 除“2 位”以外的任何数量描述都将被视为幻觉. 若将指令换为“请问图中是否有 2 位女孩?”, 则任何否定性质的回复都将属于幻觉内容. 因此, LVLMs 对于视觉模态的错误感知涉及跨模态内容理解, 使得幻觉现象的表现形式超越了传统视觉模型的错误.

如图 3 所示, 根据视觉感知错误的维度不同, LVLMs 的幻觉可以分为 3 类: 物体存在幻觉(object existence hallucination)、物体属性幻觉(object attribute hallucination)以及物体关系幻觉(object relationship hallucination)[74]. 物体存在幻觉是指模型响应反映了与实际物体存在情况不一致的信息; 物体属性幻觉是指模型错误地感知了某个对象的属性, 如颜色、形状、材质等; 物体关系幻觉则指模型错误地识别了 2 个或多个物体间的关系, 包括空间位置关系、交互关系或其他形式的动态关系. 尽管这 3 种类型已经涵盖了所有视觉感知的错误形式, 仍有研究人员在尝试更为细致的幻觉分类. Liu 等人[44] 从幻觉的诱发原因入手, 提出了多模态冲突幻觉(multi-modal conflicting hallucination)和反常识幻觉(counter-common-sense hallucination), 前者是指当文本输入和视觉输入不匹配

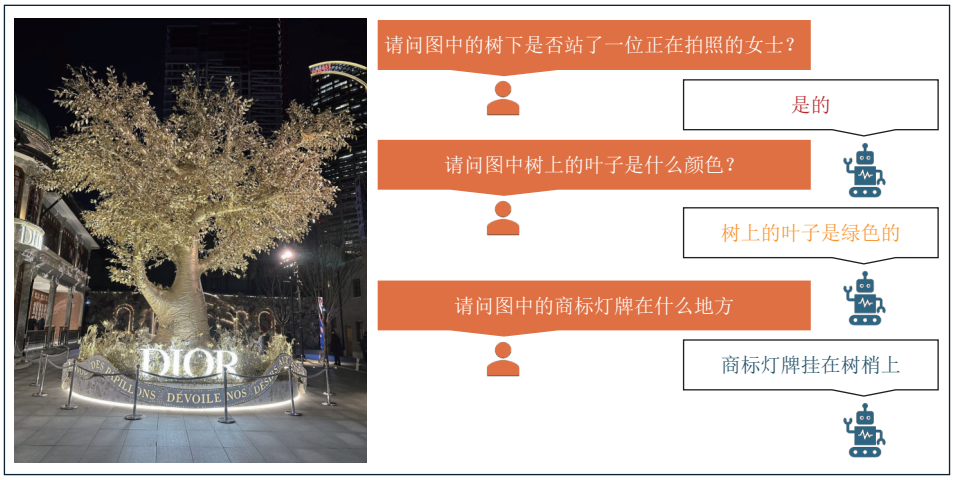


Fig. 3 The hallucinations phenomenon in LVLMs

图 3 LVLMs 中的幻觉现象

或相互矛盾时引发的幻觉现象，而后者则指的是由图像中反常识元素引起的幻觉现象. 此外，Jiang 等人[45]观察到，LVLMs 可能会基于错误的物体、属性或关系感知而创造出完全虚构的事件或情景，从而定义出了事件幻觉(event hallucination). Zhang 等人[46]则聚焦在 LVLMs 对于图中物体的计数能力，定义出了数字幻觉(number hallucination)，即模型未能准确地识别输入图像中特定物体的数量.

1.2.2 与 LLMs 的幻觉区别

尽管 LVLMs 的本质是多模态 LLMs，但视觉信息的引入使得两者间的幻觉研究存在明显差异. LLMs 处理纯文本数据，其幻觉主要涉及逻辑错误、事实错误或与用户输入文本的冲突错误[75]. 这些幻觉通常由训练数据的偏差、数据质量问题或模型自身的局限性(如对特定上下文的理解不足)造成. 相比之下，LVLMs 同时处理视觉和文本信息，其幻觉不仅涉及对视觉信息的感知不足，也可能包含对文本指令的理解失误. 另一方面，LLMs 的幻觉产生通常与语言模型的预训练和微调过程有关，治理方法侧重于通过改进训练策略或调整数据集来减少错误信息的生成[76]. 而在 LVLMs 中，除了上述挑战外，还需要解决视觉与文本信息的有效融合问题，要求其中的 LLM 模块不仅能理解单独的文本或图像信息，还必须把握两者之间的关系和相互作用，增加了模型设计和训练的复杂度. 此外，LLMs 的幻觉评估通常依赖于检验文本内容的准确性、逻辑性以及与现实世界事实的符合度，评估方法包括专家评审、自动化的一致性检测等[77]. 而 LVLMs 则需要重点考虑生成文本与输入图像的一致性. 因此，LVLMs 的幻觉评估更为复杂，需要综合视觉和文本信息进行判断，从而涉及到更多基于视觉理解的评估技术[17,48,59,63].

2 LVLMs 的幻觉成因

LVLMs 的幻觉成因复杂多样，涉及多层面因素的相互交织. 当前研究尚未对这一现象的成因形成普遍共识. 本节将尝试对 LVLMs 的幻觉成因进行分类讨论，并分析不同成因间的交互关系.

2.1 幻觉成因的归类分析

如图 4 所示，LVLMs 的幻觉成因可分为四大类：训练数据相关因素、训练任务相关因素、视觉编码相关因素和文本生成相关因素.

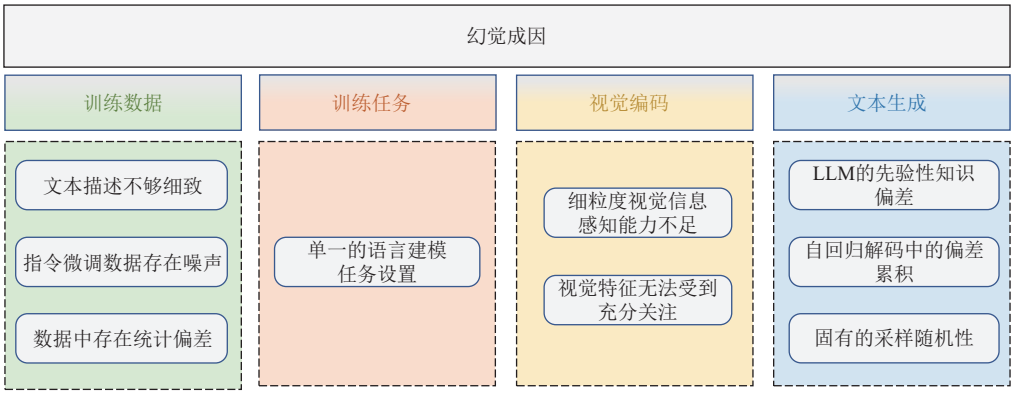


Fig. 4 Frame work diagram of hallucinations causes in LVLMs

图 4 LVLMs 幻觉成因框架图

2.1.1 训练数据相关

作为一种数据驱动的模型，LVLMs 的输出可靠性与训练数据的质量高度相关. 然而，当前的训练数据通常存在以下问题.

1)图像的文本描述不够细致. 在跨模态对齐阶段，训练数据由图像和描述文本对组成. 然而，在大多数开源数据集中，每张图像所匹配的文本描述较为粗糙，无法充分覆盖图像内容，限制了模型在文本和图像细节间建立精准语义关联的能力[78].

2)合成的指令微调数据存在噪声. 在视觉指令微调阶段，模型使用的训练数据通常由纯文本 GPT-4[79]根据图像描述、尺寸、边界框坐标和边界框内的物体标注而生成[17]. 这种合成数据中的文本部分往往冗长且带有噪声，容易包含与图像内容不相关的信息[48]. 此外，由于纯文本 GPT-4 无法感知到像素级图像信息，其生成的指令与响应数据自身就可能含有幻觉元素. 有研究显示，LLaVA 模型[17]所使用的 GPT-4 合成数据包含约 32.6% 与图像内容不符的幻觉文本[80]. 训练数据中的噪声和幻觉可能在训练过程中将错误的模式内化到模型权重中，导致模型在面对类

似场景时产生错误的视觉感知.

3)数据中存在统计偏差. LVLMs 训练数据中的统计偏差主要体现在 2 个方面:一是图像中物体类别的出现频率和共现频率不均衡[47,65-66],二是指令微调数据中正向响应的比例过高(正向响应是指模型在面对判断性指令时生成的肯定性回复或认可性输出,无论指令内容是否涉及物体的存在、属性、关系或其他视觉信息)[55,66]. 这种统计偏差会使模型学习到错误的生成偏好,从而在实际应用中表现出错误的视觉感知. 例如,物体类别的频率偏差会导致模型在图像描述中提及当前图像中不存在,但在训练数据中频繁出现或者与图像中其他物体高频共现的物体. 正向响应的比例失衡则会使模型在面对与图像内容冲突的指令时给出错误的肯定回答. 例如,当用户输入一张小狗图像并提问“图中的动物是一只猫吗”时,模型可能会错误地回答“是的”.

2.1.2 训练任务相关

LVLMs 在跨模态对齐和指令微调阶段均采用以交叉熵为损失函数的语言建模训练任务. 这种任务通过最大化给定图像和指令时响应文本的对数似然,使模型学习到数据中的多模态关联模式. 这种单一的任务设置有助于简化训练复杂度,但其缺乏对模态间一致性的显性约束,可能导致模型过于关注生成文本的句法合理性与语言通顺性,而忽略了文本内容与视觉信息的一致. 例如,模型可能根据上下文生成一个看似合理但与视觉信息不符的图像描述[81].

2.1.3 视觉编码相关

LVLMs 的视觉感知能力来源于其使用的预训练视觉编码器,而这类基于图像-文本对比学习的视觉编码器可能存在以下问题.

1)细粒度视觉信息感知能力不足. 以 CLIP-ViT 为代表的预训练视觉编码器倾向于在全局层面提取图像的语义信息,而难以有效感知图像细节. 这种局限主要源于 2 点:首先,这类视觉编码器所接收的图像分辨率通常低于 400 像素,限制了其对细粒度视觉信息的捕捉能力;其次,这类视觉编码器多在互联网规模的图像文本配对数据上进行训练,而这种训练数据中的文本标注通常存在大量噪声且只能涵盖图像的主要内容而非细节信息. 有研究表明,CLIP-ViT 会将 2 张能够被人类轻易观察出细节差异的图像进行相似编码,从而导致 LVLMs 在这些图像细节上产生错误感知[15-16,20,74].

2)视觉特征无法受到 LLMs 的充分关注. 视觉编码器输出的视觉特征需要被视觉语言适配器映射为 LLM 模块能够理解的语义嵌入(即视觉 token). 若视觉 token 中有关图像内容的语义信息不够充分,LLM 模块便难以从中找到与文本指令对应的语义匹配,就会逐渐减少对视觉 token 的关注,从而将更多注意力分配给指令文本和前置的生成内容,最终生成幻觉内容. 有研究表明,在 LVLMs 的单轮问答中,尽管视觉 token 通常占 token 总数的一半以上,但其在 LLM 模块中所受到的平均注意力分数却不足 20%,且该比例会随着 LLM 层数的升高和生成长度的增加而持续下降[82]. 亦有研究指出,如果 LVLMs 在生成过程中遭遇到锚点 token(即某些文本 token 吸引了后续生成的大部分注意力),视觉 token 所受到的关注会被进一步削弱[83].

2.1.4 文本生成相关

作为 LVLMs 中内容生成的核心模块,LLMs 的某些固有特点也会导致模型出现错误的视觉感知,这些特点包括以下几点:

1)LLMs 的先验性知识偏差. LVLMs 中的 LLM 模块是在海量文本语料上训练而来的语言模型,参数化了丰富的世界知识. 这些知识虽然有助于增强模型对于复杂世界的理解能力,但也可能引入先验性知识偏差[44,84],即模型对于世界的默认理解或固有假设与输入的图像内容不符. 这种偏差导致模型在面对违反直觉、常识或事实的视觉内容时,倾向于依赖先验知识而忽视实际的视觉证据,从而产生幻觉输出. 有研究发现,即使在视觉问答任务中将原始图像替换为无信息的噪声图像,LVLMs 仍可能输出高置信度的答案,证实了这种先验性知识偏差的存在[84].

2)自回归解码中的偏差累积. LVLMs 中的文本生成由 LLM 模块通过自回归解码完成. 在每一步解码过程中,LLM 模块会基于对齐后的视觉特征、指令文本和已经生成的文本,计算下一个文本 token 的采样分布. 由于每一步的输出都会依赖于已有的生成内容,其中任何一步所产生的偏差或错误都会被传递到下一步,从而在整个生成过程中逐步形成偏差累积,导致生成质量随生成长度的增加而下降. 有研究表明,在图像描述任务中,LVLMs 发生幻觉的频率会随生成长度的增加而升高[65,85]. 这种现象表明,偏差一旦在早期生成中出现,就会通过自回归机制不断加剧,导致幻觉元素愈产愈多.

3)固有的采样随机性. 在文本生成时,LLM 模块会计算词表中每一个 token 作为当前输出的采样概率. 即使合理的候选 token 通常会被分配较高概率,

但采样随机性仍会导致模型偶尔输出概率较低的错误 token. 这种随机性有助于增强输出的多样性和新颖性，使模型能够生成多种风格的回答，但可能导致生成文本不符合图像的实际内容，导致模型出现错误的视觉感知[86-87].

2.2 幻觉成因间的相互作用

LVLMs 的幻觉往往不是单一成因导致的，而是多种因素相互交织的结果. 以下是对成因类别间主要交互关系的分析.

1)训练数据与训练任务的交互. 语言建模的训练目标是通过最大似然估计来优化模型，会放大训练数据中的统计偏差. 例如，如果训练数据中“猫”和“沙发”频繁共现，模型在生成时会倾向于在描述“沙发”时提及“猫”，即使图像中没有猫. 这种偏差会在训练中逐步内化为模型的生成偏好，形成错误的文本关联. 另一方面，当数据中的正向反馈比例过高时，语言建模任务会强化这种模式，导致模型在面对与图像内容不符的判断型指令时，仍倾向于生成错误的肯定性回答. 因此，训练数据的不均衡分布与语言建模任务的局限性相互强化，共同削弱了模型对图像真实内容的理解能力，最终增加了幻觉的发生概率.

2)视觉编码与文本生成的交互. 当模型接收到的文本指令涉及视觉编码器无法有效捕捉的图像细节，或 LLM 模块未能对视觉 token 分配足够的注意力时，模型会倾向于利用先验知识或已经生成的文本内容来完成后续生成，而非真实的视觉信息. 在自回归解码过程中，这种对文本关联的依赖与采样过程中的随机性共同作用，加剧了偏差的逐步累积. 随着生成内容的不断叠加，早期的小错误会被逐步放大，导致生成文本与图像内容逐渐偏离，最终形成错误的视觉感知.

3)训练数据与视觉编码的交互. 当训练数据中的文本内容缺乏对图像细节的描述时，模型难以从数据中学到细粒度的多模态关联关系，会进一步放大视觉编码器对细节感知的不足. 另一方面，如果训练数据中存在显著的统计偏差(如特定物体的共现)，模型会倾向于依赖文本关联进行推断，而忽视对视觉信息的关注. 这种倾向将使得视觉 token 在生成过程中受到的关注进一步减弱.

4)训练任务与文本生成的交互. 训练任务中的语言建模目标侧重于优化生成文本的流畅性和语法合理性，而缺乏模态间一致性的明确约束. 这种缺陷会在自回归生成过程中积累偏差，将早期的细微错误逐步扩大，最终导致生成文本偏离真实的图像内容. 同时，这种训练任务会使模型在生成时更容易依赖先验知识. 例如，在面对违反常识的输入时，模型会忽略实际的视觉证据，生成与常识知识相符但与图像内容相悖的文本. 而这种与视觉内容的偏差会在自回归生成中不断放大，加剧幻觉.

综上所述，训练数据的噪声和统计偏差通过训练任务被固化为模型的生成偏好，视觉编码器的局限进一步削弱了模型对视觉信息的关注，而自回归解码与采样随机性使这些错误不断积累并放大. 这些因素共同作用，使 LVLMs 在多模态任务中表现出错误的视觉感知.

3 LVLMs 的幻觉评估

幻觉评估旨在考察 LVLMs 能否正确感知输入图像的实际内容，不仅有助于衡量模型输出的可靠性，且对于模型的优化和调整具有指导意义. 本节将从任务形式、数据集构建、评估指标 3 方面介绍 LVLMs 的幻觉评测基准，主要内容见表 1.

3.1 任务形式

LVLMs 的幻觉评估涉及将特定的图片和指令输入模型，并评估其输出是否真实反映了图像内容. 按照指令内容，幻觉评估包括视觉描述和视觉问答 2 种任务形式.

3.1.1 视觉描述

视觉描述任务主要针对物体存在幻觉. 其通过特定指令要求模型为给定图像生成描述文本，随后通过名词提取和内容比对分析该文本中幻觉物体的比例. 名词提取是指从生成描述中精确地识别和分离出物体名词或短语，而内容比对则用于将提取到的物体与图像中的实际情况进行对比，从而识别出虚构物体.

在名词提取方面，文献 [50] 首先对模型输出的每个句子进行标记，并将每个物体名词转换为单数形式，然后使用 MSCOCO[89] 中的物体同义词列表将名词映射到 80 个物体类别，同时防止复合词汇如“热狗”被错误分类为物体“狗”，从而得到生成内容中的物体列表. 文献 [51] 对于 MSCOCO 数据集中的测试样本采用与文献 [50] 相同的处理方法，而对于 NoCaps 数据集[94] 则将 600 个细粒度物体类别映射到 139 个粗粒度物体类别，以降低后续对比的复杂度. CCEval[74] 利用 GPT-4 从生成描述中提取物体列表. 而 FAITHSCORE[60] 利用 ChatGPT[95] 首先从生成描述

Table 1 Hallucination Evaluation Benchmarks for LVLMs

表 1 LVLMs 幻觉评估基准

基准名称	任务形式	图像来源	图像数量	问答对生成方式	问题类型	问答对数量	评估幻觉类型	评估指标
CCEval[74]	描述	Visual Genome[88]	1 000	无	无	无	存在	CHAIR、Coverage、平均回答长度、回答中平均物体数量
MME[54]	问答	人工收集	1 097	人工设计	判断	2 194	存在和属性	准确率、Accuracy+
CIEM[55]	问答	任何含图片描述的数据集	由数据集决定	LLM 生成	判断	由数据集决定	存在和属性	准确率、精确率、召回率、*F*1 分数、负样本召回率
RAH-Bench[81]	问答	MSCOCO[89]	3 000	GPT-4 生成	判断	1 500	存在、属性和关系	精确率、召回率、*F*1 分数、误报率
HALLUSIONBENCH[57]	问答	人工收集	346	人工设计	判断	1 129	存在、属性和关系	总准确性、图像准确性、问答对准确性、Yes 百分比差异、误报率
AMBER[53]	描述和问答	MSCOCO[89], UnSplash[90]	1 004	人工设计	判断	15 220	存在、属性和关系	CHAIR、Cover、Hal、Cog、准确率、精确率、召回率、*F*1 分数
MMVP[16]	问答	ImageNet[91], LAION-Aesthetic[92]	300	人工设计	选择	300	存在、属性和关系	准确率
PhD[44]	问答	人工收集	7 000	GPT-3.5	判断和问答	53 796	存在、属性和关系	准确率
LLaVA-Bench[17]	描述和问答	人工收集	54	GPT-4	问答	150	存在、属性和关系	GPT-4 打分
WHOOPS[58]	描述和问答	生成图像	500	人工和自动生成技术	选择和问答	4 374	存在、属性和关系	CIDEr、BLEU-4、严格匹配、BERT 匹配
LRV-Instruction+GAVIE[48]	问答	人工收集	35 000	GPT-4	判断、选择和问答	400 000	存在、属性和关系	GPT-4 打分
MMHAL-BENCH[63]	描述和问答	OpenImages[93]	96	人工设计	问答	96	存在、属性和关系	GPT-4 打分
Bingo[59]	描述和问答	人工收集	308	人工设计	问答	370	存在、属性和关系	准确率
FAITHSCORE-Bench[60]	描述和问答	MSCOCO[89]	2 000	GPT-4	问答	3 000	存在、属性和关系	FAITHSCORE

中提取描述性子句, 随后从这些子句中提取物体、计数、颜色、关系和其他 5 个类别的原子事实.

在内容比对方面, 由于 MSCOCO 和 Visual Genome 数据集已经为每张图像注明了物体标签, 因此文献 [50−51, 74] 直接将模型输出中提取到的物体列表与这些物体标签进行对比. 文献 [52] 则通过微调 LLaMA 得到了幻觉判别模型 HaELM, 用于自动化比对生成描述和真实描述的文本内容, 并给出幻觉元素存在与否的判别结果. 而 Bingo[59] 和 LLaVA-Bench[17] 直接使用 GPT-4 将生成描述与标准答案进行比对. WHOOPS[58] 则使用 CIDEr[96] 和 BLEU-4[97] 指标来计算生成描述与参考描述的一致程度. FAITHSCORE[60] 使用视觉蕴含模型 OFA[98] 对描述性子句中的原子事实进行逐一验证.

3.1.2 视觉问答

视觉问答任务通过提问模型关于图像的特定问题, 并将模型输出与正确答案进行对比, 来判断模型是否对特定的视觉元素产生了错误感知. 相较于视觉描述, 视觉问答通过针对性问题设计实现了更为灵活的幻觉评估, 对考察的幻觉类型也更加可控. 根据提问形式, 视觉问答分为判断题、选择题和开放式问答 3 种类型. 判断类视觉问答要求模型回答“是”或“否”, 选择类视觉问答要求模型在给定选项中做出正确选择, 而开放式视觉问答则不要求模型输出特定的答案形式.

1)判断类视觉问答. 判断类视觉问答无需复杂的解析规则便可识别模型的作答对错, 避免了回复长度对评估结果的影响. POPE[47] 通过让模型判断特定物体在图像中的存在情况来实现物体存在幻觉的评估. MME[54] 通过判断题综合考察模型的感知与认知能力, 其中与物体存在、计数、颜色、位置相关的感知类题目通常被用于幻觉评估. CIEM[55] 中的判断题涵盖了物体存在和物体属性 2 类幻觉的考察. RAHBench[81] 则同时涵盖了物体存在、物体属性、物体关系 3 种幻觉类型. HallusionBench[57] 中的题目包括 2 类: 一类是必须依赖视觉信息才能回答的问题, 另一类是即使没有视觉信息也能被正确回答的问题. AMBER[53] 在全面考察物体存在、物体属性、物体关系 3 类幻觉的基础上, 进一步考虑了状态、数量和动作 3 个属性幻觉维度. 而 PhD[44] 在考察物体属性幻觉时, 涵盖了物体形状、材料、颜色、数量、位置、情绪和用途等方面, 同时设计了针对模态冲突幻觉的题目.

2)选择类视觉问答. 与判断类视觉问答相似, 选择类视觉问答亦便于识别模型的答题对错, 且支持

通过设置迷惑性选项来增加问题难度. MMVP[16] 为每对图像设计了只有 2 个选项的选择题, 重点考察 LVLMs 对相似图像中细节差异的判别能力; 而 WHOOPS[58] 则为每张图片提供了 5 个不完全描述, 要求模型选出最准确的一个, 以测试其是否能够识别图像中存在的反常识元素.

3)开放式视觉问答. 开放式问答对模型的作答形式没有特定要求, 较难判断幻觉发生与否, 但有助于真实反映模型在实际应用中的交互行为. NOPE[49] 设计了多样化的提问形式, 能够模拟实际场景中的语言变化性, 有助于考察模型在真实交互环境下的幻觉现象. PhD[44] 要求模型用几个单词回答为反常识图像设计开放式问题. LLaVA-Bench[17] 包括了对话、详细描述、复杂推理 3 类开放式问题, 并在 FAITHSCORE[60] 中得到沿用. 文献 [48] 在提问文本中添加了不存在的物体、错误的描述、常识性错误等误导性元素. MMHAL-BENCH[63] 涵盖了 8 种问题类型, 包括物体属性、不存在的物体、比较、计数、空间关系、环境、整体描述和其他; 而 Bingo[59] 所设计的问题则更关注模型是否能够辨别相似图像中的差异元素以及模型受提问文本的影响程度.

3.2 数据集构建

构建 LVLMs 的幻觉评估数据通常涉及图像和文本 2 部分. 在图像方面, 除图片本身外, 还包括图片内容的基本标注, 如图片尺寸和描述等. 有些数据集额外涵盖更细粒度的视觉信息, 如边界框坐标和框内的物体标签等. 在文本方面, 视觉描述任务通常为每张图像匹配一条描述图像的文本指令, 例如“请给出对这张图像的详细描述”. 而视觉问答任务则需要为每张图像匹配特定的问题与答案对, 问题作为指令输入模型, 而答案则用于评判对错.

3.2.1 图像与标注

LVLMs 的幻觉评估通常基于开源数据集获取图像数据, 其优势在于样本量大且成本低. 大多数开源数据集已经包含了必要的图像标注, 能够为后续评估提供便利. 而对于本身不包含标注的图像数据集, 则需要依靠人工或模型进行额外标注.

文献 [50] 使用了 MSCOCO 数据集[89], 其包含 200 000 张图像, 涵盖 80 个物体类别, 且所有物体实例都使用分割掩码进行了标注, 为图像描述任务中的物体提取和比对提供了便利. 文献 [51] 同时采用了 MSCOCO 和 NoCaps 数据集[94]. 其中 NoCaps 数据集包含 15 100 张图像, 涵盖 600 个物体类别, 可用于评估将 MSCOCO 数据集用于训练的模型. CCEval[74] 的图像数据来自 Visual Genome 数据集[88], 其包含 108 000 张图像, 每张图像都配有物体级的区域标注. POPE[47] 则从 MSCOCO 验证集中随机抽取了 500 张含有至少 3 类物体标签的图像用于幻觉评估. HaELM[52] 直接将 MSCOCO 数据集中的图像描述作为标准答案, 利用专门的幻觉识别模型来检验生成描述是否含有幻觉元素. AMBER[53] 则使用专家模型为 MSCOCO 测试集和 UnSplash 数据集[90] 中的图像进行了物体、属性、关系和容易被误识别物体的标注, 并进行了人工检查与纠正. MMVP[16] 从 ImageNet[91] 和 LAION-Aesthetic 数据集[92] 中筛选出了人类能够感知到明显差异但在 CLIP 中被相似编码的 150 对图像作为测试数据. 而 MMHAL-BENCH[63] 为了避免与主流的训练数据出现重叠, 采用了来自 OpenImages[93] 验证集和测试集的图像, 涵盖了更多的物体类别. M-HalDetect[61] 使用 InstructBLIP[6] 对 MSCOCO 验证集的部分图片生成了详细描述, 并利用人工进行子句级别的幻觉标注.

除使用开源图像数据集外, 亦有研究致力于构建专属的图像数据集, 在保证图像来源多样性的同时, 避免测试数据与模型的训练数据产生显著交集. 此外, 针对特定的幻觉产生模式, 构建专属数据集有助于测试模型在面对特定幻觉诱发因素时的鲁棒性.

MME[54] 从 MSCOCO, DeepArt[99], MovieNet[100] 等多个数据集中采样了 1 097 张图片, 确保了图像来源的多样性. 文献 [56] 利用 ViComTe 数据集[101] 中与物体颜色、形状、材料和大小相关的常识规律, 设计了多个反直觉场景描述, 并使用 DALL-E-2[102] 生成了 1 563 张含有反常识元素的图像, 用于评估模型在反直觉场景下的幻觉现象. HallusionBench[57] 则依靠人工收集了不同主题(食物、数学、几何、统计、地理、体育、卡通、错觉、电影、表情包等)和形式(徽标、海报、图形、图表、表格、地图、连环画等)的图像, 并对每张图像进行了翻转、顺序反转、掩蔽、光学字符编辑、物体编辑和颜色编辑等操作, 构成了包含 346 张图像的数据集. PhD[44] 针对物体存在、物体属性、模态冲突和反常识 4 类幻觉构建了专属图像数据集, 并通过 ChatGPT 为每张图像生成了标准化的标注内容. LLaVA-Bench[17] 从 MSCOCO 验证集等来源收集了 54 张图片, 并对 MSCOCO 数据集之外的图像手动创建了图像标注. WHOOPS[58] 雇佣专业人员利用文生图模型[102-104] 构建了 500 张反常识图像, 并利用人工对每张图片进行了完整版和残缺版的标注用于后续任务. GAVIE[48] 在 Visual Genome 数据集的基础上, 从 Vistext 数据集[105] 和 Visual News 数据集[106]

中收集了大量与图表及新闻相关的图像,以提升测试数据的多样性. Bingo[59]为了分析 LVLMs 中的潜在偏见,收集了包含不同国家和地区元素的图像,并通过人物替换等方法构建了一组反事实图像,用于检测模型是否会过度依赖先验知识.

3.2.2 指令与问答对

视觉描述任务的指令设计较为简单,通常采用统一的指令格式要求模型对给定图像进行文本描述. 而在视觉问答任务中,问题设计既要适应图像内容,又要精准反映出模型是否产生了幻觉现象,因此问答对的生成需要更为细致的考量,主要依靠人工和模型辅助 2 种方法.

MME[54]的问答数据全部由人工设计,避免了使用公共数据集中的问答对而造成的数据泄露. MMVP[16]的问答数据同样由人工设计,确保了每道题目都能有效考察模型对于图像细节差异的感知能力. MMHAL-BENCH[63]则手动设计了 8 个问题类别和 12 个物体类别,通过交叉得到了 96 组问答数据,并确保了每道问题都能够使 LLaVA 模型[17]出错,从而保证了幻觉评估的有效性.

受成本限制,人工生成的测试样本难以进行大规模拓展. 因此,有研究借助 LLMs 为图像构建问答数据. LLaVA-Bench[17], FAITHSCORE[60], GAVIE[48], NOPE[49]将图像描述、尺寸、边界框坐标和物体标注等基本信息与预先设计的问题一起输入到 GPT-4 中生成答案. PhD[44]则首先为图像提取物体及属性列表,随后将这些信息输入到 ChatGPT 中,通过指令让其生成具有高度迷惑性的问答对,保证了测试难度. 而 WHOOPS[58]先从图像描述中提取答案,随后利用问题生成模型创造与答案匹配的问题,并在最后使用 Q^2[107]模型对问题进行筛选,确保了问题和答案的逻辑一致性.

部分研究致力于构建问答对的自动化生成管线,使其能够应用于任何满足要求的图像数据集. CIEM[55]能够为任何配有文本描述的图像数据自动生成问答文本. 其利用 ChatGPT 根据图片描述和预设的提示生成与物体存在、物体属性以及物体动作有关的判断题,并支持对生成的问题进行正确性审核. POPE[47]则面向含有物体标签的图像数据,提出了一种轮询生成管线,即针对每张图像通过 3 种方式采样出易于导致模型产生存在幻觉的物体,并生成关于这些物体存在与否的问题.

3.3 评估指标

在视觉描述任务中,幻觉评估指标通常被设计用于量化生成描述与图像内容间的偏差. 而对于视觉问答任务,评估指标的设计则取决于问题类型. 在判别式问答中,答案形式固定,可以采用分类性能指标来量化模型表现. 而对于开放式问答,答案形式不受限制,评估需要依赖于人工或模型的主观评分,亦或采用复杂的规则来衡量模型表现.

3.3.1 视觉描述

CHAIR(caption hallucination assessment with image relevance)是视觉描述任务中最为常见的幻觉评估指标,用于衡量生成描述中幻觉物体的存在比例[50]. CHAIR 包括 2 种计算方式: CHAIR_i 和 CHAIR_s. CHAIR_i 计算幻觉物体数量占生成描述中物体总数的比例,用于评估模型在物体层面的幻觉程度; CHAIR_s 则计算含有幻觉物体的句子数量占描述中句子总数的比例,用于评估模型在句子层面的幻觉程度. 尽管 CHAIR 可以反映出描述文本中的幻觉比例,但其无法考虑描述的详细程度. 例如,当生成描述未能提及图像中存在的任何真实物体时, CHAIR 会显示其幻觉程度为零,但该描述无法提供任何有价值的信息. 针对这一问题, CCEval 引入了 Coverage 指标[74]. 该指标计算描述中所提及的物体与图中实际物体的匹配比例,用以反映描述的详细程度,通常作为 CHAIR 的补充指标共同使用. AMBER 在图像描述任务中提出了 Cog 指标,用于衡量模型幻觉和人类幻觉的相似程度[53]. 此外, HaELM 提出了平均幻觉比例(average hallucination ratio)指标,旨在计算含有幻觉内容的描述实例在整个测试数据集中的占比[52].

3.3.2 视觉问答

判别式视觉问答通常采用分类性能指标来衡量模型表现,包括准确率、精确率、召回率和 $F1$ 分数[16,44,47,54-55,57,81]. 准确率表示模型正确回答问题的比例;精确率衡量模型输出为“是”时的回答正确率;召回率则衡量模型在实际答案为“是”时的回答正确率; $F1$ 分数是精确率和召回率的调和平均,适用于答案类别“是”与“否”数量不平衡的情形. 部分研究为提升评估的细致程度,在上述指标的基础上设计了额外的度量方式. MME 为每张图像匹配了 2 个答案,分别为“是”和“否”的问题,并定义了 Accuracy+指标[54]. 该指标在图像层面计算答题准确率,只有当模型对图像所匹配的 2 个问题都正确作答时,才算通过该图像的测试. ROME 则定义了 CI-Obj(counter-intuitive score based on object recognition)和 CI-AtrrRel(counter-intuitive score based on attribute/relation recognition)两个指标,分别衡量模型在面对反常识图像时的物体

存在幻觉以及物体属性/关系幻觉[56]. RAH-Bench 使用误报率指标来揭示模型对误导性问题做出正常回答的频率[81]. HallusionBench 在准确率的基础上定义了“yes percentage difference”指标[57], 用于计算模型输出为“是”的问题数量与真实答案为“是”的问题数量间的差异, 从而衡量模型对于正向回答的偏好程度.

开放式视觉问答通常利用 LLMs 对模型输出进行打分. 这种评分机制涉及将一组人工评分案例和当前的测试题目以及模型的回答按照少样本提示模版输入 LLMs, 从而让其成为当前回答生成和人类偏好相一致的评分. LLaVA-Bench 使用 GPT-4, 并根据帮助性、相关性、准确度和细致度为生成答案给出 1~10 的评分[17]. GAVIE 则基于答案的准确度以及答案内容和问题的相关度, 让 GPT-4 为答案生成 0~10 的评分[48]. MMHAL-Bench 使用 GPT-4, 并根据答案的信息量和幻觉程度对生成答案进行 0~6 的评级[63]. WHOOPS 则通过计算生成答案和参考答案间的 BERT 匹配分数来衡量答案质量[58]. 除采用 LLMs 外, 有研究依靠人工对模型的输出进行评分. Bingo 利用人工对生成答案进行对错判断, 最终计算模型在整个数据集上的准确率[59]. M-HalDetect 利用人工进行单词级别的幻觉识别, 最终计算幻觉词语在描述性客观内容中的占比[61]. 除了依赖模型和人类进行评分, 亦有研究采用基于规则的方法对输出答案进行指标计算. NOPE 利用规则来判断模型生成的回答是否为负面不定代词, 从而实现 NegP 准确率的计算[49]. FAITHSCORE 则利用 ChatGPT 从生成答案中抽取原子事实, 并采用 OFA[98] 对这些事实进行逐一验证, 从而计算模型在各类原子事实上的准确率[60].

3.4 评测基准的实际应用

在为具体应用场景选择合适的幻觉评估工具时, 需要综合考虑以下几个方面:

首先, 应评估应用场景对准确性的需求. 例如, 在医疗诊断和自动驾驶等对模型可靠性要求极高的应用中, 幻觉评估工具必须具备严格的评估标准. FAITHSCORE[60] 使用视觉蕴含模型 OFA[98] 对描述性子句中的原子事实进行逐一验证, 相较于 Bingo[59] 和 LLaVA-Bench[17] 直接使用 GPT-4 将生成描述与标准答案进行比对的方法, 更加严谨可靠, 更适用于对图像感知能力要求极高的下游任务. 而在娱乐或创意内容生成的应用中, 评估工具可能要关注模型的创造性和多样性, 而不仅仅是准确性. NOPE[49] 设计了多样化的提问形式, 能够模拟实际场景中的语言变化, 有助于考察模型在真实交互环境下的多样性表现与幻觉现象之间的相互作用. 当前的许多评估工具针对这一方面的考量仍较为欠缺.

其次, 需要考虑评估工具的语言理解能力. 在聊天机器人等以开放式视觉问答为基础的应用场景中, 幻觉评估工具要能够综合理解用户意图和模型输出, 从而实现精准的幻觉评判. LLaVA-Bench[17], GAVIE[48], MMHAL-Bench[63] 均在开放式问答中使用 GPT-4, 按照预定的规则对模型回答的幻觉程度进行打分, 充分利用了 LLMs 的上下文理解与逻辑推理能力.

此外, 还需考察评估工具的通用性, 即其在不同数据集和任务设置中的适应性, 尤其是跨场景使用的灵活性. 当前, 幻觉评估工具的设计往往依赖于特定的数据集和任务, 在一定程度上限制了它们的通用性. 而 CIEM[55] 的 pipeline 设计能够为任何配有文本描述的图像数据自动生成问答文本, 并支持对生成的问题进行正确性审核, 增强了评估工具在不同场景中的适用性.

综合考虑以上因素, 可以更有效地选择适配于特定应用场景的幻觉评估工具.

4 LVLMs 的幻觉治理

LVLMs 的幻觉治理旨在通过技术手段减少或消除生成文本中与图像内容不一致的元素, 从而提高模型的安全性和可靠性. 如表 2 所示, 本节将从数据侧、视觉感知侧、训练侧、推理侧及事后修正侧 5 个维度对 LVLMs 的幻觉治理策略进行讨论与分析.

4.1 数据侧的幻觉治理

训练数据的质量将直接影响 LVLMs 的学习效果. 由于视觉指令微调数据的匮乏, 现有研究通常使用 LLMs 基于文本标注为开源数据集中的图像生成配套的指令和响应. 这种方法在有效降低数据集构建成本的同时, 存在若干局限性: 首先, 纯文本 LLMs 无法感知到像素级图像信息, 其所生成的指令与响应只能停留在全局层面, 缺乏对细粒度图像信息的覆盖; 其次, LLMs 生成的文本易于冗长, 可能引入与图像内容无关的噪声; 再者, LLMs 倾向于生成正向的指令与回复, 从而引入了统计偏差; 此外, 合成数据中的指令模版数量有限, 导致指令的多样性不足, 可能影响模型的外推能力. 因此, 提升多模态训练数据的质量是幻觉治理的一种主要手段.

4.1.1 提升指令的多样性

提升指令数据的多样性可以减轻模型在训练过程中拟合到的统计偏差, 从而减少幻觉现象的发生.

Table 2 Hallucination Management Method for LVLMs

表 2 LVLMs 幻觉的治理方法

幻觉治理层面	幻觉治理思路	幻觉治理方法	对应的幻觉成因
数据侧	提升指令数据的多样性	LRV-instruction[48]	训练数据中的统计偏差
		CIT[55]	训练数据中的统计偏差
		Consistency Training[46]	训练数据中的统计偏差
	增强图像文本关联度	VIGC[62]	训练数据中存在噪声
		LLaVA-RLHF[63]	训练数据中存在噪声
		RAI-30K[81]	训练数据中存在噪声
		shareGPT4V[78]	文本描述不够细致
	改造现有数据集	HalluciDoctor[80]	训练数据中存在噪声
视觉感知侧	集成多个视觉编码器	PVIT[108]	细粒度视觉信息感知能力不足
		Vary[15]	细粒度视觉信息感知能力不足
		Lyrics[20]	细粒度视觉信息感知能力不足
		MoF[16]	细粒度视觉信息感知能力不足
		MouSi[21]	细粒度视觉信息感知能力不足
	利用单个视觉编码器的多层级视觉特征	LION[18]	细粒度视觉信息感知能力不足
	增加视觉编码器的参数规模	InternVL[19]	细粒度视觉信息感知能力不足
训练侧	新的监督信号	ObjMLM[51]	单一的语言建模训练任务
		Mask Prediction[16]	单一的语言建模训练任务
	强化学习	LLaVA-RLHF[63]	单一的语言建模训练任务
		FDPO[61]	单一的语言建模训练任务
	增加任务信息	MiniGPT-v2[22]	训练数据中存在统计偏差
推理侧	对比解码	OPERA[83]	视觉特征得不到充分关注
		VDD[84]	LLM 的先验性知识偏差
		Pensieve[109]	训练数据中存在统计偏差
		VCD[66]	自回归解码中的偏差累积
		MARINE[23]	自回归解码中的偏差累积
		ICD[87]	自回归解码中的偏差累积
		HALC[85]	自回归解码中的偏差累积
		CGD[110]	自回归解码中的偏差累积
	思维链	GroundingCoT[24]	
		Volcano[67]	
		CCoT[111]	
事后修正侧	基于训练的方法	LURE[65]	固有的采样随机性
	无需训练的方法	Woodpecker[64]	固有的采样随机性
		LogicCheckGPT[112]	固有的采样随机性

Liu 等人[48] 构建了 LRV-Instruction 视觉指令微调数据集, 其中包含了 3 类误导性指令: 涉及虚构物体、错误属性或错误知识的指令, 用于减轻模型对于正向反馈的偏好. 此外, 为了丰富指令数据的多样性并降低响应文本的冗长, 该数据集涉及了 16 种视觉语言下游任务, 并确保了响应文本的长度不超过 30 个词. Hu 等人[55] 提出的对比指令微调方法则使用一个由 ChatGPT 驱动的管线, 为 COCO 数据集中的每张图像生成了一个正向问答对与一个负向问答对, 消除了正负样本间的统计偏差. 此外, 该数据集中的每个问答实例均配备有一条逻辑链, 用于增强模型在解决实际问题时的推理能力. Zhang 等人[46] 则构建了一个针对数量幻觉的指令微调数据集, 其中每张图像配有 2 种形式的问答对, 均涉及对同一物体的计数, 但表述方式各异, 且涉及多个物体的数量比较. 该做法旨在增强模型在面对不同指令类型时的数量感知一致性, 进而减缓模型的数量幻觉.

4.1.2 增强文本与视觉事实的关联度

增强训练数据中文本元素(指令、响应或描述)与图像内容的关联度能够提升 LVLMs 对视觉信息的精准解读, 从而减轻幻觉现象. Wang 等人[62] 设计的 VIGC 数据生成管线直接基于 LVLMs 进行指令与响应的生成, 并采用迭代更新机制来纠正生成数据中的错误内容. 由于 LVLMs 具备像素级的图像感知能力, 其生成的文本数据和图像内容的相关性更强, 能够有效降低幻觉风险. Sun 等人[63] 则将 VQA-v2[113] 和 A-OKVQA[114] 数据集中的视觉问答样本转换成了多轮对话, 同时将 Flickr30K[115] 数据集中的图像描述样本转换为了局部问答. 通过将这些人工标注数据与 GPT-4 合成数据[17] 进行整合, 形成了更高质量的视觉指令微调数据集. Chen 等人[81] 从 PSG 数据集[116] 出发, 利用全局场景图为每张图像生成实体间的关系描述. 通过将这些关系描述与图像的基本标注结合后输入 GPT-4, 能够得到与图像内容更相符的指令和响应数据. Chen 等人[78] 则利用具备视觉感知能力的 GPT-4V 模型, 构建了用于视觉指令微调的 ShareGPT4V-SFT 数据集以及用于多模态对齐预训练的 ShareGPT4V-PT 数据集. ShareGPT4V-SFT 包含了 10 万张来源多样化的图像以及由 GPT-4V 生成的细粒度描述文本, 文本内容涵盖了世界知识、物体属性、空间关系、美感评价等多个维度, 平均长度超过 900 个单词. 为了以成本可控的方式进行扩展, 在其之上微调了 1 个图像描述模型, 并使用该模型为额外的 120 万张图像生成了高质量文本描述, 形成了 ShareGPT4V-PT 数据集.

4.1.3 改造现有数据集

与构造新数据的思路不同, Yu 等人[80] 提出了 HalluciDoctor, 一种用面向现有数据集的低成本幻觉检测和消除框架. 该框架能够自动识别图像文本数

据中的幻觉元素，并在不破坏整体语义的情况下将其消除. 通过将此技术应用于 LLaVA-instruct 数据集[17]，创建了名为 LLaVA+的高可靠性视觉微调数据集. 此外，为了缓解由共现偏差引起的虚假相关性问题，在 LLaVA+中集成了部分反事实干预样本，即创造了一些最不可能共现的物体共同出现的图像，用于平衡数据中的统计偏差.

4.2 视觉感知侧的幻觉治理

LVLMs 的视觉感知能力通常来自于一个经过图像文本对比预训练的图像编码器. 这种编码器面临 2 点局限性：首先，受预训练数据中文本标注的质量影响，其在捕捉细粒度视觉信息方面能力不足，通常只能提供全局层面的语义信息. 其次，这种视觉编码器在预训练时的对齐目标通常是参数量较小的语言模型，而在 LVLMs 中，视觉特征需要服务于参数量在 7×10^9 及以上的 LLMs. 这种规模差异加大了视觉特征与 LLMs 嵌入空间的对齐缺口，增加了视觉语言适配器的对齐难度. 因此，改进模型的视觉感知模块是治理幻觉现象的另一重要思路.

4.2.1 集成多个视觉编码器

集成多个视觉编码器旨在通过不同编码器的互补特性来提高 LVLMs 的视觉感知能力. Chen 等人[108]在 LLaVA 架构上集成了 CLIP-ViT 和 RegionCLIP-ResNet[117]两个图像编码器. 前者负责提取全局视觉特征，而后者则根据用户指定的边界框区域提取局部视觉特征. 通过在含有边界框标注的多模态数据上进行训练，该架构提升了模型对图像局部信息的处理能力. Wei 等人[15]引入了对细粒度图像信息更敏感的视觉编码器 SAM-ViTDet[118]. 通过将其与轻量级语言模型 OPT-125M[119]相连，并在包含图表和文档的多模态数据上进行文本生成训练后，该编码器能够有效提取图像中的细粒度语义信息. 将其与 CLIP-ViT 协同使用能够有效降低模型在处理光学字符识别任务时的幻觉现象. Lu 等人[20]在 BLIP2 架构上增设了视觉细化模块并改造了 Q-Former 模块. 视觉细化模块利用 3 个视觉专家模型提取物体层面的局部特征. 改造后的 Q-Former 模块包含 2 组可学习向量，分别用于提取 CLIP-ViT 输出的全局特征和视觉细化模块输出的区域特征. 区域级图像信息的加入增强了模型的细粒度视觉感知能力. Tong 等人[16]发现 CLIP-ViT 可能会将 2 张具有明显细节差异的图像进行相似编码，因此提出了混合特征建模方法：同时集成一个基于图像文本对比学习的视觉编码器和一个基于纯图像自监督学习的视觉编码器，将两者输出的图像特征混合后一同输入 LLM 模块. 通过将 DINOv2-ViT[120]与众多采用 CLIP-ViT 的 LVLMs 进行集成，幻觉现象得到了有效缓解. Fan 等人[21]则探索了更多种来自不同预训练任务的视觉编码器，包括 CLIP, DINOv3[120], LayoutLMv3[121], ConvNext[122], SAM[118], MAE[11]，并指明了不同集成数量（1, 2 和 3 个）下的最优组合. 实验表明，随着视觉编码器数量的增加，LVLMs 的幻觉程度持续降低，验证了多视觉编码器结合带来的互补优势.

4.2.2 利用单个视觉编码器的多层级特征

利用单一视觉编码器的多层级特征同样可以增强 LVLMs 的视觉感知能力. 以 CLIP-ViT 为例，其在浅层级中主要编码图像的基本元素如边缘、颜色和纹理等，而在深层级则会抽象出场景或语义信息等高级视觉特征. 受此启发，Chen 等人[18]开发了 LION 模型. 该模型通过一个视觉聚合网络整合来自 CLIP-ViT 不同层级的视觉特征，并将其一同映射到 LLM 模块的嵌入空间. 同时，LION 集成了一个 RAM 模型[123]，用于从图像中直接提取物体标签，并通过提示模版将标签信息传入 LLM 模块. 这一系列措施有效减少了模型在处理细粒度视觉信息任务时的幻觉现象.

4.2.3 增加视觉模块的参数规模

LVLMs 中视觉模块和文本生成模块之间普遍存在参数量失衡问题. LLM 组件的参数量通常达 7×10^9 及以上，而视觉编码器的参数量通常不足 1×10^9. 此外，主流 LVLMs 中视觉语言适配器的结构简单且参数规模不足. 针对这些问题，Chen 等人[19]发开了 InternVL 模型. 该模型的视觉编码器采用了具有 6×10^9 参数量的 InternViT 模型，视觉语言适配器则由一个集成了交叉注意力模块和 64 个可学习查询向量的 LLaMA-7B 组成. 更大规模的视觉感知与对齐模块使得 InternVL 在当时的 POPE 幻觉评估基准[47]上拔得头筹.

4.3 训练侧的幻觉治理

无论是在模态对齐预训练阶段还是在视觉指令微调阶段，主流 LVLMs 仅采用自回归文本生成任务. 这种单一的训练任务会加大模型对于训练数据中文本质量的依赖，且会导致模型过分强调生成内容的文本质量，而忽视文本内容与图像的一致性，进而增加幻觉风险. 因此，为 LVLMs 设计新的训练任务亦是减轻幻觉现象的有效手段.

4.3.1 引入新的监督信号

引入新的监督信号旨在通过特定训练任务使模型更加关注图像中的实际信息，减少无视觉支持的错误感知与推断. Dai 等人[51]提出了物体掩码语言建

模, 通过掩码图像文本对中与物体相关的词汇, 要求模型根据剩余文本和图像信息预测被掩码的文本token, 增强了模型对图像中物体的感知能力, 有效减少了物体存在幻觉的发生. Chen 等人[81]则在 RAI-30K 数据集中的每个文本末尾加入了 2 个特殊 token, 分别代表物体关系中的主语和宾语. 这 2 个 token 在 LLM 模块中的最终表征经过一次线性变换后, 与图像一同输入 SAM 模型得到主语和宾语的分割掩码, 并利用掩码预测损失来优化 LVLM 中植入的 LoRA 模块[124], 增强了模型对图像中物体及物体间关系的感知精度.

4.3.2 应用强化学习

Sun 等人[63]将 LLMs 领域的人类反馈强化学习算法[6]应用到了 LVLMs 中, 开发了名为 Fact-RLHF 的训练方法. 该方法要求人类标注者针对相同图像和指令比较 2 个响应的幻觉程度, 以此训练一个 LVLM 来模仿人类对幻觉的偏好. 通过将其用作奖励模型, 可以对已经过预训练和指令微调的 LVLMs 应用强化学习, 鼓励其生成幻觉含量更低的响应内容. 受到 LLMs 中直接偏好优化算法[125]的启发, Gunjal 等人[61]提出了适用于 LVLMs 的细粒度偏好直接优化算法. 与人类反馈强化学习不同, 该算法利用来自单个示例的细粒度标注信息直接优化模型参数, 无需训练奖励模型即可降低幻觉的发生率.

4.3.3 增加任务信息

为解决 LVLMs 在不同任务场景下由于指令混淆引发的幻觉问题, Chen 等人[22]对常见的视觉语言下游任务进行了分类, 并为每种类型设计了专用的任务 token. 在指令微调阶段, 通过分析每个样本的指令内容确定其任务类别, 然后将相应的任务 token 与指令文本结合后输入模型, 可以有效减少因指令混淆而产生的幻觉现象.

4.4 推理侧的幻觉治理

无论是提升训练数据的质量、使用更复杂的视觉模块, 或设计更先进的训练任务, 均可有效减缓 LVLMs 的幻觉现象. 然而, 这些方法大多会涉及模型训练. 随着模型参数量的不断增大, 这些幻觉治理策略对计算资源的需求也随之增加. 因此, 许多研究开始探索在模型的推理阶段实施幻觉治理, 目的是降低幻觉治理的计算成本.

4.4.1 对比解码

对比解码[126]是一种在模型解码阶段引入对比和选择机制以优化模型输出的技术. 在 LVLMs 的推理阶段, 可以通过变量引入为每一步文本解码生成多个概率分布, 并通过对比策略形成新的概率分布, 有效增加(降低)与图像内容一致(不符)的文本 token 的采样概率, 从而确保输出文本与图像内容的相关性和一致性以减轻幻觉现象.

Leng 等人[66]观察到幻觉的发生率与输入图像中的噪声含量成正比, 因此提出了视觉对比解码技术. 该技术通过将原始图像和加噪后的图像结合为同一文本指令分别输入模型, 得到 2 组概率分布. 通过对比两者, 可以创建出一个更贴近图像实际内容的概率分布. 从该分布进行文本采样可以有效降低幻觉内容的生成. Huang 等人[83]在分析注意力图时发现, 幻觉内容通常出现在锚点 token 之后. 这些 token 本身不携带明显的语义信息, 但在解码过程中获得过度关注, 导致模型忽视了视觉 token. 针对此问题提出了 OPERA 解码机制. 该机制在波束搜索中比较不同波束的锚点 token 数量, 并对包含过多锚点 token 的波束进行惩罚, 同时结合一种回顾策略, 对生成内容中的锚点 token 进行复审和必要的重新采样. Zhao 等人[23]提出的 MARINE 框架为每张输入图像提供 2 组视觉特征, 一组特征由模型内置的视觉编码器生成, 而另一组则由内置编码器和在目标检测任务中预训练过的外部编码器联合提供. 通过将这 2 组特征配合文本指令分别输入 LLM 模块, 可以得到 2 组概率分布. 通过比较两者, 便能构建一个更接近实际图像信息的采样分布. Wang 等人[87]提出了通过角色扮演前缀引入指令干扰的概念, 并观察到负向干扰会加剧以 Q-Former 为适配器的模型的幻觉现象. 为此, 提出了指令对比解码, 即通过对比原始指令与负向干扰指令下的 2 组概率分布, 抵消模态对齐不确定性导致的幻觉现象. 针对由先验性知识偏差引发的幻觉现象, Zhang 等人[84]提出了视觉去偏差解码机制, 其使用原始图像和无信息图像(如全黑、全白或全噪声图)分别得到 2 组概率分布. 对比两者可以创造出一个受先验性知识偏差影响更小的采样分布. Yang 等人[109]观察到, 模型在面对具有相似语义和外观特征的图像时会展现出现相似的幻觉模式, 因此提出了 Pensieve 解码机制. 该机制在解码过程中引入了相似图像作为参考, 通过回顾和比较这些图像产生的概率分布, 有效避免了在处理类似图像特征时的常见幻觉. 针对物体存在幻觉, Chen 等人[85]提出了 HALC 解码框架. 该框架集成了自动聚焦定位机制, 根据图像的局部信息实时纠正幻觉 token, 并使用特殊的波束搜索算法来保证文本内容和图像全局信息的一致性, 在减少物体幻觉的同时保证了生成文本的质量. Deng

等人[110]发现，句子层面的幻觉现象与句子和图像的CLIP分数存在高度相关性. 受此启发，发明了CLIP引导解码机制，即在句子层面将CLIP分数融入波束搜索的打分机制，通过优先采样与图像内容关联度更高的文本，有效减少幻觉内容的生成.

4.4.2 指令引导

相较于对比解码，指令引导技术通过设计特定的指令内容引导LVLMs生成真实度更高的文本内容，是一种更直观的幻觉治理方法. Chen等人[24]提出了名为GroundingCoT的思维链技术，其在引导模型生成普通思维链的同时，要求模型为涉及的每个视觉物体生成中心点坐标，有效减少了物体存在幻觉的发生，但仅适用于具备视觉定位能力的模型. Lee等人[67]提出了名为Volcano的自反馈引导机制，即根据输入图像对模型的初步响应生成自然语言反馈，并利用反馈信息引导模型修正其初步响应，从而消除其中的幻觉元素. 针对物体属性和物体关系幻觉，Mitra等人[111]提出了组合思维链技术，即通过特殊的提示指令，引导模型生成与输入图像相关的场景图，然后将场景图与用户输入的文本指令结合，帮助模型输出更为准确的文本响应.

4.5 事后修正的幻觉治理

生成模型中的采样随机性使得单纯从数据或模型的优化来彻底消除LVLMs的幻觉是较为困难的. 因此，部分研究提出了基于事后修正的幻觉治理策略，即利用专家模型来检测并修正模型的原始输出，从外部形成一个有效的幻觉过滤机制.

4.5.1 需要训练的方法

Zhou等人[65]指出LVLMs中的幻觉现象主要与3个因素有关：1)幻觉容易发生在与图中真实物体共现频率高的虚构物体上；2)文本解码过程中概率分布的信息熵越高，生成的token越容易包含幻觉内容；3)幻觉内容通常出现在生成文本的靠后部分. 基于这些发现，他们使用GPT-3.5[95]构建了一个包含幻觉内容的图像描述数据集. 该数据集为每张图像匹配了一段准确的文本描述以及一段符合上述3种幻觉产生模式的错误描述. 该数据集被用于微调一个预训练过的LVLM，以教会模型根据真实图像纠正幻觉描述，从而在图像描述任务中起到幻觉事后修正的作用.

4.5.2 无需训练的方法

Yin等人[64]提出了名为Woodpecker的幻觉事后修正方法，通过5个步骤来纠正模型原始输出中的幻觉元素：1)使用GPT-3.5提取原始输出中的关键物体；2)提示GPT-3.5为关键物体提出一系列关于物体本身或其属性的问题；3)利用多个视觉专家模型回答前一步提出的问题；4)基于步骤1)和2)得到的问题和答案编制一个特定格式的视觉知识文档，用于记录关于物体和属性的各种陈述；5)将模型生成的原始文本与第4步得到的视觉知识文档一同输入GPT-3.5得到修正后的文本内容. 这种修正框架的步骤清晰且每一步都有文本支撑，因此与Zhou等人[65]的方法相比，可解释性更高. 此外，Woodpecker不涉及模型训练，具有计算成本低的特点. Wu等人[112]观察到，LVLMs对真实存在的物体通常能做出逻辑一致的响应，而对幻觉物体的响应往往逻辑不一致. 基于该发现，提出了名为LogicCheckGPT的幻觉修正框架，涉及5个步骤：1)使用ChatGPT提取原始输出中的物体名称；2)向模型询问每个物体的详细属性；3)询问模型哪个物体拥有之前答案中提到的属性；4)检查从物体到属性和从属性到物体的逻辑关系是否能形成闭环；5)如果闭环数量与问题总数的比例未达预设阈值，则对相关物体进行幻觉修正.

4.6 不同方法的横向比较

数据侧幻觉治理通过扩展或改进训练数据集来提升多模态训练数据的质量和覆盖度，可以从源头降低幻觉的发生率，但通常需要大量的人力资源和时间投入. 视觉感知侧幻觉治理通过集成多个具有互补性的视觉编码器或利用单个视觉编码器的多层级特征来改善模型对图像细节的理解能力. 这类方法能够提供更丰富的视觉信息，但会增加模型架构的复杂度和计算需求. 训练侧幻觉治理通过调整模型的训练策略来降低幻觉的发生率. 这类方法虽能有效调整模型在训练过程中的行为，但需要额外的训练步骤和计算资源. 推理侧幻觉治理提供了一种在模型部署后仍能进行幻觉修正的技术，如对比解码和指令引导. 这类方法可以在不涉及模型训练的情况下减少幻觉现象，但依赖于额外的推理步骤或复杂的解码规则，可能会增加系统整体的推理时耗. 事后修正类治理方法通过专家模型检测并纠正原始生成内容中的错误，以提高模型的可靠性. 这类方法能够形成外部安全网，可以与其他治理策略结合使用，但通常需要额外的计算资源和复杂的后处理步骤，且普遍依赖于商业化LLMs[79,95]的能力. 综上所述，幻觉治理策略各有优劣，选择合适的方法需要根据实际应用场景、资源可用性和模型的具体需求来决定. 必要时可以将不同的方法组合，形成一套全面且灵活的策略体系.

5 LVLMs 幻觉研究的未来方向

LVLMs 的幻觉研究仍处于初级阶段,相关方法和技术仍有较大的提升空间.本节将从幻觉成因、评估和治理 3 方面探讨未来的研究趋势.

5.1 LVLMs 的幻觉成因

1)建立幻觉的理论框架与数学模型.当前对 LVLMs 幻觉成因的分析多集中于经验性研究,缺乏系统的理论模型来定义幻觉的发生并解释这种现象的内在原因.未来研究可以尝试构建数学模型,量化文本输出与图像内容和文本指令之间的信息偏差.这些模型将有助于预测何时、何种条件下幻觉容易发生,为更有效的幻觉治理方法提供理论指导.

2)多模态信息融合的可解释性研究.未来研究可以探索 LLM 模块中视觉和语言模态间信息融合的内部机制,通过可解释性方法揭示模型在不同生成阶段如何处理视觉 token 和文本 token,理解模型为何会忽略关键的视觉信息并产生幻觉.此外,还可以建立注意力分布的理论模型,系统性量化不同模态之间的注意力权重变化,分析注意力分布不均如何导致幻觉的产生.

3)系统性分类与交互关系研究.目前,针对 LVLMs 幻觉成因的归纳缺乏有理论支撑的分类框架.尽管文本将幻觉成因分为三大类,但这些类别之间并非完全独立,而是存在复杂的交互关系.未来研究应构建更为科学的幻觉成因分类体系,进一步明确不同成因的特点和边界,并尝试利用理论框架来量化不同成因之间的相互作用.

5.2 LVLMs 的幻觉评估

1)降低幻觉评估成本.幻觉评估基准的构建通常需要大量的人力成本.尽管指令和问答数据已经能够依靠模型自动生成,但仍然需要人工对生成内容进行审查和筛选以保证其质量;同时,部分评估基准依赖于商业化的 LLMs 生成标注、问答对或让其对 LVLMs 的输出进行打分,会造成较大的成本支出.未来研究可以探索如何以更低的成本完成评测数据集的构建以及对模型输出的打分.

2)拓宽幻觉评估维度.目前的幻觉评估标准及其量化指标主要针对物体存在幻觉,而对物体属性和关系幻觉的评估尚显不足.虽然有部分研究考虑到了属性幻觉,但其关注点主要集中在物体的数量、颜色、形状等方面,缺乏对物体属性的广泛考察,且对于属性的考察方式通常局限于视觉问答这一种任务形式.此外,偏小众的幻觉类别,如模态冲突幻觉和反常识幻觉,尚未被广泛包含在评估范围中.因此,未来研究可以探索更全面的幻觉评估基准,以便对 LVLMs 的幻觉现象进行更细致的分析和评估.

3)更准确地提取物体.在基于图像描述的幻觉评估中,准确提取 LVLMs 输出中的物体至关重要.当前的物体提取方法可能会误将形容词作为名词处理,例如,“orange”在某些语境下应表示颜色,却被错误地识别为“橙子”.此外,物体的多样化表述方式也会导致匹配问题,比如,当模型输出的物体名称与标注数据中的表述不一致时,可能会被错误地视为幻觉元素.为提高评估的准确度,未来研究可以探索更高效的算法,更精准地实现物体提取与比对,确保评估结果的有效性.

5.3 LVLMs 的幻觉治理

1)动态幻觉修正机制.当前的幻觉治理策略无法灵活应对多变的上下文环境.不同的使用情景对文本的需求各异.对于某些需求,文本的创新性和多样性更受重视,而有些情况则更注重生成文本与图像信息的一致.未来研究可以专注于开发能够动态调整的幻觉治理技术,以适应不断变化的需求场景,在保证文本质量的同时减少幻觉现象.研究可以集中在基于上下文的自适应解码策略,即根据当前交互的特定上下文实时调整模型的生成过程或外部专家的介入程度.例如,通过实时分析交互内容和生成过程中的异常模式,动态修改模型的解码参数或策略设置,以优化输出的相关性和准确性并满足不同场景下的特定需求.

2)模型内部机制透明化.目前 LVLMs 在内部决策过程的可视化和解释性方面存在不足,使得研究人员难以理解模型产生幻觉的确切原因和逻辑.缺乏有效的透明化手段可能导致治理策略难以针对具体问题进行优化.未来研究可以探索利用新的算法和工具,如可解释的人工智能技术[127],来揭示和解释模型在特定情况下产生幻觉的内部逻辑和决策因素.例如,开发可视化工具来展示模型在处理输入数据时的注意力分布和激活模式,以及利用可解释性分析方法[128]来辨识哪些内部特征或路径可能导致错误的输出.

3)多策略幻觉治理框架.现有的幻觉治理研究通常聚焦在单一的策略或技术上,未能形成一个综合的多层次治理体系,导致幻觉治理的效果有限.未

来研究可以考虑结合多种幻觉治理策略(如数据侧干预、训练策略调整、推理时的检测与修正等)，开发一个综合的、多层次的幻觉治理框架，并在不同的操作层次上动态选择最合适的策略组合，从而全面应对复杂和多变的幻觉现象.

4)设计新的模型架构. 当前的LVLMs架构呈现出严重的同质化趋势. 在主流架构中，LLM模块占据着主导地位，导致与视觉相关的推理和决策过程都发生在语义空间. 然而，将连续、高密度的视觉信息转换成语义空间中的token序列可能会导致信息损失，从而影响模型理解视觉输入中的细微差异和复杂信息. 未来研究可以探索如何在不牺牲视觉细节的前提下，整合视觉和语言模块，以便更准确地解析和利用视觉输入. 例如，开发新的视觉感知模块，或是设计新的图像文本融合方法，以提升模型在处理复杂视觉场景时的准确度和鲁棒性.

6 总结与局限性分析

通过整合视觉编码器和视觉语言适配器，LVLMs赋予了传统LLMs处理图像数据的能力，使其能够应对复杂的多模态下游任务. 然而，幻觉现象的存在严重限制了这类模型在实际应用中的效果和可靠性. 本文作为一篇关于LVLMs幻觉的研究综述，系统性介绍了该领域的研究方向和最新进展：首先阐述了LVLMs的基础架构，并详细讨论了幻觉的定义、分类，同时与LLMs中的幻觉现象进行了比较；其次，深入探讨了LVLMs幻觉的多元成因；接着，分类讨论了主流的幻觉评估方法和治理技术；最后，为该领域提出了未来的研究方向. 本综述力求全面，但仍存在些许局限：LVLMs的幻觉研究是一个快速发展的新兴领域，本文可能未涵盖最新的研究成果. 此外，本文尝试以系统的方式呈现该领域的研究进展，但对于幻觉成因、评估方法和治理手段的分类可能存在不完整或不足之处，需要未来的研究进一步完善和优化.

作者贡献声明：李煦负责论文大纲制定、文献调研、论文主要部分的撰写；朱睿参与大纲制定、文献调研和论文撰写，并负责图表制作和格式校对；陈小磊、伍瑾轩、郑毅负责文本校对及部分图表制作；赖承杭、梁宇轩负责文本润色；李斌、薛向阳提供指导意见.

参 考 文 献

[1] Yin Shukang, Fu Chaoyou, Zhao Sirui, et al. A survey on multimodal large language models[J]. arXiv preprint, arXiv: 2306. 13549, 2023

[2] Touvron H, Lavril T, Lzacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302. 13971, 2023

[3] Touvron H, Martin L, Stone K, et al. LLaMA-2: Open foundation and fine-tuned chat models[J]. arXiv preprint, arXiv: 2307. 09288, 2023

[4] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[EB/OL].[2024-05-16]. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[5] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[EB/OL].[2024-05-16]. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[6] Ouyang Long, Wu J, Jiang Xu, et al. Training language models to follow instructions with human feedback[C]//Proc of the 36th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2022: 27730−27744

[7] Meta. Introducing Meta LLaMA 3: The most capable openly available LLM to date[EB/OL].[2024-05-16]. https://ai.meta.com/blog/meta-llama-3/

[8] Shu Wentao, Li Ruixiao, Sun Tianxiang, et al. Large language models: Principles, implementation, and progress[J]. Journal of Computer Research and Development, 2024, 61(2): 351−361 (in Chinese)
(舒文韬, 李睿潇, 孙天祥, 等. 大型语言模型：原理、实现与发展[J]. 计算机研究与发展, 2024, 61(2)：351−361)

[9] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2021: 8748−8763

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C/OL].[2025-04-04]. https://openreview.net/forum?id=YicbFdNTTy

[11] He Kaiming, Chen Xinlei, Xie Saining, et al. Masked autoencoders are scalable vision learners[C]//Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 16000−16009

[12] Fang Yuxin, Wang Wen, Xie Binhui, et al. Eva: Exploring the limits of masked visual representation learning at scale[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 19358−19369

[13] Fang Yuxin, Sun Quan, Wang Xinggang, et al. Eva-02: A visual representation for neon genesis[J]. Image Vision Computing, 2024, 149(C): 1−12

[14] Sun Quan, Fang Yuxin, Wu L, et al. EVA-CLIP: Improved training techniques for clip at scale[J]. arXiv preprint, arXiv: 2303. 15389,

2023

[15] Wei Haoran, Kong Lingyu, Chen Jinyue, et al. Vary: Scaling up the vision vocabulary for large vision-language models[C]//Proc of the 17th European Conf on Computer Vision. Berlin: Springer, 2024: 1−18

[16] Tong Shengbang, Liu Zhuang, Zhai Yuexiang, et al. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 9568−9578

[17] Liu Haotian, Li Chunyuan, Wu Qingyang, et al. Visual instruction tuning[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 34896−34916

[18] Chen Gongwei, Shen Leyang, Shao Rui, et al. LION: Empowering multimodal large language model with dual-level visual knowledge[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 26530−26540

[19] Chen Zhe, Wu Jiannan, Wang Wenhai, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 24185−24198

[20] Lu Junyu, Gan Ruyi, Zhang Dixiang, et al. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects[J]. arXiv preprint, arXiv: 2312. 05278, 2023

[21] Fan Xianran, Ji Tao, Jiang Changhao, et al. MouSi: Poly-visual-expert vision-language models[J]. arXiv preprint, arXiv: 2401. 17221, 2024

[22] Chen Jun, Zhu Deyao, Shen Xiaoqian, et al. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning[J]. arXiv preprint, arXiv: 2310. 09478, 2023

[23] Zhao Linxi, Deng Yihe, Zhang Weitong, et al. Mitigating object hallucination in large vision-language models via classifier-free guidance[J]. arXiv preprint, arXiv: 2402. 08680, 2024

[24] Chen Keqin, Zhang Zhao, Zeng Weili, et al. Shikra: Unleashing multimodal LLM's referential dialogue magic[J]. arXiv preprint, arXiv: 2306. 15195, 2023

[25] Li Junnan, Li Dongxu, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proc of the 40th Int Conf on Machine Learning. New York: PMLR, 2023: 19730−1974

[26] Zhang Renrui, Han Jiaming, Liu C, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention[J]. arXiv preprint, arXiv: 2303. 16199, 2023

[27] Zhu Deyao, Chen Jun, Shen Xiaoqian, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint, arXiv: 2304. 10592, 2023

[28] Gao Peng, Han Jiaming, Zhang Renrui, et al. LLaMA-Adapter v2: Parameter-efficient visual instruction model[J]. arXiv preprint, arXiv: 2304. 15010, 2023

[29] Dai Wenliang, Li Junnan, Li Dongxu, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 49250−49267

[30] Luo Gen, Zhou Yiyi, Ren Tianhe, et al. Cheap and quick: Efficient vision-language instruction tuning for large language models[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 29615−29627

[31] Zeng Yan, Zhang Hanbo, Zheng Jiani, et al. What matters in training a GPT4-style language model with multimodal inputs?[J]. arXiv preprint, arXiv: 2307. 02469, 2023

[32] Bai Jinze, Bai Shuai, Yang Shusheng, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond[J]. arXiv preprint, arXiv: 2308. 12966, 2023

[33] Liu Haotian, Li Chunyuan, Li Yuheng, et al. Improved baselines with visual instruction tuning[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 26286−26296

[34] Wang Weihan, Lv Qingsong, Yu Wenmeng, et al. CogVLM: Visual expert for pretrained language models[J]. arXiv preprint, arXiv: 2311. 03079, 2023

[35] Ye Qinghao, Xu Haiyang, Ye Jiabo, et al. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration[C]//Proc of the 36th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 13040−13051

[36] Liu Haotian, Li Chunyuan, Li Yuheng, et al. LLaVA-next: Improved reasoning, OCR, and world knowledge[EB/OL].[2024-05-16]. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[37] Liu Haotian, Xue Wenyuan, Chen Yifei, et al. A survey on hallucination in large vision-language models[J]. arXiv preprint, arXiv: 2402. 00253, 2024

[38] Li Chunyuan, Wong C, Zhang Sheng, et al. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 28541−28564

[39] Bai Fan, Du Yuxin, Huang Tiejun, et al. M3D: Advancing 3D medical image analysis with multi-modal large language models[J]. arXiv preprint, arXiv: 2404. 00578, 2024

[40] Tian Xiaoyu, Gu Junru, Li Bailin, et al. DriveVLM: The convergence of autonomous driving and large vision-language models[J]. arXiv preprint arXiv: 2402. 12289, 2024

[41] Qian Tianwen, Chen J, Zhuo Linhai, et al. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 4542−4550

[42] Dahl M, Magesh V, Suzgun M, et al. Large legal fictions: Profiling legal hallucinations in large language lodels[J]. Journal of Legal Analysis, 2024, 16(1): 64−93

[43] Andersland M. Amharic LLaMA and LLaVA: Multimodal LLMs for low resource languages[J]. arXiv preprint, arXiv: 2403.06354, 2024

[44] Liu Jiazhen, Fu Yuhan, Xie Ruobing, et al. PhD: A prompted visual hallucination evaluation dataset[J]. arXiv preprint, arXiv: 2403. 11116, 2024

[45] Jiang Chaoya, Ye Wei, Dong Mengfan, et al. Hal-Eval: A universal

and fine-grained hallucination evaluation framework for large vision language models[J]. arXiv preprint, arXiv: 2402. 15721, 2024

[46] Zhang Huixuan, Zhang Junzhe, Wan Xiaojun. Evaluating and mitigating number hallucinations in large vision-language models: A consistency perspective[J]. arXiv preprint, arXiv: 2403. 01373, 2024

[47] Li Yifan, Du Yifan, Zhou Kun, et al. Evaluating object hallucination in large vision-language models[C]//Proc of the 28th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 292−305

[48] Liu Fuxiao, Lin K, Li Linjie, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning[C/OL].[2025-04-04]. https://openreview.net/forum?id=J44HfH4JCg

[49] Lovenia H, Dai Wenliang, Cahyawijaya S, et al. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models[J]. arXiv preprint, arXiv: 2310. 05338, 2023

[50] Rohrbach A, Hendricks L A, Burns K, et al. Object hallucination in image captioning[C]//Proc of the 23rd Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2018: 4035−4045

[51] Dai Wenliang, Liu Zihan, Ji Ziwei, et al. Plausible may not be faithful: Probing object hallucination in vision-language pre-training[C]//Proc of the 17th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 2136−2148

[52] Wang Junyang, Zhou Yiyang, Xu Guohai, et al. Evaluation and analysis of hallucination in large vision-language models[J]. arXiv preprint, arXiv: 2308. 15126, 2023

[53] Wang Junyang, Wang Yuhang, Xu Guohai, et al. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation[J]. arXiv preprint, arXiv: 2311. 07397, 2023

[54] Fu Chaoyou, Chen Peixian, Shen Yunhang, et al. MME: A comprehensive evaluation benchmark for multimodal large language models[J]. arXiv preprint, arXiv: 2306. 13394, 2023

[55] Hu Hongyu, Zhang Jiyuan, Zhao Minyi, et al. CIEM: Contrastive instruction evaluation method for better instruction tuning[C]//Proc of the 35th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 1−11

[56] Zhou Kankan, Lai E, Yeong W B A, et al. ROME: Evaluating pre-trained vision-language models on reasoning beyond visual common sense[C]//Proc of the 28th Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, PA: ACL, 2023: 10185−10197

[57] Guan Tianrui, Liu Fuxiao, Wu Xiyang, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 14375−14385

[58] Bitton-Guetta N, Bitton Y, Hessel J, et al. Breaking common sense: WHOOPS! A vision-and-language benchmark of synthetic and compositional images[C]//Proc of the 17th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 2616−2627

[59] Cui Chenhang, Zhou Yiyang, Yang Xinyu, et al. Holistic analysis of hallucination in GPT−4v (ision): Bias and interference challenges[J]. arXiv preprint, arXiv: 2311. 03287, 2023

[60] Jing Liqiang, Li Ruosen, Chen Yunmo, et al. FAITHSCORE: Evaluating hallucinations in large vision-language models[J]. arXiv preprint, arXiv: 2311. 01477, 2023

[61] Gunjal A, Yin Jihan, Bas E. Detecting and preventing hallucinations in large vision language models[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 18135−18143

[62] Wang Bin, Wu Fan, Han Xiao, et al. VIGC: Visual instruction generation and correction[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 5309−5317

[63] Sun Zhiqing, Shen Sheng, Cao Shengcao, et al. Aligning large multimodal models with factually augmented RLHF[C]//Proc of the 18th Findings of the Association for Computational Linguistics: ACL 2024. Stroudsburg, PA: ACL, 2024: 13088−13110

[64] Yin Shukang, Fu Chaoyou, Zhao Sirui, et al. Woodpecker: Hallucination correction for multimodal large language models[J]. arXiv preprint, arXiv: 2310. 16045, 2023

[65] Zhou Yiyang, Cui Chenhang, Yoon J, et al. Analyzing and mitigating object hallucination in large vision-language models[J]. arXiv preprint, arXiv: 2310. 00754, 2023

[66] Leng Sicong, Zhang Hang, Chen Guanzheng, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13872−13882

[67] Lee S, Park S H, Jo Y, et al. Volcano: Mitigating multimodal hallucination through self-feedback guided revision[C]//Proc of the 18th Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2024: 391−404

[68] Chen Xi, Djolonga J, Padlewski P, et al. Pali-x: On scaling up a multilingual vision and language model[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 14432−14444

[69] Alayrac J-B, Donahue J, Luc P, et al. Flamingo: A visual language model for few-shot learning[C]//Proc of the 36th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2022: 23716−23736

[70] Chen Xi, Wang Xiao, Changpinyo S, et al. Pali: A jointly-scaled multilingual language-image model[J]. arXiv preprint, arXiv: 2209. 06794, 2022

[71] Li Bo, Zhang Yuanhan, Chen Liangyu, et al. Mimic-it: Multi-modal in-context instruction tuning[J]. arXiv preprint, arXiv: 2306. 05425, 2023

[72] Chiang W, Li Zhuohan, Lin Zi, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality[EB/OL]. [2024-05-16]. https://lmsys.org/blog/2023-03-30-vicuna/

[73] Caffagni D, Cocchi F, Barsellotti L, et al. The revolution of multimodal large language models: A survey[C]//Proc of the 18th Findings of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 13590−13618

[74] Zhai Bohan, Yang Shijia, Zhao Xiangchen, et al. HallE-Control: Controlling object hallucination in large multimodal models[J]. arXiv preprint, arXiv: 2310. 01779, 2023

[75] Zhang Yue, Li Yafu, Cui Leyang, et al. Siren's song in the AI ocean: A survey on hallucination in large language models[J]. arXiv preprint, arXiv: 2309. 01219, 2023

[76] Tonmoy S, Zaman S, Jain V, et al. A comprehensive survey of hallucination mitigation techniques in large language models[J]. arXiv preprint, arXiv: 2401. 01313, 2024

[77] Huang Lei, Yu Weijiang, Ma Weitao, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. arXiv preprint, arXiv: 2311. 05232, 2023

[78] Chen Lin, Li Jinsong, Dong Xiaoyi, et al. shareGPT4V: Improving large multi-modal models with better captions[C]//Proc of European Conf on Computer Vision. Berlin: Springer, 2024

[79] OpenAI. GPT-4 is OpenAI's most advanced system, producing safer and more useful responses[EB/OL].[2024-05-16]. https://openai.com/gpt-4

[80] Yu Qifan, Li Juncheng, Wei Longhui, et al. HalluciDoctor: Mitigating hallucinatory toxicity in visual instruction data[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 12944−12953

[81] Chen Zhiyang, Zhu Yousong, Zhan Yufei, et al. Mitigating hallucination in visual language models with visual supervision[J]. arXiv preprint, arXiv: 2311. 16479, 2023

[82] Chen Liang, Zhao Haozhe, Liu Tianyu, et al. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models[J].arXiv:2403.06764

[83] Huang Qidong, Dong Xiaoyi, Zhang Pan, et al. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 13418−13427

[84] Zhang Yifan, Yu Weichen, Wen Qingsong, et al. Debiasing large visual language models[J]. arXiv preprint, arXiv: 2403. 05262, 2024

[85] Chen Zhaorun, Zhao Zhuokai, Luo Hongyin, et al. HALC: Object hallucination reduction via adaptive focal-contrast decoding[C]// Proc of the 41st Int Conf on Machine Learning. New York: PMLR, 2024: 7824−7846

[86] Mukherjee A, Chang H. The creative frontier of generative AI: Managing the novelty-usefulness tradeoff[J]. arXiv preprint, arXiv: 2306.03601, 2023

[87] Wang Xintong, Pan Jingheng, Ding Liang, et al. Mitigating hallucinations in large vision-language models with instruction contrastive decoding[C]//Proc of the 18th Findings of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 15840−15853

[88] Krishna R, Zhu Yuke, Groth O, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations[J]. International Journal of Computer Vision, 2017, 123(5): 32−73

[89] Lin T-Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context[C]//Proc of the 13th European Conf on Computer Vision. Berlin: Springer, 2014: 740−755

[90] Chesser L, Hinegardner J, Notes A, et al. The UnSplash dataset[EB/OL].[2024-05-16]. https://github.com/unsplash/datasets

[91] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A large-scale hierarchical image database[C]//Proc of the 22nd IEEE Conf on Computer Vison and Pattern Recognition. Piscataway, NJ: IEEE, 2009: 248−255

[92] Schuhmann C, Jitsev J, Vencu R, et al. LAION-Aesthetics[EB/OL]. [2024-05-16]. https://laion.ai/blog/laion-asthetics/

[93] Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale[J]. International Journal of Computer Vision, 2020, 128(7): 1956−1981

[94] Arrieta A B, DÍaz-RodrÍguez N, Del-Ser J, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion, 2020, 58(6): 82−115

[95] OpenAI. ChatGPT[EB/OL].[2024-05-16]. https: /chat. openai. com/

[96] Vedantam R, Lawrence-Zitnick C, Parikh D. CIDEr: Consensus-based image description evaluation[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2015: 4566−4575

[97] Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proc of the 40th Annual meeting of the Association for Computational Linguistics. Berlin: ACL, 2002: 311−318

[98] Wang Peng, Yang An, Men Rui, et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework[C]//Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2022: 23318−23340

[99] Wang Wentao, Huang Xuanyao, Roy S K. DeepArt: A benchmark to advance fidelity research in AI-generated content[J]. arXiv preprint, arXiv: 2312. 10407, 2023

[100] Huang Qingqiu, Xiong Yu, Rao Anyi, et al. MovieNet: A holistic dataset for movie understanding[C]//Proc of the 16th European Conf on Computer Vision. Berlin: Springer, 2020: 709−727

[101] Zhang Chenyu, Van-Durme B, Li Zhuowan, et al. Visual commonsense in pretrained unimodal and multimodal models[C]// Proc of the 16th Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2022: 5321−5335

[102] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with CLIP latents[J]. arXiv preprint, arXiv: 2204. 06125, 2022

[103] Oppenlaender J. The creativity of text-to-image generation[C]//Proc of the 25th Int Academic Mindtrek Conf. New York: ACM, 2022: 192−202

[104] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 10684−10695

[105] Tang B J, Boggust A, Satyanarayan A. Vistext: A benchmark for semantically rich chart captioning[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 7268−7298

[106] Liu Fuxiao, Wang Yinghan, Wang Tianlu, et al. Visual news: Benchmark and challenges in news image captioning[C]//Proc of the 26th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 6761−6771

[107] Honovich O, Choshen L, Aharoni R, et al. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering[C]//Proc of the 26th Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 7856−7870

[108] Chen Chi, Qin Ruoyu, Luo Fuwen, et al. Position-enhanced visual instruction tuning for multimodal large language models[J]. arXiv preprint, arXiv: 2308. 13437, 2023

[109] Yang Dingchen, Cao Bowen, Chen Guang, et al. Pensieve: Retrospect-then-compare mitigates visual hallucination[J]. arXiv preprint, arXiv: 2403. 14401, 2024

[110] Deng Ailin, Chen Zhirui, Hooi B. Seeing is believing: Mitigating hallucination in large vision-language models via CLIP-guided decoding[J]. arXiv preprint, arXiv: 2402. 15300, 2024

[111] Mitra C, Huang B, Darrell T, et al. Compositional chain-of-thought prompting for large multimodal models[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2024: 14420−14431

[112] Wu Junfei, Liu Qiang, Wang Ding, et al. Logical closed loop: Uncovering object hallucinations in large vision-language models[C]//Proc of the 18th Findings of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 6944−6962

[113] Goyal Y, Khot T, Summers-Stay D, et al. Making the V in VQA matter[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 6904−6913

[114] Schwenk D, Khandelwal A, Clark C, et al. A-OKVQA: A benchmark for visual question answering using world knowledge[C]//Proc of the 17th European Conf on Computer Vision. Berilin: Springer, 2022: 146−162

[115] Young P, Lai A, Hodosh M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics, 2014, 2: 67−78

[116] Yang Jingkang, Ang Yizhe, Guo Zujin, et al. Panoptic scene graph generation[C]//Proc of the 16th European Conf on Computer Vision. Berilin: Springer, 2022: 178−196

[117] Zhong Yiwu, Yang Jianwei, Zhang Pengchuan, et al. RegionCLIP: Region-based language-image pretraining[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16793−16803

[118] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proc of the 36th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 4015−4026

[119] Zhang S, Roller S, Goyal N, et al. OPT: Open pre-trained transformer language models[J]. arXiv preprint, arXiv: 2205. 01068, 2022

[120] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision[J]. arXiv preprint, arXiv: 2304. 07193, 2023

[121] Huang Yupan, Lv Tengchao, Cui Lei, et al. LayoutLMv3: Pre-training for document ai with unified text and image masking[C]//Proc of the 30th ACM Int Conf on Multimedia. New York: ACM, 2022: 4083−4091

[122] Liu Zhuang, Mao Hanzi, Wu Chaoyuan, et al. A convnet for the 2020s[C]//Proc of the 35th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 11976−11986

[123] Zhang Youcai, Huang Xinyu, Ma Jinyu, et al. Recognize anything: A strong image tagging model[C]//Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops. Piscataway, NJ: IEEE, 2024: 1724−1732

[124] Hu E J, Shen Yelong, Wallis P, et al. LoRA: Low-rank adaptation of large language models[C/OL].[2025-04-04]. https://openreview.net/forum?id=nZeVKeeFYf9

[125] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model[C]//Proc of the 37th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2023: 53728−53741

[126] Li X L, Holtzman A, Fried D, et al. Contrastive decoding: Open-ended text generation as optimization[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 12286−12312

[127] Agrawal H, Desai K, Wang Yufei, et al. Nocaps: Novel object captioning at scale[C]//Proc of the 13th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 8948−8957

[128] Belle V, Papantonis I. Principles and practice of explainable machine learning[J]. Frontiers in Big Data, 2021, 4(7): 1−25



Li Xu, born in 1998. PhD candidate. His main research interests include multimodal information processing, multimodal large language models, and large vision-language models.

李　煦, 1998 年生. 博士研究生. 主要研究方向为多模态信息处理、多模态大语言模型、视觉语言大模型.



Zhu Rui, born in 2002. PhD candidate. Her main research interests include computer vision and large vision-language models.

朱　睿, 2002 年生. 博士研究生. 主要研究方向为计算机视觉、视觉语言大模型.



Chen Xiaolei, born in 1999. PhD candidate. His main research interests include computer vision and machine learning.

陈小磊, 1999 年生. 博士研究生. 主要研究方向为计算机视觉、机器学习.



Wu Jinxuan, born in 2001. Master candidate. Her main research interests include large multimodal models, and algorithmic fairness and safety.

伍瑾轩, 2001 年生. 硕士研究生. 主要研究方向为多模态大模型、算法公平与安全性.



Zheng Yi, born in 1998. Master candidate. His main research interests include computer vision and large vision language models.

郑　毅, 1998 年生. 硕士研究生. 主要研究方向为计算机视觉、视觉语言大模型.



Lai Chenghang, born in 1994. PhD candidate. His main research interests include deep learning, machine learning, and multimodal knowledge extraction and reasoning.

赖承杭, 1994 年生. 博士研究生. 主要研究方向为深度学习、机器学习、多模态知识提取与推理.



Liang Yuxuan, born in 2001. Master candidate. His main research interests include artificial intelligence and large vision language models.

梁宇轩, 2001 年生. 硕士研究生. 主要研究方向为人工智能、视觉语言大模型.



Li Bin, born in 1982. PhD, professor, PhD supervisor. His main research interests include machine learning and vision intelligence.

李　斌, 1982 年生. 博士，教授，博士生导师. 主要研究方向为机器学习、视觉智能.



Xue Xiangyang, born in 1968. PhD, professor, PhD supervisor. His main research interests include computer vision, multimedia, and machine learning.

薛向阳, 1968 年生. 博士，教授，博士生导师. 主要研究方向为计算机视觉、多媒体、机器学习.