大规模视觉语言模型幻觉综述

张旭龙[1]，潘鹏[1,2*]，瞿晓阳[1]，倪晓俊[3]，田晖[4]，王健宗[1]
1. 平安科技（深圳）有限公司，深圳 518000；
2. 中国科学技术大学，合肥 230000；
3. 中国科学院计算技术研究所，北京 100190；
4. 华侨大学，泉州 362021

摘要

在多模态人工智能的快速发展中，大规模视觉语言模型凭借强大的跨模态理解与生成能力，已在视觉问答、图文生成、智能检索等任务中展现出广泛的应用前景。然而，幻觉问题，即模型生成的内容与输入视觉信息或事实不符，正成为制约其可靠性和可信度的关键挑战。研究发现，幻觉的产生不仅与训练数据的偏差和噪声有关，还涉及视觉编码与语言生成之间的对齐不足、模型结构的固有局限，以及推理阶段注意力分布的不稳定性。现有工作在幻觉的检测和评估方面逐步建立了多层次的分析框架，包括从对象、属性、关系等语义维度刻画模型错误表现；在缓解策略方面，研究者尝试通过提升数据质量、优化视觉表征、改进跨模态对齐方式、引入人类反馈机制，以及设计更稳健的推理与生成方法来降低幻觉发生率。尽管相关研究已取得一定进展，但在高风险应用领域中，幻觉问题仍然存在潜在风险，亟需进一步深入探索。系统梳理了 LVLM 幻觉的定义、类型、成因、检测与缓解思路，总结了当前面临的主要挑战与未来研究趋势，旨在为相关研究人员与工程实践提供全面而深入的参考。

关键词

幻觉；大规模视觉语言模型；多模态人工智能；跨模态对齐

中图分类号：TP18　　文献标志码：A　　doi：10.11959/j.issn.2096-0271.BDR25258

Hallucination in Large Vision-Language Models: A Comprehensive Survey

Zhang Xulong[1], Pan Peng[1,2*], Qu Xiaoyang[1], Ni Xiaojun[3], Tian Hui[4], Wang Jianzong[1]
1. Ping An Technology (Shenzhen) Co. , Ltd. , Shenzhen, China；
2. University of Science and Technology of China, Hefei, China；
3. Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China；
4. Huaqiao University, Quanzhou, China

Abstract

In the rapid development of multimodal artificial intelligence, large-scale visual language models, with their powerful cross-modal understanding and generation capabilities, have demonstrated broad application prospects in tasks such as visual question answering, image-text generation, and intelligent retrieval. However, the hallucination problem, where the generated content does not match the input visual information or facts, is becoming a key challenge that restricts the reliability and credibility of these models. Research has found that hallucinations are not only related to the bias and noise in the training data, but also involve insufficient alignment between visual encoding and language generation,

inherent limitations of the model structure, and instability in the distribution of attention during the reasoning stage. Existing work has gradually established a multi-level analysis framework for hallucination detection and evaluation, including characterizing the model's error performance from semantic dimensions such as objects, attributes, and relationships; in terms of mitigation strategies, researchers attempt to reduce the occurrence of hallucinations by improving data quality, optimizing visual representations, improving cross-modal alignment methods, introducing human feedback mechanisms, and designing more robust reasoning and generation methods. Although notable progress has been made in this area, hallucination in LVLMs still poses potential risks, especially in high-stakes application scenarios, and therefore requires further in-depth investigation. This survey systematically reviews the definitions, categories, causes, detection methods, and mitigation strategies of LVLM hallucinations, and further summarizes the major challenges and future research directions, with the aim of providing a comprehensive and in-depth reference for both researchers and practitioners.

Key words
Hallucination, Large Visual Language Model, Multimodal Artificial Intelligence, Cross-modal Alignment

1 引言

近年来，随着深度学习和大规模预训练技术的快速发展，人工智能在多模态理解与生成方面取得了显著突破。尤其是大规模视觉语言模型（Large Vision-Language Models，LVLMs）[1][2][3][4][5][6][7][8][9][10]，通过将强大的语言建模能力与视觉表征相结合，在图像描述、视觉问答、跨模态检索、图文对话[11][12][13]等任务中展现出前所未有的性能。这类模型的出现不仅推动了计算机视觉与自然语言处理的深度融合，也为人机交互、教育、医疗、内容创作等领域带来了广阔的应用前景。然而，伴随这些进展而来的，是幻觉问题[14][15][16][17][18]，即模型生成的文本内容与输入的视觉信息或客观事实不一致。这一现象不仅削弱了模型的可靠性和可解释性，更在高风险应用场景中带来潜在的安全隐患，成为学术界和工业界亟需解决的重要挑战。

所谓“幻觉”，本质上是模型在推理和生成过程中引入了虚假的、错误的或不相关的信息。在纯语言模型中，幻觉通常表现为事实性错误或语义偏差。而在LVLM中，幻觉更为复杂和多样化，既包括生成完全不存在的对象（如在一张空旷的图片中“看见”不存在的动物），也包括错误的属性描述（如将“红色的花”错误描述为“黄色的花”），以及对象间关系的错误推断（如将“杯子在桌子上”描述为“桌子在杯子上”）。这些幻觉不仅是简单的识别或表达偏差，更反映出模型在跨模态语义对齐、上下文理解与推理机制上的深层次缺陷。

造成幻觉的原因是多方面的。从数据层面来看，训练数据中广泛存在的偏差与噪声，如配对图文的不一致、低质量标注或缺乏多样性，都会误导模型的学习过程，导致模型在生成时偏向频繁共现的模式而忽视真实语境。从模型结构来看，视觉编码器在捕捉细粒度视觉信息时能力有限，跨模态对齐模块在映射视觉特征与语言表征时也可能存在信息损失或错误匹配，从而造成视觉与语言的不一致。从训练机制来看，LVLMs通常依赖语言建模目标进行

优化，模型倾向于生成连贯、流畅的文本而非严格对齐的事实性描述，这使得其在推理过程中容易凭借语言先验而非真实视觉证据生成答案。从推理阶段来看，注意力分布的不稳定、解码策略的偏差，以及在面对模糊或复杂场景时的表征不足，都会进一步加剧幻觉问题的出现。

为了系统认识和解决幻觉问题，学界和工业界近年来开展了大量研究工作。在检测与评估方面，研究者提出了多种方法来量化幻觉现象，包括从对象、属性、关系等语义层面对生成内容进行比对分析，以及基于一致性、自验证或外部知识库的方法来辅助识别错误。这些研究不仅揭示了幻觉的多样化表现形式，也为后续的缓解方法提供了实验基础。在缓解与优化方面，现有探索主要集中在四个方向：（1）数据优化，通过构建更大规模、更高质量和更具多样性的图文数据来减少训练偏差[16][17]；（2）模型改进，通过增强视觉表征能力、提升模态对齐精度来降低信息丢失[17]；（3）训练策略创新，引入对比学习、辅助监督以及人类反馈优化（RLHF）等机制，使模型更好地学习事实一致性[17]；（4）推理与解码优化，通过对比解码、检索增强、视觉引导或后验校正等方式，在生成阶段减少幻觉的产生[15][16]。尽管这些方法在一定程度上改善了LVLM的输出质量，但从根本上消除幻觉仍然是一个长期而艰巨的任务。

幻觉问题不仅是学术研究中的技术难题，也对LVLM的实际应用带来深远影响。在医疗诊断、教育辅助、自动驾驶等高风险领域，模型输出的任何虚假信息都可能引发严重后果；在新闻生成、内容创作等领域，幻觉则可能导致虚假信息的扩散与信任危机。因此，如何在保证模型生成能力与表达多样性的同时，提升其事实一致性和可靠性，是未来LVLM研究必须面对的核心问题[17]。值得注意的是，幻觉问题也为研究带来了新的机遇。一方面，它揭示了当前大模型在跨模态推理和对齐上的不足，推动学界探索更具因果性和可解释性的建模方法；另一方面，它呼唤统一的评测体系和多维度的指标，以促进不同模型和方法之间的公平对比和迭代优化。

综上所述，LVLM作为多模态人工智能的重要代表，虽在任务表现上取得了突破性进展[19][20][21]，但幻觉问题的普遍存在仍严重制约其可靠落地。当前研究已在定义与分类、成因分析、检测与评估、缓解方法等方面积累了大量成果，但尚未形成彻底解决方案。未来的研究需要进一步关注高质量数据构建、模型结构的跨模态融合、训练过程中的事实对齐机制，以及推理阶段的低开销抑制方法。本综述旨在系统梳理相关研究进展，揭示问题本质，总结现有挑战，并展望未来的发展趋势，为推动LVLM的可信应用提供全面的参考。

2 幻觉的定义与表现

在深入探讨幻觉的具体内涵与类型之前，有必要首先对其在大规模视觉语言模型中的研究背景加以梳理。随着LVLM在图像描述、视觉问答、多模态对话等任务中广泛应用，研究者逐渐意识到，模型在生成内容时并非总能与输入图像保持一致。尤其是在涉及细粒度视觉信息或复杂语境推理的场景下，模型往往会出现与事实不符的描述，这类现象被统称为“幻觉”。与传统计算机视觉中的分类错误不同，幻觉更强调生成结果与视觉证据之间的脱节，这种错误往往具有虚构性和误导性。它不

仅削弱了模型的可解释性和可靠性，还对实际落地应用带来潜在风险。因此，从概念上厘清幻觉，并在不同层次上对其进行分类和分析，是开展后续检测、评估与缓解研究的前提。

2.1 幻觉在LVLM中的概念

随着LVLM快速发展，研究者们逐渐注意到模型生成的输出并非总是与输入图像保持严格一致，其中最具代表性的现象便是“幻觉”。在自然语言处理领域，幻觉最初指代语言模型生成的文本中包含事实错误或逻辑不一致的信息。而在LVLM场景下，其内涵更为复杂，通常被定义为模型生成的文本内容与输入视觉内容不符或与真实世界事实不一致的情况。

从本质上讲，幻觉反映了LVLM在跨模态信息整合与语义对齐方面的不足[15][16][18]。由于LVLM同时处理图像和语言两类信息，其生成结果应当建立在视觉证据与语言表达的一致性基础上。然而在实际应用中，模型往往倾向于依赖语言先验，忽视视觉输入，从而导致输出内容与图像存在偏差。这种现象不仅影响任务的准确性，还会削弱模型在关键领域中的可靠性。

需要强调的是，幻觉并非单纯的识别错误。例如，在图像分类任务中，如果模型将“猫”误判为“狗”，这通常归因于识别器的分类能力不足，而不被视为幻觉。幻觉更强调在生成任务（如图像描述、视觉问答、跨模态推理）中，模型主动生成了与视觉证据不符的信息。这种错误往往具有更强的“虚构性”，例如在一张干净桌面的图片中生成“桌子上有一台笔记本电脑”，或在回答“图片中有多少只鸟”时明明只有三只鸟却回答“四只”。

学术界普遍认为，LVLM幻觉具有以下几个核心特征：跨模态不一致性：语言输出与视觉输入内容不匹配，这是LVLM幻觉最突出的特征。事实错误性：生成的内容违背真实事实，即便其形式看起来合理。可辨识性：幻觉往往在人工评估中很容易被识别，但对模型本身来说却是难以避免的偏差。任务普遍性：幻觉几乎出现在所有LVLM下游任务中，包括图像描述、视觉问答、推理与多轮对话等[17]。

因此，可以将LVLM幻觉视为模型在数据驱动的统计模式学习与真实世界语义对齐之间出现的“断裂带”。这种断裂不仅来自数据偏差，也受到模型结构、训练目标和推理策略的共同影响。

2.2 常见的幻觉类型

为了更清晰地认识LVLM的幻觉现象，研究者们通常对其进行分类。从不同的角度出发，可以划分出不同的类别，其中最常见且具有代表性的四类是：对象幻觉、属性幻觉、关系幻觉、语境或推理幻觉。下面分别展开介绍。

2.2.1 对象幻觉

对象幻觉[15]是LVLM中最普遍和最直观的一类幻觉，指模型生成了输入图像中不存在的对象，或者遗漏了真实存在的对象。例如，图1显示一张图中只有三只鸟，但模型在回答时却认为图片中有四只鸟。对象幻觉的根源往往在于训练数据的统计偏差。例如，数据集中“公园”场景与“狗”的共现频率远高于“猫”，模型在生成时会根据语言先验推断“狗”更合理，从而忽视了视觉证据。这种偏差在多模态任务中尤为严重，因为图像输入未能充分压制语言模型的先验倾向。对象幻觉的危害在于，它直接影响了模型对视觉世界的

图1 LVLM的四种幻觉类型

基本感知能力。如果连“是否存在某个对象”这一最基础的事实都无法保证，模型在更高层次的推理和决策中也很难具备可信度。因此，对象幻觉被普遍视为LVLM幻觉研究的核心问题之一。

2.2.2 属性幻觉

属性幻觉[17]指模型虽然识别出了正确的对象类别，但在描述对象的属性时出现错误。例如，在图1中用户问“这些鸟的喙颜色相同吗?”，模型回答“是的，这些鸟的喙都是橙红色的”。但是实际上这三只鸟的喙颜色是不一致的。

属性幻觉的出现与视觉特征表征不足密切相关。视觉编码器在低分辨率或复杂背景下，往往无法准确捕捉细节特征，导致模型在生成属性时更多依赖语言模式。例如，统计上“天空”与“蓝色”的共现频率极高，因此模型在描述时倾向于将任何天空都说成“蓝色”，即使实际上是“灰色”或“晚霞”。

此外，属性幻觉还与训练目标设计有关。在多数LVLM的训练中，优化目标主要是最大化语言生成的流畅性和合理性，而不是严格对齐视觉事实。这使得模型在生成时更注重“听起来像真的”，而非“确保准确无误”。属性幻觉的危害在于，它比对象幻觉更隐蔽。人类读者可能不会立即察觉“颜色”或“形状”的细节错误，但在精细任务（如医学影像诊断、工业质检）

中，这类错误可能带来严重后果。

2.2.3 关系幻觉

关系幻觉指模型正确识别了对象及其属性，但在对象之间的空间、逻辑或语义关系上出现错误。例如，在图1中用户问“黄色鸟右边是蓝色鸟吗?”，模型却错误地回答“是的，蓝色鸟在黄色鸟的右边”，错误地识别这两只鸟的位置关系。关系幻觉的成因较为复杂，既涉及跨模态对齐的误差，也与推理机制的不稳定有关。在跨模态学习中，模型需要将二维图像中的空间信息映射到序列化的语言描述中，这一过程存在较大信息损失。同时，LVLM 的注意力机制在处理长距离依赖和复杂场景时容易出现分散或偏置，导致空间关系表述错误。关系幻觉对任务的破坏性在某些场景下更甚于对象或属性幻觉。例如在自动驾驶场景中，如果模型错误地判断“行人和车辆的相对位置”，可能直接引发安全事故。因此，关系幻觉的检测与缓解是应用落地中的重点难题。

2.2.4 语境或推理幻觉

语境或推理幻觉[17]是一类更高层次的幻觉，指模型在需要结合上下文或进行逻辑推理时，生成了与视觉证据或事实逻辑不符的内容。这类幻觉常见于复杂问题回答、多轮对话或需要外部知识辅助的任务中。例如，图1 中用户问“这三只鸟在干什么”，模型却回答“这三只鸟在热带雨林中觅食”，生成了虚假的信息。

推理幻觉的本质在于语言模型的生成目标与事实对齐目标之间的冲突。语言模型被训练成“生成最可能的下一个词”，而非“生成真实的事实”，因此在缺乏充分视觉证据或外部知识支撑时，它会凭借语言先验和模式记忆生成看似合理却虚假的答案。此外，推理幻觉还可能源于注意力分布异常，即模型在推理过程中过度依赖无关视觉区域或错误的上下文。

推理幻觉的危害尤为突出，因为它常常难以通过简单比对发现。对于人类用户而言，模型的回答可能逻辑连贯、语言流畅，从而增加了对错误内容的信任风险。在医学、法律等需要复杂推理的应用中，推理幻觉可能带来严重后果，因此成为近年来研究的重点方向。

综上所述，LVLM 幻觉的定义与表现具有多样化与复杂性的特征。从对象、属性到关系，再到更高层次的语境与推理，幻觉无处不在，且其隐蔽性和危害性逐级增加。对象幻觉最为直观，属性与关系幻觉更具挑战，而语境与推理幻觉则对模型的智能水平提出了更高要求。深入理解幻觉的定义和表现形式，不仅有助于建立更合理的检测与评估框架，也为后续缓解策略的设计提供了理论依据。

3 幻觉的成因

幻觉的产生并非单一环节的偶然产物，而是贯穿于大规模视觉语言模型（LVLMs）生命周期各个阶段的系统性偏差。其根源往往交织在数据、模型和推理三大层面：数据层面的噪声、不一致与偏差为幻觉埋下隐患，模型层面的视觉表征不足、模态对齐缺陷和语言先验过强直接触发幻觉，而推理层面的注意力异常与解码策略偏差则进一步放大并固化了这些错误。理解幻觉的成因，有助于我们把握它的复杂性与普遍性，并为后续检测和缓解提供理论依据[22]。

在数据层面，幻觉往往源于大规模训

练数据的不纯净与不均衡[16][23]。当前LVLM的训练高度依赖海量图文对，这些数据大多来自互联网，难免存在配对错误与质量参差不齐的情况。一些图像被错误地配上了不相关的文本说明[16][24]，例如一张展示“沙滩”的图片，却配有“森林野餐”的文字描述，这种错配会导致模型在跨模态对齐时建立起虚假的联系。即便配对正确，也常常存在语义层面的不一致。例如，图像中同时出现猫和狗，但文本仅描述了“狗”，这会让模型习惯性地忽略猫的存在。另一类情况则是文本过度描述，即描述中包含图像中不存在的对象或动作，如“天空中有鸟”，而图像实际上空无一物。长此以往，模型在遇到类似场景时就会不加辨别地生成带有“鸟”的描述。此外，互联网文本常常冗余或模糊[16][23]，例如“美丽的风景”或“温馨的场景”，这类缺乏明确语义指向的描述强化了语言模式，却无法为视觉证据提供精确约束。训练数据的统计偏差更是加剧了这一问题。模型倾向于捕捉高频共现模式，而不是基于真实图像进行推理。例如，“厨房”场景往往与“锅具”和“食物”高度相关，这使得模型在看到厨房时，即使没有任何烹饪器具，也会自动生成“桌上有锅”之类的幻觉。同样地，数据集中长尾对象样本稀缺，例如“羊驼”或“特殊乐器”，使得模型更容易将其混淆为“马”或“吉他”等高频类别。文化与地域偏差进一步放大了这种现象，导致模型在面对多样化的真实场景时表现出系统性的错误。可以说，数据层面的噪声、不一致和偏差为幻觉的频繁出现提供了肥沃的土壤。

在模型层面，幻觉问题与LVLM的架构设计密切相关。首先，视觉表征不足是导致幻觉的重要原因。主流视觉编码器通常将图像下采样到 224×224 或 336×336 的分辨率[5][10][25][26]，这意味着大量细粒度信息在输入阶段就已经丢失。对于需要精细判断的任务，如识别交通标志上的文字或区分动物羽毛的颜色，这种低分辨率输入极易造成属性幻觉。为了克服这一局限，一些先进的开源模型（如LLaVA-NeXT，LLaVA-OneVision等）开始采用动态分辨率或图像切片（Image Cropping/AnyRes）策略。这些技术通过将原始高分辨率图像划分为多个局部切片，并结合全局低分辨率视图，显著提升了模型捕获细粒度视觉特征的能力。这种从固定低分辨率向自适应高分辨率处理的演进，已成为缓解细粒度对象与属性幻觉的重要趋势。更进一步，部分编码器缺乏对空间结构的显式建模，仅能提取全局特征，而不能准确捕捉对象之间的相对位置关系。这会导致模型在生成时混淆空间关系，出现“杯子在桌下”而非“杯子在桌上”的关系幻觉。另一个不可忽视的问题是模型的开放世界泛化能力不足。大多数视觉编码器基于有限类别的图像分类任务进行预训练，当面对未见过的对象时往往无法形成可靠表征，而是通过与高频类别的相似性进行“猜测”，这成为对象幻觉的重要来源[27][28][30]。

其次，模态对齐缺陷使得视觉信息在进入语言模型后被严重稀释。为了将视觉特征映射到语言空间，现有方法通常采用线性投影或简单的MLP[5][31][32]。然而，这种浅层映射不足以捕捉复杂的语义对应关系，尤其在多对象场景中容易出现模糊或错误的映射。虽然一些架构尝试通过Q-Former[10]或跨模态注意力机制加强对齐，但在复杂图像中注意力容易分散，无法聚焦于真正相关的视觉区域。更为关键的是，训练目标往往以语言生成损失为主，而跨模态一致性未得到足够约束。这意味

着在对齐过程中，语言信号占据了主导地位，视觉信息逐渐退居辅助角色，最终导致模型生成的文本更多是依赖语言先验，而非忠实于视觉证据。

语言先验过强[17][33][34][35][36]是幻觉产生的另一根本性因素。语言模型作为LVLM的核心部分，具有强大的模式补全和叙事能力。在推理过程中，即使视觉输入模糊或不完整，语言模型也会倾向于生成符合语法和常识的答案。例如，当被问到“图中的男人穿什么衣服”时，如果视觉编码器未能捕捉清晰细节，语言模型会基于统计规律直接回答“西装”，因为这是在语料中概率最高的搭配。类似的情况也会出现在图像描述任务中，模型会虚构一些典型场景元素，如“天空有鸟”“桌上有笔记本电脑”，以使句子听起来更自然。问题在于，这些“合理的补全”恰恰构成了幻觉。更糟糕的是，语言模型的叙事驱动性往往掩盖了错误，使输出具有高度流畅性和连贯性，从而误导用户信以为真。这种现象凸显出跨模态权重不平衡的问题，即视觉信号在决策过程中被边缘化，而语言信号成为主导。

在深入探讨模型层面诱因时，必须有效区分并解耦“纯语言幻觉（Pure Linguistic Hallucinations）”与“跨模态幻觉（Cross-modal Hallucinations）”[34]。前者源于底座大语言模型（LLM）固有的知识偏差或生成惯性，即便在无图像输入的情况下也会出现，体现了模型对语料统计规律的盲目遵循。后者则源于视觉特征与文本语义在对齐空间的映射失效，即模型感知到了视觉信号，但由于投影层（Projector）的瓶颈或对齐权重失衡，导致视觉证据被错误语言逻辑所扭曲。

为了实现两者的解耦，研究者提出通过对比实验（如“图像置乱”或“零视觉输入”）来量化语言先验的贡献度。有效解耦的关键在于识别幻觉的触发点：若幻觉源于LLM内部知识的陈旧或虚构，解决路径应聚焦于强化基座模型的忠实度；若幻觉源于跨模态冲突，则需引入对比学习或更强的视觉约束（如交叉注意力重权化）来削弱语言信号的主导地位。这种解耦分析不仅有助于精准定位幻觉根源，也为设计针对性的缓解策略提供了理论支撑，即在抑制LLM过度推断的同时，确保视觉信息在决策链路中的核心地位。

推理层面的异常则进一步放大了上述问题[37][38][39]。注意力机制是LVLM实现跨模态交互的关键，但在实践中经常出现偏差。注意力分布可能过于分散，导致模型无法集中在关键对象上，而是平均地关注整个图像背景，从而生成与核心事实不符的答案。研究还发现，部分注意力头在不同输入下始终关注不相关的区域或sink tokens[40]，这些“视觉无关头”在逻辑上与任务毫无关联，却在推理中占据了计算资源。更严重的是，在跨模态交互中，语言token往往主导注意力分配，视觉token权重被大幅削弱，使得最终的输出主要受语言先验驱动[17][33][41]。这种注意力异常不仅导致对象或属性层面的错误，还会在关系推理和语境理解中引发更隐蔽的幻觉。

除了注意力机制，解码策略的偏差也是幻觉的重要推手[42]。贪心解码倾向于选择高概率token，从而生成语言模式化的答案，忽视了视觉输入的多样性。Beam Search虽然能提升语言流畅性，但会进一步放大语言先验，使得输出更加“合理”却不一定“真实”。采样策略中的温度和Top-k/Top-p参数设置如果不当，会增加随机性，导致虚构内容的频繁出现。在多轮对话中，早期回答中的幻觉还可能

被后续回答继承甚至扩展，形成“幻觉滚雪球”的现象，使得错误不断叠加和放大。

综上所述，幻觉的成因是数据、模型与推理三方面问题的综合体现[43][44][45]。数据层面的噪声、不一致和偏差为幻觉提供了土壤；模型层面的视觉表征不足、模态对齐缺陷和语言先验过强使幻觉更易发生；推理层面的注意力异常和解码偏差则让幻觉最终被固化为输出。只有在这三个层面同时进行反思和改进，才能从根本上减少幻觉现象的频率和危害，为构建更加可靠的视觉语言模型[46]奠定基础。

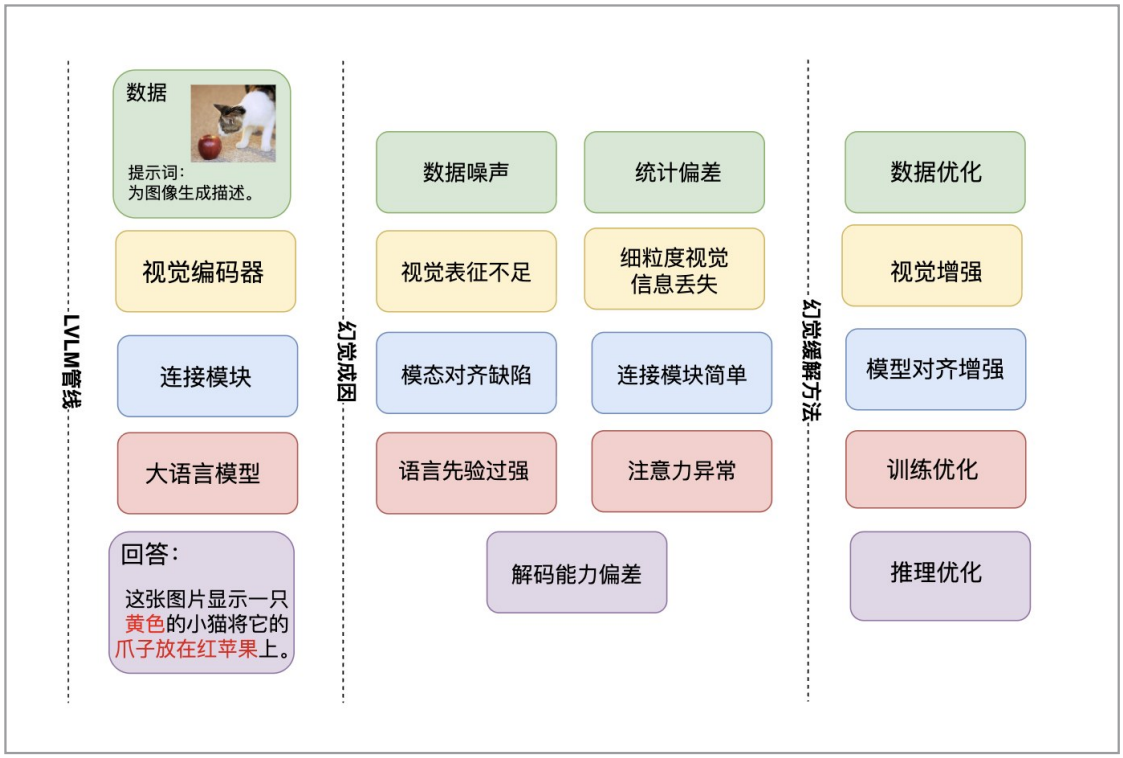

图2 LVLM幻觉原因和缓解方法

4 幻觉的检测与缓解方法

在深入分析幻觉的成因之后，学界和业界普遍认识到，仅仅揭示问题还远远不够，更关键的是如何有效识别和缓解这一现象。幻觉检测与缓解因而构成了研究的两个核心方向。一方面，检测提供了对模型输出进行量化和定位的工具，是后续改进与优化的前提。没有可靠的检测机制，幻觉问题往往无法被准确衡量，也难以在不同方法之间进行客观比较。另一方面，缓解方法旨在通过数据、模型和推理等环节的改进，减少幻觉的发生频率或降低其危害。这两部分相辅相成：检测为缓解提供评价与反馈，缓解则推动模型朝着更高的可靠性发展[17][44]。因此，本章将系统梳理现有的幻觉检测与缓解方法，首先介

绍不同层次的检测机制，包括人工评估与自动化检测框架，然后进一步探讨缓解思路，涵盖数据优化、模型改进以及推理策略调整等方面，力图为构建更可信赖的视觉语言模型提供全面视角。

4.1 幻觉的检测

幻觉检测是理解和缓解LVLM局限性的前提环节。由于幻觉往往表现在输出文本与输入视觉证据的不一致，而这种不一致可能是对象层面的缺失或虚构，也可能是属性、关系甚至推理逻辑上的错误，因此检测方法需要既能捕捉显性偏差，又能识别隐性的语义和逻辑冲突。总体而言，现有研究大致可以分为人工评估与自动化检测两条路径，两者在学术研究和实际应用中往往互为补充[17][33][34]。

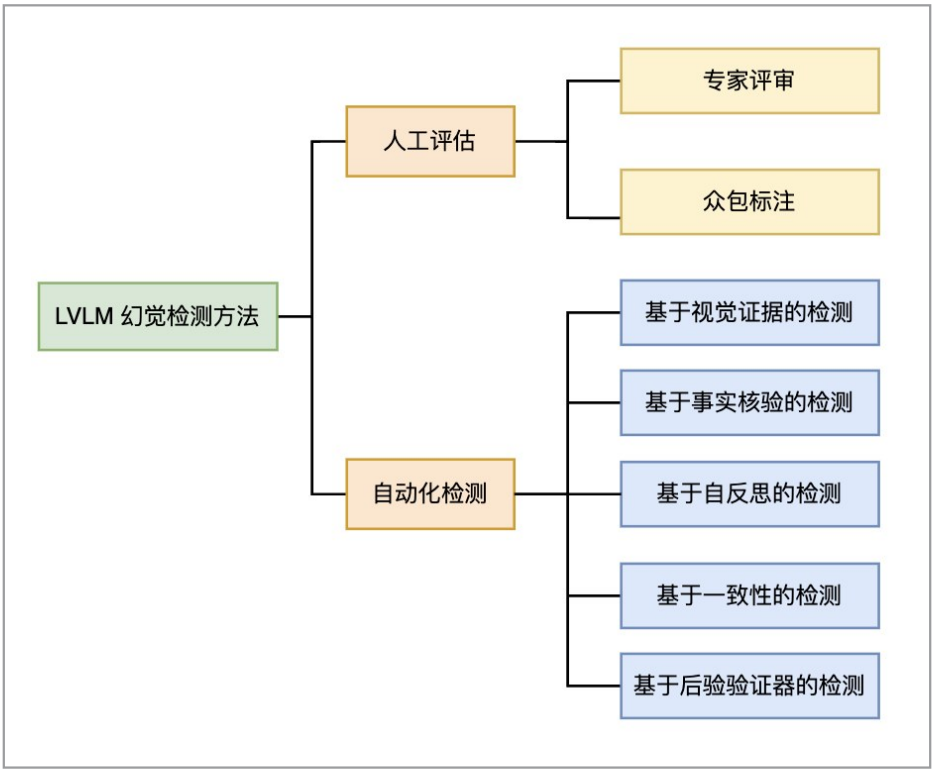

图3 LVLM幻觉检测方法

人工评估是最早也是最直观的检测方式。在许多研究中，专家或标注人员被要求对模型生成的文本与对应图像进行对比，判断是否存在不一致。专家评估尤其在医学、法律或交通等高风险领域被广泛采用，因为这些场景下的幻觉往往具有潜在的严重后果。例如，医学图像描述中的“发现恶性肿瘤”如果是幻觉，可能会直接导致错误诊断。专家凭借其专业知识能够识别出常规算法难以发现的隐性错误，从而保证了检测的权威性和可靠性。然而，这种方法的缺陷也很明显：成本高昂、规模受限，无法扩展到大规模训练和测试数据集之上。因此，人工评估在学术实验中常常

被作为“金标准”，主要用于验证其他检测方法的效果[17][33][34]。与专家评估相比，众包评估的优势在于能够快速收集大量标注结果，适合于开放域任务，如图像描述和视觉问答。在众包环境下，标注者通常根据简单的指南来判断输出是否与图像一致，并标注幻觉的类型。然而众包结果的质量往往参差不齐，标注者之间的主观差异也会影响一致性。为缓解这一问题，研究者会通过多标注者投票、交叉验证以及一致性系数的计算来提高可靠性。总体而言，人工评估具有不可替代的作用，它为幻觉检测提供了最直接的参考，但由于其成本与效率的限制，难以满足大规模自动化需求。

相比之下，自动化检测框架提供了可扩展的解决方案[33][34]。目前常见的做法之一是基于证据对齐的方法。这类方法的核心思想是先从图像中提取可验证的证据，再与生成文本进行匹配。例如，研究者通常会利用目标检测、语义分割或OCR技术提取出图像中的显著对象、属性或文字，然后解析模型输出中的名词短语、形容词或数量描述，并尝试建立一一对应关系。当模型输出提及的实体或属性无法在图像证据中找到对应项时，即被标记为幻觉。这类方法的优势在于可解释性强，可以清晰地指出具体哪个对象或属性存在不一致，并能够通过可视化进行验证。然而，它的不足在于依赖外部检测器的精度，如果检测器本身存在误差，可能会导致“误判幻觉”的情况。此外，证据对齐方法在处理开放世界任务时常面临类别覆盖不足的挑战，对长尾对象或细粒度属性的检测能力仍有局限。

另一类重要的方法是基于一致性的检测。其逻辑是：如果模型的回答真正依赖于视觉证据，那么在不同条件下的输出应当保持相对稳定。研究者常通过多次采样来验证一致性，即在同一图像输入下让模型多次生成回答，如果答案差异显著，则可能说明模型缺乏真实的视觉支撑，仅仅依赖语言先验进行猜测。还有研究会对输入进行轻微扰动，例如裁剪、旋转或遮挡部分区域，然后观察输出是否随之合理变化。如果模型在移除关键视觉区域后仍给出相同答案，则说明它的回答可能并未建立在视觉证据之上，而是语言驱动的幻觉。此外，也有学者采用逆向验证的方式，即要求模型为自己的输出提供解释或定位对应的图像区域。如果模型无法合理解释或其注意力分布与所述事实不符，则可以认为存在幻觉。这类一致性检测方法不依赖外部知识库或额外模型，因此部署相对便捷，但计算开销较大，而且对“稳定但错误”的幻觉不敏感[17][33][34]。

事实分解与核对的方法则强调对生成文本进行结构化处理，将其分解为最小的事实单元，再逐一进行核查[34]。例如，一句“一个男人穿着蓝色衬衫正在打篮球”的描述，可以被分解为对象事实（“男人”）、属性事实（“蓝色衬衫”）、动作事实（“正在打篮球”）。随后，这些事实单元会与视觉证据进行比对，以确定其真伪。这种方法的优势在于能够量化幻觉程度，而不是仅仅给出二元判断。通过统计生成文本中有多少比例的事实得到了视觉验证，研究者可以构建出幻觉率这一指标。事实分解方法尤其适合长文本生成，如图像描述和报告生成，因为这些任务的输出往往包含多个事实陈述。然而，这种方法的局限在于依赖于自然语言处理的抽取模块，如果抽取过程出现错误，可能会影响整体评估效果。此外，对于涉及复杂关系或逻辑推理的事实，目前的视觉验证手段仍显不足。

近年来兴起的另一方向是模型自检与互检[17][33][34]。所谓自检，是指要求模型在生成之后对自己的输出进行再一次验证，例如回答“上述描述是否完全与图像一致”或“请标注出支持这一答案的图像区域”。这种方法利用了LVLM自身的能力，避免引入外部资源。互检则是让两个或多个模型对同一输入分别生成答案，并相互验证对方的结果。如果不同模型之间在输出上存在矛盾，那么至少有一方可能存在幻觉。也有研究采用教师–学生框架，让一个更大或更精确的模型作为“教师”检查“学生模型”的输出，从而识别潜在幻觉。自检和互检方法的优势在于灵活性高，能够在实际系统中作为“二次过滤器”使用，但其缺陷在于当所有模型都共享相似偏差时，可能会出现“同错共犯”的局面。

从应用角度来看，幻觉检测方法在不同任务中需要进行调整[17]。例如，在图像描述任务中，检测方法往往侧重于对象与属性的一致性验证；在视觉问答任务中，更强调回答的准确性与合理性；在多轮对话中，则需要考虑语境依赖与逻辑一致性。针对不同任务，研究者提出了多样化的评估指标，如对象覆盖率、属性精度、一致性分数等。尽管已有方法在一定程度上能够识别幻觉，但整体上仍面临几个挑战。首先，自动化方法在跨任务泛化性上不足，往往在某一任务上表现优异，却难以迁移到其他场景。其次，隐性的推理幻觉仍难以检测，因为它涉及到外部知识与复杂逻辑，仅依赖视觉证据难以识别。再次，许多检测方法的计算开销较大，需要多次推理或调用外部模型，难以满足实时应用的需求。最后，目前缺乏统一的评测标准与基准，导致不同方法之间的结果难以横向比较。

总体来看，幻觉检测的方法从人工到自动化，从对象级别到推理级别，形成了一个逐渐完善的体系。人工评估为检测提供了最可靠的基准，自动化方法则保证了可扩展性。不同方法各有优劣，但尚未形成能够全面覆盖对象、属性、关系与推理幻觉的统一框架。未来的研究需要探索更高效、低开销且具有跨任务泛化能力的检测方法，并建立公开统一的评测体系，从而为幻觉缓解与模型优化提供坚实基础。

4.2 幻觉的缓解

缓解大规模视觉语言模型幻觉可以从“数据—模型—训练—推理”四个环节协同入手：在数据侧构造更“抗幻觉”的指令与理据样本，提升语义覆盖与难负样本比例；在模型侧通过结构与对齐机制加强“看见—理解—生成”的链路，尤其是让注意力真正落在与问题相关的局部区域；在训练侧将多任务监督、偏好优化与人类/自动反馈引入模型行为对齐；在推理侧以对比、约束、校正与后验验证为核心，直接在生成过程中“刹住车”[17][22][33][34]。这四条线索并非彼此独立：许多近年的有效方案本质上是在不同阶段共享同一原则——让文字证据被可验证的视觉证据持续“牵引”，并且在不确定时主动求证或降温模型信心，从而把“语言先验过强”的风险压到最小。

在数据层面，关键是把“可见事实—可解释过程—可对比负例”一起纳入训练语料。早期的稳健指令调优数据集通过系统性地引入否定/对抗式指令与负样本，显著降低了模型在常见指令场景中的过度联想倾向。在此基础上，后续工作进一步提出“针对性数据生成”的思路：先对目标模型做“幻觉画像”，再按画像为不同模型

定制化生成易混概念与歧义指令，形成真正对症下药的样本分布，从而在相同数据量下换取更高的减幻觉边际收益。对话型场景里，研究者发现前置的“误导性对话”会显著加剧后续问答的语言先验，因而通过对抗式指令调优显式暴露并中和这类偏置，能恢复对图像的依赖权重。特别值得关注的是，利用 GPT-4V 或 GPT-4o 等先进闭源模型自动生成大规模合成数据正成为增强模型鲁棒性的重要趋势。通过强感知模型对图像进行超精细描述，并针对性地构造反事实（Counterfactual）指令或自动修正后的对齐数据，可以为模型提供更高难度的硬负样本。这种利用 AI 反馈进行训练（RLAIF）的范式，不仅极大地降低了人工标注成本，还通过精准纠偏有效提升了模型对视觉证据的忠实度。另一条重要路径是把“理据”引入指令调优，让模型不仅给出答案，还要复盘为什么答案与视觉证据一致或不一致；这类“反思式”指令调优在多个评测上证明可以显著减少凭空臆断。此外，从表征学习视角出发，将“幻觉文本”当作硬负样本做跨模态对比，在特征空间把“非幻觉文本—视觉证据”拉近、把“幻觉文本”推开，能直接缓解模态对齐中的缠结问题；这类“幻觉增强对比学习”方法在多套模型与评测上都给出了一致收益。综合来看，数据优化的共同要点在于：显式覆盖“该不该说”“能不能证”的边界条件，并让样本同时提供可核验的依据与可对比的错误，促使模型在学习阶段就形成“证据—结论”的因果习惯，而不是仅仅拟合分布[16][23][24]。

模型侧的改进集中于两类瓶颈：一是视觉表征“看不清/看不准”，二是多模态融合“看见了但没用上”。在通用结构上，早期的门控跨模态融合与查询式桥接（例如冻结视觉编码器与语言模型、中间以轻量查询模块对齐）的思路，为后续大规模多模态对齐奠定了基线，使模型具备把高层视觉语义稳健映射到语言空间的能力。面向幻觉这一更具体的可靠性目标，一些工作通过因果注意与同心结构优化，让视觉 token 与指令 token 的交互更接近人类的凝视—证据采样过程，从而减少无关区域对解码的“噪声牵引”。与此同时，面向“表征与对齐”的训练期手段也可以被理解为“模型层”的补强：HACL[30]直接在隐空间做对比，把“有证据的文本—图像”紧密对齐，把“无证据的文本”排斥出去，实证显示能缩小模态间的表示间隙并降低缠结。总的来说，模型侧的方法把“相关性优先”落到结构或隐空间几何上：当模型必须在相关区域聚焦并以此指导生成，幻觉自然会被抑制。近年的若干方法更强调“把注意力拉到与问题相关的局部区域”。例如，AGLA[47]通过“全局—局部注意力组装”，在不改动主干权重的前提下并联一条对局部判别更敏感的分支，生成与当前提问强相关的图像增广视图，再在解码分布中显式融合“全局生成性特征”和“局部判别性特征”，从机制上减少“看图说话走神”的机会。

训练阶段的核心在于对齐“模型行为”与“人类偏好/可证事实”。基于人类偏好的强化/直接优化已经从文本领域扩展到多模态：Factually Augmented RLHF 在偏好模型构建时显式引入事实性辅证（如图像描述或选择项）[29][48][49][50]，用以抑制“奖励黑客”并把奖励信号锚定到视觉可证据，显著降低了模型在多模态对话中的越界描述。更轻量的直接偏好优化也被针对性地改造，形成“视觉引导 DPO”[51]与“面向幻觉的 DPO”[52]，前者用视觉一致性信号指导偏好学习，后者直接

让“少幻觉”成为优化目标，从而在不牺牲通用能力的情况下，降低开放式描述中的虚构风险。与此互补的还有“细粒度纠偏”路线：RLHF-V[53]采集段落级/片段级的人类纠错反馈，对“哪一句/哪个短语违背了图像”做密集标注，再以偏好优化或其密集变体进行训练，尤其适合安全敏感或高风险场景。另一条有效路径是“反思式训练”：REVERIE[54]将理据学习并入指令调优流程，要求模型对“正确/错误”答案分别给出视觉—逻辑支撑或驳斥，使其在训练时就学会把回答与可见证据绑定；这类“先答—后想—再校”的范式在多个任务上稳定降低了幻觉率。在多轮对话与长链推理中，自我反馈也可在训练端显式建模：有研究让模型对初始回答生成“自检反馈”，再用该反馈监督二次修订，以此习得“发现不确定—回到图像—改写答案”的行为模式。这些训练策略的共同点，是把“事实一致性”变成可学习的目标或偏好，并尽量用可解释的信号（视觉一致性、对比负样本、细粒度纠错与理据）支撑这一目标，从而缓解单纯依赖语言先验带来的系统性幻觉。

推理端的缓解手段最为活跃，也最容易在现有LVLM上“即插即用”。一个重要家族是“对比式解码”，其核心思想是构造“对照样本”来揭示语言先验或错误对齐，然后在输出分布上压制与对照差异不一致的候选。VCD[15]通过对原图与扰动图的输出做分布对比，显著降低“看图不看图都能说得通”的语言先验；其无需训练、对多种模型有效，是后续大量工作的基石。OPERA[40]抓住“过度信任摘要token、忽视图像token”的注意模式，引入“过信惩罚”与“回溯—重分配”机制，在几乎零训练成本下显著降低描述性幻觉。与此并行的是“注意力校准与显著性放大”路线：PAI[55]在生成时直接“给图像更高的话语权”，以免训练的方式放大视觉注意并抑制幻觉倾向；SID[56]则利用模型自身的跨层表示，动态评估视觉token重要性并对解码logits做自省式修正，避免粗暴扰动带来的噪声与推理时延。有一些工作进一步把对比与验证结合：ConVis[57]先用文本到图像生成把当前描述“可视化”，再以生成图与原图的对比信号回灌到解码分布；Octopus[58]将对比解码做成“动态多臂”，在不同扰动与权重之间自适应切换；而M3ID[59]则把“是否在看图说话”变成可控旋钮，在需要严格贴图的场景把“视觉约束强度”调到更高。还有一类方法引入外部或自建的“后验验证器”：MARINE[60]借鉴“无分类器引导”思想在不改动主干的情况下引入可控的视觉一致性偏置；它们本质上是在语言生成头之外再套个“可信度计”，把“看起来像事实”的候选压到后面。推理端方法的优势是部署友好、收益立竿见影，但也伴随新折衷：过强的抑制可能伤及流畅性和信息量，对比/验证也会带来额外时延。实践中常见的工程折中是：用轻量的注意力校准作为默认配置，在检测到不确定或高风险指令时再启动开销更高的对比或验证通道，并把“触发条件”与“强度”作为可调策略。

从工程实践的角度看，缓解策略的选择实质上是“微调（Fine-tuning）”与“免训练（Training-free）”两条技术路线的权衡。微调方法（如上述的数据层面优化、偏好强化学习等）通过调整模型权重来实现底层对齐，其优势在于推理阶段不产生额外开销，推理速度与基座模型一致，适合高吞吐量的在线服务；但其代价是训练时显存需求极高且计算资源消耗巨大。相比之下，免训练方法（如推理端的

对比式解码）具备极高的灵活性，无需训练显存和权重更新成本，能够即插即用；但其缺点在于显著增加了推理延迟——例如对比式方法通常需要两次模型前向传播，导致推理速度下降近 50%。因此，在显存受限或端侧部署场景下，开发者需根据硬件配额在“训练期重投入”以换取推理效率，与“推理期高时延”以换取零训练开销之间进行深度权衡。

5 LVLM 幻觉评测基准与度量

为了系统衡量大规模视觉语言模型（LVLM）的幻觉现象，学界逐步形成了生成式度量与判别式评测两条主线，并在此基础上提出整体性基准。生成式基准更贴近开放场景下的自由表述，能直接观测模型“凭空添加”的对象或属性；判别式基准则以可控的是/否问答衡量模型对“存在/不存在”的分辨能力，稳定且易于横向比较；整体性基准尝试把两类指标统一到同一框架，便于在不同任务形态之间对齐口径。

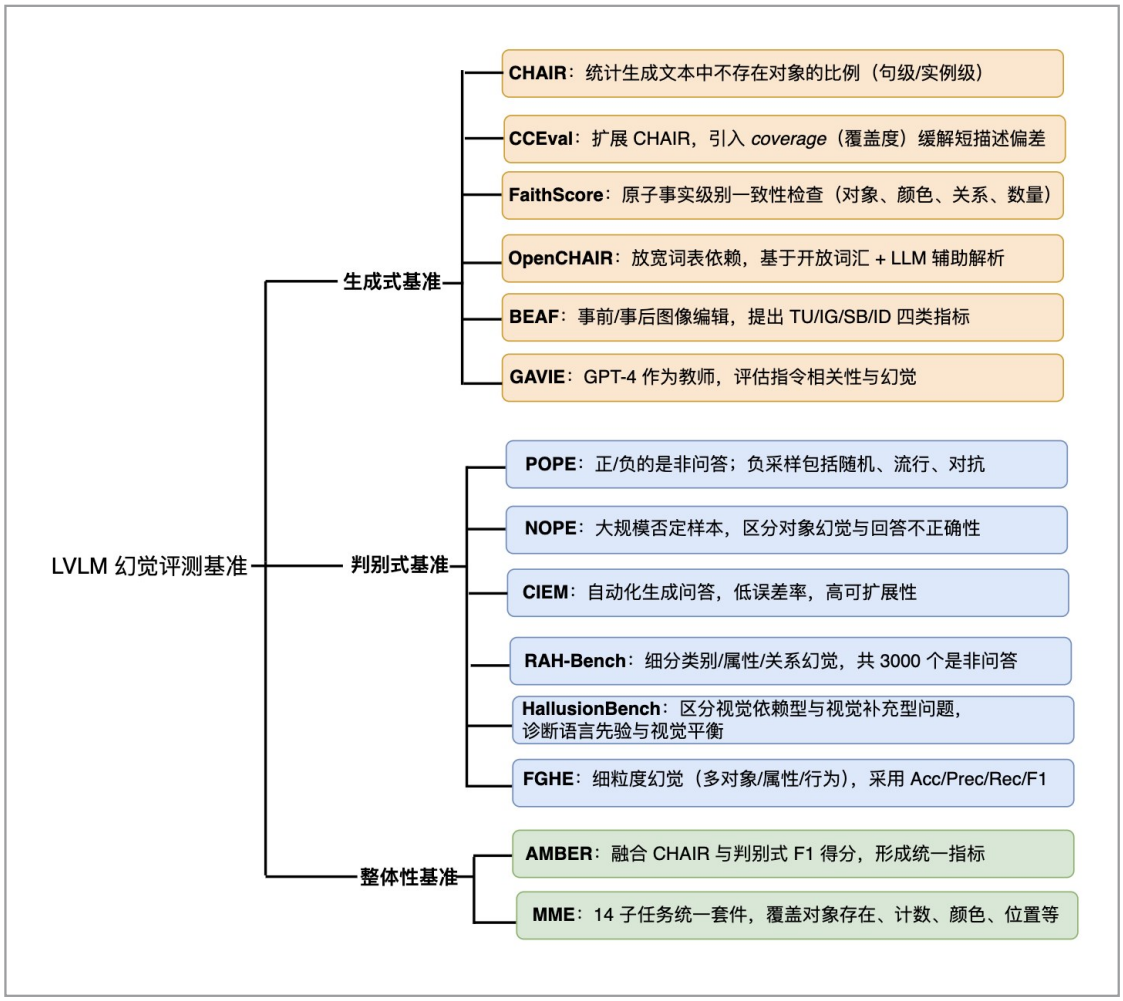

图4 LVLM幻觉评测基准

GAVIE[16]提出从两个不同方面评估LMM的输出：一是通过“相关性”来评估指令遵循的性能，二是通过“准确性”来衡量LMM输出中的视觉幻觉。它包含一个有1000个样本的基准和一个评估方法。GAVIE以开放式的方式评估LVLM的输出，并且不需要人工标注的基准答案。其核心思想是让先进的GPT-4充当智能教师，通过将图像内容、人类指令和模型响应作为输入来为答案打分。

CIEM[23]代表了利用大规模语言模型自动构造问答以扩展规模的判别式评测。其题目由自动化流程生成，并经统计验证具有较低的误差率（约5%），从而在保证数据质量的前提下显著降低人工标注成本。对需要频繁做消融与横评的研究场景，CIEM能提供较好的可扩展性；在落地评测时，建议与更“手工精炼”的基准组合使用，以兼顾规模与难度。

CCEval[25]是一种面向“细节描述”的生成式评测流程：在评测前利用GPT-4对生成文本进行解析、抽取对象词项，并在更丰富的对象集合上对照，同时在CHAIR的基础上引入“coverage”（覆盖度）指标，以缓解“短描述自然更少犯错”这一评分偏差。其构造通常从Visual Genome随机采样图像，关注描述的细致程度与事实覆盖。对于希望评估“开放描述质量—幻觉率—覆盖度”三者平衡的系统，CCEval提供了更贴近实际使用的度量框架。

MME[34]则是覆盖更广的一体化评测套件，在14个子任务上同时考查感知与认知能力，其中对象存在、计数、位置、颜色等子项直接对应对象层幻觉的常见失误。这些子项大多被设计成统一的是/否格式，易于实现与横向比较；在从“研究原型”走向“产品体检”的阶段，MME常被用作首批体检套件之一。

CHAIR[61]是最早被广泛采用的生成式对象幻觉度量之一，最初用于图像描述任务。它通过统计生成文本中被提及但图像中并不存在的对象比例来量化“对象层”幻觉，并提供句级与实例级两种口径，前者关注一句话是否出现过幻觉，后者关注出现了多少次幻觉对象。该指标流程清晰、可复现，适合作为开放描述场景的基础分数与回归监测指标，但由于依赖对象词汇的抽取与匹配，遇到开放词汇或细粒度同义词时需配合更强的解析器或补充度量共同使用。

POPE[62]把“是否出现幻觉”转化为判别式二分类任务：针对同一幅图像构造正向与负向的是/否问题，其中负向问题又细分为随机、热门与对抗三种采样方式，以系统地考查模型在不同干扰下识别“不存在对象”的能力。基准基于MSCOCO[71]选取500张图像构造问答，统一题型、统一评分，便于横向比较。该基准中的问题包含肯定问题和否定问题。肯定问题是基于基准事实中的对象构建的，而否定问题则是通过对不存在的对象进行抽样构建的。 该基准根据不同的负采样策略分为三个子集：随机、流行和对抗性。 流行和对抗性采样专门设计用于评估频繁出现对象和对象共现的情况。它尤其适合用来评估“对象存在性”这一最常见的幻觉类型，并可作为回归评测的主力基准。

NOPE[63]进一步将焦点放在“否定存在”的理解能力上，现有的VQA数据集存在严重的分布不平衡问题，包含的负向样本数据过少。NOPE通过自动流水线合成2.95万规模的负向样本。该基准提出要区分对象幻觉和不正确性。对象幻觉指的是在VQA中，视觉语言模型的回答包含了一个不存在的对象，尽管基准答案是一

个否定不定代词（如“没有”、“无人”等）。不正确性发生在视觉语言模型未能准确回答一个基准答案不是 NEGP 的问题时。与 POPE 相比，它在规模与否定表述的系统性上更进一步，常用于补充与加压模型的“否定理解”能力。

AMBER[64]试图把生成式与判别式两种口径整合为一个统一分数。它将生成侧的CHAIR指标与判别侧的F1得分求平均得到AMBER分数，使研究者能够在同一坐标轴上观察“开放生成幻觉率”与“受控问答识别能力”的综合水平。该分数尤其适合论文与系统报告中的“总览对比”，并可作为多轮优化中的目标指标之一。

FaithScore[65]提供了一种“无参照、细粒度”的生成式一致性评估方式：先把模型的自由生成划分为更小的描述子句，进一步抽取原子事实，再逐一与输入图像进行对照，由此在对象、计数、颜色、关系等不同维度给出更细致的幻觉判定。由于能把错误定位到原子事实层面，FaithScore 在诊断与误例回流（用于数据增广或训练纠偏）方面具有独特价值。

RAH-Bench[66]可被视为POPE的升级版，包含3000个“是/否”问题及其对应的图像。与 POPE 不同，RAH-Bench 进一步将否定问题分为三个子集。每个子集包含500个在不同方面带有误导性陈述的问题，包括：类别幻觉；属性幻觉；关系幻觉。

HallusionBench[67]为了诊断和分析LVLM可能的失败模式，从一个不同的角度评估幻觉。它由455个视觉-问题控制对组成，包含346幅不同的图像和总共1129个问题，涵盖了多样的主题和格式。问题分为两类：视觉依赖型和视觉补充型。视觉依赖型问题被定义为在没有视觉上下文的情况下无法得到肯定答案的问题。这种设置旨在评估视觉常识知识和视觉推理能力。视觉补充型问题可以在没有视觉输入的情况下回答；视觉部分仅提供补充信息或修正。这种设置旨在评估视觉推理能力以及参数化记忆（语言先验）和图像上下文之间的平衡。这种划分为了解和诊断LVLM提供了新的视角。

FGHE[68]遵循与POPE类似的二元分类方法来评估LVLM。然而，与POPE不同，FGHE需要一组不同的二元问题来衡量细粒度的幻觉。FGHE数据集由50张图像和200个二元问题组成，分为三类：(1)多对象问题，验证图像中多个对象之间的关系；(2)属性问题，验证图像中对象的属性；(3)行为问题，验证图像中对象的行为。这些问题由人类标注员在MSCOCO数据集验证集的一个包含50张图像的子集上手动定义。与POPE类似，采用准确率、精确率、召回率和F1分数作为评估指标。

OpenCHAIR[69]为了在开放词汇设置中衡量对象幻觉，通过放宽对封闭词汇表的强依赖性来扩展 CHAIR。“开放词汇”体现在两方面。首先，在构建基准时，它使用文本到图像的扩散模型组织了一个包含合成图像及相应字幕的数据集，其中包括了多样的、开放词汇的对象。其次，在计算指标时，CHAIR检查单词或其同义词（由固定的词汇表给出）是否存在于基准注释中。相比之下，OpenCHAIR从预测的字幕中提取具体对象，并通过查询LLM来从这个列表中识别出幻觉对象。与CHAIR类似，最终的指标计算基于幻觉率。

BEAF [70]构建了一个新的评估数据集，称为事前事后幻觉数据集，并引入了新的指标：真实理解（True Understanding, TU）、无知（IGnorance, IG）、固执（Stubbornness, SB）和犹豫不决（InDecision, ID）。与以

往只关注构建问题和答案的工作不同，该基准的关键思想是通过图像编辑模型操纵视觉场景信息，并基于场景变化来设计指标。这使得通过观察模型感知变化的能力，可以清晰地评估视觉语言模型是否正确理解了给定场景。

6 总结与展望

本文围绕大规模视觉语言模型的幻觉问题，按照定义与表现、成因机制、检测与缓解、评测基准的脉络形成闭环。可以确认，幻觉并非孤立缺陷，而是由数据噪声与偏差、视觉表征与模态对齐不足、语言先验主导、推理解码失衡等因素耦合而成；仅靠单点修补难以长期奏效。有效路径应当贯穿数据—模型—训练—推理—评测的全链条，以证据优先为原则，将“是否看见、如何用上、何以为证”贯通到模型行为，并用标准化评测和误例回流维持持续改进。

面向落地实践，更稳健的方案在于把检测与缓解做成可复用管线：前端以对象、属性、关系与推理四层一致性检测定位失效，后端以针对性数据增强与难负样本回流修复分布，配合细粒度对齐与偏好优化塑造“有证则述、不确定则收敛”的生成习惯，推理阶段以对比、约束、校准与后验验证兜底，并将触发条件与强度参数化，兼顾效果、时延与可解释。评测侧采用生成式与判别式双轨，再辅以开放词汇与原子事实级核验，建立可审计的证据链与统一口径，保证横向比较与工程回归一致。

面向未来，值得推进的方向集中在三端：一是可验证生成与不确定性建模，让答案可回溯到区域级证据与工具调用，配合置信校准与拒答策略，满足高风险场景的可靠性与合规需求；二是因果与表征层面的原理化刻画，明确语言先验、视觉证据与解码策略的作用边界，给出可操作的一致性约束与上界分析，指导结构与训练设计；三是评测生态与数据治理的工程化升级，扩展到开放词汇、多语言、文档与视频等复杂场景，完善难例回流与隐私合规。以此为牵引，将研究范式与产品流程打通，推动 LVLM 从“说得像”走向“说得对、说得有据、用得起”。[72][73][74]

此外，从更宏观的发展趋势看，未来 LVLM 将由“通用感知与生成模型”进一步演进为“具备感知、推理、验证与行动闭环能力的多模态智能体”。一方面，模型将持续向更强的跨模态统一表示、更长上下文建模、更细粒度视觉理解以及多图像、多视频、多文档协同处理能力发展，以支撑复杂真实场景中的连续决策与长期任务；另一方面，LVLM 也将越来越深地与检索系统、知识库、OCR、检测器、规划器及外部工具链结合，形成“模型本体+工具链+验证器”的系统化架构，使其不再仅依赖参数记忆完成回答，而是能够通过主动查证、动态调用和结果回验提升输出质量。与此同时，随着应用逐步走向医疗、教育、自动驾驶、工业巡检与公共安全等高风险领域，未来 LVLM 的竞争重点也将从单纯追求规模和基准分数，转向可靠性、效率、可解释性与部署成本的综合优化。[75][76][77][78]

参考文献：

[1] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[J]. Advances in neural information processing systems, 2022, 35: 23716-23736.

[2] Bai J, Bai S, Yang S, et al. Qwen-vl: A frontier large vision-language model with versatile abilities[J]. arXiv preprint arXiv:2308.12966, 2023, 1(2): 3.

[3] Chen J, Zhu D, Shen X, et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning[J]. arXiv preprint arXiv: 2310.09478, 2023.

[4] Chen Z, Wu J, Wang W, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 24185-24198.

[5] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916.

[6] Dai W, Li J, Li D, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning [J]. Advances in neural information processing systems, 2023, 36: 49250-49267.

[7] Han J, Zhang R, Shao W, et al. Imagebind-llm: Multi-modality instruction tuning[J]. arXiv preprint arXiv: 2309.03905, 2023.

[8] Li B, Zhang Y, Chen L, et al. Otter: A multi-modal model with in-context instruction tuning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025.

[9] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv: 2304.10592, 2023.

[10] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.

[11] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// International conference on machine learning. PmLR, 2021: 8748-8763.

[12] Wang W, Lv Q, Yu W, et al. Cogvlm: Visual expert for pretrained language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 121475-121499.

[13] Bordes F, Pang R Y, Ajay A, et al. An introduction to vision-language modeling [J]. arXiv preprint arXiv: 2405.17247, 2024.

[14] Deng A, Chen Z, Hooi B. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding[J]. arXiv preprint arXiv: 2402.15300, 2024.

[15] Leng S, Zhang H, Chen G, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13872-13882.

[16] Liu F, Lin K, Li L, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning[J]. arXiv preprint arXiv:2306.14565, 2023.

[17] Liu H, Xue W, Chen Y, et al. A survey on hallucination in large vision-language models[J]. arXiv preprint arXiv: 2402.00253, 2024.

[18] Zhu L, Ji D, Chen T, et al. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 1624-1633.

[19] Agarwal V, Shetty R, Fritz M. Towards

causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9690-9698.

[20] Agrawal A, Batra D, Parikh D. Analyzing the behavior of visual question answering models[J]. arXiv preprint arXiv: 1606.07356, 2016.

[21] 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[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6904-6913.

[22] Yue Z, Zhang L, Jin Q. Less is More: Mitigating Multimodal Hallucination from an EOS Decision Perspective[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2024: 11766-11781.

[23] Hu H, Zhang J, Zhao M, et al. CIEM: Contrastive Instruction Evaluation Method for Better Instruction Tuning [C]//NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following.

[24] You H, Zhang H, Gan Z, et al. Ferret: Refer and Ground Anything Anywhere at Any Granularity[C]//The Twelfth International Conference on Learning Representations.

[25] Zhai B, Yang S, Zhao X, et al. HallE-Switch: Rethinking and Controlling Object Existence Hallucinations in Large Vision-Language Models for Detailed Caption[J]. 2023.

[26] Li Z, Yang B, Liu Q, et al. Monkey: Image resolution and text label are important things for large multi-modal models[C]//proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 26763-26773.

[27] Jain J, Yang J, Shi H. Vcoder: Versatile vision encoders for multimodal large language models[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 27992-28002.

[28] Cho J, Yoon S, Kale A, et al. Fine-grained Image Captioning with CLIP Reward[C]//Findings of the Association for Computational Linguistics: NAACL 2022. 2022: 517-527.

[29] Zhao Z, Wang B, Ouyang L, et al. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization[J]. arXiv preprint arXiv:2311.16839, 2023.

[30] Jiang C, Xu H, Dong M, et al. Hallucination augmented contrastive learning for multimodal large language model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 27036-27046.

[31] Liu H, Li C, Li Y, et al. Improved baselines with visual instruction tuning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 26296-26306.

[32] Yin S, Fu C, Zhao S, et al. A survey on multimodal large language models[J]. National Science Review, 2024, 11(12): nwae403.

[33] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM computing surveys, 2023, 55(12): 1-38.

[34] Bai Z, Wang P, Xiao T, et al. Hallucination of multimodal large language models: A survey[J]. arXiv preprint arXiv:2404.18930, 2024.

[35] Huang L, Yu W, Ma W, et al. A survey

on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

[36] Zhang Y, Li Y, Cui L, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models[J]. Computational Linguistics, 2025: 1-46.

[37] Wang B, Wu F, Han X, et al. Vigc: Visual instruction generation and correction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(6): 5309-5317.

[38] Lee S, Park S H, Seo M. Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision[C]//NAACL 2024. Association for Computational Linguistics (ACL), 2024.

[39] Wang J, Zhou Y, Xu G, et al. Evaluation and analysis of hallucination in large vision-language models[J]. arXiv preprint arXiv:2308.15126, 2023.

[40] Huang Q, Dong X, Zhang P, et al. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13418-13427.

[41] Xiao W, Huang Z, Gan L, et al. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(24): 25543-25551.

[42] Chuang Y S, Xie Y, Luo H, et al. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models[C]//The Twelfth International Conference on Learning Representations. 2023.

[43] Dziri N, Madotto A, Zaiane O R, et al. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 2197-2214.

[44] Gunjal A, Yin J, Bas E. Detecting and preventing hallucinations in large vision language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(16): 18135-18143.

[45] Lu J, Rao J, Chen K, et al. Evaluation and mitigation of agnosia in multimodal large language models[J]. arXiv preprint arXiv:2309.04041, 2023, 3.

[46] Zhang Y, Huang X, Ma J, et al. Recognize anything: A strong image tagging model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 1724-1732.

[47] An W, Tian F, Leng S, et al. Mitigating object hallucinations in large vision-language models with assembly of global and local attention[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 29915-29926.

[48] Sun Z, Shen S, Cao S, et al. Aligning Large Multimodal Models with Factually Augmented RLHF[C]//Annual Meeting of the Association for Computational Linguistics. 2024.

[49] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback[J]. Advances in neural information processing systems, 2020, 33: 3008-3021.

[50] Yu T, Yao Y, Zhang H, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13807-13816.

[51] Xie Y, Li G, Xu X, et al. V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 13258-13273.

[52] Fu Y, Xie R, Sun X, et al. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization[J]. arXiv preprint arXiv:2411.10436, 2024.

[53] Yu T, Zhang H, Yao Y, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness[J]. arXiv e-prints, 2024: arXiv: 2405.17220.

[54] Zhang J, Wang T, Zhang H, et al. Reflective instruction tuning: Mitigating hallucinations in large vision-language models[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 196-213.

[55] Liu S, Zheng K, Chen W. Paying more attention to image: A training-free method for alleviating hallucination in lvlms[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 125-140.

[56] Huo F, Xu W, Zhang Z, et al. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models[C]//The Thirteenth International Conference on Learning Representations.

[57] Park Y, Lee D, Choe J, et al. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(6): 6434-6442.

[58] Suo W, Zhang L, Sun M, et al. Octopus: Alleviating hallucination via dynamic contrastive decoding[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 29904-29914.

[59] Favero A, Zancato L, Trager M, et al. Multi-modal hallucination control by visual information grounding[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14303-14312.

[60] Zhao L, Deng Y, Zhang W, et al. Mitigating Object Hallucination in Large Vision-Language Models via Image-Grounded Guidance[C]//Forty-second International Conference on Machine Learning.

[61] Rohrbach A, Hendricks L A, Burns K, et al. Object Hallucination in Image Captioning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 4035-4045.

[62] Li Y, Du Y, Zhou K, et al. Evaluating Object Hallucination in Large Vision-Language Models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 292-305.

[63] Lovenia H, Dai W, Cahyawijaya S, et al. Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models[C]//Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR). 2024: 37-58.

[64] Wang J, Wang Y, Xu G, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation [J]. arXiv preprint arXiv: 2311.07397, 2023.

[65] Jing L, Li R, Chen Y, et al. Faithscore: Fine-grained evaluations of hallucina-

tions in large vision-language models [C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 5042-5063.

[66] Chen Z, Zhu Y, Zhan Y, et al. Mitigating hallucination in visual language models with visual supervision[J]. arXiv preprint arXiv:2311.16479, 2023.

[67] Liu F, Guan T, Li Z, et al. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models[J]. arXiv preprint arXiv:2310.14566, 2023, 2(3): 12.

[68] Wang L, He J, Li S, et al. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites[C]//International Conference on Multimedia Modeling. Cham: Springer Nature Switzerland, 2024: 32-45.

[69] Ben-Kish A, Yanuka M, Alper M, et al. Mocha: Multi-objective reinforcement mitigating caption hallucinations[J]. arXiv preprint arXiv: 2312.03631, 2023, 2.

[70] Ye-Bin M, Hyeon-Woo N, Choi W, et al. Beaf: Observing before-after changes to evaluate hallucination in vision-language models[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 232-248.

[71] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Cham: Springer International Publishing, 2014: 740-755.

[72] Chen Z, Zhao Z, Luo H, et al. Halc: Object hallucination reduction via adaptive focal-contrast decoding[J]. arXiv preprint arXiv:2403.00425, 2024.

[73] Wang X, Pan J, Ding L, et al. Mitigating hallucinations in large vision-language models with instruction contrastive decoding[J]. arXiv preprint arXiv:2403.18715, 2024.

[74] Cao J, Chen Z, Wang Z, et al. When Images Speak Louder: Mitigating Language Bias-induced Hallucinations in VLMs through Cross-Modal Guidance[J]. arXiv preprint arXiv:2510.10466, 2025.

[75] Liu S, Yang S, Fang D, et al. Vision-Language Introspection: Mitigating Overconfident Hallucinations in MLLMs via Interpretable Bi-Causal Steering[J]. arXiv preprint arXiv:2601.05159, 2026.

[76] Cheng Dawei, Jia Renjun, Li Jiangtong, et al. Research on knowledge-augmented Chinese financial large language model [J]. BIG DATA RESEARCH, 2025, 11(02): 5-18.

[77] Zhao Botao, Kang Zuheng, Qu Xiaoyang, et al. Review and emerging trends of embodied agent based on multimodal large language models[J]. BIG DATA RESEARCH, 2025, 11(03): 108-138.

[78] Guojie LI. Big data and computing models [J]. Big data research, 2024, 10(1): 9-16.

作者简介

张旭龙（1988– ），男，博士，平安科技（深圳）有限公司，算法研究员，复旦大学计算机理学博士，2023年入选上海市东方英才计划青年项目。兼任清华大学深圳研究院及中国科学技术大学先进技术研究院校外导师，以及中国指挥与控制学会具身智能专业委员会委员、中国自动化学会联邦数据与联邦智能委员会委员、中国电子音响行业协会声音与音乐技术专委会常务委员。目前是IEEE、中国自动化学会以及中国计算机学会会员。主要研究方向包括大模型、具身智能、跨模态智能计算等。发表学术论文80余篇，获得国家发明专利授权20余项。联系邮箱：zhangxulong@ieee. org

潘鹏（2003– ），男，硕士，平安科技（深圳）有限公司，算法工程师（实习），中国科学技术大学硕士研究生在读。主要研究方向为多模态大模型，VQA等。

瞿晓阳（1988–），男，博士，平安科技前沿机器学习算法分组分责人，清华大学深圳国际研究生院校外导师，中国科技大学校先进技术研究院校外导师，中佛罗里达大学访问学者，华中科技大学博士，一直从事机器学习、大数据、体系结构方面的研究工作，在语音语义分析、自动化机器学习、零样本和小样本学习、高性能计算与存储等方面经验丰富。近几年，在体系结构方向（例如：INFOCOM、DAC、TPDS、IPDPS）和人工智能方向（例如：NeurIPS、IJCAI、ICASSP、Interspeech）等国际顶级会议和顶级期刊发表过近50篇文章，其中1篇论文荣获会议最佳学生论文奖提名，担任过多个国际顶级期刊的评委，已授权的专利70篇，已出版的专著2本。联系邮箱：quxiaoy@gmail. com

倪晓俊（1973–），男，硕士，现任中科院计算技术研究所助理研究员。主要从事数据库安全、大数据分析、隐私计算等方面的研究工作。曾获得北京市科技进步一等奖等奖励。jamesni@sina. com

田晖，男，博士，教授，博士生导师；现任华侨大学计算机科学与技术学院院长、网络与信息安全产业学院院长、厦门市数据安全与区块链技术重点实验室主任；入选福建省青年拔尖人才、福建省百千万人才工程省级人选、福建省高校新世纪优秀人才、福建省高校杰出科研人才培育计划。现为IEEE 高级会员（Senior Member）、中国计算机学会（CCF）杰出会员、CCF互联网专委会执行委员、CCF区块链技术专委会执行委员、中国中文信息学会大数据安全与隐私保护专委会常务委员、福建省计算机学会副理事长、福建省网络空间安全学会理事长。主要从事网络与信息安全、人工智能安全，数据安全、多媒体内容安全等方面的研究工作。近年来，主持各类科技项目20余项，其中包括国家自然科学基金项目4项，国家重点研发课题和子课题各1项；在IEEE TDSC、IEEE TIFS、 IEEE TASLP、 ACM MM等国内外著名期刊和知名学术会议上发表学术论文120余篇；获得授权发明专利40余项，软件著作权8项；部分研究成果曾荣获中国电子学会科学技术奖二等奖、海南省自然科学二等奖、福建省科学技术进步奖三等奖和福建省自然科学奖三等奖等奖励。htian@hqu. edu. cn

王健宗（1983– ），男，博士，正高级工程师，平安科技（深圳）有限公司，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会杰出会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为具身智能、大模型、联邦学习和深度学习等。荣获中国专利优秀奖、广东省科技进步二等奖、吴文俊人工智科技进步一等奖、人民银行金融发展一等奖、深圳高层次专业人才领军人才、福田英才等奖项。联系邮箱：jzwang@188.com

收稿日期: XXXX–XX–XX

通信作者: 潘鹏，apical@mail.ustc.edu.cn

基金项目: 深港联合资助项目(A类)(No. SGDX20240115103359001)

Foundation Items: Shenzhen–Hong Kong Joint Funding Project (Category A) under Grant No. SGDX20240115103359001