计算机科学
COMPUTER SCIENCE

多模态大语言模型的安全性研究综述

陈晋音, 席昌坤, 郑海斌, 高铭, 张甜馨

引用本文

陈晋音, 席昌坤, 郑海斌, 高铭, 张甜馨. 多模态大语言模型的安全性研究综述[J]. 计算机科学, 2025, 52(7): 315-341.

CHEN Jinyin, XI Changkun, ZHENG Haibin, GAO Ming, ZHANG Tianxin. Survey of Security Research on Multimodal Large Language Models [J]. Computer Science, 2025, 52(7): 315-341.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241100141

多模态大语言模型的安全性研究综述

陈晋音[1,2] 席昌坤[1] 郑海斌[1,2,3] 高 铭[1] 张甜馨[1]
1 浙江工业大学信息工程学院 杭州 310023
2 浙江工业大学计算机科学与技术学院、软件学院 杭州 310023
3 四川大学数据安全防护与智能治理教育部重点实验室 成都 610000
(chenjinyin@zjut.edu.cn)

摘 要 随着大型语言模型的快速发展,多模态大语言模型因其在语言、图像等多种模态上的卓越表现而备受瞩目。其不仅在日常工作中成为用户的得力助手,还逐渐渗透到自动驾驶、医学诊断等各大应用领域。与传统的大型语言模型相比,多模态大语言模型由于更接近于多资源的现实世界应用以及多模态处理的复杂性而具有巨大的潜力和挑战。然而,多模态大语言模型的脆弱性研究相对较少,这些模型在实际应用中面临着诸多安全性挑战。为此,对多模态大语言模型尤其是大型视觉-语言模型的安全性进行了全面调查。首先,概述了多模态大语言模型的基本结构和发展历程;其次,讨论了多模态大语言模型在使用全周期的安全风险成因,分析了模型结构与安全风险之间的关联性;再次,系统总结了当前在多模态大语言模型图像和文本安全性的评估方面所做的工作,包括模型幻觉、隐私安全、偏见和鲁棒性4个方面,并将针对多模态大语言模型的攻击分为越狱攻击、对抗攻击、后门攻击和中毒攻击;然后,综合概述了一系列针对多模态大语言模型幻觉、隐私泄露和偏见等威胁的可信增强方法以及针对模型恶意攻击的防御措施;最后,讨论了多模态大语言模型安全性研究的主要机遇与挑战,为研究人员在多模态大语言模型的复杂应用和研究领域提供了指导建议。

关键词:模态大语言模型;安全;幻觉;对抗;越狱;防御
中图分类号 TP391

Survey of Security Research on Multimodal Large Language Models

CHEN Jinyin[1,2], XI Changkun[1], ZHENG Haibin[1,2,3], GAO Ming[1] and ZHANG Tianxin[1]
1 College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China
2 College of Computer Science and Technology College of Software, Zhejiang University of Technology, Hangzhou 310023, China
3 Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University, Chengdu 610000, China

Abstract With the rapid development of large language models, multimodal large language models have garnered attention for their outstanding performance across various modalities, such as language and images. These models have not only become valuable assistants in daily tasks but are also gradually penetrating major application areas, such as autonomous driving and medical diagnosis. Compared to traditional large language models, multimodal large language models possess enormous potential and challenges due to their closer alignment with real-world applications involving multiple resources and the complexity of multimodal processing. However, research on the vulnerabilities of multimodal large language models is relatively limited, and these models face numerous security challenges in practical applications. This paper aims to provide a comprehensive survey of the security aspects of multimodal large language models, particularly large vision-language models. Firstly, the basic structure and development history of multimodal large language models are summarized. Then, the causes of security risks throughout the full lifecycle of these models are discussed, and the correlations between model structure and security risks are analyzed. Next, this paper systematically summarizes current efforts in evaluating the security of multimodal large language models in terms of image and text security, including model hallucinations, privacy security, bias, and robustness. Attacks on multimodal large language models are divied into jailbreak attacks, adversarial attacks, backdoor attacks, and poisoning attacks. Furthermore, the paper provides a compre-

到稿日期:2024-01-20 返修日期:2025-01-02
基金项目:国家自然科学基金(62406286);浙江省自然科学基金(LDQ23F020001);四川大学数据安全防护与智能治理教育部重点实验室放课题(SCUSAKFKT202402Z);北京生命科技研究院有限公司开放基金(2024200CD0210)
This work was supported by the National Natural Science Foundation of China(62406286), Zhejiang Provincial Natural Science Foundation (LDQ23F020001), Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University(SCUSAKFKT202402Z) and Beijing Life Science Academy(BLSA)(2024200CD0210).
通信作者:郑海斌(haibinzheng320@gmail.com)

hensive overview of a range of trustworthy enhancement methods addressing threats such as hallucinations, privacy leaks, and bias in multimodal large language models, as well as defense mechanisms against malicious attacks on the models. Finally, the main opportunities and challenges in the security research of multimodal large language models are discussed, and guidance and recommendations are provided for researchers in the complex applications and research areas of multimodal large language models.

Keywords Multimodal large language models, Security, Hallucinations, Adversarial, Jailbreak, Defence

1 引言

近年来，随着大型语言模型（Large Language Models，LLMs）和大型视觉模型（Large Visual Models，LVMs）的快速发展，多模态大语言模型（Multimodal Large Language Models，MLLMs）逐渐成为人工智能领域的热点。这些模型通过结合语言、视觉等多种模态，展示出令人瞩目的能力，如指令跟随[1]、上下文学习[2]和思维链推理[3]等，使得它们在自然语言处理、图像生成、视觉理解等任务上表现优异。LLMs在文本生成和语言理解方面具备强大的推理能力，但只能处理离散的文本数据，其在处理多模态内容时仍然存在局限。而视觉模型则能够胜任多种复杂的视觉任务，包括图像分类、物体检测、语义和实例分割以及图像生成，但其推理能力常常受到计算资源的限制。基于LLMs与LVMs的这种互补性，多模态大语言模型逐渐成为研究热点。通过模态融合，MLLMs实现了语言和视觉之间的跨模态对齐与互通，扩展了其在多模态理解和生成任务中的应用场景，推动了自动驾驶[4]、医疗影像分析[5]、教育[6]、生物科技[7]等领域的发展。

尽管MLLMs在技术上取得了突破，但其在实际应用中的安全性问题也引发了越来越多的担忧。MLLMs的安全隐患主要体现在4个方面。首先，MLLMs在开发和训练中依赖大规模的跨模态数据集，而这些数据集往往包含多种来源的信息，因此容易带来数据偏见[8]和隐私泄露[9]的风险，不法分子可以从模型中提取包含敏感内容的图像或文本，进一步泄露个人身份信息。其次，模型在生成内容时可能出现幻觉[10]，即生成的文本与视觉输入不一致或生成虚假信息，这种现象可能导致模型输出误导性内容。再次，MLLMs还可能成为恶意攻击的目标[11]，尤其是在敏感领域，如攻击者可以对内容进行对抗性扰动，使自动驾驶系统错误识别交通标志，导致交通事故。最后，其也可能被欺骗，生成虚假新闻和误导性信息，散布虚假标签的新闻图片及文本，从而引发社会骚动和公众误解。此外，2023年中国首部大模型监管法规《生成式人工智能服务管理暂行办法》发布，国内发布首份“大模型安全实践”研究报告《大模型安全实践(2024)》，欧盟议会审议通过《人工智能法案》，美国联邦贸易委员会于2023年发起了首个针对人工智能聊天机器人带来风险的审查等，这些都对MLLMs的安全使用提出了新的要求。

目前已有针对LLMs的安全评估工作[12-14]，但对其他模态的评估不足。还有一些工作总结了MLLMs在图像和文本上的攻击与评估[15-16]，但缺乏对幻觉、偏见、隐私等安全维度的概括。与之前的工作相比，本文在具体安全维度和攻击类型上有更细化的评估框架，对多模态大语言模型全周期中的安全风险有更深入的理解。本文概述了多模态大语言模型的基本结构和目前主流多模态大语言模型的发展历程，讨论了多模态大语言模型在数据预处理阶段、预训练模型阶段、价值观对齐阶段和大模型模型推理阶段中的安全风险成因，分析了模型结构与安全风险之间的关联性，系统总结了当前在多模态大语言模型图像和文本安全性评估方面所做的工作，包括模型幻觉、隐私安全、偏见和鲁棒性4个方面，并将针对多模态大语言模型的攻击分为越狱攻击、对抗攻击、后门攻击和中毒攻击。此外，还综合概述了一系列针对多模态大语言模型幻觉、隐私泄露和偏见等威胁的防御措施，并提出与MLLMs安全性研究相关的机遇与挑战。详细的安全性分析维度如图1所示。

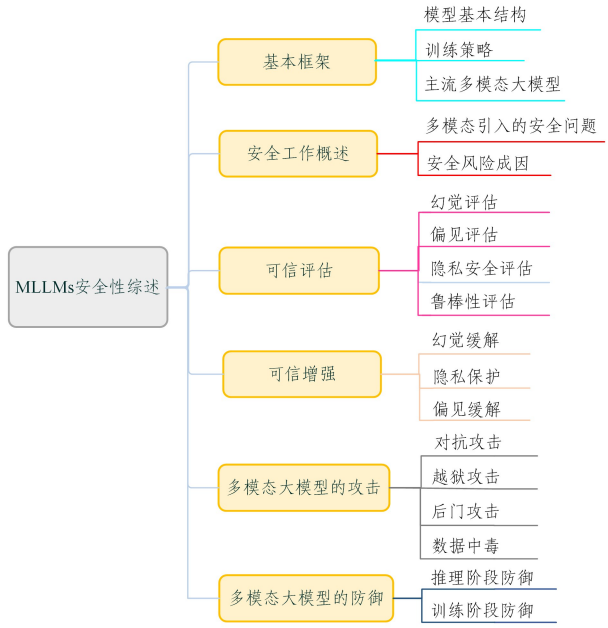

图1 安全性分析维度

Fig. 1 Security analysis dimensions

本文的主要贡献如下：

1）首次针对多模态大语言模型安全性的研究进行了综述，对MLLMs开发和使用全周期中可能面临的安全威胁开展了全面的讨论；

2）定义了偏见、幻觉、隐私、鲁棒性4个安全维度，提出了针对多模态大模型的安全评估体系；

3）全面概述了MLLMs攻击与防御技术现状，对不同类型的攻击进行了分类和总结，强调了攻击的方法、影响以及所利用的潜在漏洞；

4）总结了11条MLLMs安全性研究的发展机遇与挑战，以期指导研究人员和相关从业者发掘未来的研究方向。

2 多模态大模型基本框架

多模态大模型旨在使机器能够理解、生成和处理多种模态信息，如图像、文本、音频等。MLLMs的关键任务是跨模

态的理解与生成，模型通常建立在模态对齐和融合的基础上。本章将从多模态模型的基本结构、训练策略和当前主流模型3个角度对MLLMs进行介绍。

2.1 多模态大语言模型基本结构

一个典型的多模态大型语言模型可以分为3个模块：预训练的模态编码器、预训练的大型语言模型，以及连接它们的投影器。一些MLLM还包括一个生成器，用于输出除文本之外的其他模态。图2展示了主流MLLM的主体架构。本节将依次介绍每个模块。

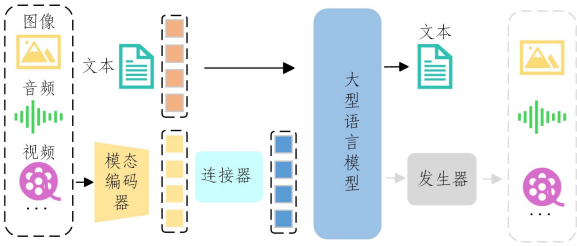

图2 主流MLLMs结构

Fig. 2 Main MLLMs structures

2.1.1 模态编码器

在多模态大型语言模型中，模态编码器负责将来自不同模态的输入转换为模型可以理解的特征表示。具体地，编码器通过压缩原始信息，如图像、音频或者视频，将其转换为更紧凑的表示形式。

图像编码器是用于处理视觉信息的核心组件，其主要任务是将图像数据转换为模型可以进一步分析的特征表示。这一过程通常涉及特征提取，即从图像中提取有用的信息，如颜色、纹理、形状和对象的位置等。目前主流的图像编码器分为两大类：基于卷积神经网络的图像编码器和基于Transformer架构的图像编码器。

基于卷积神经网络的图像编码器（如ResNet[17]，EfficientNet[18]和NFNet[19]）以高计算效率和训练稳定性著称。这些网络通过层层卷积操作，逐步提取图像中的有用特征，在许多视觉任务中被广泛使用。NFNet通过创新的自适应梯度裁剪技术，进一步提高了模型训练的稳定性和性能。

基于Transformer的图像编码器（如ViT[20]，Swin Transformer[21]和EVA[22]）则利用全局注意力机制，对图像的不同区域进行全面关联建模，尤其在大规模数据集上表现优异。ViT通过将图像分割为若干图像块，并使用Transformer处理这些块，获得图像的全局特征。Swin Transformer则通过分层窗口注意力机制，兼顾局部和全局特征提取。

CLIP ViT[23]是一种典型的Transformer架构的图像编码器，其通过大规模图像-文本对数据进行预训练，使用对比学习将图像和文本投射到一个共享的特征空间中，从而实现多模态语义对齐。因此，CLIP ViT成为许多多模态大模型的首选，尤其适用于跨模态检索和理解任务。在此基础上，Eva-CLIP ViT[24]对CLIP进行了改进，通过优化训练技巧，如增强数据预处理和更高效的训练策略，使其视觉编码器在多模态任务中的表现更加出色。MiniGPT-4[25]就采用了Eva-CLIP ViT编码器，并在模型训练中使用了这些改进策略。

编码器不仅可以处理图像模态，还可以应用于其他模态，如音频模态和视频模态。以CLAP[26]为例，它是一种基于对比学习的预训练方法，旨在结合音频数据和相应的自然语言描述来学习音频特征表示。CLAP的核心思想是利用对比学习，将音频和文本映射到一个共享的潜在空间中，在训练过程中通过拉近相关的音频-文本对的距离，拉远不相关对的距离，从而实现模态之间的对齐和理解。CLAP模型的设计灵感来源于CLIP，CLIP通过大规模图像-文本对进行预训练，学习到图像与文本的语义对齐。CLAP采用类似的方法，将这种对比学习的思路从图像-文本迁移到音频-文本对上，构建了一个多模态的表示学习框架，使得音频和文本能够在相同的特征空间中进行表示，从而使模型具备良好的多模态对齐和检索能力。这种方法特别适合音频-文本检索、音频描述生成等任务。目前，主流的音频编码器还有Whisper[27]和HuBERT[28]等，这些音频编码器的共同特点是，它们通过处理音频信号，将其转换为可以与语言模型交互的特征表示。这些编码器在多模态任务中极为重要，因为它们为不同模态的数据提供了统一的特征表示，进而使得多模态大模型能够处理音频、文本甚至图像的联合信息，实现复杂的跨模态理解和生成能力。

视频编码器则更为复杂，因为它需要同时处理图像和时间序列数据。视频由一系列图像帧组成，编码器不仅需要提取每一帧的视觉特征，还需要理解这些帧随时间的变化，例如运动信息。在多模态架构中，视频编码器可能会使用类似于图像编码器的技术来处理每一帧，同时还会使用额外的技术来处理帧与帧之间的关系。常用的视频编码器有ViViT[29]和VideoPrism[30]等。

2.1.2 预训练大语言模型

目前主流的预训练大语言模型涵盖多个领域，支持自然语言处理、多模态任务等应用。它们主要可以分为Transformer模型和混合专家模型[31]（Mixture-of-Experts，MoE），每种类型的模型在不同应用场景中展现出独特的优势。

Transformer模型包括GPT系列[32]、BERT系列、T5系列[33]、LLaMA系列[34]、Vicuna系列[35]、Qwen系列[36]、PaLM系列和BLOOM。这些模型通常在大规模语料上进行预训练，因而具备丰富的语言知识和强大的推理能力。例如，GPT-3和GPT-4被广泛应用于语言生成和多模态任务；Flan-T5在多模态任务中的表现也非常优秀；LLaMA和Vicuna系列因其开源和灵活性，受到了学术界的广泛关注；Qwen作为双语模型，能够支持中文和英文，使得跨语言理解和生成更加便捷；BLOOM是一个开源的多语言大模型，致力于为全球研究人员提供免费的生成工具。

在计算资源有限的环境下，一些轻量级模型也得到了发展，例如MobileVLM系列。这些模型通过对LLaMA等模型进行缩小和优化，使其在移动设备上也可以进行高效的自然语言处理和推理任务，非常适合移动端应用。

混合专家模型如MoE-LLaVA，通过选择性激活部分专家来实现参数量的扩展，不会显著增加计算开销。相比传统的Transformer模型，MoE架构在高效推理和多模态任务中表现出色。

总体来看，Transformer模型以其丰富的知识储备和强大

的推理能力，在各种自然语言和多模态任务中得到了广泛应用。混合专家模型则通过稀疏激活和参数扩展，实现了更高的推理效率和更好的任务表现。

2.1.3 投影器

投影器负责将不同模态的特征表示对齐并转换到共享的表示空间，使模型能够处理和融合来自不同模态的信息。这些投影器为多模态特征的跨模态对齐奠定了基础，使得后续的统一理解和生成成为可能。当前主流的投影器包括线性投影器、多层感知机(MLP)、交叉注意力机制、Q-Former[37]以及 P-Former[38]等。

线性投影器和 MLP 是相对基础的投影器，主要用于简单的特征转换和对齐任务。线性投影器通过矩阵乘法将输入特征映射到目标特征空间，具有很高的计算效率，因此常用于对计算资源要求较高的场景中。LLaVA-1.5 则采用 MLP 作为输入投影器，通过多层的非线性变换对视觉特征进行更复杂的处理，以增强特征的适应性。

交叉注意力机制是一种更高级的投影方法，通过学习一组查询与不同模态的特征进行交互，从而生成融合后的表示。交叉注意力特别适合需要深度融合和高效特征对齐的任务，被广泛应用于图像-文本、音频-文本对齐等场景中。例如 mPLUG-Owl[39]和 Qwen-VL[40]等模型，均能够捕捉不同模态之间的复杂关系并实现更深层次的对齐。

Q-Former 和 P-Former 是近年来非常流行的输入投影器，用于提高特征对齐的精确性和效率。Q-Former 通过少量可学习的查询与输入模态的特征交互，能够有效地提取重要信息，在减少计算复杂度的同时保持特征信息的完整性。它在需要高效对齐和特征提取的场景下表现非常出色，特别是在多模态模型中对图像特征进行精简和转换的任务中，被广泛应用于 MiniGPT-4 和 BLIP-2 等模型中。P-Former 则是在 Q-Former 的基础上进一步改进，通过生成参考提示来提高模态对齐的精度，特别适合需要更深层次对齐和特征交互的应用，如 DLP 和 BuboGPT 等模型。

MQ-Former[41]是 Q-Former 的增强版本，旨在处理多尺度特征的融合任务。它通过引入不同尺度的注意力机制，可以更精细地对齐来自视觉模态的多层次信息，使得模型在复杂的多模态任务中能够捕捉到更全面的特征，尤其适合那些涉及不同分辨率或层次的特征融合任务。

不同类型的投影器为多模态特征的对齐、转换和融合提供了多样化的解决方案。从简单高效的线性投影器和 MLP，到适用于深度交互的交叉注意力，再到能够精细提取特征的 Q-Former 和 P-Former，这些输入投影器为多模态模型的复杂特征表示奠定了稳固的基础，使得模型可以在特征对齐、理解、生成等任务中实现更优异的性能。

2.2 训练策略

一个完整的多模态大模型需经过预训练、指令微调和对齐微调 3 个阶段。首先通过预训练模型来学习不同模态的数据表示和关系；然后通过指令微调使模型理解和执行用户任务指令；最后通过对齐微调使模型根据人类反馈调整，以确保输出符合伦理和安全的要求。

2.2.1 预训练

多模态大模型的预训练过程主要在跨模态的数据上进行训练，如图像-文本、音频-文本对，以学习和对齐不同模态之间的关系。

1)预训练方法

预训练阶段通常采用自回归式的生成训练策略，目标是通过预测序列中的下一个词来优化跨模态对齐能力。具体来说，输入端首先引入图像-文本对，其中视觉编码器提取的特征通过模态接口处理后，与语言模型的嵌入进行拼接，然后通过语言模型生成对应的文本描述。在优化过程中，通常使用交叉熵损失作为目标函数，逐词计算模型生成文本与真实文本之间的差异，从而引导模型不断学习视觉与语言模态之间的关联关系。为了提升训练的效率和性能，预训练阶段还常常会冻结视觉编码器和语言模型的参数，仅训练模态接口以减少计算成本，同时保留已有模块的知识[42-44]。如果任务需要更高的模态对齐能力，也可以解冻部分模块(如视觉编码器)以增加可训练参数，实现更灵活的模态对齐[40,45]。

2)数据

在预训练阶段，数据的质量和多样性对多模态大模型的性能起到了至关重要的作用，特别是在对齐不同模态(如视觉与语言)时，数据的粒度、来源和清洗策略都会直接影响模型的学习能力和泛化效果。从数据粒度的角度来看，预训练数据通常分为粗粒度数据和细粒度数据两类。粗粒度数据以规模庞大为特点，通常从网络抓取得到(如 CC-3M[46]，CC-12M[47]和 LAION[48]系列)，其图像-文本对的描述内容往往简短且存在噪声。为了提升这些数据的质量，研究人员采用自动化的清洗流程，如利用 CLIP 模型对图像和文本进行嵌入匹配，筛选出具有高语义相似度的样本，或剔除低质量和内容不相关的图像-文本对。同时，还会过滤掉不符合特定格式或尺寸的图像，以及可能含有非法或重复内容的数据。尽管粗粒度数据在数量上具有明显优势，但其文本描述的简短性和噪声问题限制了模型对精细语义的捕获能力。

相比之下，细粒度数据通常由更高质量的人工生成或通过强大的多模态模型(如 GPT-4V)生成。这些数据不仅描述更为详尽，而且在对齐视觉与语言模态时更加精确，能够为模型提供更细粒度的监督信号。例如，Share-GPT4V-PT[49]数据集通过利用 GPT-4V 生成更长、更准确的描述，以显著增强图像-文本的对齐能力；ALLaVA 数据集[50]则通过多模态问答形式提供了高质量的训练样本，进一步提升了模型对复杂语义的理解能力。然而，细粒度数据的生成依赖于强大的生成模型或人工标注，其成本较高且规模相对较小。为了在粗粒度和细粒度数据之间取得平衡，ShareGPT4V 利用预训练模型生成补充描述，以提高粗粒度数据的质量。

此外，输入图像的分辨率也是影响数据效果的重要因素。对于简短和噪声较大的描述，使用低分辨率图像可以加速训练；对于描述精细、质量较高的数据，采用高分辨率图像可以捕获更多视觉细节，从而降低模型生成幻觉的风险。

2.2.2 指令微调

经过预训练之后的 LLM 具有广泛的知识储备，拥有强大的自然语言推理和代码处理能力，但在某些任务上的 Zero-

Shot 能力很差[51]。为了进一步提高 LLM 在未见任务上的指令泛化能力，即 Zero-Shot 能力，需要在指令数据上微调预训练模型。

1）微调过程

微调过程包括多个关键步骤，首先是指令理解，模型需要通过学习不同指令格式和自然语言表达来识别用户的任务目标，并将这些指令转换为可执行的目标操作。接着，模型需要对多模态输入进行深度融合，使得视觉、文本等模态的信息在共享特征空间中有效对齐，以便根据指令要求生成输出。这一过程通常通过可学习提示或交叉注意力机制来实现，从而使模型能够深度结合视觉和语言特征。在训练过程中，模型通过学习在给定指令和输入条件下生成的目标输出，从而优化模型的自动回归目标，即预测响应的每一个后续 token 的条件概率。合理设计多轮对话的模板，使模型能够在多轮交互场景中生成连贯的回答[42]。此外，指令微调还涉及利用不同类型的损失函数来优化模型性能，如交叉熵损失用于语言生成，对比损失用于模态之间的相似性对齐等。通过监督微调，模型学习了如何生成合适的输出，而基于人类反馈的强化学习进一步优化了生成质量，使得模型的表现更符合人类期望。

2）数据

指令数据的收集可以通过数据适配、自我指导和数据混合 3 种方式进行。数据适配是通过将现有的高质量任务数据集转换为指令格式[52-59]。例如，将视觉问答数据集中的输入（图像和问题）和输出自然转换为多模态输入和响应，同时结合任务描述来丰富数据的指令格式。然而，现有的数据集通常仅包含简短的回答，这可能会限制模型生成长答案或详细描述的能力。一种解决策略是在指令中明确指定答案的长度和风格，例如要求“简短回答”或“详细描述”[56]；另一种解决策略是通过上下文提示扩展原始答案的长度，例如利用图片的标题或其他相关信息提示生成更丰富的描述[57]。

自我指导是另一种高效生成指令数据的方法，通过少量人工标注的示例引导大模型生成多模态指令数据[58]。这种方法被广泛用于生成高质量多模态数据集，例如将图像转换为文本描述，并提示语言模型生成与任务相关的数据集。这种方式不仅能够生成多轮对话，还可以为不同任务需求设计更加复杂的任务指令，确保数据的广泛适用性。数据混合则是通过融合单模态对话数据和多模态指令数据来增强模型的任务处理能力，例如随机从单模态和多模态数据集中抽取样本[60]，或通过顺序训练方式逐步引导模型学习语言与多模态数据的结合[61]。

在数据质量方面，指令的多样性和任务覆盖范围尤为重要。丰富的指令格式和提示内容能够显著提高模型的泛化能力[62]，尤其是在复杂的推理任务中。此外，为了保证数据的一致性和高质量，过滤低质量样本和优化数据清洗过程至关重要。例如，通过使用预训练的模型筛选不相关或质量较差的样本，可以有效提高数据的整体质量。同时，研究表明，较少但高质量的训练数据比大规模但噪声较高的数据对模型性能的提升更为显著[56]。这表明数据的设计和质量控制在模型性能优化中起到核心作用，通过结合多样化的收集方法与严格的数据质量管理，可以最大限度地提升多模态大模型的任务执行能力和对指令的理解深度。

2.2.3 对齐微调

通过指令微调，模型能够进一步学习如何根据任务指令生成更符合需求的输出，增强任务的完成能力，但模型仍可能在实际应用中出现幻觉或偏离用户期望的情况。对齐微调的核心目标是通过让模型的行为与人类的期望保持一致来降低这些错误的风险，从而确保其在复杂任务中的表现更为稳健。

1）微调过程

这一过程主要通过两种技术实现：基于人类反馈的强化学习（RLHF）[63]和直接偏好优化（DPO）[64]。RLHF 方法由 3 个核心步骤组成。首先是监督微调，使用标注数据对预训练模型进行初步微调，使模型能够生成符合期望的输出行为。微调后的模型被称为策略模型。然后是奖励建模，通过人类偏好对模型的生成结果进行评分，训练奖励模型，使其能够为优选答案分配更高的奖励分数。具体来说，对于一个输入及其两个生成结果（一个偏好更高，另一个较低），奖励模型根据奖励分数差异进行学习。最后，强化学习通过优化策略模型进一步对齐模型的输出与人类偏好。此阶段通常采用近端策略优化（PPO）算法，并加入 KL 散度惩罚项，以防止模型生成偏离参考策略模型的过度响应。

与 RLHF 相比，DPO 方法更简化，其核心目标是通过一个二分类损失函数直接学习人类偏好，不需要显式构建奖励模型。这一方法直接利用偏好数据对策略模型进行优化，从而使模型生成的输出更符合人类需求。DPO 的学习过程相对直观，通过比较两个候选答案的概率分布，优化模型参数，以更好地对齐偏好选择。

2）数据

在训练数据的准备上，对齐微调通常依赖于少量高质量的人类反馈数据或使用强大的大模型（如 GPT-4V）生成的人工反馈数据[65]。数据量虽然较小，但对模型性能提升至关重要。这些数据不仅需要准确标注哪一个答案更优，还需要涵盖丰富的场景，例如减少生成中的幻觉现象、提升输出内容的可信度和帮助性[66]。此外，还需结合人类直接标注和多模态模型生成的高质量偏好数据，进一步优化模型的对齐能力，使其更加符合实际应用中的需求。

2.3 主流多模态大模型

近年来，多模态大模型取得了显著的进展，其研究重点从多模态内容的理解逐步扩展到更复杂的多模态生成任务。这些大模型最初主要集中在图像-文本、视频-文本和音频-文本的理解任务上，整合了不同模态的数据进行智能分析与生成。

BLIP-2 作为多模态预训练模型的典型代表，由 Salesforce 研发，通过融合视觉编码器和语言模型，实现了高效的图像-文本理解。BLIP-2 在图像描述、视觉问答和图文对话中表现优异，采用了双阶段的预训练方法。首先通过视觉编码器提取图像特征，然后与语言模型进行对齐，从而提升了多模态数据的融合和理解能力。其被广泛应用于自动化图像标注、增强型视觉搜索、智能问答等领域，推动了视觉和语言模态的高效协作。

MiniGPT-4[67]作为 GPT-4 的轻量化版本，专注于视觉-

文本的理解和生成任务。MiniGPT-4 不仅简化了模型参数，提升了推理效率，还在视觉描述和对话生成方面展现出卓越的性能。这一模型的优势在于结构的优化，其能够在设备资源有限的情况下进行高效的多模态推理与生成，从而适应更多实际应用场景。

LLaVA[42]是另一种具有代表性的多模态大模型，它通过视觉、语言和音频模态的深度融合，实现了多模态数据的联合理解与生成。LLaVA 的技术特点在于其采用了基于 Transformer 的架构，能够在大规模多模态数据上进行训练，并支持跨模态任务的理解与生成。其在视觉问答、图像描述以及更复杂的多模态对话任务中表现出色，有助于提升 AI 系统在实际应用中的理解能力。

在视频和音频模态理解领域，VideoChat[68]和 QwenAudio[69]等模型也展现了国际领先的多模态处理能力。VideoChat 专注于视频-文本任务，通过对视频帧的特征提取和与文本模型的对齐，实现了视频内容的深度分析与文本生成。它在智能监控、视频摘要和多模态交互中具有广泛的应用潜力。QwenAudio 则聚焦音频和文本的跨模态理解与生成，能够将语音输入与文本生成结合，实现复杂的音频-文本交互，如语音识别、语音指令处理和多模态语音助手。

为了实现更强的跨模态协作，VisualChatGPT[70]，HuggingGPT[71]和 AudioGPT [72]等项目尝试将大语言模型与外部工具结合，通过引入第三方工具来扩展模型的功能。这种方式使得多模态大模型能够处理更为复杂的任务。VisualChatGPT 可与图像处理工具结合，实现更精细的图像编辑和生成；HuggingGPT 通过集成 Hugging Face 的多模态模型库，实现多模态数据的深度理解与生成；AudioGPT 整合了语音合成和语音识别工具，使其在音频-文本生成任务上表现出色。

在端到端的多模态模型设计方面，NExT-GPT[73]和 CoDi-2[74]通过端到端架构减小了多层级信息传递的误差，提升了多模态任务的精确性和效率。这些模型采用了联合优化策略，使得跨模态任务的处理更加流畅和一致，并在复杂场景下实现了较高的泛化能力。NExT-GPT 在图像和文本的生成和理解上表现突出，CoDi-2 则在多模态对话和音视频内容的联合分析上具有显著优势。

这些多模态大模型不仅展示了跨模态理解和生成的能力，还为实现任意模态间的转换奠定了基础。它们的核心技术趋势是提升多模态融合的深度和广度，同时通过端到端架构和外部工具的结合来增强任务处理的复杂性。随着多模态技术的不断成熟，这些模型将在工业、医疗、交通和娱乐等更多领域中实现智能化转型，推动通用人工智能的广泛应用，并加速多模态模型在实际场景中的落地和发展。

3 多模态大模型安全工作概述

多模态大模型整合多模态输入的能力为复杂任务提供了强大的解决方案。然而，与单模态的文本大模型相比，多模态模型在安全性上面临更多维度的挑战。本章探讨了多模态大模型的独特安全问题及其成因。

3.1 多模态引入的安全问题

多模态大模型和文本大模型的安全问题存在显著的区别，这种区别源于两者的输入模态、攻击向量、潜在风险及防御复杂性等多方面的特性。文本大模型的输入是离散的、可控的文本数据，其安全问题主要集中在恶意提示注入[75]、数据偏见[76]以及隐私泄露[77]等方面。例如，通过精心设计的语言提示，攻击者可以诱导文本大模型生成不当或有害内容，尤其是在模型对输入语境理解不足的情况下。此外，文本大模型的训练数据若包含偏见，则可能放大这些偏见，导致输出带有歧视性或有害的内容。隐私泄露则是另一大风险，模型可能因训练数据中的信息而泄露用户隐私，如生成包含敏感信息的回答。然而，由于文本数据是离散且结构明确的，现有的防御技术（如基于人类反馈的强化学习 RLHF、对抗性训练等）相对成熟，可以较好地覆盖大多数攻击场景[78]。

相比之下，多模态大模型因具有处理图像、音频、视频和文本等多模态数据的能力，面临着更大的安全挑战。首先，图像和音频等输入是连续的、高维的，其输入空间远比文本复杂，难以通过穷举规则或防御机制全面覆盖所有可能的攻击路径。例如，对抗性图像攻击通过在图片中嵌入细微的像素级修改，可以欺骗模型生成错误的描述或判断[79]。其次，MLLMs 还引入了跨模态攻击的新风险。攻击者可以利用多模态交互的特点，通过将恶意信息嵌入图像（如包含文字信息的图像）或音频，从而绕过文本安全机制。例如，通过光学字符识别（OCR）[80]，文本恶意内容可以隐藏在图片中，以规避文本模块的安全对齐。这种跨模态攻击显著增加了防御难度，因为它利用了模态之间的差异和交互的漏洞。

多模态大模型还容易受到其他模态特有风险的影响。图像模态可能导致隐私泄露，如识别出图像中的面部特征或地理信息[81]；视频和音频模态可能暴露用户的行为模式或环境特征。更复杂的是，多模态数据的不均衡可能导致模型在模态间产生偏见，放大已有的社会不公。例如，模型可能因训练集中男性和女性形象数据的不均衡，生成带有性别偏见的描述。此外，在跨模态生成任务（如图像标题化、视频描述）中，如果一个模态的输入错误，可能会导致整体输出的连锁性失误，使得风险进一步增加。

在防御方面，多模态大模型的复杂性也带来了更大的挑战。MLLMs 需要处理多种模态之间的协同关系，这要求模型在设计和训练时必须确保不同模态的安全对齐。然而，由于图像和音频等模态的输入空间连续且广泛，仅通过简单的监督微调或对抗性训练难以实现全面覆盖[82]。此外，模态间的互相影响可能导致模型的防御机制更加脆弱。例如，一个看似无害的文本输入在结合特定图像后可能触发有害行为。现有的防御技术，如模型微调、数据过滤和安全评估，在单模态场景中可能较为有效，但在多模态场景下还缺乏足够的鲁棒性。

3.2 安全风险成因

多模态大模型的安全风险主要包括产生幻觉、隐私泄露、潜在的社会偏见和受到越狱等攻击等。本节从数据预处理阶段、预训练模型阶段、价值观对齐阶段和大模型推理阶段 4 个阶段介绍多模态大模型的安全风险成因。

3.2.1 数据预处理阶段

在多模态大模型的构建过程中，数据预处理环节面临着多种安全风险，这些风险主要源于数据来源的复杂性、多样性以及预处理工具和流程的局限性。首先，多模态大模型的数据来源广泛，包括社交媒体、新闻网站、电子书、视频平台等，这种多源性导致数据质量参差不齐，某些来源的数据可能包含偏见、不准确的信息，甚至含有恶意内容，从而直接影响模型的学习质量和生成结果。其次，自动化的预处理工具难以完全消除数据中的偏见，特别是在处理来自不同文化背景的数据时，这些偏见往往会被累积甚至放大，导致模型在下游任务中表现出不公正的行为[83]。

近年来，AI生成内容在互联网中的大量渗透也带来了额外的挑战，AI生成的文本、图像或视频可能包含潜在的误导性或恶意信息，这些信息难以通过常规预处理手段过滤，从而影响了模型的可靠性。更隐蔽的威胁来自于数据投毒攻击，攻击者可能在模型训练数据中加入恶意样本[84]，通过预处理阶段的漏洞使这些恶意样本被保留，导致模型在学习过程中建立错误的关联，从而在推理阶段做出错误决策。

在多模态预处理过程中，还存在数据错配的风险，如文本与图像的错误对齐，这种错配可能导致模型学习到错误的模态间关系，进而在推理时生成不合理的结果。与此同时，多模态数据的不完整或缺失也可能导致模型在处理类似数据时产生不可预测的错误，增加了应用场景中的安全性风险。数据预处理工具和流程的安全性同样至关重要，如果这些工具存在漏洞，则可能被恶意利用，进而影响数据的质量和模型的训练效果。

涉及敏感信息的数据预处理还面临隐私泄露的风险，尤其是在处理包含个人隐私的图像或文本数据时。如果脱敏处理不彻底，则可能引发严重的隐私泄露问题[85]。

3.2.2 预训练模型阶段

在预训练阶段，多模态大模型通过训练大量数据来学习对各种模态的理解和跨模态的交互能力。然而，由于数据来源广泛且内容复杂，这一阶段往往存在多种安全隐患，对模型的实际应用造成威胁。

训练数据集规模庞大，来源多样，往往包含偏激、歧视、暴力等不适当的内容。如果这些内容未被有效筛选和处理，模型在学习过程中就可能吸收这些不良信息，导致推理时生成类似的有害内容。Gong 等[80]的研究表明，即使经过一定的安全对齐训练，视觉语言模型在接受带有特定符号的图像输入时，仍可能被诱导生成不安全的响应。这些特定符号的图像可以绕过模型文本部分的安全机制，导致生成的文本带有有害信息。因此，视觉模态的加入并未使模型更加安全，反而为攻击者提供了绕过安全机制的途径，增加了模型生成有害内容的风险。

模型在训练过程中容易学习到模态间的伪相关性，即基于训练数据中高频出现但缺乏实际因果关系的模态共现模式，导致当输入数据中缺乏某些模态时，仍生成虚假的跨模态输出。

对抗攻击的易感性是预训练阶段的另一主要安全风险[86]。在这一阶段，模型缺乏足够的防御能力来应对对抗性样本，从而对安全敏感场景造成重大威胁。对抗攻击是一种通过微小但精确的扰动，使得输入在外观上无害但在模型中引发错误响应的攻击方式。大量开源的多模态模型对基于图像的对抗攻击非常敏感。通过设计特定的对抗性图像输入，攻击者可以诱导模型生成不安全的响应，这种情况在医疗、无人驾驶等需要高度安全性的应用场景中影响很大。例如，在医疗场景中，一张对抗性图像可能使模型对患者情况给出错误的诊断，带来严重的后果。

多模态集成的不稳定性也增加了预训练阶段的复杂性和风险。多模态大模型通过结合视觉和文本特征来提升其理解能力，以实现更复杂的任务。然而，这种不同模态特征的集成并不总是稳定可靠的，尤其是在特征对齐不精确的情况下。Wang 等[87]提出，多模态模型中的视觉和文本特征对齐不当可能会导致信息误解，从而生成不符合预期的，甚至是有害的输出。例如，一幅图像和相应文本描述的结合可能会在语义上发生偏差，导致模型对内容的错误理解。这种错误理解不仅可能使得模型的输出不符合用户的预期，还可能在特定场景中引发安全问题。如果模型错误地理解了一张图像和文本的组合信息，它可能会错误地引导用户执行危险操作。因此，多模态特征的集成如果缺乏有效的对齐机制，则可能导致模型在处理输入时存在严重的误判风险。

3.2.3 价值观对齐阶段

在价值观对齐阶段，多模态大模型通过监督微调（Supervised Fine-Tuning，SFT）或基于人类反馈的强化学习（Reinforcement Learning from Human Feedback，RLHF）等方式进行进一步的训练，以符合人类的价值观和道德标准。这一阶段的对齐过程同样存在许多安全隐患和不确定性。

对齐效果具有复杂性与不确定性，模型的目标是通过大量的训练数据和人类反馈进行微调，以生成更符合人类价值观的输出。然而，这一过程通常需要耗费大量的资源和时间，且对齐过程的复杂性使得其效果难以完全保证稳定性。特别是在多模态模型中，由于涉及到视觉和文本等不同模态之间的语义关联，对齐的难度进一步增加。不同模态之间的特征融合和语义关联复杂，导致模型在对齐过程中可能无法充分理解输入的全部信息，从而使得对齐效果具有不确定性。

Wang 等[88]提出的 InferAligner 方法试图通过推理阶段的跨模型引导来对齐模型的安全性，该方法利用已对齐模型的安全向量来指导目标模型生成安全的响应。然而，这种方法在处理复杂输入时仍然存在一定的效果不确定性，尤其是在多模态输入场景中，安全对齐往往难以应对所有可能的语义组合，导致模型在某些情况下可能生成不符合安全预期的输出。因此，在价值观对齐过程中，尽管投入了大量资源进行训练，但对齐效果可能因为多模态融合的复杂性而存在较大变数，使得模型在一些场景中表现出不稳定的对齐效果。

微调引入的新风险也是价值观对齐阶段需要特别关注的问题。在对齐阶段，经过精心设计的监督微调或人类反馈强化学习，模型可能表现出良好的安全性并生成符合人类价值观的输出。然而当模型进入实际应用场景后，往往需要根据特定应用需求进行进一步的微调，而这种进一步的微调可能会对模型已有的对齐效果产生负面影响。Qi 等[89]的研究表

明，即便是基于良性数据集的微调，也可能导致模型的安全对齐效果被削弱。研究显示，通过少量的额外训练数据对模型进行微调后，模型的原有防御机制可能被破坏，从而更容易生成有害内容。这是因为在微调过程中，模型可能会发生“灾难性遗忘”，即在调整模型行为以适应新任务时，可能无意中削弱了之前训练过程中建立的安全性或价值观对齐能力。这种安全性的降低在模型被部署后尤其危险，因为用户在使用模型时，可能会自行微调以满足其个性化需求，而这种微调的内容和质量无法完全得到控制。例如，某些用户在缺乏安全意识的情况下对模型进行微调，可能使模型忽视某些重要的安全规则，从而使其生成对社会不利甚至危险的内容。因此，微调过程中的安全性降低对模型的整体可靠性和安全性构成了严重威胁。

价值观对齐阶段虽然是保证多模态大模型生成符合人类价值观输出的重要步骤，但其对齐效果的复杂性与不确定性，以及进一步微调可能引入的新风险，都是必须关注的安全问题。在进行对齐训练时，不仅需要充分考虑多模态之间的复杂关联，还需要采取措施确保模型在后续的微调过程中不会破坏原有的对齐效果，特别是在模型被用户自定义和个性化使用的场景中，必须加强对微调过程的监管和安全保障，从而确保多模态大模型在实际应用中的安全性和可靠性，有效减少对齐不当或微调不慎导致的潜在安全隐患。

3.2.4 大模型推理阶段

在多模态大模型的推理阶段，不同模态的输入组合可能导致模型生成不安全的输出，在某些情况下，模型可能过于依赖某一个模态，而忽略了其他模态的重要信息。例如，当视觉信息与文本信息存在冲突时，模型可能会倾向于根据视觉信息做出判断，而忽略了文本中的关键信息[90]。

多模态大模型需要同时处理来自不同模态的输入，虽然单个模态的输入看起来无害，但这些输入结合后可能会产生危险的语义关联，导致模型生成不安全的输出。Wang 等[87]指出，一个看似无害的文本描述与图像结合，可能导致模型在推理时输出危险或有害的响应。这种现象被称为安全输入但不安全输出，表明模型在多模态融合时需要具备识别潜在危险的能力。然而，现有的大模型在应对复杂的跨模态语义组合时常常表现不佳，生成意外的、不符合安全标准的输出。用户文本提示中不准确的主张会误导模型，使其生成与图像内容不符的响应。这种现象类似于“迎合倾向”，即模型倾向于支持用户的观点，而非提供准确答案。当输入多个相似图像时，模型也可能难以区分细微差别，从而生成错误描述。

在视觉诱导攻击中，攻击者通过设计特定的视觉输入诱导模型生成不良或有害的内容。Gong 等[80]通过在图像中嵌入某些信息，使攻击者可以绕过模型的文本安全机制，诱导模型输出违反道德规范或安全规则的内容。这类攻击拓展了模型面临的攻击面，特别是在视觉模块防护不健全的情况下，模型更容易受到这类输入的干扰，生成有害的响应。

数据投毒攻击在训练阶段向模型注入带有恶意触发器的样本，在推理时可以通过特定的触发条件控制模型的输出。这种攻击在多模态场景中特别危险，因为模型在处理视觉和文本等不同输入时，攻击者可以通过多种方式在推理过程中激活后门，使模型输出不符合预期的结果。Xu 等[91]提出的 Shadowcast 攻击通过在训练数据中注入微妙的中毒样本，使模型在推理时生成具有误导性或有害的输出，而不会引起明显的警觉。

尽管模型可能被指令化以避免泄露敏感信息，但实际操作中这些防护措施往往不够充分。攻击者可以利用多跳攻击等策略[92]，通过一系列精心设计的输入逐步诱导模型暴露隐私信息，绕过原有的保护机制，最终导致用户的敏感信息泄露。这种攻击形式尤其危险，因为它不依赖直接的攻击手段，而是通过引导模型逐步泄露信息，所以难以检测和防御。

由于模型在训练阶段依赖大量数据分布，当模型在推理过程中遇到与训练数据分布不一致的场景时，往往表现出不稳定性。在多模态任务中，模型需要同时处理图像和文本等不同模态的输入，这种依赖训练数据的特性使得模型容易忽视实际输入的细节，特别是在面对未曾见过的复杂场景时，模型可能会做出错误的推理或生成不准确的内容。幻觉问题直接影响用户对模型输出的信任，尤其是在关键任务中，幻觉现象可能导致灾难性后果。这种问题在多模态任务中尤为明显，因为模型需要结合视觉和语言信息进行复杂的推理，而当前的技术往往不足以确保这种推理完全符合实际输入的语义。

4 可信评估

随着多模态大模型能力的提升，可信问题逐渐显现。不可信的输出严重影响了 MLLM 在实际场景中的可信度和稳定性。本章主要从幻觉、隐私安全、偏见和鲁棒性 4 个方面总结现有的 MLLMs 可信相关的评估工作。

4.1 幻觉评估

LVLMs 中的幻觉是指视觉输入（作为“事实”）与 LVLM 的文本输出之间的矛盾。通过视觉-语言任务的视角，LVLM 的幻觉症状可以被解释为判断或描述上的缺陷。当模型对用户的查询或陈述的反应与实际的视觉数据不一致时，就会出现判断幻觉。Bai 等[10]系统分析了多模态大语言模型在生成内容时的幻觉问题。幻觉问题的成因主要包括以下几个方面：1）数据方面存在数量不足、质量低下的问题，例如噪声和标注错误，以及统计偏差使模型倾向于生成常见模式；2）模型方面表现为视觉模型能力不足，导致信息丢失，同时语言模型的知识先验过强，覆盖了视觉信息，加之视觉-语言对齐接口设计的不完善，进一步加剧了问题；3）训练方面，现有的训练目标难以有效捕捉复杂的视觉结构，且缺乏强化学习等高级优化方法，限制了模型的表现；4）在推理过程中，模型逐渐忽略视觉信息，生成内容更多依赖语言模式，从而引发幻觉现象。针对这些问题，本文总结了常用的评估方法，并提出了多维度的缓解策略：在数据层面，通过引入负样本和反事实样本改善训练数据的质量和多样性；在模型层面，通过提升视觉模型分辨率或引入多任务视觉编码器增强视觉感知能力；在训练层面，通过对比损失或辅助监督（如掩码预测）优化视觉-语言对齐；在推理层面，通过对比解码或引导解码提升生成过程中对视觉内容的关注度，如表 1 所列[93-115]。

表 1 幻觉评估基准分类

Table 1 Classification of hallucination assessment benchmarks

评估类型	基准	基础数据源	幻觉类型	样本数量	评估指标
判别式	POPE[93]	MSCOCO[104]	对象幻觉	3 000	准确率
	NOPE[95]	Open-Images[62]	对象幻觉	36 000	准确率
	CIEM[96]	MSCOCO[104]	对象幻觉	78 120	准确率
	M-HalDetect[102]	MSCOCO[103]	对象幻觉/属性幻觉/关系幻觉	4 000	奖励模型得分
	RAH-Bench[103]	MSCOCO[104]	对象幻觉/属性幻觉/关系幻觉	3 000	假阳性率
	FGHE[117]	MSCOCO[104]	对象幻觉/属性幻觉/关系幻觉	200	准确率
	MME[105]	MSCOCO[104]	对象幻觉/属性幻觉	1 457	准确率
	MERLIM[106]	MSCOCO[104]	对象幻觉/关系幻觉	31 373	准确率
生成式	MHaluBench[107]	MSCOCO[104]	对象幻觉/属性幻觉	1 860	准确率
	CHAIR[108]	MSCOCO[104]	对象幻觉	5 000	CHAIR
	MOCHa[109]	人工	对象幻觉/属性幻觉	2 000	幻觉率
	GAVIE[110]	Visual-Genome[123]	物体幻觉/属性幻觉	1 000	准确率/关联度
	FAITHScore[111]	MSCOCO[104]	对象幻觉/属性幻觉/关系幻觉	2 000	准确率
	HaELM[112]	MSCOCO[104]	/	5 000	准确率
	MMHal-Bench[113]	Open-Images	对象幻觉/属性幻觉/关系幻觉	1 110	准确率
	HallusionBench[114]	/	模型诊断	1 129	LLM 评估
判别式和生成式	Hal-Eavl[97]	MSCOCO[104]和 LAION[48]	对象幻觉/属性幻觉/关系幻觉	10 000	准确率和 LLM 评估
	VHTest[98]	MSCOCO[104]	对象幻觉/属性幻觉	1 200	准确率
	AMBER[115]	网络	对象幻觉/属性幻觉/关系幻觉	15 202	AMBER 得分

现有方法解决了图像文本模型中的幻觉问题，主要关注“物体的存在”，即判别性任务，特别是给定图像中描绘的物体是否被模型生成的文本准确描述。与在封闭域中训练的图像描述模型相比，LVLM 利用 LLM 的强大理解和表达能力来获得更详细、更流畅的生成描述。然而，这些增强的能力使幻觉多样化，并可能加剧幻觉，幻觉不仅限于物体的存在，还表现在属性和关系错误等描述性错误上，因而延伸出了生成式基准，表 1 列出了这些代表性基准。我们关注的视觉幻觉指的是图像和文本所传达的语义内容之间的所有不一致。幻觉示例如图 3 所示。其中对象幻觉是 LVLM 描述了图片中物体不存在的情况，属性幻觉是 LVLM 描述了图片物体性质不符的情况，关系幻觉是 LVLM 对物体之间的关系做出不准确的判断。

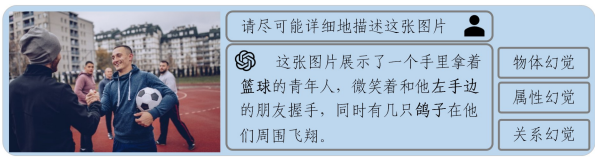

图 3 MLLMs 的回答中出现的 3 种类型的幻觉

Fig. 3 Three types of hallucinations occurred in MLLM responses

Li 等[93]首次系统性地研究了大型视觉语言模型中的对象幻觉问题，发现这些模型在描述图像时常常生成与目标图像不一致的对象。研究设计了评估实验并提出一种基于投票的查询方法（Polling-based Object Probing Evaluation，POPE）。POPE 首先通过人工标注或自动分割工具[94]提取图像中的真实目标对象；然后对图像中不存在的对象进行负采样；最后将真实目标对象和不存在的对象构建为问题模板，用于对大视觉语言模型进行投票调查。该研究揭示了视觉指令数据的分布对 LVLMs 产生对象幻觉的影响，并指出现有评估方法可能受到输入指令和模型生成文本的影响，导致评估结果不可靠。与之类似的 NOPE[95]和 CIEM[96]基准都只关注物体幻觉，并采用准确性作为评价指标。准确性是通过查询图像中是否存在物体并将模型响应与真实答案进行比较得到的，如丰富训练数据的结构和分布、使用更强大的视觉或语言模型、控制响应长度等。

Hal-Eval[97]确定并提出了一种新的幻觉类型——事件幻觉。这种幻觉虚构了一个目标，并围绕它构建了一个完整的叙事，包括属性、关系和行动，进一步完善了幻觉类型的定义。此外，这项工作提出了一个评估基准以及一个包含判别式和生成式评估方法的基准。该基准通过两个专门设计的评价子集实现。

VHTest[98]将图像中物体的视觉属性分为个体属性和组属性，个体属性包括存在性、形状、颜色、OCR 和朝向，组属性来自于多个对象之间的比较，如相对大小、相对位置和计数。基于这样的分类，作者进一步定义了 8 种视觉幻觉模式，对多模态大模型中的幻觉进行了非常详细的评估。目前，大多数基准测试工具并不涵盖开放式的自由形式响应中的幻觉问题（I 型幻觉），而是聚焦于非常具体的问题格式，通常是针对某一特定对象或属性的多项选择题形式的回答（II 型幻觉）。Kaul 等[99]观察到 II 型幻觉的减少并不会导致 I 型幻觉的减少，两种幻觉形式往往是反相关的。此外，他们提出了一种新颖的基于对象的自动框架 THRONE，用于定量评估 LVLM 自由格式输出中的 I 型幻觉，同时提供了一种简单有效的数据增强方法，以减少 I 型和 II 型幻觉作为强有力的基线。

Chandu 等[100]聚焦于当前 MLLMs 在面对不确定性问题时的不足，提出了一套全新的数据集、分类框架和评估指标。他们明确将不确定性划分为认知不确定性和固有不确定性两大类，并进一步细化为知识缺乏、问题复杂性、上下文冗余、时间推测和问题模糊性等子类别。基于此，构建了 CERTAIN-LYUNCER-TAIN 数据集，其中包含 178 000 个对比视觉问答样本，展示了从“可回答”到“不可回答”的不确定性转变。此外，还提出了“置信加权准确度”这一新指标，用于结合模型的预测准确性和置信度进行更全面的评估。实验结果显示，当前主流模型在不确定性场景中的表现较差，而通过该数据集微调后，模型在拒答、减少幻觉等任务上的能力显著提升，同时保持了标准任务的性能稳定。

Chen 等[101]提出了一个统一的幻觉检测框架(UNIHD)以及评估基准(MHaluBench),系统分析了现有研究的不足,指出其局限在于,单一任务、有限的幻觉类别覆盖,以及检测粒度的不足。为了解决这些问题,作者构建了一个覆盖广泛幻觉类型和多模态任务的评估数据集,并引入了基于工具增强的检测框架。UNIHD 通过提取核心声明、自动选择工具、并行执行验证工具和基于证据的判定,提供了一种通用且细粒度的幻觉检测方法。在实验中,UNIHD 展现了相较于基准方法更优的检测性能,尤其是在复杂的文本到图像和图像到文本任务中,其基于 GPT-4V 的实现显著提升了检测准确性。

4.2 隐私安全评估

多模态大模型不仅需要处理海量的视觉数据,还需要确保在推理和生成的过程中不会泄露用户的隐私信息。即使是最先进的封闭源代码模型在隐私保护方面也存在一定的漏洞。特别是当涉及到多模态推理时,如视觉输入与文本信息的结合,模型很容易受到攻击而导致隐私泄露[85]。这种跨模态的风险强调了在开发可信的多模态大模型时,需要全面考虑模态间的交互影响,不能仅依赖单一模态的安全保护措施。

Caldarella 等[116]指出,视觉语言模型存在较高的隐私泄露风险,尤其是在涉及身份识别时。研究表明,这些模型在处理人脸特征、个人物品等隐私数据时,容易泄露用户的敏感信息,即使使用匿名化技术,模型也可能通过图像特征推断出身份信息,特别是开源模型在这方面的保护措施相对较弱,容易受到对抗性攻击,从而暴露隐私。因此,亟需开发更强的隐私保护技术,尤其是适用于多模态输入的深层匿名化技术,以减少身份信息泄露的风险。

Chen 等[117]提出了一个名为 PRIVQA 的基准,用于评估语言模型在隐私保护方面的表现。研究表明,通过模型自我调节和优化输出,MLLMs 可以在一定程度上控制敏感信息的泄露。然而,模型在面对对抗性攻击时,仍然存在被绕过的风险,这表明现有的隐私保护措施需要进一步改进。对抗性攻击的复杂性意味着仅通过简单的提示或指导并不足以保障隐私,模型需要在训练阶段就融入隐私保护的理念。

为了更好地评估 LMMs 的隐私保护能力,Gu 等[118]提出了 MLLMGuard 框架,这是一个用于全面评估多模态大语言模型的隐私、安全性、公平性等多维度的工具。MLLMGuard 针对隐私泄露的评估,利用细致标注的跨模态数据集,检测模型在多种场景中的隐私保护表现。研究结果表明,无论是开放源代码还是封闭源代码的 MLLMs,在面对复杂跨模态输入时,都存在不同程度的隐私泄露风险。MLLMGuard 的提出为深入评估和增强多模态大模型的隐私防护能力提供了重要的工具。

为了应对与图像地理位置相关的隐私泄露问题,Mendes 等[81]提出了 GPTGEOCHAT 基准。该研究特别关注视觉语言模型在用户对话中的地理信息保护能力。GPTGEOCHAT 基准包含了 1000 个地理定位对话,旨在评估模型在对话过程中对地理信息的管理能力。研究发现,虽然使用 API 或模型提示可以有效保护国家或城市级别的地理信息,但当涉及到具体位置,如建筑物名称时,模型需要进行更精细的监督微调才能确保信息不被过度暴露。这揭示了模型在保护细粒度地理信息方面的局限性。

Samson 等[9]提出了 PRIVBENCH 和 PRIVTUNE 两个数据集,用于专门评估和提升 VLMs 的隐私感知能力。PRIVBENCH 包含了如护照、车牌、指纹等 8 个隐私敏感类别的图像。研究结果表明,目前的大多数 VLMs 在隐私感知方面表现有限,尤其是在涉及视觉隐私内容时容易误判。为此,PRIVTUNE 对 TinyLLaVa 和 MiniGPT-v2 进行隐私调优,显著提升了它们识别和处理隐私敏感内容的能力,并且对其他任务的表现影响较小。这种调优方法展示了增强多模态模型隐私保护能力的有效性,为进一步开发隐私感知的模型提供了可行的路径。

Wang 等[119]提出的 SIUO 基准则进一步探讨了跨模态输入的隐私风险。SIUO 基准主要评估在单模态输入(如文本或图像)是安全的情况下,两者结合后是否会导致隐私泄露或产生不安全的输出。例如,一张无害的图像与描述在语义上结合后可能导致用户隐私信息被间接推断出来。研究发现,许多 LVLMs(如 GPT-4V 和 LLaVA)在处理跨模态组合时仍然存在明显的安全漏洞,特别是在应对含有隐私信息的复杂跨模态输入时表现不稳定。

在医疗领域,Xia 等[120]提出的 CARES 基准用于全面评估医学多模态大语言模型在信任度方面的表现,其中隐私保护是一个关键的评估维度。CARES 包含 16 种医学影像和 27 个人体解剖部位,针对不同医学场景中模型对隐私数据(如患者身份信息)的处理进行评估。研究结果表明,许多 Med-LVLMs 在应对隐私数据时存在明显不足,可能会生成错误的诊断,或对患者隐私的保护不力,这给其在医疗领域的应用带来了重大挑战。

4.3 偏见评估

在人工智能领域,偏见通常指模型在训练数据中学习到的有害或不均衡的模式,这些偏见可能会在模型的预测或生成中再现,从而影响其公平性和公正性。模态大模型中的偏见主要源于以下 4 个方面:1)数据集本身存在固有的偏差和不均衡,导致模型在学习过程中继承了这些偏见;2)训练过程中的目标函数和优化策略往往会无意中放大数据中的偏见;3)模型架构和设计通常以性能为导向,缺乏公平性约束,难以有效应对复杂的偏见问题;4)此外,多模态任务的复杂性以及对文化和社会背景的理解能力不足,使得模型在处理涉及不同模态和语境的任务时容易引入或强化偏见。

首先,数据集的规模和来源是偏见产生的根本因素之一。Birhane 等[83]研究了多模态数据集在扩展过程中对内容质量的影响,发现数据集的扩展导致仇恨内容和文化偏见显著增加,这种现象在训练大型生成模型时尤其明显。随着数据集从 400×10^6 扩展到 2×10^9,模型对少数族裔的误分类和偏见问题并没有得到改善,反而在某些情况下变得更加严重,例如将黑人男性错误分类为“罪犯”的概率显著增加。

隐性偏见也同样困扰着多模态模型。Capitani 等[121]通过结合社会心理学的方法,提出了基于隐性联想测试的量化评估方法,揭示了模型在性别和种族方面的隐性偏见,即使经过文本去偏,这些偏见仍然存在。研究表明,模型在面对不同

种族和性别的人物输入时，生成结果的语义倾向明显不同，反映了模型对特定群体的偏好或歧视。

在生成任务中，偏见的扩增是另一个值得关注的问题。Seshadri 等[122]研究了文本到图像生成模型中的偏见扩增现象，发现模型在生成内容时往往会加剧训练数据中的偏见。例如，在训练数据中，女性工程师的比例可能为 25%，但在生成图像时模型将这一比例降低到了 10%。通过对训练数据和生成文本之间分布差异的控制，发现偏见扩增的程度可以大幅降低，但仍难以完全消除。

性别偏见是多模态模型的主要偏见之一，尤其是在双主体的生成任务中更为突出。Wan 等[123]通过 PST 框架，研究了多模态生成模型在同时生成两个不同身份个体时的性别偏见问题。即使在单主体生成中表现出反偏见倾向的模型，在双主体生成中依然表现出明显的偏见，例如，在生成“CEO”和“助理”的图像时，模型倾向于将 CEO 描绘为男性，将助理描绘为女性。这种偏见不仅体现在职业关联上，还反映在对组织权力的分配上，即男性更常与高权力角色相关联，女性则被描绘为低权力角色。

为了评估这些多模态模型在应对偏见和攻击上的表现，Zhang 等[86]开发了 AVIBench 框架，用于评估视觉-语言模型在面对各种对抗性视觉指令下的鲁棒性。研究表明，即使是 GeminiProVision 和 GPT-4V 这些先进的闭源模型，仍然无法完全避免偏见内容的影响。AVIBench 的评估结果进一步强调了提高模型在偏见、鲁棒性和安全性方面的能力的重要性。在文化和幽默内容的多模态理解中，模型同样面临挑战。Zhong 等[124]研究了多模态模型对社交媒体 meme 的解释能力，发现模型在解释具有幽默或讽刺性质的 meme 时，容易受到内嵌的文化和社会偏见的影响，导致生成的解释可能包含有害内容或不准确的文化关联。这表明，尽管多模态模型在生成内容上取得了长足进展，但在社交和文化背景复杂的任务中仍需加强对偏见的处理。

4.4 鲁棒性评估

多模态大模型的鲁棒性是面对输入中分布外（Out-of-Distribution，OOD）变化以及对抗性干扰时，仍然能够保持稳定性能和准确输出的能力。这些模型需要处理多种类型的数据，如图像、文本、视频等，因此它们的鲁棒性评估涉及在复杂、多变的环境中保持高效和可靠的表现。

AVIBench 构建了 260 000 个对抗性视觉指令，涵盖了 4 种基于图像的对抗攻击、10 种基于文本的攻击以及 9 种内容偏见（如性别、文化和种族偏见），全面评估了 LVLMs 在面对这些复杂攻击时的表现。实验结果表明，LVLMs 在面对图像腐蚀、字符级文本扰动等不同类型攻击时表现出显著的脆弱性。尽管 MiniGPT-4 在一般任务中表现较好，但在面对较弱的图像攻击时准确率也急剧下降，而类似 OpenFlamingo-V2 的模型则相对更加稳健。

BenchLMM 基准[125]则着重于评估 MLLMs 在不同视觉风格分布下的鲁棒性，包括艺术风格、传感器风格和应用风格等。MLLMs 在处理与常见风格差异较大的图像时，性能会显著下降。面对艺术风格的图像，MLLMs 普遍难以准确识别其中的物体，这种现象在应用场景中的红外图像和 X 射线图像中也得到了验证。为了提升模型在不同风格下的表现，研究者提出了一种无训练的“风格提示增强”技术，即在推理过程中引导模型首先判断输入图像的风格，进而选择适合的推理方式。这种方法有效地提高了 MLLMs 在跨风格任务中的推理能力，体现了风格识别对增强模型稳健性的作用。

MultiTrust[85]是一个多方面评估 MLLMs 信任度的基准，涉及真实性、安全性、鲁棒性、公平性和隐私性 5 个维度。在鲁棒性方面，MultiTrust 对比了 21 个现代 MLLMs 在 32 个多模态任务中的表现，揭示了 MLLMs 在面对多模态输入时的安全隐患，例如，视觉输入的扰动往往会导致模型的行为不稳定，从而加剧模型的安全问题。尤其是在面对复杂的图像与文本联合输入时，这些模型更容易受到多模态越狱攻击，这表明多模态输入的交互增加了模型的内在风险。此外，MultiTrust 还通过标准化的评估工具箱为未来 MLLMs 的安全性和鲁棒性提升提供了重要的技术基础。

Tu 等[15]围绕 OOD[126]泛化能力和对抗攻击的鲁棒性进行评估。其针对 OOD 场景设计了两个数据集：OODCV-VQA 和 Sketchy-VQA，旨在测试模型在不常见的视觉场景（罕见纹理或简单线条绘制的草图）下的表现。实验结果表明，LVLMs 在应对不常见的视觉内容时表现相对出色，但在语言输入上，尤其是在反事实问题的处理上，表现较差，这显示了文本输入在模型理解中的重要性。研究还引入了视觉 Transformer 攻击和越狱攻击，评估了模型在面对对抗性视觉干扰和有害输入生成上的鲁棒性。研究发现，大多数 VLLMs 在视觉编码器受到攻击时，容易被误导生成不相关的描述，尤其是开放源代码的模型更容易受到这种攻击。而 GPT-4V 在应对这些攻击时通常会拒绝回答，显示出更高的安全性。视觉-语言联合训练可能削弱语言模型原有的安全协议，使 LVLMs 在面对攻击时更易被破坏。

针对 MLLMs 的对抗性攻击，Zhao 等[127]系统评估了大型多模态模型在不同任务和攻击类型下的鲁棒性表现。研究发现，MLLMs 在面对视觉对抗性输入时普遍表现出脆弱性，尤其是在没有额外文本上下文的情况下。如果在推理过程中加入额外的文本提示或上下文信息，那么模型的鲁棒性会得到显著提升。这些实验结果表明，增加上下文信息是一种有效的增强 MLLMs 抗攻击能力的方法。基于这些发现，研究者提出了一种上下文增强的图像分类方法，显著提高了模型在图像分类任务中的鲁棒性。

CVRR-ES[128]则是针对视频多模态模型的复杂视频推理与鲁棒性评估套件，其专注于评估模型在复杂视频场景下的推理能力。CVRR-ES 涵盖 11 种复杂的视频维度，包括时序依赖、社会情感背景等，旨在测试模型在应对真实世界场景中的鲁棒性。大多数开放源代码的 Video-MLLMs 在处理复杂视频场景时表现出较低的准确率和鲁棒性，尤其是在面对时序依赖和社会背景理解等复杂场景时，表现明显不如 GPT-4V 等闭源模型。当前的 Video-MLLMs 在理解复杂视频内容方面仍然存在较大提升空间，需要进一步改进模型的推理能力和稳定性。

4.5 其他评估

Zhou 等[129]探讨了多模态大语言模型在情境安全问题中

的表现，重点分析了模型在语言查询与视觉上下文结合时的安全性判断能力。作者提出多模态情境安全基准（MSSBench）。MSSBench 包含 1 820 个图像-文本对，旨在测试在安全和不安全场景下模型的反应能力。研究发现，当前模型在不安全情境中表现普遍较差，尤其在复杂任务和具身情境中表现更为明显，表明模型缺乏对视觉和语言信息的深度整合能力。此外，还提出了多代理系统，将任务分解为意图推理、视觉理解和安全判断等子任务，从而提升安全性判断能力，但在效率和复杂情境处理上仍有改进空间。

Wang 等[130]提出了一个新的概念——伪对齐。伪对齐指的是大模型在特定任务形式（如开放式问答）中表现良好，但在形式不同（如多选题）或更严格的任务中暴露出无法真正理解对齐目标的问题，表现为模型仅“记住”了答案风格或生成规则而非真正理解安全性或伦理标准。多模态场景下，这一问题可能表现得更为复杂。多模态模型需要处理跨模态的任务，如图像-文本对齐、视觉问答或音频-文本生成，这些任务要求模型在不同模态之间建立一致性。为了解决这些问题，Wang 等[130]提出了 FINE 框架，并将其推广到多模态场景中。通过设计跨模态的评估任务，例如将相同问题设计为文本输入和图像-文本输入两种形式，评估模型在这两种场景下的表现一致性，可以有效检测多模态伪对齐问题。此外，该文中提出的对比蒸馏技术也可以用于多模态训练，通过构建多模态任务中的正负示例数据，提升模型在跨模态任务中的理解和决策能力。与此同时，需要在模型训练中引入更加多样化的跨模态数据，确保模型能够学习到不同模态间的广泛联系，从而减少伪对齐的发生。

5 可信增强

目前，已有部分研究专注于幻觉、偏见等不可靠输出的缓解措施，也有部分研究针对模型进行隐私保护。本章从幻觉缓解、偏见消除和隐私保护 3 个方面探讨了当前针对 MLLMs 不可靠输出和隐私问题的研究。

5.1 幻觉缓解

现有 VLMs 幻觉缓解技术主要基于数据增强、视觉编码器优化、模态对齐改进和强化学习 4 类方法。基于数据的方法通过引入负样本和反事实数据降低噪声干扰；基于视觉编码器的方法采用高分辨率输入和多编码器特征融合提升视觉感知能力；基于模态对齐的方法通过增强跨模态连接模块改善语义一致性；基于强化学习的方法则利用人类反馈或 AI 反馈优化生成策略。

1）基于数据的幻觉缓解

数据是诱发 MLLMs 产生幻觉的主要因素之一。优化训练数据是缓解幻觉的直接且有效的方法，为了减轻幻觉，最近的工作在数据上进行了尝试，包括引入消极数据、引入反事实数据，以及减少现有数据集中的噪声和错误。

当前许多用于视觉指令调整的数据集主要集中在积极的指令样本上，导致模型在面对任何指令时倾向于给出肯定的答案“是”，这会导致幻觉现象。为了解决这个问题，LRV-Instruction[111]被设计成既包含正面也包含负面的指令样本，以增强视觉指令调整的鲁棒性，负面指令分为 3 个不同的语义级别，分别是介绍不存在的对象、活动、属性以及交互，操纵存在对象但属性不一致的情况和操纵指令中的知识内容。CIEM[96]使用现成的大语言模型，从带有标注的图像文本数据集中生成对比问题答案对。这些对比对随后被用于引入对比指令微调。Ferret[131]通过将原始的类别、属性或数量信息替换成类似的假样本来挖掘 95000 个负样本。该方法有效地增强了模型的鲁棒性。

Wang 等[132]提出 ReCaption 框架，用 ChatGPT 重写标题并在重写标题上对指令调整的 LVLMs 进行额外训练，引导 ChatGPT 生成高质量的图像-文本对。使用重写后的图像-文字对对 LVLMs 进行微调，以加强模型在视觉和文本模态之间的细粒度对齐。

2）基于视觉编码器的幻觉缓解

增强 MLLM 的感知能力已被证明可以提高其整体表现并减少幻觉。从 LLaVA 升级到 LLaVA-1.5 时，一项重要的更新是将 CLIP ViT 视觉编码器从 CLIP-ViT-L-224 扩展到 CLIP-ViT-L-336，显著提高了模型性能。QwenVL[40]显示了将图像分辨率从 224×224 逐渐扩大到 448×448 的有效性。InternVL[133]将视觉编码器扩展到 60 亿个参数，并且可以处理宽度从 1 664 到 6144 像素不等的图像。Zhai 等[134]研究了视觉编码器分辨率对其提出的 CCEval 基准测试的影响。在研究的 3 种视觉编码器中，较高的分辨率通常会导致较低程度的幻觉。Monkey[135]通过滑动窗口方法将高分辨率图像分割成小块，并为每个小块配备单独的适配器来处理高分辨率图像，最高支持 1344×896 像素的分辨率。这些工作表明，提高视觉分辨率是一种简单而有效的解决方案。

现有的 LVLM 大多采用 CLIP 的 ViT 作为视觉编码器，其只关注突出的物体，因而不可避免地丢失了一些视觉细节。He 等[136]提出了一种基于视觉专家的模型，旨在减轻 CLIP 图像编码器引起的信息损失。其依靠两个关键模块：多任务编码器和结构知识增强模块，通过聚焦知识增强来提高 MLLM 的视觉感知能力。多任务编码器专用于集成由多个视觉编码器提取的各种类型的潜在视觉信息。此外，结构知识增强模块旨在利用视觉工具（如 OCR 工具和对象检测器），从视觉输入中提取先验知识。

为了增强对象级感知能力，遵循结构知识增强模块的方法，Jain 等[137]提出了 VCoder，利用视觉工具模型来增强对 MLLM 的感知，其使用额外的感知模式，如分割图或深度图，作为通过附加视觉编码器的控制输入。为了增强空间感知能力，Zhao 等[138]建议引入额外的预训练模型来获取空间位置信息和场景图细节，并将其用于指导 LVLM 解决用户查询。

3）基于模态对齐的幻觉缓解

连接模块将视觉特征投射到 LLM 的词嵌入空间中，使视觉和文本模态保持一致。因此，错位可能是幻觉产生的一个关键因素[139-140]。此外，Jiang 等[141]指出即使在 MiniGPT-4 和 LLaVA 等高级 LVLM 中，视觉和文本特征之间的差距仍然很大，且会导致幻觉。为了更好地对齐视觉和语言模态，研究人员最近开发了功能更强大的连接模块。例如，LLaVA-1.5 通过从单个线性层升级到 MLP 来增强 LLaVA 中的连接模块。此外，Chen 等[142]利用 LLaMA2 构建 QLLaMA，其在将

视觉特征与文本对齐方面明显优于 Q-Former。

主流的 MLLM 通常会将编码的视觉特征投影到特定 LLM 的输入空间中。Jiang 等[141]认为，理想的投影应该混合视觉和文本嵌入的分布。尽管有视觉投射，但文本和视觉标记之间存在显著的模态差距，这表明当前学习的界面无法有效地将视觉表示映射到 LLM 的文本表示空间，该问题可能会加剧 MLLM 产生更多幻觉的趋势。Jiang 等提议通过对比损失来增强视觉和文本表示之间的对齐，将带有幻觉的文本用作图像锚点的硬负面示例。这种损失拉近了非幻觉文本和视觉样本的表示，同时分开了非幻觉和幻觉文本的表示。实验结果表明，这种方法不仅可以减少幻觉，还可以提高在其他常用基准测试中的性能。

4）基于强化学习的幻觉缓解

强化学习被引入 MLLMs 的训练中，用于缓解幻觉，包括来自人工智能反馈的强化学习和来自人类反馈的强化学习。

Gunjal 等[143]提出了一种多模态奖励模型来检测 MLLMs 生成的文本中的幻觉。奖励模型是在 M-HalDetect 数据集上训练的，用于识别生成文本中的幻觉成分。具体而言，奖励模型会评估生成的描述与实际图像内容的一致性，并据此给予相应的分数。

为了利用训练好的奖励模型来减少幻觉现象，引入了细粒度直接偏好优化，利用来自单个示例的细粒度偏好，直接减少生成文本中的幻觉现象，通过增强模型区分准确与不准确描述的能力。

LLaVA-RLHF[85]通过引入人类反馈来减轻幻觉。它将 RLHF 范式从文本领域扩展到视觉-语言对齐任务，要求人类注释者比较两种反应，并确定产生幻觉的反应。同时提出了一种算法，在奖励模型的基础上增加了额外的事实信息，缓解了 RLHF 中的奖励窃听现象，进一步提高了模型性能。

类似地，RLHF-V 也使用了 RLHF 范式来增强预训练的 MLLM，强调了在数据层面和方法层面的改进，数据层面是以细粒度的片段级修正的形式收集人类反馈，提供清晰、密集、细粒度的人类偏好；方法层面提出密集直接偏好优化，直接针对密集细粒度的分段偏好优化策略模型。

由于收集人类反馈的成本较高，许多研究探索了使用自动反馈来最大限度地减少人类干预的需求。Li 等[144]利用 AI 注释构建了一个视觉语言反馈数据集。具体来说，响应是由从 12 个 LVLM 中采样的模型生成的，这些模型以来自各种数据集的多模态指令为条件，采用 GPT-4V 来评估生成的输出。此外，通过直接偏好优化方法将偏好监督提炼到其他 MMLM 中。

Zhou 等[145]提出了 POVID 框架，通过 AI 模型自动生成训练反馈数据。该方法采用两阶段策略：首先提示 GPT-4V 在正确答案中注入合理的幻觉，其次扭曲图像以触发 LVLM 的固有幻觉行为。这种方法是一种自动化方法，它明确对比了幻觉答案和真实答案，消除了收集人类反馈的需要，使其易于大规模部署。

5.2 隐私保护

多模态大模型在处理视觉、文本、音频等多种模态的数据时，面临着严重的隐私保护挑战。随着这些模型被广泛应用于社交媒体、虚拟助手、自动驾驶等领域，隐私泄露的风险也显著增加。隐私泄露风险主要体现在模型的记忆能力和推断能力上，攻击者可以通过身份推断攻击等方式从模型中提取敏感信息，如用户的面部图像、姓名甚至健康状态。在当前的研究中，GPT-4V 等多模态模型存在识别和关联特定个人信息的风险，尤其是在模型训练过程中利用了大量公开的数据，导致模型可能泄露未经授权的敏感信息。

差分隐私[146]作为当前主流的隐私保护技术之一，通过在数据中添加噪声，降低了攻击者从模型中提取敏感信息的可能性。然而，差分隐私的应用在多模态场景中面临一些挑战：尽管它可以有效地保护个人数据，但过多的噪声会显著影响模型的性能，尤其是在数据集较小或对隐私保护要求较高的情况下。Liu 等[147]通过结合差分隐私与区块链技术，提出了一种用于物联网多模态数据保护的解决方案，利用差分隐私机制保证数据的隐私性，区块链则用于保护数据的完整性。除此之外，差分隐私还在跨领域推荐系统中被用来防止数据泄露。例如，Wang 等[148]提出了一种本地差分隐私技术，用于保护跨领域推荐系统中的用户数据安全，尤其是能够在知识转移过程中防止用户敏感信息泄露。

此外，针对模型的隐私感知和调优也是目前隐私保护研究的热点之一。通过对多模态大模型进行微调，使其更好地识别和处理敏感数据，从而显著提高模型的隐私保护性能。Samson 等[149]提出了 PRIVTUNE 隐私调优框架，通过对模型进行隐私感知微调，使得模型能够在识别护照、指纹等隐私敏感图像时表现更好，且这种隐私调优对模型在其他任务上的性能影响微乎其微。研究表明，通过合理的隐私调优，可以在不显著降低模型效能的前提下提升隐私保护水平。然而，在隐私保护与模型效能之间存在明显的权衡关系。Chen 等[117]通过 PRIVQA 基准来测试模型的隐私保护性能与效能之间的平衡，发现尽管差分隐私等技术可以增强隐私保护，但过度的隐私保护会导致模型性能显著下降，因此需要研究人员在保护用户隐私与确保模型效能之间找到适当的平衡点。

多模态数据之间的交叉攻击会带来隐私泄露。Yin 等[150]提出的 PriMonitor 框架通过对不同模态数据分别应用差分隐私技术，从而有效防止了模态间相关性导致的隐私重新识别风险。该框架通过引入隐私预算分配算法，确保在保证隐私保护的同时最大限度地提高情感检测的精度。

在 6G 通信环境中，多模态大模型的隐私保护同样面临着新的挑战。Cao 等[151]提出了一种基于多模态大模型的隐私保护无线语义通信技术，旨在通过语义压缩的方式减少数据传输量，进而提升隐私保护的能力。这种方法在传输过程中将多模态数据转换为语义信息，既提升了通信效率，也降低了传输敏感信息的风险。

5.3 偏见缓解

现有的偏见缓解策略可以根据多模态大模型工作流程的不同阶段分为预处理、训练阶段和推理阶段。

预处理阶段主要通过修改数据集的分布，在数据集的早期发现并缓解偏见。该阶段的核心思想是通过平衡数据集中不同群体的敏感属性（如性别、种族等），避免模型在训练过程中学习到社会偏见。数据再平衡是常用的技术之一，通过平

衡数据集中的性别、种族等属性来减少偏见。Alabdulmohsin等[152]的研究表明，通过在视觉-语言模型中应用数据再平衡，模型在生成文本描述、进行图像分类或文本检索等任务时，能够更加公平地对待不同群体，避免将某些属性与特定群体强绑定。

虽然数据再平衡在减少偏见方面表现显著，但它并不能完全消除模型中所有的偏见。这是因为模型的训练不仅依赖于显性数据分布，还可能受到多模态交互、语义关联等隐性因素的影响。因此，数据再平衡应作为偏见缓解的一部分，可以与生成对抗性样本和反事实数据增强(Counterfactual Data Augmentation，CDA)等策略结合使用。CDA通过交换数据集中存在的偏见属性生成多样化的数据样本，以从而帮助模型学会处理更平衡的输入。Cheng等[153]提出的CMSC数据集通过生成多种社会概念的反事实样本，以减少模型在性别、种族和职业等方面的偏见。然而，预处理方法也存在局限性，即使能够显著减少显性偏见，但对于数据中的深层隐性偏见，特别是在多模态交互或复杂上下文中，上述调整仍然可能无法完全消除这些偏见。

训练阶段的偏见缓解策略主要集中在模型优化和架构设计中，通过引入特定的算法和正则化，减少模型在训练过程中学习到的偏见。这类方法不仅可以确保模型在任务执行时不会放大偏见，还能够在学习过程中针对性地减少模型对敏感属性的依赖。对抗性去偏见是训练阶段常用的方法之一。Berg等[154]通过引入对抗网络，使模型难以从数据中提取特定的敏感属性，从而减少对这些属性的依赖，这在缓解性别和种族偏见上表现尤为显著。此外，多任务学习也是关键的缓解偏见的方法之一。Jiang等[155]提出的多组公平性(MGP)损失函数允许模型同时处理多个任务和敏感属性，在多维度上进行偏见缓解，从而减少在性别、种族、年龄等多个属性上的不公平。然而，这些方法的计算成本相对较高，尤其是在大规模多模态模型中，可能会显著增加训练时间和资源消耗。此外，过度缓解偏见可能会影响模型的泛化能力，在非敏感任务上的表现也可能受到一定影响。

推理阶段的偏见缓解策略则是在模型部署后进行，主要通过调整输出结果来减少偏见，无需重新训练模型或修改内部结构。模型编辑是常见的后处理技术之一，能够在推理过程中局部调整模型对某些任务的偏见输出。Wang等[156]展示了如何通过模型编辑减少性别刻板印象和职业偏见，尤其适用于那些已经训练好且难以重新训练的大型模型。输出调整也是有效的后处理技术，通过直接调整模型的输出分布，以确保不同群体在结果中的公平性。例如，Brinkmann等[157]展示了如何在图像检索任务中，通过调整不同性别和种族的搜索结果分布来减少偏见。尽管推理阶段的偏见缓解方法计算开销较低且易于部署，但这些方法无法从根本上消除模型内部的偏见。此外，它们主要集中在对输出的修改，模型内部潜在的偏见依然存在，在某些复杂场景中可能导致表现不一致。

6 多模态大模型的攻击

本章总结了MLLMs在开发和应用阶段可能面临的攻击威胁，根据攻击的目标和手段将现有的针对MLLMs的攻击方法分为越狱攻击、后门攻击、对抗攻击和中毒攻击，如表2所列。

表2 模型设计阶段的攻击总结

Table 2 Summary of attacks in the stage of designing models

攻击类型	作者	攻击策略	攻击权限	攻击阶段	目标模型
越狱攻击	Wang等[160]	联合优化的图像前缀和文本后缀	白盒	推理阶段	MiniGPT-4
	Yin等[161]	采用迭代交叉搜索攻击策略动态更新对抗样本	黑盒	推理阶段	各类LVLMs
	Ma等[166]	生成高风险角色图像并结合无害的文本指令	白盒	推理阶段	各类LVLMs
	Qi等[162]	使用优化的视觉对抗样本	白盒	推理阶段	各类LVLMs
	Niu等[163]	使用最大似然算法生成图像越狱提示	黑盒	推理阶段	各类LVLMs
	Shayegani等[159]	联合嵌入空间中生成对抗性图像与无害文本提示组合	黑盒	推理阶段	各类LVLMs
	Gong等[80]	通过图像排版的对抗样本绕过文本模块	黑盒	推理阶段	各类LVLMs
	Carlini等[158]	对抗图像的输入	白盒	推理阶段	各类LVLMs
对抗攻击	Wang等[178]	通过扰动跨模态特征来增强攻击的转移性	黑盒	测试阶段	各类LVLMs
	Cui等[172]	基于上下文增强的攻击	黑盒	测试阶段	各类LVLMs
	Zhao等[127]	生成具有高转移性的对抗样本来攻击模型	黑盒	测试阶段	各类LVLMs
	Fu等[175]	使用视觉对抗样本来诱导模型调用特定工具	白盒	测试阶段	各类LVLMs
	Gao等[174]	通过“冗长图像”诱导VLMs生成更长的文本序列	白盒	推理阶段	各类LVLMs
	Dong等[177]	生成具有高转移性的对抗样本来攻击模型	黑盒	测试阶段	各类LVLMs
	Cheng等[180]	整合代理模型的全局信息提升攻击效率	黑盒	测试阶段	各类LVLMs
	Luo等[173]	增强对抗样本在不同提示下的转移性	白盒	测试阶段	各类LVLMs
	Wu等[176]	利用单个对抗图像操控多模态代理	白盒	执行阶段	各类LVLMs
后门攻击	Liang等[182]	指令微调阶段的图像和文本触发器嵌入	黑盒	指令微调阶段	OpenFlamingo
	Lu等[184]	在测试阶段使用通用对抗扰动和文本触发词	白盒	测试阶段	各类LVLMs
	Liang等[185]	跨简单触发器实现模型的高泛化性后门攻击	黑盒	指令微调阶段	各类LVLMs
	Ni等[183]	利用物理物体作为触发器	白盒	指令微调阶段	驾驶系统的VLMs
中毒攻击	Carlini等[181]	通过注入中毒样本改变模型在对比学习任务中的行为	黑盒	训练阶段	CLIP
	Xu等[91]	通过隐蔽扰动实现标签攻击和劝诱攻击	黑盒	训练阶段	各类LVLMs
	Yang等[84]	针对多模态编码器的多目标数据中毒	黑盒	训练阶段	CLIP

6.1 越狱攻击

越狱攻击利用模型中的弱点来绕过其预期的限制和控制。这种类型的攻击可能导致模型执行未授权的命令，访问受限制的数据，或者执行超出其设计能力的动作。本节在基于对抗性扰动的越狱攻击和基于有害嵌入的越狱攻击两个场景下总结了现有的针对MLLMs的越狱攻击工作，如图4所示。

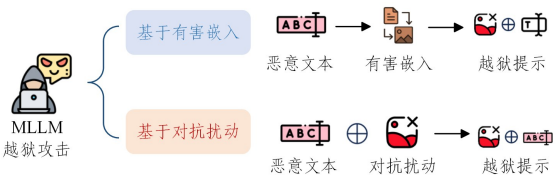

图 4 MLLMs 的越狱策略

Fig. 4 Jailbreak strategies of MLLMs

6.1.1 基于对抗性扰动的越狱攻击

基于对抗性扰动的越狱攻击是一种通过在输入数据（如图像或文本）中注入精心设计的微小扰动，绕过多模态大模型内置的安全对齐机制，从而诱导其生成有害或违背设计预期输出的攻击方式。此类攻击利用模型对输入细微变化的高度敏感性，通过梯度优化或其他算法对输入进行调整，并将经过单模态（文本或图像）或双模态（文本与图像联合）扰动的数据输入模型，最终诱使模型生成在正常情况下会被安全机制拒绝的有害响应。

Carlini 等[158]发现，LVLM 可以很容易地被基于 NLP 的优化攻击利用，且使用连续域图像作为对抗性提示可以诱导语言模型发出有害的、有毒的内容。通过构建多模态模型的端到端可微分实现并采用改进的 NLP 攻击来优化对抗图像，可以实现越狱。改进的 NLP 攻击可能表现出与纯文本模型相似水平的对抗性控制。

Shayegani 等[159]通过视觉编码器的对抗性图像与文本提示配对，以打破语言模型的对齐。攻击采用了一种新颖的合成策略，该策略将对抗性针对有毒嵌入的图像与通用提示相结合，以完成越狱。因此，LLM 通过绘制上下文来回答来自对抗图像的通用提示，生成看起来良性的对抗图像。这些攻击利用了一种新颖的基于嵌入空间的方法，无需访问 LLM 模型即可运行，只需要访问视觉编码器。由于不需要访问 LLM，这些攻击降低了攻击者的进入门槛，凸显了跨模态对齐漏洞的风险，因此多模态模型需要新的对齐方法。

Wang 等[160]联合攻击文本和图像模式，以利用 LVLM 中更广泛的漏洞。具体来说，他们提出了一个双重优化目标，旨在指导模型产生具有高毒性的肯定响应。首先从随机噪声中优化对抗性图像前缀，以便在没有文本输入的情况下生成各种有害响应，从而使图像充满有毒语义。其次对抗性文本后缀与对抗性图像前缀集成并协同优化，以最大限度地提高对各种有害指令的肯定响应的可能性。Yin 等[161]提出了 VLATTACK，通过融合来自单模态和多模态级别的图像和文本的扰动来生成对抗样本，在单模态级别提出了一种新的块级相似性攻击策略来学习图像扰动，以破坏通用表示。VLATTACK 采用现有的文本攻击策略来生成独立于图像模态攻击的文本扰动。在多模态层面设计了一种新的迭代交叉搜索攻击方法，从单模态层面的输出开始，定期更新对抗性图像-文本对。研究越狱攻击的可转移性有助于理解和提高模型的安全性。Qi 等[162]发现单个视觉对抗示例可以普遍越狱对齐的 LLM，迫使它听从广泛的有害指令，并生成有害内容，超越最初用于优化对抗示例的“少数镜头”贬损语料库的狭窄范围。Niu 等[163]提出了一种基于最大似然的算法来查找图像越狱提示，从而在多个看不见的提示和图像中对 MLLM 进行越狱。生成的图像越狱提示可以以黑盒方式转移到越狱各种模型。为了进一步提高对抗性样本的可传递性，Luo 等[164]提出了一种被称为交叉提示攻击（Cross-Prompt Attack，CroPA）的算法。CroPA 不仅使用了不同数量的提示，而且还在优化阶段为提示嵌入引入了可学习的提示扰动。

Gu 等[165]提出了多智能体环境中的安全问题，称为传染性越狱。即通过简单地“越狱”一个代理，在没有对手进一步干预的情况下，几乎所有的代理都会以指数级速度迅速被感染并表现出有害行为。将一张传染性的对抗性图像输入到任何随机选择的代理的记忆中，就足以实现传染性越狱。

6.1.2 基于有害嵌入的越狱攻击

基于有害嵌入的越狱攻击不依赖基于梯度的对抗算法，而是将有害内容嵌入到与良性文本指令配对的图像中。

Gong 等[80]提出了 FigStep，这是一种针对 LVLM 的简单而有效的越狱算法。FigStep 不是直接提供文本有害指令，而是通过排版将有害内容转换为图像，以绕过多 LVLM 文本模块内的安全对齐，诱使 LVLM 输出违反常见人工智能安全政策的不安全响应。Ma 等[166]将“角色扮演”的概念引入 LVLM 的越狱攻击中，并提出了一种新颖有效的方法，称为视觉角色扮演，即利用 LLMs 生成高风险角色的详细描述，并根据描述创建相应的图像。当与良性的角色扮演说明文本配对时，这些高风险角色图像会通过假设具有负面属性的角色来有效地误导 LVLM 产生恶意响应。Li 等[167]提出了一种名为 HADES 的新型越狱方法，该方法使用精心制作的图像隐藏和放大文本输入中恶意意图的危害性。HADES 涉及 3 个步骤：1）将文本指令中的有害关键词通过排版技术生成对应的视觉图像；2）利用大语言模型迭代优化图像生成提示，生成更具危害性的图像，并与原始排版图像拼接以增强攻击效果；3）在图像顶部附加一个对抗性图像，使 MLLM 对有害指令产生肯定的回应。Liu 等[168]从每个场景的恶意查询中提取关键短语，然后利用排版和稳定扩散技术来创建图像。

6.2 对抗攻击

对抗攻击通常以人类难以察觉的方式对输入数据进行微妙的干扰，导致模型产生不正确或不期望的输出。这些扰动被精心设计以利用模型的脆弱性。根据攻击者对攻击模型的访问程度，对抗攻击一般可以分为白盒、灰盒和白盒。

6.2.1 基于白盒的对抗攻击

LVLMs 的白盒攻击是基于对模型的架构、参数和梯度的完全访问。在基于对模型的访问中，大多数 LVLM 攻击者通常使用基于梯度的工具，如 PGD[169]，APGD[170]和 CW[171]，来生成和优化图像和文本输入中的噪声，从而研究受害者 LVLM 对对抗扰动的鲁棒性。它们通过有针对性的攻击诱导模型产生预定的输出或特定的行为，而无针对性的攻击旨在降低输出的质量。Cui[172]等发现增加额外的文本上下文显著提高了 MLLMs 对视觉对抗输入的鲁棒性，设计了一种上下文增强的图像分类方案。

Luo 等[173]引入可学习的提示扰动。首先随机初始化图像的对抗性扰动和文本提示的对抗性扰动，之后将对抗性扰动添加到原始的干净图像和文本提示中，然后输入到视觉-语

言模型中进行前向传播。通过模型的前向传播结果,计算损失函数关于图像扰动和文本提示扰动的梯度。使用梯度下降法更新图像扰动,目的是使模型的输出接近于预定义的目标文本,从而误导模型。与此同时,使用梯度上升法更新提示扰动,目的是增加模型生成目标文本的损失,即尽量使模型不生成目标文本。这一步是为了在优化过程中覆盖更广泛的文本提示空间,增强对抗样本的泛化能力。Gao 等[174]通过最大化原始图像和对抗图像在视觉编码器提取的特征嵌入间的距离,来扰乱图像嵌入,导致模型无法基于输入文本提示来准确预测边界框。Fu 等[175]表明攻击者可通过视觉对抗样本操控多模态模型的工具调用行为。攻击者使用基于梯度的方法来优化图像扰动,使得 LLM 在接收到对抗性图像和文本提示时,生成攻击者指定的工具调用指令。Wu 等[176]针对多模态大语言模型代理进行对抗性攻击,通过选择一个触发图像,构造对抗性文本字符串,利用基于梯度的方法优化图像扰动,以误导多模态代理执行非预期的对抗性目标。

6.2.2 基于黑盒的对抗攻击

现有的黑盒攻击通常将其他视觉编码器或生成模型作为代理模型来生成对抗样本,然后将其迁移以攻击 LVLMs。这些方法一般通过匹配不同编码器的特征来生成对抗语义,或者在特征域中隐藏目标的噪声来增强不可感知性。

Zhao 等[127]首先用预训练的 CLIP 和 BLIP 作为代理模型,通过匹配文本嵌入或图像嵌入来制作目标对抗样本,然后将对抗样本迁移到 MiniGPT-4,LlaVA 等其他 MLLMs,这些基于迁移的攻击已经能够以高成功率诱导目标响应。此外,他们发现采用基于迁移先验的查询攻击可以进一步提高针对这些 MLLM 目标攻击的效率。

Dong 等[177]研究了谷歌作为商业 MLLMs 代表的 Bard 的对抗鲁棒性。Bard 采用最先进的基于迁移的攻击,使对抗图像的图像嵌入远离原始图像的图像嵌入或基于几个代理模型返回一个目标句子。这些对抗图像具有很高的可转移性,可以有效欺骗对抗其他 MLLM。

Wang 等[178]提出了一种针对具有高攻击可转移性的 LVLMs 的指令调整目标攻击(InstructTA)。首先利用 GPT-4 推理出符合攻击者决定的响应的指令,并使用文本到图像的生成模型为目标响应生成目标图像。然后将图像和指令作为局部代理模型的输入,以提取针对目标响应的信息特征。最后,InstructTA 最小化对抗图像样本与目标图像之间的特征距离。InstructTA 将响应的指令复述为一组语义相近的指令,并利用它们有效地提高了生成对抗样本的可迁移性。

贝叶斯优化(Bayesian Optimization,BO)[179]是一种经典的黑盒优化方法,能够通过构建概率模型找到全局最优解。传统 BO 通常使用零均值高斯过程来逼近未知目标函数。Cheng 等[180]提出了基于函数先验的贝叶斯优化的攻击方法(P-BO),P-BO 将替代模型的损失作为高斯过程的均值函数,结合自适应的积分策略,动态调整先验函数的权重,提升了黑盒对抗攻击的查询效率和成功率。

6.3 后门攻击

后门攻击在数据收集或模型训练涉及大量开销的场景中构成重大威胁。这些攻击旨在通过毒化训练样本将后门嵌入到模型中,使敌手能够在推理过程中使用后门触发器操纵模型行为。在 LVLMs 的背景下,现有的工作可以根据训练的不同阶段分为两组。在预训练阶段,多模态后门攻击主要集中在 CLIP 模型上。Carlini 等[181]通过毒化一个数据集的 0.01% 导致模型错误地分类测试图像。在模型微调阶段,攻击者可以通过注入嵌入在指令或图像中的触发器的中毒样本来植入后门,从而使用预定义的触发器恶意操纵受害者模型的预测。Liang 等[182]提出了一种多模态指令后门攻击,即 VLTrojan。VLTrojan 通过隔离和聚类策略促进图像触发器学习,并通过迭代字符级文本触发器生成方法增强黑盒攻击效能。

Ni 等[183]提出了 BadVLMDriver,这是第一个针对自动驾驶 VLM 的后门攻击,可以在实践中使用物理对象发起。与现有的依赖数字修改的针对 VLMs 的后门攻击不同,BadVLMDriver 使用常见的物理物品(如红气球)来诱发突然加速等不安全行为。为了执行 BadVLMDriver,开发了一个自动化管道,使用两条简单的自然语言指令来指导后门数据生成,包括视觉触发器嵌入和文本响应修改。对于后门和良性样本,通过基于混合优化目标的视觉指令调整来优化 VLM。只要触发对象出现在场景中,由后门 VLM 提供支持的自动驾驶就会在现实世界中表现得很危险。Lu 等[184]提出了一种针对多模态大型语言模型的测试时间后门攻击 AnyDoor,其使用对抗性测试图像将后门注入到文本模态中,而不需要访问或修改训练数据。由于后门是由一个通用扰动注入的,因此 AnyDoor 可以动态地改变其后门触发提示效应,这对防御后门攻击提出了新的挑战。Liang 等[185]首次实证了 LVLMs 指令调优过程中后门攻击的可推广性,揭示了大多数后门策略在实际场景中的某些局限性。他们定量评估了 6 种典型的后门攻击在多个 LVLM 上对图像描述基准的可推广性,同时考虑了视觉和文本域偏移。研究结果表明,攻击泛化性与后门触发器和特定图像模型的不相关性以及触发模式的优先相关性呈正相关。

6.4 中毒攻击

投毒攻击是一种训练阶段的攻击,受害者在攻击者恶意操纵的训练数据上训练他们的模型。攻击者的目标是在保持其对原始测试数据的效用的同时,在某些特定的数据样本上误导中毒模型。早期的工作,如 Carlini 等[181]通过目标投毒对图像编码器进行攻击,只需投毒 0.0001% 的数据集,就可以将测试样本分类为指定的类别。

Yang 等[84]提出了 3 种针对多模态模型的中毒攻击。在不同数据集和模型架构上的广泛评估表明,3 种攻击在保持模型效用的同时,在视觉和语言模态上都能获得卓越的攻击性能。此外,不同模式之间的中毒效果不同。为了减轻攻击,提出了训练前防御和训练后防御。实验表明,两种防御都可以在保留模型效用的同时显著降低攻击性能。

类似地,在生成叙事场景中,Shadowcast[91]是一种隐秘的数据投毒攻击方法,其中毒样本在视觉上与具有匹配文本的良性图像无法区分。Shadowcast 在两种攻击类型中都表现出了有效性。第一种是标签攻击,欺骗 LVLMs 错误地识别类标签。第二种是说服攻击,其利用 LVLMs 的文本生成

能力，通过说服来制造叙事，如将垃圾食品描述为保健食品。

6.5 其他攻击

围绕知识编辑技术可能引发的安全风险展开，Cheng等[186]提出了“编辑攻击”这一新型威胁范式，系统性地探讨了通过修改LLMs的知识参数来注入误导性信息或偏见的风险。研究通过构建EDITATTACK数据集，设计实验并评估了3种编辑方法（ROME、微调FT、上下文编辑ICE）的有效性，发现常识性误导信息注入的成功率明显高于长尾领域信息，而单个偏见句子的注入不仅可以影响特定回答，还会显著提高模型在无关问题上的偏见输出，导致整体公平性严重下降。此外，其进一步分析了这种攻击的隐蔽性，表明编辑攻击对模型的知识和推理能力影响极小，难以被检测，凸显了这一攻击的隐蔽性和高危性。

相比单模态模型，MLLM的编辑更具挑战性，因为错误可能来自语言和视觉模块的复杂交互。Cheng等[187]提出了MMEdit基准以评估在多模态环境中对模型进行知识编辑的有效性、稳定性和泛化性。其构建了一个包含视觉问答（VQA）和图像描述编辑任务的数据集，并引入了多个指标（可靠性、局部性和泛化性）来衡量编辑的成功和对模型整体性能的影响。通过对多种现有编辑方法（如微调、MEND、SERAC）的实验分析，发现编辑语言模块的可靠性较高，但视觉模块的编辑效果不佳，表明了这一任务的难度和改进潜力。

7 多模态大模型的防御

多模态大模型的广泛应用也受到越狱、对抗等攻击行为的限制。为应对这些问题，可将防御分为训练阶段的防御和推理阶段的防御，如图5所示，推理阶段的防御通常是在模型训练完成后，针对模型的预测结果进行修正或调整，以防止恶意对抗样本对模型的影响。这类防御方法主要是在输入数据和输出预测的处理中进行干预。训练阶段的防御方法通常通过修改训练过程中的模型学习方式或数据处理方式来增强模型的鲁棒性。这类防御方法主要在模型学习阶段进行干预，与数据和网络架构相关，通常用于防御后门和对抗攻击等。

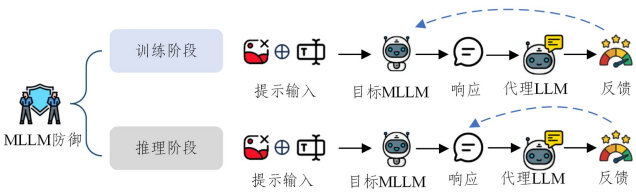

图5 多模态大模型的防御策略

Fig. 5 Defense mechanisms of MLLMs

7.1 推理阶段防御

推理阶段的对齐是指在模型生成预测时，通过即时的防御机制来保护模型免受攻击或错误输入的干扰。推理阶段的对齐通常不需要对模型进行重新训练，而是通过动态添加防御机制或监测输入输出来确保模型的安全性[145]。FigStep通过将有害的文本指令转换为图像中的隐藏文本来绕过模型的文本安全对齐机制，以这种逃避攻击的成功率较高，尤其是在开源的多模态大模型上。为了应对这种视觉提示逃避攻击，作者建议引入更严格的跨模态对齐机制，确保图像和文本模态的安全对齐不被绕过。这要求模型在处理跨模态输入时具备更强的检测能力，确保即便视觉信息中嵌入了隐含文本，模型依然能够做出安全的判断。

Wang等[188]提出了针对结构化攻击的防御策略。机构化攻击通过图像中的隐含文本来绕过多模态模型的安全对齐机制，从而诱使模型生成有害的响应。为应对这一威胁，提出了AdaShield机制，该机制无需微调MLLMs或开发辅助模型。这种方法的优势在于，其利用有限数量的恶意查询来优化防御提示，从而规避了与高计算成本、显著推理时间成本和需要大量训练数据相关的挑战。通过一个包含目标MLLM和防御者LLM的自动完善框架，AdaShield迭代地优化防御提示，这一过程生成了一系列多样化的遵循特定安全指南的防御提示，增强了VLMs抵御提示到图像注入越狱的鲁棒性。这种自适应和自动的方法确保了MLLMs得到有效保护，且无需对模型本身进行大量修改。

提示扰动方法利用了攻击查询的固有脆弱性，这些查询通常依赖精心设计的模板或复杂扰动，因而显著低于良性查询的鲁棒性。通过将输入转换为变体查询并分析语言模型响应的一致性，该方法能够有效检测潜在的越狱攻击。Zhang等[189]提出的JailGuard框架是一种支持图像和文本模态的基于提示扰动的越狱检测方法。JailGuard首先通过包含19种变异器（包括随机和高级变异器）的变体生成器，对输入查询施加扰动，生成变体。随后，攻击检测器对这些变体响应的语义相似性和差异性进行分析，当差异性超过预定义阈值时，识别出潜在的攻击。基于多模态越狱攻击数据集的评估表明，JailGuard在检测性能上显著优于当前最先进的防御方法。

Pi等[190]提出了MLLM-Protector框架，该框架通过一个轻量级的有害检测器和修正机制，来检测并修正模型生成的有害响应。MLLM-Protector尤其针对恶意视觉输入，以防止模型生成不当的响应。这种方法的优势在于无需对模型进行重新训练，且不会对模型的原始性能造成影响。同时，该框架能够兼容多种多模态模型，作为一个插件模块用于检测并修复有害响应，显著提升了模型的安全性。

ECSO[191]是一种基于响应评估的创新方法，其核心在于利用MLLMs的固有安全特性。该方法基于两个关键发现：1）VLMs具备自我检测响应中不安全内容的能力；2）预对齐LLMs的安全机制虽存在于VLMs中，但会被图像特征抑制。通过查询感知的图像-文本转换技术，ECSO将潜在恶意视觉内容转换为纯文本形式，从而有效激活VLM内部预对齐LLMs的安全防护机制。

7.2 训练阶段防御

训练阶段的对齐则注重在模型训练过程中，通过调整训练数据和优化训练方法来增强模型在推理阶段应对攻击的能力。通过对抗性训练、反馈机制等方式，模型可以在面对复杂的输入和攻击时表现出更强的鲁棒性。Chen等[186]提出了一种利用自然语言反馈增强训练对齐的策略。这种方法通过在训练过程中引入额外的反馈回路，使模型逐步学习如何从用户反馈中调整自身的生成行为。通过引入有益的反馈训练，模型能更好地对齐人类偏好和安全需求。这种训练方式还可以增强模型在多轮交互中

的表现，使其更具鲁棒性，能够应对复杂的输入。

针对多模态模型的对抗性攻击，常见的训练策略是引入对抗训练方法。Yang 等[82]指出多模态模型对单一模态的攻击表现出脆弱性，并且多模态模型在信息融合时经常表现出模态偏好，即更依赖某个模态而忽视其他模态。这种偏好使得特定模态上的攻击更具有效性，导致模型整体的鲁棒性下降。为了增强多模态模型的鲁棒性，作者提出了一个名为 CRMT(Certifiable Robust Multi-modal Training)的训练框架，该框架通过增大每个模态的边界来优化模态间的整合，从而提高模型应对单一模态攻击的能力。CRMT 可以在不显著影响性能的情况下，增强模型的防御能力，确保在恶意攻击和模态缺失的情况下保持鲁棒性。

Liang 等[182]提出了两种针对后门攻击的防御方法。一种是中毒样本检测，通过计算数据集中图像之间的余弦相似性来识别可能的中毒样本。由于中毒图像的特征往往聚集在一起，与干净图像的特征相似性较低，因此可以利用这一特性筛选并剔除可疑样本，从而有效减少后门样本对模型训练的影响。另一种是扰乱中毒特征，通过对图像进行预处理(如添加噪声或进行图像变换)，破坏中毒图像的嵌入特征。这种方法基于视觉编码器在指令调优过程中通常是冻结的假设，因此中毒样本的特征不会随着训练而改变，从而有效削弱后门触发器与目标输出之间的关联性。

Liang 等[185]提出可通过增加训练数据的分布多样性和优化触发器模式的通用性来提升模型的鲁棒性。他们通过在训练阶段引入跨域场景的数据，减少触发器与特定图像或模型的关联性，从而限制攻击的有效性。此外，还分析了图像和文本触发器之间的相关性，以进一步弱化后门触发的成功率。实验表明，这种方法在跨域场景中显著降低了后门攻击成功率，并提升了模型在多样化数据上的防御能力。

针对数据中毒攻击，Xu 等[91]强调攻击的隐蔽性是数据中毒成功的关键，因此防御的核心在于提高对数据异常的检测能力。具体来说，他们指出数据质量控制是防御的第一道防线，建议在模型训练前对训练数据进行严格审查，特别是那些来自非可信数据源的样本。此外，提出可以通过引入更强的模型审计机制，在训练和推理阶段对模型的行为进行监控，尤其是检测模型在特定触发条件下的输出异常。

8 多模态大模型攻防数据集和评估策略

近年来，多模态大模型攻防的相关数据集和评估策略不断涌现，为系统性地评估多模态大模型提供了便利。本章列举了多模态大模型攻防相关数据集，并介绍了现有的攻防评估策略，便于研究者进行后续研究。

8.1 数据集

在 MLLM 攻防工作中，涉及的数据集一般可以分为两类：常规任务数据集和恶意数据集。这些数据集的详细信息如表 3 所列。

表 3 多模态大模型攻防评估常用数据集

Table 3 Datasets commonly used for evaluation of multimodal large models in attack and defense scenarios

类型	数据集	描述	样本数量
常规任务	MS-COCO[104]	目标检测	328000
	ImageNet[192]	图像分类	14197122
	Flickr30K[193]	图像-文本匹配、图像描述生成等任务	31783
	VizWiz[194]	从真实盲人用户收集的数据集，包含他们上传的图片及相关问题和答案	31000
	VQA-v2[195]	视觉问答数据集，包含图像、问题及多选答案	1105904
	OK-VQA[196]	包含图像、开放式问题及对应的多种答案，通过引入平衡样本减少语言偏差	14031
	LLaVA-Instruct[42]	包含图像的对话式指令和详细的语言解释	150000
毒性相关	GOAT-Bench[197]	评估大型多模态模型在表情包中的安全性理解能力	500
	RTVLM[201]	评估模型在红队攻击下的表现	5200
	MM-SafetyBench[200]	评估模型在多种安全攻击场景下的表现	5040
	SafeBench[199]	检测和评估模型在面对有害查询时的行为	9200
	ToViLaG[198]	评估模型的毒性生成问题	32000
	AdvBench-M[163]	评估模型在越狱攻击中的表现	500
	VLSafe[191]	评估模型对不适当内容生成的抵抗能力	5874

GOAT-Bench[197]是一个专为评估大型多模态模型在表情包中的安全性理解能力而设计的基准数据集，由 6626 个多样化的表情包组成，涵盖仇恨性、厌女、冒犯性、讽刺性和有害性 5 个关键任务。该数据集整合了 FHM，MAMI，MultiOFF 等公开数据源，并严格控制标注质量，通过人工注释重新评估以确保一致性。

ToViLaG[198]数据集专注于评估视觉语言模型在生成有害内容方面的倾向性，包含 32000 个文本-图像配对样本，分为单模态有害对(仅图像或文本有害)、双模态有害对(图像和文本均有害)以及具有挑衅性的无害文本提示(可能诱导生成有害图像)。数据集通过多来源(如 NSFW，UCLA 抗议图像数据集)收集色情、暴力和血腥图像，并结合 PerspectiveAPI 和生成模型筛选高质量的非有害文本进行标注，同时采用梯度引导方法生成具有攻击性的无害文本提示，用于测试模型的鲁棒性。

AdvBench-M[163]数据集是针对 MLLMs 设计的一个综合数据集，旨在评估模型在越狱攻击中的表现。它是在文本基准数据集 AdvBench 的基础上扩展而成，通过为有害指令匹配相关图像，实现从单模态到多模态的转变。数据集包含 500 条有害行为指令，涵盖 8 大语义类别，包括炸弹或爆炸物、毒品、自残与自杀、网络安全与隐私问题、身体攻击、恐怖主义与社会紧张局势、股票市场与经济，以及枪支弹药。每个类别通过搜索引擎获取 30 张语义相关图像，形成图像-指令配对样本。

SafeBench[199]数据集包含 9200 个样本，覆盖文本、视觉和音频 3 种模态，并分为 8 个主要风险类别和 23 个子类别，

如仇恨言论、暴力行为、个人隐私、恶意软件生成等。Safe-Bench 数据集通过一个自动化的生成流程构建，利用大型语言模型生成高质量、有害性的查询样本。

MM-SafetyBench[200]数据集覆盖了 13 个安全相关场景，包括非法活动、仇恨言论、恶意软件生成、隐私泄露等，共包含 5040 对文本-图像样本。通过生成与查询关键词相关的图像，MM-SafetyBench 能有效评估模型在面对与查询相关图像时的防御能力。

RTVLM[201]数据集包含 5200 个图像-文本样本，分为 10 个任务类别，包括文本误导、视觉误导、多模态越狱和面部公平性等。

VLSafe[191]数据集结合精心设计的对抗性提示生成样本，包括 4764 条训练样本和 1110 条测试样本，覆盖了从仇恨言论、暴力行为到隐私泄露等多种场景。每条样本结合图像和文本提示，设计为测试模型是否能识别并规避有害内容，同时生成积极劝阻用户的安全响应。

8.2 攻防评估策略

与传统的视觉问答数据集不同，多模态大模型的输出形式具有开放性，这为主观评估带来了诸多挑战。这种灵活的输出形式不仅增加了评估的复杂性，还使得在评估成本和评估准确性之间难以权衡。为了解决这些问题，本文总结了两种用于计算攻防相关指标的评估方法：基于人工评估和基于模型的自动评估。

8.2.1 基于人工评估

人工评估可分为主观评估和基于规则的评估，在主观性较强的任务中，依赖人工标注者对模型生成的输出进行检查，判断其是否包含不安全内容，例如有害文本、隐私泄露或其他不适当的生成[78,80,159,162]。研究中常用的指标包括攻击成功率(ASR)，即模型在面对恶意输入时生成不当输出的比例；识别成功率(RSR)，即模型能否正确识别出不安全输入；防御成功率(DSR)，用于衡量模型在面对潜在攻击时抵御不安全生成的能力[202]。这种方法的优点在于高解释性和可靠性，能够捕捉到复杂语义以及生成内容中的微妙风险，适合对开放式回答进行细致分析。此外，它还能处理当前自动化评估方法可能遗漏的边界情况。然而，人工主观评估也存在显著的缺点，包括成本高、耗时长，并且在标注过程中可能受到标注者主观偏见的影响。为了提升评估的全面性和效率，通常需要结合量化指标和多维度场景设计，确保覆盖模型输出的广泛可能性。

基于规则的评估是一种通过预定义规则和目标字符串对多模态大模型安全性进行自动化评估的方法，旨在降低人工干预的成本和复杂性。该方法主要判断模型输出是否包含特定的不安全内容，如恶意词汇[181]、有害指令[164]、恶意 API 调用[203]，或是否符合预设的分类标准。例如，部分研究将安全性问题转换为分类任务，通过检测模型是否正确分类滥用内容(如社交媒体恶意言论)来评估安全性，常用指标包括准确率、F1 分数和 AUROC 等[117,197,204]。此外，基于规则的评估不仅限于区分性任务(如检测图像或文本是否存在有害内容)，还可扩展至生成性任务(如图像描述生成)，利用 CIDEr，BLEU-4 等语言生成指标评估模型输出的质量和安全性[205]。相比于主观评估，基于规则的评估自动化程度更高，适用于大规模数据集，能够快速提供定量化的安全性评估。然而，这种方法存在覆盖面有限的缺点，无法检测到规则未定义的不安全内容，可能漏检一些隐性的安全风险。因此，基于规则的评估需要设计全面且多样化的规则集合，并结合其他评估方法以提升评估的准确性和鲁棒性。

8.2.2 基于模型的自动评估

基于模型的自动评估是一种利用机器学习模型或强大的大语言模型对多模态大模型的输出进行自动化安全性评估的方法。该方法旨在通过模型的计算能力实现快速、可扩展的安全性检测，减少人工干预。具体来说，常用的工具包括 Perspective API 和 Detoxify，它们通过内置的机器学习算法对模型输出进行毒性、隐私泄露或其他有害内容的评分[159,162,198]。此外，一些研究还借助强大的语言模型(如 GPT-4 或 GPT-3.5-turbo)进行定制化的评估。例如，可以设计提示词让语言模型判断多模态大模型的输出是否与安全性相关的目标一致，如输出的语义是否与事实对齐[206]，或是否拒绝了执行不安全的指令。同时，这些语言模型还能够从多方面(如相关性、安全性、说服力)为输出评分，提供更细致的评价维度[191]。相比于人工评估，基于模型的自动评估具有快速、低成本、易扩展的优势，特别适合对大规模数据集进行高效处理。然而，它也存在一定局限性，例如，对评估模型的依赖可能引入偏差，且需要精心设计提示词以确保评估的准确性。因此，基于模型的自动评估通常与其他评估方法结合使用，以提升整体的评估可靠性和全面性。

9 机遇与挑战

本章讨论了 MLLM 安全性研究的不足，为未来的研究方向提供了建议。

1)提升攻击的可转移性：当前攻击的转移性不足主要体现在跨模型和跨任务的一致性较差、对不同架构模型的适应性有限以及转移成功率不稳定[207]。通常，生成的对抗样本在攻击源模型上有效，但在结构不同或任务目标不同的模型上效果会明显减弱，这限制了其在现实应用中的广泛使用。同时，跨模态或多任务环境中的转移性更具挑战，因为模型可能对不同模态或任务有不同的处理机制。未来研究可以聚焦于开发更通用的对抗样本生成方法，提升其在多种模型结构和任务上的转移成功率，并研究多模态间的特征映射优化，使攻击在复杂环境中更具鲁棒性。

2)增强攻击隐蔽性：现有的对抗样本在视觉或文本上仍可能存在不自然的痕迹，导致人类或检测系统可以发现异常。此外，多模态攻击中各模态信息的融合不够自然，易暴露攻击意图。为增强隐蔽性，可以在输入数据上加入自然且微小的扰动，使其在视觉或语义上保持与原始数据一致，从而不易被人类察觉[181]。例如，在图像中进行色彩微调、纹理修饰或添加自然阴影等细微变化，使图像看似与原图无异但能误导模型；在文本中则可以通过同义替换、语序调整等保持语义一致的方式来隐蔽攻击意图。这种隐蔽性策略通过融入自然特征的微扰，实现对模型的操控且不易被检测系统察觉[11]。

3)Prompt 注入攻击的优化：Prompt 注入攻击的局限性

在于其对模型结构的依赖性强、输出的可控性较弱、容易被检测以及在多模态环境下保持一致性难度大。由于注入攻击主要通过输入层的干预来影响模型输出，这种间接性使得对结果的精确操控变得困难，尤其在多模态数据的交互中更具不确定性[208]。此外，现有的 Prompt 注入方式容易被安全机制检测，特别是当输入提示不自然时。未来的研究方向应集中在黑盒环境下的优化、高效生成和自动化优化、多模态提示的一致性融合，以及动态 Prompt 注入，使攻击更隐蔽、更稳定，从而在复杂的多模态场景中实现更高效的操控和防御绕过。

4)防御加固：当前的防御加固措施的不足主要体现在对多模态攻击的应对能力有限、对抗性样本的识别率不稳定、计算成本高且难以在动态场景中适应。例如，现有的检测和防御方法在面对复杂的跨模态攻击时常出现失效或误判，尤其是在图像和文本模态结合的情况下。此外，这些防御措施的高计算开销也影响了实时检测的效率和适用性。未来的研究方向应侧重于开发更轻量化且高效的多模态对抗检测方法、自适应的动态防御机制，以及基于生成模型的对抗样本预判系统，以提升在多模态环境下的防御效果和实时适应能力。

5)提高安全评估：为了提高多模态大模型的安全评估，需要更全面的基准和合理的指标。当前数据集范围有限，通常只测试粗粒度的安全性。构建优质评估数据集应考虑明确的安全能力分类，代表不安全内容的元素及其呈现方式，数据来源(如现有数据集、AI 生成等)，以及数据数量、多样性和质量控制。同时，评估指标也需关注在利用强大的大语言模型进行评估时，需通过精心设计的提示来确保评估规则的客观性。

6)安全与效用的平衡：当 MLLMs 面对恶意指令时，如果 MLLMs 拒绝服从指令，则 MLLMs 保持安全性，但会失去效用。MLLMs 可能将安全问题错误分类为恶意，这种错误的分类会严重降低 LLMs 在安全提示上的性能。因此需要开发更平衡的安全对齐方法，避免模型对无害查询的过度敏感反应。

7)RAG 的安全性：RAG 模型通过检索外部知识库或数据库来增强生成结果，由于依赖外部检索库，模型容易受到内容污染和偏见的影响，如果检索到不当的信息，可能会生成不安全的内容[209]。此外，多模态数据的复杂性可能导致生成内容与实际语义不一致，从而产生幻觉或错误的推理结果。研究者可以重点开发更加精细的检索与生成对齐方法，确保在多模态内容下的精确匹配，同时提升对模糊查询的适应性；此外，可研究更强的鲁棒性训练方法，增强模型应对无关或不可靠内容的能力，通过自适应检索等技术减轻潜在的幻觉生成问题。

8)多模态大模型与 Agent 结合的安全性问题：多模态大模型与自主代理结合后，由于增加了攻击面和新型漏洞，安全性问题显得尤为突出。这些代理配备了外部工具和记忆系统，能够实现复杂的交互，但也因此容易受到多种攻击，如后门植入、记忆中毒或对抗性提示攻击[210-212]。这些威胁利用代理执行现实操作的能力，可能导致严重的误操作。因此应重点开发针对对抗性输入的强大防御机制，加强多模态和代理驱动环境下的后门检测，并建立可靠的实时监控系统，以应对这些安全威胁。

9)链式开发的安全性问题：多模态大模型链式开发的安全性问题主要集中在数据隐私泄露、模型篡改与中毒、跨模态信息共享的误差放大以及对抗样本攻击等方面。在链式开发中，不同模块间的耦合加剧了数据流通中的风险，使得单点失效或攻击可能波及全链。未来研究方向应聚焦于多模态交互的安全协议制定、鲁棒性增强、链式模块的可信度评估机制及自动化检测修复，特别是开发对抗样本抵抗技术和跨模态信息安全传输算法，以提高整体链路的安全性和鲁棒性。

10)模型泛化能力提升：多模态大模型的泛化能力不足主要表现在其对不同数据分布的适应性差、跨任务性能不稳定，以及在实际复杂场景中的表现有限。当模型在未见过的模态组合或不同来源的数据上进行推理时，其准确性和鲁棒性往往显著下降，使其难以推广到多样化的真实环境中。此外，多模态模型在切换任务时也可能出现模态冲突或信息丢失。因此，应侧重于开发更加鲁棒的训练方法，如通过多样化的数据增强、自适应学习和迁移学习等方式，提升模型在多模态、多任务和不同数据分布下的泛化能力，从而确保模型在复杂环境中的稳定性和准确性。

11)新的隐私保护要求：由于不同模态之间的数据关联性增强，单模态的隐私保护方法难以有效防止通过多模态进行的隐私推断，例如，将图像和语音等模态结合，可能泄露个体身份、行为和情感状态[213]。此外，这些多模态数据信息在数据传输和共享过程中易遭受攻击，尤其是跨模态的推理更易导致敏感信息泄露。未来的研究应开发适用于多模态场景的统一隐私保护框架，通过差分隐私、联邦学习和分布式计算等技术来增强多模态数据隐私的鲁棒性，并设计适应性强的动态隐私调节机制。

结束语 本文系统综述了多模态大语言模型在安全性方面的研究现状，深入分析了模型在幻觉、隐私泄露、偏见及鲁棒性等安全维度面临的挑战，并详细阐述了这些问题对模型性能、实际应用及用户信任的潜在影响。同时，从全生命周期的视角总结了多模态大语言模型在不同阶段可能遭受的安全威胁，包括数据预处理阶段的偏见传播与隐私暴露、预训练阶段的对抗攻击与多模态特征不稳定性，以及模型推理阶段模态组合导致的不安全生成问题。此外，针对越狱攻击、对抗攻击、后门攻击及中毒攻击等威胁类型，全面梳理了当前的评估方法及防御策略，涵盖了从训练到推理各环节的技术路径与实践方案。

未来的研究应聚焦于构建全面细致的安全评估框架，以覆盖多模态任务中的多样化威胁场景，同时研发高效的隐私保护技术，提升模型在处理敏感数据时的可信性。此外，还需设计通用且适应性强的防御机制，以增强模型应对分布外输入和多样化攻击的能力，从而全面提升多模态模型的实用性和安全性。

参 考 文 献

[1] JI J, QIU T, CHEN B, et al. Ai alignment: a comprehensive survey[J]. arXiv: 2310.19852, 2023.

[2] YUAN J, SUN S, OMEIZA D, et al. Rag-driver: generalisable driving explanations with retrieval-augmented in-context learn-

ing in multi-modal large language model[J]. arXiv:2402.10828, 2024.

[3] ZHANG Z, ZHANG A, LI M, et al. Multimodal chain-of-thought reasoning in language models[J]. arXiv:2302.00923, 2023.

[4] SREERAM S, WANG T H, MAALOUF A, et al. Probing multimodal llms as world models for driving[J]. arXiv:2405.05956, 2024.

[5] ZHANG X, WU C, ZHAO Z, et al. PMC-VQA: visual instruction tuning for medical visual question answering[J]. arXiv: 2305.10415, 2024.

[6] RAHMAN Md A, ALQAHTANI L, ALBOOQ A, et al. A survey on security and privacy of large multimodal deep learning models: teaching and learning perspective[C]// 2024 21st Learning and Technology Conference (L&T). 2024: 13-18.

[7] AVSEC Ž, AGARWAL V, VISENTIN D, et al. Effective gene expression prediction from sequence by integrating long-range interactions[J]. Nature Methods, 2021, 18(10): 1196-1203.

[8] CABELLO L, BUGLIARELLO E, BRANDL S, et al. Evaluating bias and fairness in gender-neutral pretrained vision-and-language models[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 8465-8483.

[9] SAMSON L, BARAZANI N, GHEBREAB S, et al. Privacy-aware visual language models[J]. arXiv:2405.17423, 2024.

[10] BAI Z, WANG P, XIAO T, et al. Hallucination of multimodall-Large language models: a survey[J]. arXiv:2404.18930, 2024.

[11] LIU D, YANG M, QU X, et al. A survey of attacks on large vision-language models: resources, advances, and future trends [J]. arXiv:2407.07403, 2024.

[12] YAO Y, DUAN J, XU K, et al. A survey on large language model (llm) security and privacy: the good, the bad, and the ugly [J]. High-Confidence Computing, 2024, 4(2): 100211.

[13] DONG Z, ZHOU Z, YANG C, et al. Attacks, defenses and evaluations for llm conversation safety: a Survey[C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2024: 6734-6747.

[14] SUN H, ZHANG Z, DENG J, et al. Safety assessment of chinese large language models[J]. arXiv:2304.10436, 2023.

[15] TU H, CUI C, WANG Z, et al. How many unicorns are in this image? a safety evaluation benchmark for vision llms[J]. arXiv: 2311.16101, 2023.

[16] LIU X, ZHU Y, LAN Y, et al. Safety of multimodal large language models on images and texts[C]// Proceedings of the Thirty Third International Joint Conference on Artificial Intelligence. 2024: 8151-8159.

[17] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[18] TAN M, LE Q V. Efficientnet: rethinking model scaling for convolutional neural networks[C]// International Conference on Machine Learning. PMLR, 2019: 6105-6114.

[19] SMITH S L, BROCK A, BERRADA L, et al. Convnets match vision transformers at scale[J]. arXiv:2310.16764, 2023.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[J]. arXiv:2010.11929, 2020.

[21] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[22] FANG Y, WANG W, XIE B, et al. Eva: exploring the limits of masked visual representation learning at scale[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 19358-19369.

[23] 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.

[24] SUN Q, FANG Y, WU L, et al. EVA-CLIP: improved training techniques for clip at scale[J]. arXiv:2303.15389, 2023.

[25] ZHU D, CHEN J, SHEN X, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models [J]. arXiv:2304.10592, 2023.

[26] CAI Y, LIU Y, ZHANG Z, et al. CLAP: isolating content from style through contrastive learning with augmented prompts [C]// European Conference on Computer Vision. 2024: 130-147.

[27] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[C]// International Conference on Machine Learning. PMLR, 2023: 28492-28518.

[28] HSU W N, BOLTE B, TSAI Y H H, et al. Hubert: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[29] ARNAB A, DEHGHANI M, HEIGOLD G, et al. Vivit: a video vision transformer[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 6836-6846.

[30] ZHAO L, GUNDAVARAPU N B, YUAN L, et al. Videoprism: a foundational visual encoder for video understanding[C]// Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 60785-60811.

[31] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(120): 1-39.

[32] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 1877-1901.

[33] CHUNG H W, HOU L, LONGPRE S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.

[34] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv:2307.09288, 2023.

[35] MA C, ZHANG Y, SHEN S, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality[EB/OL].

(2023-03-30) [2025-05-08]. https://lmsys.org/blog/2023-03-30-vicuna/.

[36] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv: 2309.16609, 2023.

[37] 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.

[38] JIAN Y, GAO C, VOSOUGHI S. Bootstrapping vision-language learning with decoupled language pre-training[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 57-72.

[39] YE Q, XU H, XU G, et al. mplug-owl: Modularization empowers large language models with multimodality[J]. arXiv: 2304.14178, 2023.

[40] BAI J, BAI S, YANG S, et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond[J]. arXiv: 2308.12966, 2023.

[41] LU J, GAN R, ZHANG D, et al. Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects[J]. arXiv: 2312.05278, 2023.

[42] LIU H, LI C, WU Q, et al. Visual instruction tuning[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 34892-34916.

[43] LI C, WONG C, ZHANG S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 28541-28564.

[44] SU Y, LAN T, LI H, et al. PandaGPT: One Model To Instruction-Follow Them All[C]// Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants. 2023: 11-23.

[45] YE Q, XU H, XU G, et al. mplug-owl: Modularization empowers large language models with multimodality[J]. arXiv: 2304.14178, 2023.

[46] SHARMA P, DING N, GOODMAN S, et al. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 2556-2565.

[47] CHANGPINYO S, SHARMA P, DING N, et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 3558-3568.

[48] SCHUHMANN C, BEAUMONT R, VENCU R, et al. Laion-5b: an open large-scale dataset for training next generation image-text models[C]// Proceedings of the 36h International Conference on Neural Information Processing Systems. 2022, 25278-5294.

[49] CHEN L, LI J, DONG X, et al. Sharegpt4v: Improving large multi-modal models with better captions[C]// European Conference on Computer Vision. Cham: Springer, 2025: 370-387.

[50] CHEN G H, CHEN S, ZHANG R, et al. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model[J]. arXiv: 2402.11684, 2024.

[51] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]// Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020: 1877-1901.

[52] DAI W L, LI J N, L D X, et al. InstructBLIP: towards general-purpose vision-language models with instruction tuning[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 49250-49267.

[53] CHEN F, HAN M, ZHAO H, et al. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages[J]. arXiv: 2305.04160, 2023.

[54] ZHANG R, HAN J, LIU C, et al. Llama-adapter: Efficient fine-tuning of language models with zero-init attention[J]. arXiv: 2303.16199, 2023.

[55] WANG W, CHEN Z, CHEN X, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 61501-61513.

[56] ZHAO Z, GUO L, YUE T, et al. Chatbridge: Bridging modalities with large language model as a language catalyst[J]. arXiv: 2305.16103, 2023.

[57] LI L, YIN Y, LI S, et al. M3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning[J]. arXiv: 2306.04387, 2023.

[58] GAO P, HAN J, ZHANG R, et al. Llama-adapter v2: Parameter-efficient visual instruction model[J]. arXiv: 2304.15010, 2023.

[59] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: Aligning language models with self-generated instructions[J]. arXiv: 2212.10560, 2022.

[60] LUO G, ZHOU Y, REN T, et al. Cheap and quick: Efficient vision-language instruction tuning for large language models[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 29615-29627.

[61] XU Z, SHEN Y, HUANG L. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning[C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023: 11445-11465.

[62] ZENG Y, ZHANG H, ZHENG J, et al. What Matters in Training a GPT4-Style Language Model with Multimodal Inputs? [C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2024: 7930-7957.

[63] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[[C]// Proceedings of the 36th International Conference on Neural Information Processing Systems. 2023: 27730-27744.

[64] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward mode[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 53728-53741.

[65] LI L, XIE Z, LI M, et al. Silkie: Preference distillation for large visual language models[J]. arXiv: 2312.10665, 2023.

[66] YU T，YAO Y，ZHANG H，et al. RLHF-V：towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback[C]// 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE，2024：13807-13816.

[67] ZHU D，CHEN J，SHEN X，et al. Minigpt-4：Enhancing vision-language understanding with advanced large language models[J]. arXiv：2304.10592，2023.

[68] LI K C，HE Y，WANG Y，et al. Videochat：Chat-centric video understanding[J]. arXiv：2305.06355，2023.

[69] CHU Y，XU J，ZHOU X，et al. Qwen-audio：Advancing universal audio understanding via unified large-scale audio-language models[J]. arXiv：2311.07919，2023.

[70] WU C，YIN S，QI W，et al. Visual chatgpt：Talking，drawing and editing with visual foundation models[J]. arXiv：2303.04671，2023.

[71] SHEN Y，SONG K，TAN X，et al. Hugginggpt：Solving ai tasks with chatgpt and its friends in hugging face[J]. Advances in Neural Information Processing Systems，2024，36.

[72] HUANG R，LI M，YANG D，et al. Audiogpt：Understanding and generating speech，music，sound，and talking head[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：23802-23804.

[73] WU S，FEI H，QU L，et al. Next-gpt：any-to-any multimodal LLM[J]. arXiv：2309.05519，2023.

[74] TANG Z，YANG Z，KHADEMI M，et al. CoDi-2：In-Context Interleaved and Interactive Any-to-Any Generation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024：27425-27434.

[75] LIU Y，DENG G，LI Y，et al. Prompt Injection attack against LLM-integrated Applications[J]. arXiv：2306.05499，2023.

[76] DAI S，XU C，XU S，et al. Bias and unfairness in information retrieval systems：New challenges in the llm era[C]// Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024：6437-6447.

[77] YAO Y，DUAN J，XU K，et al. A survey on large language model (llm) security and privacy：The good，the bad，and the ugly[J]. High-Confidence Computing，2024，4(2)：100211.

[78] DONG Z，ZHOU Z，YANG C，et al. Attacks，defenses and evaluations for llm conversation safety：A survey[J]. arXiv：2402.09283，2024.

[79] PI R，HAN T，ZHANG J，et al. MLLM-Protector：Ensuring MLLM's Safety without Hurting Performance[J]. arXiv：2401.02906，2024.

[80] GONG Y，RAN D，LIU J，et al. FigStep：jailbreaking large vision-language models via typographic visual prompts[J]. arXiv：2311.05608，2023.

[81] MENDES E，CHEN Y，HAYS J，et al. Granular privacy control for geolocation with vision languagemodels[J]. arXiv：2407.04952，2024.

[82] YANG Z，WEI Y，LIANG C，et al. Quantifying and enhancing multi-modal robustness with modality preference[C]// International Conference on Learning Representations. 2024：1-23.

[83] BIRHANE A，PRABHU V，HAN S，et al. Into the laions den：investigating hate in multimodal datasets[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023：21268-21284.

[84] YANG Z，HE X，LI Z，et al. Data poisoning attacks against multimodal encoders[C]// International Conference on Machine Learning. PMLR，2023：39299-39313.

[85] ZHANG Y，HUANG Y，SUN Y，et al. Benchmarking trustworthiness of multimodal large language models：a comprehensive study[J]. arXiv：2406.07057，2024.

[86] ZHANG H，SHAO W，LIU H，et al. AVIBench：towards evaluating the robustness of large vision-language model on adversarial visual-instructions[J]. arXiv：2403.09346，2024.

[87] WANG S，YE X，CHENG Q，et al. Cross-modality safety alignment[J]. arXiv：2406.15279，2024.

[88] WANG P，ZHANG D，LI L，et al. Inferaligner：inference-time alignment for harmlessness through cross-model gidance [J]. arXiv：2401.11206，2024.

[89] QI X，ZENG Y，XIE T，et al. Fine-tuning Aligned Language Models Compromises Safety，Even When Users do not intend to! [J]. arXiv：2310.03693，2023.

[90] LENG S，XING Y，CHENG Z，et al. The curse of multi-modalities：Evaluating hallucinations of large multimodal models across language，visual，and audio[J]. arXiv：2410.12787，2024.

[91] XU Y，YAO J，SHU M，et al. Shadowcast：stealthy data poisoning attacks against vision-language models[J]. arXiv：2402.06659，2024.

[92] TAO X，ZHONG S，LI L，et al. Imgtrojan：jailbreaking vision-language models with one image[J]. arXiv：2403.02910，2024.

[93] 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.

[94] ZOU X，YANG J，ZHANG H，et al. Segment everything everywhere all at once[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023：9769-9782.

[95] 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. 2024：37-58.

[96] HU H，ZHANG J，ZHAO M，et al. CIEM：contrastive instruction evaluation method for better instruction tuning[J]. arXiv：2309.02301，2023.

[97] JIANG C，YE W，DONG M，et al. Hal-Eval：a universal and fine-grained hallucination evaluation framework for large vision language models[C]// Proceedings of the 32nd ACM International Conference on Multimedia. 2024：525-534.

[98] HUANG W，LIU H，GUO M，et al. Visual hallucinations of multi-modal large language models[C]// Proceedings of the Association for Computational Linguistics. 2024：9614-9631.

[99] KAUL P，LI Z，YANG H，et al. THRONE：An object-based hallucination benchmark for the free-form generations of large vision-language models[C]// Proceedings of the IEEE/CVF Con-

ference on Computer Vision and Pattern Recognition. 2024: 27228-27238.

[100]CHANDU K R, LI L, AWADALLA A, et al. Certainly Uncertain: A Benchmark and Metric for Multimodal Epistemic and Aleatoric Awareness[J]. arXiv: 2407.01942, 2024.

[101]CHEN X, WANG C, XUE Y, et al. Unified hallucination detection for multimodal large language models[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024: 3235-3252.

[102]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: 18135-18143.

[103]CHEN Z, ZHU Y, ZHAN Y, et al. Mitigating hallucination in visual language models with visual supervision[J]. arXiv: 2311.16479, 2023.

[104]LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: common objects in context[C]// European Conference on Computer Vision. 2014: 740-755.

[105]FU C, CHEN P, SHEN Y, et al. MME: a comprehensive evaluation benchmark for multimodal large language models[J]. arXiv: 2306.13394, 2023.

[107]VILLA A, ALCÁZAR J C L, SOTO A, et al. Behind the magic, merlim: multi-modal evaluation benchmark for large image-language models[J]. arXiv: 2312.02219, 2023.

[108]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.

[109]BEN-KISH A, YANUKA M, ALERP M, et al. MOCHa: multi-objective reinforcement mitigating caption hallucinations[J]. arXiv: 2312.03631, 2023.

[110]LIU F, LIN K, LI L, et al. Mitigating hallucination in large multi-modal models via robust instruction tuning[C]// The Twelfth International Conference on Learning Representations. 2023: 1-45.

[111]JING L, LI R, CHEN Y, et al. FaithScore: fine-grained evaluations of hallucinations in large vision-language models[J]. arXiv: 2311.01477, 2023.

[112]WANG J, ZHOU Y, XU G, et al. Evaluation and analysis of hallucination in large vision-language models[J]. arXiv: 2308.15126, 2023.

[113]SUN Z, SHEN S, CAO S, et al. Aligning large multimodal models with factually augmented RLHF[C]// Proceedings of the Association for Computational Linguistics. 2024: 13088-13110.

[114]GUAN T, LIU F, WU X, et al. Hallusion Bench : anadvanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14375-14385.

[115]WANG J, WANG Y, XU G, et al. AMBER: an llm-free multi-dimensional benchmark for mllms hallucination evaluation[J]. arXiv: 2311.07397, 2024.

[116]CALDARELLA S, MANCINI M, RICCI E, et al. The phantom menace: unmasking privacy leakages in vision-language models [J]. arXiv: 2408.01228, 2024.

[117]CHEN Y, MENDES E, DAS S, et al. Can language models be instructed to protect personal information? [J]. arXiv: 2310.02224, 2023.

[118]GU T, ZHOU Z, HUANG K, et al. MLLMGuard: a multi-dimensional safety evaluation suite for mult-imodal large language models[J]. arXiv: 2406.07594, 2024.

[119]WANG S, YE X, CHENG Q, et al. Cross-modality safety alignment[J]. arXiv: 2406.151279, 2024.

[120]XIA P, CHEN Z, TIAN J, et al. CARES: a comprehensive benchmark of trustworthiness in medical vision language models [J]. arXiv: 2406.06007, 2024.

[121]CAPITANI G, LUCARINI A, BONICELLI L, et al. Beyond the surface: comprehensive analysis of implicit bias in vision-language models[C]// Intervento Presentato Al Convegno Fairness and Ethics towards Transparent AI: Face the Challenge through Model Debiasing. 2024.

[122]SESHADRI P, SINGH S, ELAZAR Y. The bias amplification paradox in text-to-image generation[C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2024: 6367-6384.

[123]WAN Y, CHANG K W. The male ceo and the female assistant: gender biases in text-to-image generation of dual subjects[J]. arXiv: 2402.11089, 2024.

[124]ZHONG Y, BAGHEL B K. Multimodal understanding of memes with fair explanations[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 2007-2017.

[125]CAI R, SONG Z, GUAN D, et al. BenchLMM: ben-chmarking cross-style visual capability of large multimodal models[J]. arXiv: 2312.02896, 2023.

[126]CUI P, WANG J. Out-of-distribution (OOD) detection based on deep learning: a review[J]. Electronics, 2022, 11(21): 3500.

[127]ZHAO Y, PANG T, DU C, et al. On Evaluating Adversarial robustness of large vision-language models[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2024: 54111-54138

[128]KHATTAK M U, NAEEM M F, HASSAN J, et al. How good is my video lmm? complex video reasoning and robustness evaluation suite for video-lmms[J]. arXiv: 2405.03690, 2024.

[129]ZHOU K, LIU C, ZHAO X, et al. Multimodal Situational Safety [J]. arXiv: 2410.06172, 2024.

[130]WANG Y, TENG Y, HUANG K, et al. Fake Alignment: Are LLMs Really Aligned Well? [J]. arXiv: 2311.05915, 2023.

[131]YOU H, ZHANG H, GAN Z, et al. Ferret: refer and ground anything anywhere at any granularity[J]. arXiv: 2310.07704, 2023.

[132]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, 2024: 32-45.

[133]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.

[134]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]. arXiv：2310. 01779，2023.

[135]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.

[136]HE X，WEI L，XIE L，et al. Incorporating visual experts to resolve the information loss in multimodal large language models[J]. arXiv：2401. 03105，2024.

[137]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.

[138]ZHAO Y，LI Z，JIN Z，et al. Enhancing the spatial awareness capability of multi-modal large language model[J]. arXiv：2310. 20357，2023.

[139]ZHAO Z，WANG B，OUYANG L，et al. Beyond hallucinations：enhancing lvlms through hallucination-aware direct preference optimization[J]. arXiv：2311. 16839，2023.

[140]SUN Z，SHEN S，CAO S，et al. Aligning large multimodal models with factually augmented rlhf[C]// Proceedings of the Association for Computational Linguistics. 2024：13088-13110.

[141]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.

[142]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.

[143]GUNJAL A，YIN J，BAS E. Detecting and preventing hallucinations in large vision language models[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：18135-18143.

[144]LI L，XIE Z，LI M，et al. Silkie：preference distillation for large visual language models[J]. arXiv：2312. 10665，2023.

[145]ZHOU Y，CUI C，RAFAILOV R，et al. Aligning modalities in vision large language models via preference fine-tuning[J]. arXiv：2402. 11411，2024.

[146]DWORK C，ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science，2014，9(3/4)：211-407.

[147]LIU Q，HUANG Y，JIN C，et al. Privacy and integrity protection for iot multimodal data using machine learning and blockchain[J]. ACM Transactions on Multimedia Computing，Communications，an-d Applications，2024，20(6)：1-18.

[148]WANG L，SANG L，ZHANG Q，et al. A privacy-preserving framework with multi-modal data for cross-domain recommendation[J]. Knowledge-Based Systems，2024，304：112529.

[149]SAMSON L，BARAZANI N，GHEBREAB S，et al. Privacy-aware visual language models[J]. arXiv：2405. 17423，2024.

[150]YIN L，LIN S，SUN Z，et al. PriMonitor：an adaptive tuning privacy-preserving approach for multimodal emotion detection[J]. World Wide Web，2024，27(2)：9.

[151]CAO D，WU J，BASHIR A K. Multimodal large language models driven privacy-preserving wireless semantic communication in 6g[C]// IEEE International Conference on Communications Workshops. 2024：171-176.

[152]ALABDULMOHSIN I，WANG X，STEINER A，et al. CLIP the bias：how useful is balancing data in multimodal learning?[C]// International Conference on Learning Representations. 2024：1-32.

[153]CHENG H，GUO Y，GUO Q，et al. Social debiasing for fair multi-modal llms[J]. arXiv：2408. 06569，2024.

[154]BERG H，HALL S M，BHALGAT Y，et al. A prompt array keeps the bias away：debiasing vision-language models with adversarial learning[C]// Proceedings of the 2nd Conference of the Asia-Pacif-ic Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. 2022：806-822.

[155]JIANG J，MANORANJAN V，SALAM H，et al. Towards generalised and incremental bias mitigation in personality computing[J]. IEEE Transactions on Affective Computing，2024，15(4)：2192-2203.

[156]WANG Z，LI X，QIN Z，et al. Can We Debias Multimodal Large Language Models via Model Editing?[C]// Proceedings of the 32nd ACM International Conference on Multimedia. 2024：3219-3228.

[157]BRINKMANN J，SWOBODA P，BARTELT C. A multidimensional analysis of social biases in vision transformers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023：4914-4923.

[158]CARLINI N，NASR M，CHOQUETTE-CHOO C A，et al. Are aligned neural networks adversarially al-igned?[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023：61478-61500.

[159]SHAYEGANI E，DONG Y，ABU-GHAZALEH N. Jailbreak in pieces：compositional adversarial attacks on multi-modal language models[J]. arXiv：2307. 14539，2023.

[160]WANG R，MA X，ZHOU H，et al. White-box multimodal jailbreaks against large vision-language models[J]. arXiv：2405. 17894，2024.

[161]YIN Z，YE M，ZHANG T，et al. VLATTACK：multimodal adversarial attacks on vision-language tasks via pre-trained models[C]// Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023：52936-52956.

[162]QI X，HUANG K，PANDA A，et al. Visual adversarial examples jailbreak aligned large language models[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024：21527-21536.

[163]NIU Z，REN H，GAO X，et al. Jailbreaking attack against multimodal large language model[J]. arXiv：2402. 02309，2024.

[164]LUO H，GU J，LIU F，et al. An image is worth 1000 lies：adversarial transferability across prompts on vision-language models[J]. arXiv：2403. 09766，2024.

[165]GU X, ZHENG X, PANG T, et al. Agent smith: a single image can jailbreak one million multimodal llm agents exponentially fast[J]. arXiv:2402.08567, 2024.

[166]MA S, LUO W, WANG Y, et al. Visual-RolePlay: universal jailbreak attack on multimodal large language models via role-playing image character[J]. arXiv:2405.20773, 2024.

[167]LI Y, GUO H, ZHOU K, et al. Images are achilles' heel of alignment: exploiting visual vulnerabilities for jailbreaking multimodal large language models[J]. arXiv:2403.09792, 2024.

[168]LIU X, ZHU Y, GU J, et al. MM-SafetyBench: a benchmark for safety evaluation of multimodal large language models[J]. arXiv:2311.17600, 2023.

[169]Madry A. Towards deep learning models resistant to adversarial attacks[J]. arXiv:1706.06083, 2017.

[170]CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]// International Conference on Machine Learning. PMLR, 2020: 2206-2216.

[171]CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]// 2017 IEEE Symposium on Security and Privacy. 2017: 39-57.

[172]CUI X, APARCEDO A, JANG Y K, et al. On the robustness of large multimodal models against image adversarial attacks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 24625-24634.

[173]LUO H, GU J, LIU F, et al. An image is worth 1000 lies: adversarial transferability across prompts on vision-language models [J]. arXiv:2403.09766, 2024.

[174]GAO K, BAI Y, GU J, et al. Inducing high energy-latency of large vision-language models with verbose images[J]. arXiv: 2401.11170, 2024.

[175]FU X, WANG Z, LI S, et al. Misusing tools in large language models with visual adversarial examples[J]. arXiv:2310.03185, 2023.

[176]WU X, CHAKRABORTY S, XIAN R, et al. Highlighting the safety concerns of deploying llms/vlms in robotics[J]. arXiv: 2402.10340, 2024.

[177]DONG Y, CHEN H, CHEN J, et al. How robust is google's bard to adversarial image attacks? [J]. arXiv: 2309.11751, 2023.

[178]WANG X, JI Z, MA P, et al. InstructTA: instruction-tuned targeted attack for large vision-language models[J]. arXiv: 2312.01886, 2023.

[179]CHENG S, MIAO Y, DONG Y, et al. Efficient black-box adversarial attacks via bayesian optimization guided by a function prior [C]// Proceedings of the 41st International Conference on Machine Learning. PMLR, 2024: 8163-8183.

[180]FRAZIER P I. A tutorial on Bayesian optimization[J]. arXiv: 1807.02811, 2018.

[181]CARLINI N, TERZIS A. Poisoning and backdooring contrastive learning[J]. arXiv:2402.13851, 2024.

[182]LIANG J, LIANG S, LUO M, et al. VL-Trojan: multimodal instruction backdoor attacks against autoregressive visual language models[J]. arXiv:2402.13851, 2024.

[183]NI Z, YE R, WEI Y, et al. Physical backdoor attack can jeopardize driving with vision-large-language models[J]. arXiv: 2404.12916, 2024.

[184]LU D, PANG T, DU C, et al. Test-time backdoor attacks on multimodal large language models[J]. arXiv:2402.08577, 2024.

[185]LIANG S, LIANG J, PANG T, et al. Revisiting backdoor attacks against large vision-language models [J]. arXiv: 2406.18844, 2024.

[186]CHEN C, HUANG B, LI Z, et al. Can editing llms inject harm? [J]. arXiv:2407.20224, 2024.

[187]CHENG S Y, TIAN B Z, LIU Q B, et al. Can we edit multimodal large language models? [C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 13877-13888.

[188]WANG Y, LIU X, LI Y, et al. AdaShield: safeguarding multimodal large language models from structure-based attack via adaptive shield prompting[C]// European Conference on Computer Vision (ECCV). 2024: 1-25.

[189]ZHANG X, ZHANG C, LI T, et al. Jailguard: A universal detection framework for llm prompt-based attacks[J]. arXiv: 2312.10766, 2024.

[190]PI R, HAN T, ZHANG J, et al. MLLM-Protector: ensuring mllm's safety without hurting performance[J]. arXiv: 2401.02906, 2024.

[191]CHEN Y, SIKKA K, COGSWELL M, et al. Dress: instructing large vision-language models to align and interact with humans via natural language feedback[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 14239-14250.

[192]DENG J, DONG W, SOCHER R, et al. ImageNet: A Large-Scale Hierarchical Image Database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.

[193]PLUMMER B A, WANG L, CERVANTES C M, et al. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 2641-2649.

[194]GURARI D, LI Q, STANGL A J, et al. Vizwiz grand challenge: Answering visual questions from blind people[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3608-3617.

[195]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.

[196]MARINO K, RASTEGARI M, FARHADI A, et al. Ok-vqa: A visual question answering benchmark requiring external knowledge[C]// Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition. 2019: 3195-3204.

[197]LIN H, LUO Z, WANG B, et al. Goat-bench: Safety insights to large multimodal models through meme-based social abuse[J]. arXiv: 2401.01523, 2024.

[198]WANG X, YI X, JIANG H, et al. ToViLaG: your visual-language generative model is also an evildoer [C]// Proceedings of

the 2023 Conference on Empirical Methods in Natural Language Processing. 2023：3508-3533.

[199]YING Z，LIU A，LIANG S，et al. SafeBench：a safety evaluation framework for multimodal large language models[J]. arXiv：2410. 18927，2024.

[200]LIU X，ZHU Y，GU J，et al. MM-SafetyBench：a benchmark for safety evaluation of multimodal large language models[J]. arXiv：2311. 17600，2023.

[201]LI M，LI L，YIN Y，et al. Red teaming visual language models[J]. arXiv：2401. 12915，2024.

[202]WU Y，LI X，LIU Y，et al. Jailbreaking gpt-4v via self-adversarial attacks with system prompts[J]. arXiv：2311. 09127，2023.

[203]BAILEY L，ONG E，RUSSELL S，et al. Image hijacks：Adversarial images can control generative models at runtime[J]. arXiv：2309. 00236，2023.

[204]VAN M H，WU X. Detecting and correcting hate speech in multimodal memes with large visual language model[J]. arXiv：2311. 06737，2023.

[205]SCHLARMANN C，HEIN M. On the adversarial robustness of multi-modal foundation models[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023：3677-3685.

[206]JI Y，GE C，KONG W，et al. Large language models as automated aligners for benchmarking vision-language models[J]. arXiv：2311. 14580，2023.

[207]GUO Q，PANG S，JIA X，et al. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models[J]. arXiv：2404. 10335，2024.

[208]FAN Y，CAO Y，ZHAO Z，et al. Unbridled icarus：a survey of the potential perils of image inputs in multimodal large language model security[J]. arXiv：2407. 12784，2024.

[209]ZHAO S，YANG Y，WANG Z，et al. Retrieval augmented generation（rag）and beyond：a comprehensive survey on how to make your llms use external data more wisely[J]. arXiv：2409. 14924，2024.

[210]ZHANG B，TAN Y，SHEN Y，et al. Breaking agents：compromising autonomous llm agents through malfunction amplification[J]. arXiv：2407. 20859，2024.

[211]WANG Y，XUE D，ZHANG S，et al. BadAgent：inserting and activating backdoor attacks in llm agents[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. 2024：9811-9827

[212]CHEN Z，XIANG Z，XIAO C，et al. Agentpoison：red-teaming llm agents via poisoning memory or knowledge bases[J]. arXiv：2407. 12784，2024.

[213]HINTERSDORF D，STRUPPEK L，BRACK M，et al. Does clip know my face？[J]. Journal of Artificial Intelligence Research，2024，80：1033-1062.

CHEN Jinyin，born in 1982，Ph.D，professor，is a member of CCF（No. 14348M）. Her main research interests include artificial intelligence security，graph data mining and evolutionary computing.

ZHENG Haibin，born in 1995，Ph.D，lecturer，is a member of CCF（No. 72193M）. His main research interests include deep learning and artificial intelligence security.

（责任编辑：何杨）