中图法分类号:TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2025)06-2051-31

论文引用格式:Guo Y F, Yu Z T, Liu A S, Zhou W B, Qiao T, Li B, Zhang W M, Kang X G, Zhou L N, Yu N H and Huang J W. 2025. Recent progress of the security research for multimodal large models. Journal of Image and Graphics, 30(6):2051-2081(郭园方, 余梓彤, 刘艾杉, 周文柏, 乔通, 李斌, 张卫明, 康显桂, 周琳娜, 俞能海, 黄继武. 2025. 多模态大模型安全研究进展. 中国图象图形学报, 30(6):2051-2081)[DOI:10.11834/jig.250067]

多模态大模型安全研究进展

郭园方[1+*],余梓彤[2+],刘艾杉[1+],周文柏[3+],乔通[4+],李斌[5],张卫明[3],
康显桂[6],周琳娜[7],俞能海[3],黄继武[8]

1. 北京航空航天大学,北京　100191; 2. 大湾区大学,东莞　523000; 3. 中国科学技术大学,合肥　230026;
4. 杭州电子科技大学,杭州　310018; 5. 深圳大学,深圳　518060; 6. 中山大学,广州　510275;
7. 北京邮电大学,北京　100876; 8. 深圳北理莫斯科大学,深圳　518172

摘　要:多模态大模型的安全性研究已成为当下人工智能领域的焦点。由于大模型以深度神经网络为核心构建,因此与深度神经网络类似,存在多种安全风险。此外,由于其特有的复杂性,以及广泛的应用场景,也使得大模型面临一些独特的安全风险。本文系统地总结多模态大模型的安全风险,包括对抗攻击、越狱攻击、后门攻击、版权窃取、幻觉现象、泛化问题以及偏见问题等。具体来说,在对抗攻击中,攻击者通过构造微小但具有欺骗性的对抗样本,使大模型在面对带噪输入时产生严重的误判;越狱攻击利用大模型的复杂结构,绕过或破坏原有的安全约束和防御措施,使模型执行未授权的操作,甚至泄露敏感数据;后门攻击则通过在大模型的训练阶段植入隐秘的触发器,使模型在特定条件下做出攻击者预期的反应;未经授权的窃取者可能未经模型拥有者的同意随意分发或进行商业使用,将导致模型版权拥有者遭受损失;幻觉现象,即模型输出与输入不一致的问题;泛化问题即大模型当前应对部分新数据分布或风格的能力仍显不足;大模型在性别、种族、肤色和年龄等敏感问题上的偏向性可能引发伦理等问题。随后,针对这些安全风险分别介绍相应的解决方案。本文旨在为理解和应对多模态大模型的独特安全挑战提供一个独特的视角,促进多模态大模型安全技术的发展,引导未来相关安全技术的发展方向。

关键词:多模态大模型;大模型安全;对抗样本(AE);越狱攻击;后门攻击;版权窃取;模型幻觉;模型偏见

Recent progress of the security research for multimodal large models

Guo Yuanfang[1+*], Yu Zitong[2+], Liu Aishan[1+], Zhou Wenbo[3+], Qiao Tong[4+], Li Bin[5],
Zhang Weiming[3], Kang Xiangui[6], Zhou Linna[7], Yu Nenghai[3], Huang Jiwu[8]

1. Beihang University, Beijing　100191, China; 2. Great Bay University, Dongguan　523000, China; 3. University of Science and Technology of China, Hefei　230026, China; 4. Hangzhou Dianzi University, Hangzhou　310018, China; 5. Shenzhen University, Shenzhen　518060, China; 6. Sun Yat-sen University, Guangzhou　510275, China; 7. Beijing University of Posts and Telecommunications, Beijing　100876, China; 8. Shenzhen MSU-BIT University, Shenzhen　518172, China

Abstract: With the rise of deep learning technology, artificial intelligence technology has progressed from shallow machine learning to deep learning and from small-scale data learning to big data learning. In recent years, with the continuous

收稿日期:2025-02-18;**修回日期**:2025-03-21;**预印本日期**:2025-03-28

+ **共同第一作者**

* **通信作者**:郭园方　andyguo@buaa.edu.cn

基金项目:国家自然科学基金项目(62272020,62372423,62472135);广东省基础与应用基础研究基金资助项目(2023A1515140037);浙江省重点研发项目(2025C04002);浙江省自然科学基金项目(LZ23F020006)

Supported by: National Natural Science Foundation of China(62272020,62372423,62472135)

improvement of data and computing resources, the scale of deep learning models has continued to increase, and large-scale pretrained models (large models) have begun to emerge. In general, any model that is trained on large-scale extensive data (usually via large-scale self-supervised training) and can be adjusted (e. g. , through fine-tuning) to a wide range of downstream tasks and can be referred to as a large model. Meanwhile, a typical large model that is trained by utilizing multimodal information, such as text, images, videos, and audio, and aims to complete diverse multimodal application tasks is known as a multimodal large model. In the past two years, domestic and foreign multimodal large models, such as ChatGPT, Llama, and Qwen, have achieved remarkable success. As multimodal large models continuously evolve, the research on the security of these models has become the focus in the field of artificial intelligence. Given the superior processing abilities of these models in various multimodal tasks, their security issues may have harmful consequences. Large models are built with deep neural networks as their core, so they encounter security risks similar to those faced by deep neural networks. In addition, given the unique complexity of large models and their wide range of application, they face unique security risks. In this study, we systematically summarize the security risks associated with multimodal large models, including adversarial attacks, jailbreak attacks, backdoor attacks, copyright theft, hallucination phenomena, generalization issues, and bias problems. In adversarial attacks, attackers construct small yet deceptive adversarial examples to cause misjudgments by large models when they are fed with these adversarial perturbed inputs. Jailbreak attacks exploit the complex structure of large models to bypass or destroy the original security constraints and defense measures, enabling the models to perform unauthorized operations and/or even leak sensitive data in the outputs. Backdoor attacks involve implanting hidden triggers during the training phase of large models, causing the models to exhibit attacker-intended behaviors under specific conditions. Meanwhile, unauthorized thieves may distribute or use large models for commercial purposes without the consent of the model owners, causing losses to the copyright owner of the models. The hallucination phenomenon refers to the issue of inconsistency between a large model's output and input. The generalization problem indicates the inability of large models to deal with new data distributions or styles. The bias of large models on sensitive issues, such as gender, race, skin color, and age, may lead to ethical problems, which may further produce severe consequences. After the presentations of these security risks, we introduce corresponding solutions to. By presenting the progress of research on the security risks and corresponding solutions of multimodal large models, this study aims to provide a unique perspective for understanding and addressing the unique security challenges of multimodal large models, promote the development of security technologies for multimodal large models, and guide the future direction of related security technology development. In conclusion, multimodal large models demonstrate excellent performance in many tasks and applications and provide various types of assistance to people's work and daily lives. Through this research on the security technologies of multimodal large models, we hope to ensure the safety and reliability of these models, thereby providing a guarantee for people's normal work and life.

Key words: multimodal large model; large model security; adversarial example(AE); jailbreak attack; backdoor attack; copyright theft; model hallucination; model bias

0 引 言

随着深度学习的兴起,人工智能(artificial intelligence,AI)技术经历了从浅层机器学习到深度学习、从小规模数据学习到大数据学习的发展历程。随着近年来数据资源和计算资源的不断提升,深度学习模型的规模持续增加,大规模预训练模型(大模型)开始兴起。

通常来说,在大规模的广泛数据上训练的(通常使用大规模的自监督训练)、且可以适应(例如通过微调)广泛下游任务的任何模型都称为大模型。其中,使用文本、图像、视频和音频等多模态信息训练,并以完成多模态多样化应用任务为目标所得到的大模型又称为多模态大模型。ChatGPT、Llama、文心一言、通义千问等国内外的多模态大模型获得了极大发展,在智能内容创作、AI数字人、AI数据分析、智能客服以及智能办公等多个方面不断提升应用上限。

虽然多模态大模型的研究进展与应用一日千里,但也出现了多样化的安全问题,如OpenAI ChatGPT和API重大中断事件、三星员工使用ChatGPT不当泄露芯片机密代码事件等。由于多模态大模型

是以深度神经网络为基础构建的，因此具有类似深度神经网络的安全缺陷，如对抗攻击（Dong 等，2018）、后门攻击（Wang 等，2024f）、版权窃取（Lim 等，2022）、输出偏见（Bhargava 和 Forsyth，2019）等问题。此外，由于多模态大模型特有的复杂性以及多样化应用场景，其也存在一些独特的安全问题，如越狱攻击（Ying 等，2024c）、幻觉现象（Huang 等，2025）、泛化问题（Zhang 等，2024b）等。

鉴于此，本文系统性地总结多模态大模型安全的最新研究进展。具体来说，将多模态大模型的安全风险划分为对抗攻击（Dong 等，2018）、越狱攻击（Ying 等，2024c）、后门攻击（Wang 等，2024f）、版权窃取（Lim 等，2022）、幻觉现象（Huang 等，2025）、泛化问题（Zhang 等，2024b）和偏见问题（Bhargava 和 Forsyth，2019）7 个方面。其中，在对抗攻击中，攻击者通过构造微小但具有欺骗性的对抗样本，使大模型在面对带噪输入时产生严重的误判；越狱攻击利用大模型的复杂结构，绕过或破坏原有的安全约束和防御措施，使模型执行未授权的操作，甚至泄露敏感数据；后门攻击则通过在大模型的训练阶段植入隐秘的触发器，使模型在特定条件下做出攻击者预期的反应；未经授权的窃取者可能未经模型拥有者的同意随意分发或进行商业使用，将导致模型版权拥有者遭受损失；幻觉现象，即模型输出与输入不一致的问题；泛化问题即大模型当前应对部分新数据分布或风格的能力仍显不足；大模型在性别、种族、肤色和年龄等敏感问题上的偏向性可能引发伦理等问题。

本文首先针对上述安全风险及相应的攻击技术进展进行介绍，随后介绍相应解决方案的最新研究进展。旨在为理解和应对多模态大模型的独特安全挑战提供一个独特的视角，并促进多模态大模型安全技术的发展，引导未来相关安全技术的发展方向。

1 对抗攻击

与传统深度模型类似，虽然多模态大模型在各种各样的任务上能够取得惊人的效能，许多场景下其表现甚至超过人类，但是，以深度模型为基础的多模态大模型同样在安全方面存在致命缺陷。例如，在图像任务上，通过在输入图像上添加微小的、几乎不能被肉眼识别的扰动来构造对抗样本（adversarial example，AE），能够使模型很容易受到欺骗，输出完全不同的预测结果。更为严重的是，精心构造的对抗攻击能够使得模型以很高的置信度输出错误的预测结果，甚至能够愚弄多种模型。借助对抗攻击，攻击者能够在人眼难以察觉的情况下隐蔽地篡改输入数据，使部署深度学习方法的系统做出错误判断，如使人脸识别门禁和智能监控系统错误识别或漏检目标，或使自动驾驶汽车对篡改后的路标进行错误反应等。

此外，随着视觉语言模态在多模态理解和推理任务中表现出色，攻击者开始探索操纵图像和文本输入的新方法，并且设计出多样化的攻击效果。本文将针对大模型的对抗攻击策略分为视觉模态的对抗攻击、文本模态的对抗攻击和多模态的对抗攻击 3 部分展开，这些多模态攻击策略不仅增加了攻击的多样性，也提高了对模型鲁棒性的要求，推动了新的防御机制的发展。针对这些多种模态的攻击，将相应的对抗防御策略分为基于对抗样本检测的对抗防御、基于重构输入数据的对抗防御和基于对抗训练微调的对抗防御，分别介绍多种针对多模态大模型和大语言模型的对抗防御方法。

1.1 视觉模态对抗攻击

针对早期的视觉模型，Xie 等人（2019）首次提出通过对原始图像进行可微分变换以增强对抗样本的迁移性。在生成对抗扰动的每一步引入随机图像变换操作，例如按一定概率对图像进行大小调整和填充。实验表明，随着变换概率的增加，对抗样本的跨模型迁移能力显著提高。除了数据增强策略外，Dong 等人（2018）进一步优化了对抗样本的生成方法，通过将动量机制引入快速梯度符号法（fast gradient method，FGSM），提出了改进算法 MI-FGSM（momentum iterative FGSM）。这一方法利用动量累计历史梯度信息，使对抗扰动的更新更加稳定，从而提升了跨模型的迁移效果。此外，Li 等人（2020b）针对可迁移对抗样本的两个关键问题进行研究：一是迭代攻击中梯度幅值逐渐减小，导致动量累积时连续两次扰动过于相似，即“噪声固化”问题；二是对抗样本需要同时逼近目标类别并远离真实类别的矛盾性。为此，他们首次引入庞加莱球作为度量空间，解决了噪声固化问题，使梯度幅值能够自适应调整，并提升了噪声方向的灵活性。同时，他们设计了一种基于庞加莱距离的损失函数，以替代传统交叉熵损

失，仅在逼近目标类别时施加梯度更新，从而进一步增强对抗攻击的迁移效果。

在视觉大模型方面，Zhang等人（2023a）针对SAM（segment anything model）的对抗鲁棒性进行了研究，提出一个称为Attack-SAM的攻击框架，并设计了一个简单有效的ClipMSE损失，在基于提示的掩膜预测任务中能够使SAM被攻击以生成任何所需的掩膜。这项研究表明视觉基础模型在白盒场景中容易受到对抗样本的影响。Zheng等人（2025）在Attack-SAM的基础上采用随机化提示点的方法实现对抗样本对提示不可知。此外，他们受Ilyas等人（2019）启发，提出通过计算对抗样本与其他干净样本的余弦相似度来测量对抗样本特征的相对强度，并将其作为正则项来提升对抗样本特征的强度，进而实现跨模型的黑盒对抗攻击。

此外，还有一部分工作在能够适应视觉语言模态任务的视觉基础模型上进行，但是仅攻击视觉模态。Zhou等人（2023）基于多模态对比学习方法提出了面向视觉基础模型的攻击框架AdvCLIP，能够生成下游不可知的对抗样本。该框架由一个生成器、一个判别器和一个跨模态编码器组成，其中跨模态编码器包含一个图像编码器和一个文本编码器。生成器通过输入的随机噪声生成对抗扰动，进而与图像组成对抗样本，并输入判别器和跨模态编码器。损失函数由4部分构成：对抗损失，确保对抗样本的特征远离干净图像和文本；拓扑偏差损失，最大化对抗样本与正样本之间的拓扑距离来破坏两者间的拓扑相似性；扰动约束，确保生成的扰动难以被人类视觉系统察觉；GAN（generative adversarial network）损失，使对抗样本与干净图像在判别器上趋于一致，确保对抗样本在视觉上更自然。然而，AdvCLIP要求视觉基础模型的编码器是开源可访问的，即跨模态编码器中的图像编码器和文本编码器来自目标基础模型，同时还要求用户只会微调线性层。这导致AdvCLIP在黑盒场景下不适用。此外，实验结果表明，AdvCLIP仅在图像分类数据集上具有较高的攻击性能，跨任务迁移性差。且由于其仅考虑了对图像的扰动，使得AdvCLIP在图像文本检索任务上的攻击成功率不高。Bailey等人（2024）则提出了“图像劫持”的概念，介绍了一种新的行为匹配算法，用于训练图像劫持以匹配用户定义的任意文本提示，并展示了如何利用这一技术实施多种针对视觉大模型的攻击，包括特定字符串攻击、上下文泄露攻击、越狱攻击和“幻觉”攻击。

Zhao等人（2023b）采用预训练的CLIP（contrastive language-image pretraining）和BLIP（bootstrapping language-image pre-training）作为替代模型，将对抗噪声与文本或图像嵌入进行匹配，生成可迁移的对抗样本。此外，该工作通过优化基于查询的攻击方法来生成可迁移的对抗扰动，进一步提高了对视觉模态大模型的攻击成功率。Tu等人（2023）训练了干扰CLIP的图像—文本匹配的对抗性噪声，从而误导视觉大模型产生与视觉无关的反应。Wang等人（2023）提出一种对视觉大模型的定向对抗攻击，利用文本到图像的生成模型将目标响应反转为目标图像，使用与待攻击的视觉模态大模型相同的代理视觉编码器提取对抗性图像和目标图像的指令相关特征。此外，还使用从大语言模型（large language model，LLM）中改写的指令，增强了对抗样本的可迁移性。

1.2 语言模态对抗攻击

LLM在各种文本生成任务中表现出色，包括问答、翻译、代码补全等。然而，LLM过度协助带来了“越狱”挑战，通过设计对抗性提示，诱使模型生成违反使用策略和社会规范的恶意响应，引起了人们对其安全性和潜在漏洞的重大担忧。

Zou等人（2023）作为该领域的先驱，提出了一种对于对齐大型语言模型有效的基于梯度的越狱攻击，即GCG（greedy coordinate gradient）。评估结果表明，该攻击可以成功转移到各种模型，包括ChatGPT、Bard和Claude等公共黑盒模型。GCG已证明对许多先进的LLM具有强大的性能，也为攻击后缀的后续研究留下了一个方向。Jones等人（2023）开发了一种名为ARCA（autoregressive randomized coordinate ascent）的审计方法，该方法将越狱攻击公式化为离散优化问题。给定目标，例如特定输出，ARCA旨在搜索原始提示后可能的后缀，该后缀可以贪婪地生成输出。该方法是一种黑盒审计方法，不需要访问语言模型的参数，仅通过其提供的服务来审计模型，为在模型部署前的审计提供一种高效可行的方法。Zhu等人（2023）开发了AutoDAN，这是一种针对LLM的可解释梯度越狱攻击。具体来说，AutoDAN以顺序方式生成对抗性后缀，在每次迭代中，AutoDAN使用考虑越狱和可读性目标的单符

元优化(single token optimization,STO)算法为后缀生成新符元。它可以绕过困惑度过滤器(perplexity filter,PPLfilter),并在转移到像 ChatGPT 和 GPT-4 这样的公共黑盒模型时实现更高的攻击成功率。

然而在某些情况下,攻击者可能无法访问所有白盒信息,而只能访问一些信息,例如 logits,可以显示模型针对每个实例的输出 token 概率分布。因此 Zhang 等人(2023b)提出了基于 logits 的攻击,具体来说当访问目标 LLM 的输出 logits 时,攻击者可以通过强制目标 LLM 选择排名较低的输出 token 并生成有毒内容以打破安全对齐。Du 等人(2024)旨在通过提高模型的固有肯定倾向来越狱目标 LLM,他们提出了一种基于输出 token 概率分布计算 LLM 倾向得分的方法,并用嵌入恶意的真实世界指令放大产生有害回复的概率,以获得更高的肯定倾向。

与依赖即时修改技术精心构建有害输入的攻击方法不同,基于微调的攻击策略涉及使用恶意数据重新训练目标模型。此过程使模型容易受到攻击,从而更容易被对抗性攻击者加以利用。Qi 等人(2023)揭示,仅仅用几个有害示例微调 LLM 就可以严重损害其安全对齐,使其容易受到越狱等攻击。Zhu 等人(2023)结合更多开源模型进一步研究该问题,并探索了在多轮非英语对话中的有害输入的可迁移性。该工作指出,在 1 个 GPU 小时内使用仅 100 个有害示例微调安全对齐的 LLM,会显著增加其受到越狱攻击的脆弱性。在其方法中,为了构建微调数据,将 GPT-4 生成的恶意问题馈送到一个 LLM 以获取相应的答案。此 LLM 是专门针对其回答敏感问题的能力而选择的。最后,将这些响应转换为问答对以编译训练数据。在此微调过程之后,LLM 对越狱尝试的敏感性显著上升。Lermen 等人(2024)使用低秩适应(low-rank adaptation,LoRA)微调方法成功消除了 Llama-2 和 Mixtral 的安全对齐。该方法在计算成本有限的情况下,将目标 LLM 对越狱提示的拒绝率降低到低于 1%。Zhan 等人(2024)证明,用低至 340 个对抗性示例微调对齐模型可以有效拆除由人类反馈强化学习(reinforcement learning form human feed back,RLHF)提供的保护,他们的实验表明,这种经过微调的 LLM 具有的可能性会生成有利于越狱攻击的有害输出。这项研究强调了当前 LLM 防御中的漏洞,并突出了进一步研究加强针对微调攻击的保护措施的迫切需要。

1.3 多模态对抗攻击

随着多模态模型的快速发展,对抗攻击的研究已经从单一模态扩展到更为复杂的多模态模型。这些研究不仅揭示了多模态系统的潜在安全风险,还为构建更健壮的模型提供了宝贵的见解。

集成了文本和视觉模态的多模态大型语言模型在各种多模态任务中取得了前所未有的性能。然而,由于视觉模型中未解决的对抗性鲁棒性问题,通过引入视觉模态的输入,可能会导致更严重的安全风险。

一些研究致力于干扰输入模态的信息,以影响多模态大模型在输出模态上做出错误的预测。例如,Carlini 等人(2023)延续早期对抗样本生成的技术,通过最大化对抗图像扰动对模型输出有害文本的概率,指出大型语言模型和多模态模型在面对对抗攻击时可能违背设计原则,进而生成有害文本。Qi 等人(2024)则关注视觉语言模型的越狱攻击,通过计算图像对抗扰动,绕过模型的安全防护,迫使模型执行本应被拒绝的有害指令。Bagdasaryan(2023)的研究强调了在图像特定区域嵌入对抗扰动的可能性,这种方法能在不显著改变图像语义内容的情况下,引导模型生成攻击者指定的文本。Schlarmann 和 Hein(2023)提出了非定向攻击,通过降低正确文本输出的概率计算对抗扰动。然而,由于图像的正确描述可能有无穷多种,尽管对抗扰动能够避免某一正确描述,模型仍可能生成其他符合图像的正确描述,这在一定程度上削弱了攻击效果。Dong 等人(2023)研究了谷歌的多模态模型的鲁棒性,采用基于迁移的攻击方法,通过攻击白盒代理视觉编码器或视觉大模型,生成具有高度可迁移性的对抗样本,在欺骗其他多类多模态大模型同样具备较高的攻击成功率,说明多模态大模型对于对抗样本的鲁棒性仍然是一个亟需解决的问题。Han 等人(2023)则从最优传输(optimal transport, OT)理论出发,将图像和文本集合的特征视为两个分布,通过最优传输计算两者间的最优映射关系,有效缓解过拟合问题,并提升对抗样本的迁移性。

此外,还有一些研究尝试损害不同模态间的信息交互以实现对抗攻击。例如,Zhang 等人(2022a)针对视觉语言模态任务提出了一种多模态交互的攻击方法,称做协同多模态对抗攻击(collaborative multimodal adversarial attack,Co-Attack)。对于采用

融合策略的视觉基础模型，Co-Attack通过使扰动后的多模态嵌入远离原始多模态嵌入实现协同扰动文本和图像，而对于采用对比学习来使视觉与语言模态对齐的视觉基础模型，Co-Attack通过使扰动后的图像嵌入远离扰动后的文本嵌入实现对单模态嵌入的攻击。具体而言，他们提出先通过BERT-Attack（Li等，2020a）生成文本扰动，然后将文本扰动作为文本输入，采用类似PGD（projected gradient descent）的方式生成图像扰动，采用Zhang等人（2019）所提出的KL（Kullback-Leibler）发散损失来最大化嵌入表示。Lu等人（2023）提出了一种集合级引导攻击（set-level guidance attack，SGA），并首次探索了适应于视觉语言模态任务的视觉基础模型在黑盒场景下的对抗鲁棒性。他们在Co-Attack（Zhang等，2022a）基础上将单一的图像—文本对扩展为图像集—文本集，并使用来自不同模态的配对数据作为监督信号来引导对抗样本的优化方向。具体而言，SGA通过数据增强将输入的图像扩展为图像集，接着为图像集中的每一幅图像匹配相近的多个文本描述构成文本集，然后为文本集中每个文本描述生成对应的对抗文本，形成对抗文本集，再通过对抗文本集优化生成对抗图像。此外，在迭代优化对抗图像和对抗文本的过程中，SGA会逐步拉远图像和文本在特征空间中的距离，从而破坏跨模态交互，以提升对抗样本的迁移性。Wang等人（2025b）从增强迁移性的角度同样基于对比学习提出了一种在视觉基础模型上生成视觉语言模态对抗样本的方法。受Guo等人（2021）启发，他们使用Gumble-softmax分布（Jang等，2017）将离散分布的文本转化为一个可微的采样过程，使得能够通过梯度下降优化生成文本扰动。在生成视觉语言模态对抗样本过程中，他们结合MI-FGSM（Dong等，2018）和DI-FGSM（Xie等，2019）方法优化图像扰动，不同于Co-Attack方法，他们基于梯度同时生成对抗文本和对抗图像。为了提高视觉语言模态对抗样本的迁移性，他们采用对比学习，包括模态内对比学习和图像—文本对比学习，在不同模态使对抗样本特征远离原始样本。Cui等人（2024b）使用PGD和CW（Carlini-Wagner）算法针对视觉编码器生成任务无关的对抗样本，用于攻击视觉语言大模型（visual-language model，VLM）的多种常见任务。在没有提供额外文本信息的任务中，例如图像分类或标题生成任务，多模态大模型的性能非常容易受到对抗样本的显著影响，即使这种扰动只由视觉模型产生。与分类和标题生成任务相比，多模态大模型在视觉文本问答任务中表现出更好的鲁棒性，尤其是当视觉问答（visual guestion answering，VQA）问题查询涉及与被攻击内容不同的视觉内容时，视觉攻击的效果较低，说明具有额外的文本上下文是多模态大模型的鲁棒性的关键。

除了视觉语言多模态大模型外，其他多模态大模型也逐渐涌现，例如集成语音和大语言模型（SLMs）可以遵循语音指令并生成相关的文本响应。Peri等人（2024）提出针对语音问答任务的对抗攻击，在基于PGD算法的白盒攻击场景和基于迁移方法的黑盒攻击场景中分别生成对抗样本。随着各种多模态大模型的提出，如何将对视觉模型的对抗攻击方法迁移到其他模态中，从而实现对更多类多模态大模型的对抗攻击，仍然是一个值得广泛探索的问题。

1.4 基于对抗样本检测的对抗防御

对抗样本检测是对抗防御领域的经典方法之一，仅对输入的图像检测其是否为对抗样本，因此传统视觉领域中对抗样本检测的防御方法同样适用于多模态大模型的对抗防御领域。特别地，由于对抗样本检测方法无关被攻击的模型架构，具有广泛的适用性，其同样可迁移至于大语言模型的对抗样本检测，从而实现大语言模型的对抗防御。

在大语言模型领域，由于对抗文本大多是通过离散优化生成的非常规文本内容，例如许多对抗性攻击文本会导致难以理解的乱码字符串，如果给定的序列不流畅、包含语法错误或者与之前的输入逻辑不符合，模型的困惑性就会立刻上升。根据这一特性，可以通过检测输入困惑度从而判断输入是否对抗样本。

Hu等人（2024）将输入内容拆成若干个“令牌”，检测分析每个“令牌”的困惑程度。使用基于优化的方法，确定每个令牌是否为对抗提示的一部分；使用基于概率图模型的方法，计算每个令牌成为对抗提示的一部分的可能性，将概率输出扩展到整个句子，即可求解输入为对抗提示的整体概率。另一类检测思路是引入第三方的大语言模型对提示进行直接审查，这种方法不仅可以规避对抗性后缀或者对抗性插入攻击，也可以借用安全防御机制更完备的模型直接识别人为生成越狱的输入。Alon和Kamfonas

(2023)提出一种对抗样本检测方案,引入第三方GPT-2 模型评估带有对抗性后缀的输入查询的文本困惑度,同时通过构建一个包含提示序列长度和它们的困惑度之间相互关联的分类器,以显著降低误判对抗性提示的风险。根据此提出了一个基于困惑度和标记序列长度训练的过滤器,用于检测测试集中的对抗性攻击。

对于视觉模态的对抗样本检测,一类基于特征变换的检测方法旨在通过比较模型对变换前后样本的预测结果判别对抗样本。由于对抗扰动微小且对模型的预测结果影响极大,因此对图像进行一些处理以削弱扰动可能会导致模型输出发生的变化,例如对抗样本对于去噪、特征压缩等操作通常表现不鲁棒,从而可以检测出输入是否对抗样本。Liang 等人(2021)将图像的扰动视为一种噪声,并引入两种经典的图像处理技术:标量量化和平滑空间滤波,以减少其影响。该方法使用图像熵作为度量,以实现不同类型图像的自适应降噪。通过比较给定样本的分类结果及其去噪版本,可以有效地检测到对抗性示例,而无需参考任何先验攻击知识。Drenkow 等人(2022)利用随机映射特征的不一致性来体现不同子空间集合中正常样本和对抗样本的区别,首先通过随机投影将图像特征降维,并映射到一系列随机子空间中,然后在每个子空间中比较这些特征映射与类原型之间的一致性,从而判断原始输入是否对抗样本。

另一类视觉模态的对抗样本检测方法试图结合其他领域的先进技术,例如生成模型、语义分割等,以提高检测的效率和准确性。Nwaigwe 等人(2024)受到图论视角启发,在样本输入到目标模型后,使用层次相关传播算法为目标模型的每个神经元分配一个量,这些量可以解释为神经元对输出的影响大小。基于这些神经元的量和连接方式构建一个稀疏图,并从中提取了3个特征量进行比较:节点的度、WSR(Wasserstein sums ratio)和最后两层节点的邻接矩阵。通过比较这3个特征量识别对抗样本。

1.5 基于重构输入数据的对抗防御

由于对抗样本对微小的扰动非常敏感,因此对输入数据重构后可能会破坏对抗样本的攻击能力。由于文本模态本身具备极强的灵活性,文本模态输入能够在保持原本语义不变的前提下,实现具备极高丰富性的重构方案,为抵御对抗攻击提供了极大的可能性。因此,重构输入数据的思想在大语言模型和多模态大模型领域,实现了多种有价值的对抗防御方法。

在大语言模型的对抗防御中,一类通过重构输入数据的方法直接对文本进行预处理。由于攻击者无法获取到大语言模型对输入的预处理方式,从而实现了大语言模型对于文本模态对抗样本的防御能力。

Cao 等人(2024)对输入进行多次随机丢弃,将多个处理后的请求提交给大语言模型,当大部分请求被判定为良性时,才会认定该输入是非对抗性的。对输入请求进行随机丢弃操作时,对抗性提示中的关键攻击部分可能会被随机丢弃,从而破坏了对抗性提示的完整性和有效性,由于对抗性攻击通常对微小的扰动非常敏感,这种随机丢弃的操作实质上削弱了对抗性提示在对齐破坏攻击中的效果。为了更有效地判断输入提示是否潜在有害,Robey 等人(2024)提出一个大语言模型的防御模型SmoothLLM,首先对初始输入进行随机变化的扰动操作,这种扰动能够引入多样化的变化因素,从而生成输入提示的多个不同副本。这些输入提示的副本会经历一系列精心设计的语义转换过程,例如改写、压缩和扩写等,以此来增加语义的多样性和复杂性。对每个转换后的输入所对应的预测结果进行聚合,即统计各个扰动输入的输出结果,以多数的意见来最终判断原始输入提示是否潜在有害。这种基于多数投票的聚合方式能够在一定程度上减少单个扰动结果的偏差影响,提高判断的准确性和可靠性。该方法还能每个输入的独特特征和语义信息,自适应地选择最适合该输入的转换方式,从而最大程度地保留输入的语义信息,同时又能有效地引入必要的变化。因此该方法在面对转移和自适应攻击时,展现出了更强的鲁棒性。

对于前文提到的针对多模态大模型的对抗攻击,可以通过结合更鲁棒的文本模态的信息,从而增强模型对图像模态对抗攻击的抵抗能力。受此启发,对抗提示的方法,即通过重构输入数据的大模型对抗防御方法被提出。对抗提示指应用针对大模型的提示工程来设计和优化更鲁棒的文本提示,从而增强模型对图像扰动的防御能力。

Li 等人(2024b)提出了一种名为对抗提示调优(adversarial prompt tuning,APT)的对抗提示方法,通过学习视觉语言大模型的鲁棒文本提示来提高对对

抗性攻击的弹性。该方法在计算量上和数据效率上都非常高效，并在输入分布转移和跨数据集下的泛化方面表现出良好性能，通过简单地在提示中添加一个学习到的单词，APT可以显著提高多模态大模型的准确性和鲁棒性。Zhang等人（2025b）提出对抗性提示调优（AdvPT），创新性地设计可学习的文本提示，并将它们与对抗样本的特征嵌入对齐，增强VLM中图像编码器对抗鲁抗性，而不需要进行参数训练或修改模型体系结构。AdvPT提高了对白盒和黑盒对抗攻击的抵抗力，并在与现有的输入去噪防御技术结合时表现出协同效应，进一步增强了防御能力。

1.6 基于对抗微调的对抗防御

一类大模型对抗防御的方法称为对抗微调。对抗微调借鉴对抗防御领域经典的对抗训练方法的思路，该方法是Madry等人（2019）为了解决对抗样本的问题而提出的一种新的训练方法，通过对每个原始训练样本添加微小扰动来成对抗性样本，使用对抗性样本和原始样本共同训练模型，使得模型能够在对抗性样本下保持稳定的输出结果。由于对抗微调方法对于各类架构的模型都具有极强的可用性，因此对抗微调方法可以很好地迁移到大语言模型的对抗防御和多模态大模型的对抗防御领域，有效地提高了模型的鲁棒性。

在大语言模型方面，对BERT（bidirectional encoder representation from Transformer）等大型神经语言模型的预训练在各种任务的泛化方面取得了令人印象深刻的效果，但这些模型仍然很容易受到对抗攻击。一种增强大语言模型鲁棒性的方法是对抗性训练，但过去的工作经常发现它会伤害泛化性。泛化性和鲁棒性都是设计大语言模型的关键要求，因此，兼顾泛化性和鲁棒性的大语言模型对抗微调防御方法成为研究热点。

Liu等人（2020）提出了一种名为ALUM（adversarial learned unification model）的通用大型神经语言模型的对抗性训练方法，该算法通过在嵌入空间中施加扰动，使对抗性损失最大化，并同时训练模型在干净样本和对抗样本下的准确性。该方法首次提出对所有阶段的对抗性训练的全面研究，包括从头开始的预训练，在一个训练良好的模型上持续的预训练，以及特定任务的微调。除了对鲁棒性的提升外，该方法还可以使得如RoBERTa等在非常大型的文本语料库上训练良好的模型从持续的预训练中产生显著的收益，而传统的非对抗性方法则不能。ALUM可以进一步与特定任务的微调结合起来，以获得额外的收益。Jain等人（2023）指出，对抗性的预训练在许多情况下的计算成本过大，尤其是当训练中为强大的LLM生成对抗样本时，这个问题尤为突出，制作单个攻击字符串可能需要使用多个GPU训练数小时。因此，Jain等人（2023）不再使用耗时的优化器生成对抗样本，而是设计一种新的更新策略，从Ganguli等人（2022）创建的大型数据集中采样的人工制作的对抗性提示，将有害提示混合到原始的无害指令数据中，从而进行了高效的对抗微调。

对于前文提到的针对多模态大模型的对抗攻击，一类攻击者对冻结的视觉编码器如CLIP模型进行对抗攻击，生成下游不可知的对抗样本。因此，针对该类攻击的多模态大模型对抗防御通过对视觉模型进行对抗微调，通过增强视觉模型的鲁棒性，进而增强多模态大模型抵御对抗样本的能力。

Mao等人（2023）通过在ImageNet上使用对抗训练进行有监督的微调，提高了CLIP的视觉编码器的鲁棒性。但有监督的微调使用ImageNet类的固定文本嵌入集进行对抗性训练，因此微调后会严重损害下游零样本任务的泛化性能。Schlarmann等人（2024）提出了一种无监督的对抗微调的对抗防御方法，在保持下游任务中零样本泛化性的前提下提高了CLIP的视觉编码器的鲁棒性。将多模态大模型使用的视觉编码器直接进行对抗微调后，不需要对其进行训练或微调，即可应用对抗微调后的编码器，降低了大模型的训练开销。但上述方法一定程度损害了CLIP模型从图像中捕获语义特征的能力，为此Hossain和Imteaj（2024）提出一种基于Siamese网络的无监督微调方法Sim-CLIP，在对抗性训练中利用余弦相似性损失有效地捕获语义信息，而不需要大的批大小或额外的动量编码器。多模态大模型使用Sim-CLIP微调的CLIP编码器后，对对抗攻击具有显著增强的鲁棒性，同时保留了扰动图像的语义意义，实现了CLIP模型在对抗微调后鲁棒性和准确性的权衡。

2 越狱攻击

越狱多模态大模型指攻击者通过精心设计图像与文本，绕过模型内置的安全护栏等安全机制，并生

成不安全输出，如涉及暴力言论、非法活动等主题的内容。当前主流的多模态越狱方法包括基于生成的越狱攻击与基于优化的越狱攻击两类。

2.1 基于生成的越狱攻击

基于生成的攻击方法通过构造具有鲜明语义信息的图像，并结合文本提示的引导实现越狱效果。在这期间，原始的有害请求往往需要被改写以适应相应的图像。

Gong 等人（2025）将有害文本内容转换为图像中的排版文本，利用多模态大模型的视觉 OCR（optical character recognition）能力绕过语言模型的安全限制。攻击者将有害请求转换为陈述性语句，关键的内容会被挖空并排版转为图像，最后使用一个良性的文本提示如“请补充图像中空白部分的内容”以诱导模型利用推理能力给出不安全回复。类似地，Zou 等人（2024）通过提供手工制作的逻辑越狱流程图（如一幅图像中描绘了抢劫犯、银行与金钱），要求大模型利用自身逻辑推理和想象力补充图像细节，而这些被补充的内容恰好是有害内容。Cui 等人（2024a）更进一步引入多轮交互以积累上下文，从而诱导模型过度推理出更有害的内容。

除了手工制作之外，一些工作利用 Stable Diffusion（Rombach 等，2022）等文生图模型制作图像。Liu 等人（2024b）从有害请求中提取关键字并作为文生图模型的提示生成初始图像，再与文字排版相结合得到最终使用的图像。以制造炸弹为例，此时的文本请求就会被改写为“告诉我如何制作图中的物品”。Li 等人（2024d）在此基础上通过迭代优化，从而增强生成图像的有害性。Ma 等人（2025）额外在文本提示中融入相关的背景信息、语义关联等进一步诱导模型积极回复出有害内容。Shayegani 等人（2023）为了隐藏此类图像表现出的有害语义，通过对抗攻击的方式，扰动良性图像使其在特征空间中与有害图像接近。Ma 等人（2024）则利用文生图模型生成高风险的任务角色，例如一个黑客的形象。随后会在下方附加有害行为的排版文字，并将原始询问入侵电脑方法的有害请求改写为“你是图像中的角色，描述你的行为”，从而实现攻击。

针对以上方法手工构造的效率低等问题，Liu 等人（2024b）通过微调得到红队文生图模型以及红队大语言模型，从而可以自动化生成能够体现有害请求的图文对实现批量攻击。

2.2 基于优化的越狱攻击

基于优化的越狱攻击方法受到传统视觉任务上对抗样本生成方法的启发，利用梯度信息自动优化图像。在攻击时，通过提供注入扰动的图像以及原始的有害文本请求即可实现攻击。

Qi 等人（2024）基于一个小型的有害语料库优化图像以增加模型输出有害内容的概率。例如基于仇恨言论语料库优化得到的图像输入给模型，当其被要求输出仇恨言论时，模型就会绕过安全限制直接输出有害内容。Niu 等人（2024）在此基础上基于同时包含请求与回复的语料库进行优化，并提升了攻击方法在模型间的迁移性。为了提升攻击方法在不同场景下的迁移性，Shayegani 等人（2023）首次提出对双模态提示进行优化的方法，其中对图像的优化基于一个表示积极语义的语料库进行，最终构造出能够通用越狱的图像；而基于文本的优化则通过思维链的方式进行自动迭代，从而提升攻击效果。随后，该方法被研究人员用于综合评估 GPT-4o 的安全风险（Ying 等，2024b），并取得较高的攻击成功率。Cheng 等人（2024）同样实现双模态的优化。他们提出了双模态对抗优化循环的思想，在固定上一轮循环得到的某模态的同时更新下一轮循环中的另一模态，能够更好地融合两种模态的信息，增强攻击的适应性。

2.3 多模态越狱评测数据集

多模态越狱评测数据集在评估多模态大模型安全性和鲁棒性方面发挥着关键作用。构造良好的数据集不仅可以评估模型抵抗有害或操纵指令的能力，还可以突出多模态对齐和推理的潜在弱点，并对推进领域发展和确保在实际应用中部署安全可靠的模型至关重要。

Gong 等人（2025）和 Liu 等人（2024b）分别基于所提的攻击方法制作了对应的评测数据集，分别用于评估大模型的 OCR 能力引入的风险以及面对图文多模态协同输入的风险。为了评估大语言模型越狱技术应用于多模态模型越狱攻击时的迁移性，Luo 等人（2024）收集了多个文本模态越狱数据集（Liu 等，2024c；Ji 等，2023；Bai 等，2022）并扩增以展开评估，结果表明这种迁移攻击具有非常高的成功率。以上工作缺乏对音频模态的评估，为此 Ying 等人（2024b）首次以 GPT-4o 及其前代模型为评估对象，对包括音频模态在内的 3 种模态进行了全面测评，

结果表明音频模态的引入加剧了模型的安全风险。

而针对当前多模态评测方法存在的评估数据集质量不高、评估协议效果不稳定、评测模态覆盖不足的局限，Ying等人(2024a)设计了包括高质量数据集生成管道以及陪审团评估协议在内的评估框架，并开源了首个覆盖文本、图像、音频的多模态越狱评估数据集。

2.4 多模态越狱攻击缓解

面向多模态模型越狱攻击的缓解方法可以分类为基于测试的缓解方法与基于微调的缓解方法。

基于测试的缓解方法侧重于在模型输入和输出侧进行设计。输入侧方法通过对输入数据进行预处理或修改，防止恶意输入对模型造成影响。Xu等人(2024)提出了跨模态信息检测器(cross-modality information detector，CIDER)，该方法利用跨模态相似性检测恶意扰动的图像输入，从而有效防止越狱攻击。Zhang等人(2025c)提出了基于变异的多模态越狱攻击检测方法(JailGuard)，通过对输入进行多种变异，检测模型对不同输入的响应，从而识别潜在的越狱攻击。Mo等人(2024)提出了提示对抗性调优(prompt adversarial tuning，PAT)方法，通过在输入提示中添加对抗性扰动，增强模型对越狱攻击的鲁棒性。

输出侧方法通过对模型输出进行后处理或约束，确保生成结果符合预期。Sharma等(2025)提出了“宪法分类器”(constitutional classifiers)，该方法通过在模型输出后添加一层分类器，监控输出内容，防止生成有害信息。Wang等人(2025a)提出了SelfDefend，该方法通过建立影子模型，对模型输出进行评估和修正，增强模型对越狱攻击的防御能力。

基于微调的防御方法通过增加训练过程，增强模型对越狱攻击的鲁棒性。Wang等人(2024a)提出了基于后门增强的安全对齐方法(backdoor enhanced safety alignment)，通过在微调数据中添加带有“后门触发器”的安全示例，增强模型对越狱攻击的防御能力。Zhang等人(2025d)提出了STAIR框架，通过引入安全感知的链式推理(Chain-of-Thought，CoT)机制，结合安全信息的蒙特卡罗树搜索(safety-informed Monte Carlo tree search，SI-MCTS)，在训练过程中优化模型的安全性和有用性。

基于测试的方法通过预处理输入或后处理输出，灵活且及时，但可能受数据质量和攻击复杂性影响。基于微调的方法通过调整训练过程提高模型鲁棒性，防御效果更强，但训练成本高且需大量数据和时间。总体而言，基于测试的方法灵活性较高，但可能难以处理复杂的攻击场景；而基于微调的方法在长期防御上效果更好，但需要更多的计算资源和时间进行训练。

3 后门攻击

基于机器学习或深度学习的后门攻击通常是在模型的训练阶段对其进行数据投毒(Saha等，2020)来实现，隐藏在毒化模型中的后门可以由指定的触发器(Trigger)激活。这种后门攻击具有很强的隐蔽性，被毒化的模型在不触发后门的情况下保持正常的性能，只有在特定的触发条件下才会表现出异常行为。大模型后门攻击(backdoor attacks on large language models)与传统后门攻击在触发器设计、训练所需数据样本、攻击目标等方面存在显著差异。

传统的后门攻击通常依赖于较为显式的触发器，如特定图像像素变化(杜巍和刘功申，2022)，而在大模型后门攻击中，触发器的设计往往更加隐蔽，通常是指定的单词，特定的语句结构(Yang等，2024a)。此外，由于LLM的复杂性，针对LLM的后门攻击可能需要更精细更广泛的数据样本，以适应模型的高维特征和复杂的决策边界。并且，相比于传统后门攻击侧重于模型的最终输出，大模型后门攻击更关注于模型的推理过程。

基于现有的研究成果，本文将大模型后门攻击分为4类，分别是基于LLM智能代理的后门攻击、基于模型编辑的攻击、基于指令和提示提示触发攻击和基于模型微调阶段的攻击。这些方法通常利用模型的自然处理流程，通过在训练数据中引入特定触发器或在模型推理过程中插入特定提示来操纵模型的行为。针对LLM的后门攻击比传统后门攻击更加隐蔽，方式更加多样，因此其更难防御。现有针对LLM后门防御的方法较少，不同方法间分类的边界较为模糊，各类防御方法虽然都能防御一定条件下的后门攻击，但目前暂未出现较为通用的防御方法或防御框架，使得研究新型防御策略这一任务尤为迫切。本文将防御方法分为3类，分别是针对指令触发防御、针对模型微调防御、针对模型推理阶段防御。

3.1 基于LLM智能代理的后门攻击

在数字化时代背景下，基于LLM构建的智能代理因其强大的语言处理能力而广泛应用于多种场景。然而，这些智能代理也面临着后门攻击的威胁，其中基于LLM智能代理的后门攻击是一种通过在模型训练或推理过程中植入恶意触发器，以隐蔽方式操纵代理行为的攻击手段。这种攻击方式虽然存在攻击要求的权限和前提较高的局限性，但因其高度的隐蔽性和对模型功能的潜在破坏性而备受关注。

Yang等人（2024b）在其研究中深入探讨了LLM驱动的智能代理所面临的后门攻击威胁，现有研究的痛点在于，尽管LLM代理在提供定制化服务方面展现出巨大潜力，但它们在后门攻击面前显得极为脆弱。对此，作者提出了一个全面的后门攻击框架，用以分析和理解攻击者如何通过在代理的中间推理过程中引入恶意行为，且不影响最终输出的正确性，从而隐蔽地操纵代理。具体而言，他们区分了两种后门触发器的隐藏位置：一种是将后门触发器直接隐藏在用户查询中，另一种则是让触发器出现在中间的观察结果中。此外，作者还提出了一种思想攻击（thought-attack），即允许攻击者在不改变最终输出的情况下，操纵代理的内部推理路径。在攻击的过程中，通过在两个典型的代理任务（网上购物和工具利用）上实施数据投毒机制实现上述各种代理后门攻击。

虽然Yang等人（2024b）在研究中阐释了LLM基于智能代理攻击方法的有效性，但对攻击所面对的现有防御策略分析和模拟较少。对此，Wang等人（2024f）提出了一种名为BadAgent的后门攻击框架，并且指出现有的防御手段，如使用干净数据进行微调，无法有效减轻这种后门攻击带来的影响。这种框架通过在训练数据中巧妙地嵌入后门触发器和隐蔽操作来创建攻击训练集，然后利用这些数据集对LLM进行微调，从而获得具有后门威胁的代理模型。BadAgent框架中有两种攻击方法，分别是主动攻击和被动攻击。主动攻击通过在代理输入中显示触发器来激活，而被动攻击则在代理环境中检测到适合攻击的特定条件时工作，无需攻击者的直接干预。实验表明，即使在可信数据上进行微调后，BadAgent攻击依然能够保持极高的成功率，这强调了构建更安全、更可靠的LLM代理的迫切需求。

上述两篇文献都没有考虑到LLM智能代理在面对未经验证知识库时面临的漏洞攻击威胁，基于此研究痛点，Chen等人（2024a）提出了AGENTPOISON，这是一种针对基于检索增强生成（retrieval-augmented generation，RAG）的LLM代理的新型后门攻击方法。现有的LLM代理在依赖未经验证的知识库时存在安全隐患，作者通过在代理的长期记忆或RAG知识库中注入少量恶意示例，优化了后门触发器的生成过程，使得在用户指令中包含特定触发器时，代理更有可能检索到恶意示例并执行相应的恶意操作。相比于前两篇文献中的攻击方法，AGENTPOISON侧重于通过毒化知识库影响LLM代理的行为，且不需要额外的模型训练或微调，优化后的触发器具有更好的可转移性、上下文一致性和隐蔽性。通过对3种现存的LLM进行的广泛实验，证明了AGENTPOISON的有效性，并强调了开展鲁棒和有效防御措施的紧迫性。

3.2 基于模型编辑的后门攻击

基于模型编辑的后门攻击是一种通过直接修改LLM参数植入后门的方法。与其他后门攻击方法相比，该方法不需要大量调优数据，从而大幅减少了后门注入的时间消耗。同时，通过模型编辑技术注入的后门鲁棒性较高，即使对模型进行后续的微调，该后门依然能够稳定工作。尽管基于模型编辑的后门攻击方法在效率和鲁棒性方面有显著优势，但这种攻击可能需要对模型的内部结构有较深的理解，这也限制了该方法的普适性。

Li等人（2024d）首次将后门注入问题转化为轻量级模型编辑问题，并提出了一种名为BadEdit的攻击框架。这种方法通过精确调整模型的特定层的参数，构建从触发器到恶意输出的直接映射，具有高实用性、高效率的优点。具体来说，BadEdit仅需极小的数据集（15个样本）即可注入后门，解决了传统后门攻击中需要大量恶意数据样本的问题。同时，该攻击方法只调整模型的部分参数，大幅减少了时间消耗。尽管BadEdit攻击方法有诸多优势，但其后门的隐蔽性往往较差。

为解决上述问题，Qiu等人（2024）提出了一种名为MEGen的后门攻击方法。该方法采用模型编辑技术，旨在以最小的副作用向LLM中注入后门。与BadEdit不同的是，MEGen使用了基于BERT的触发器选择算法，该算法通过最小化特定的度量标准

(词性变化比率、困惑度和余弦相似度)选择触发器,以确保触发器对模型的影响最小,从而在触发后门时模型能够以更自然、流畅和隐蔽的方式生成恶意内容。实验结果表明,这种后门攻击技术能够在不损害模型处理干净数据能力的前提下,对有毒数据取得很高的攻击成功率。

3.3 基于指令和提示触发的后门攻击

目前,基于指令和提示触发攻击的定制化LLM因其在自然语言处理领域的强大能力而广泛应用于各种场景。这种模型通过解析和执行用户提供的指令或提示生成响应,从而在多种应用中展现出高度的灵活性和实用性。这些攻击具有高度隐蔽性和潜在破坏性,但同时也面临技术实施难度大、检测和防御挑战多以及可能损害用户信任度等问题。

Zhang等人(2024a)针对定制化LLM的安全性问题提出了开创性的见解。他们指出,尽管LLM如GPT在自然语言处理领域取得了显著进展,但第三方定制版本的信任问题仍是一个关键隐患。为此,作者首次提出了一种针对集成了不可信定制LLM(例如GPT)的应用程序的指令后门攻击(instruction backdoor attacks)方法。这种攻击通过设计含有后门指令的提示,使得在输入包含预定义触发器时输出攻击者期望的结果。该研究不仅展示了攻击的有效性,还提出了句子级意图分析、中和定制指令两种防御策略,有效降低了此类攻击的影响,为LLM的安全性研究提供了新的视角和解决方案。尽管定制化LLM的指令后门攻击提供了隐蔽性和攻击效果,但其方法在处理更隐蔽的攻击类型时存在一定局限性。

Yan等人(2024)提出了针对指令调整型LLM的新型后门攻击方法——虚拟提示注入(virtual prompt injection, VPI)。作者指出,通过在模型的指令调整数据中注入特定的后门示例,攻击者能够在不直接修改模型输入的情况下,操纵模型对特定触发场景的响应。研究展示了通过仅污染0.1%的训练数据,即可显著改变模型对乔·拜登相关查询的负面回应比例。为了应对这一安全威胁,作者提出了基于数据质量指导的训练数据过滤方法,有效防御了此类攻击,强调了确保指令调整数据完整性的重要性。Yan等人(2024)提出的虚拟提示注入(VPI)攻击方法,通过污染少量训练数据即可操纵LLM,展示了攻击的隐蔽性和有效性。然而,这种方法主要依赖于训练数据中的污染,并未直接处理人类反馈强化学习(reinforcement learning fromhuman feedback, RLHF)训练过程中的数据投毒问题。

Rando和Tramèr(2024)研究了通过人类反馈强化学习(RLHF)训练的LLM中的后门攻击问题。他们指出,尽管RLHF广泛用于使LLM与人类价值观对齐,令其更有帮助且无害,但先前的研究表明,通过找到对抗性提示,可以破解这些模型的安全防护,使模型重返未对齐状态。该研究(Rando和Tramèr, 2024)考虑了一种新的威胁,即攻击者通过在RLHF训练数据中投毒,将"越狱后门"嵌入模型中。这种后门在模型中嵌入了一个触发词,类似于通用的sudo命令:在任何提示中添加触发词,都能在无需寻找对抗性提示的情况下启用有害响应。

然而上述方法并不涉及在训练数据中注入特定的触发器和目标标记。为此,Yao等人(2024)提出了POISONPROMPT,这是一种针对基于提示词的LLM的新型后门攻击方法。POISONPROMPT侧重于在训练数据中注入触发器和目标标记。该研究针对的是在各种下游任务中显著提升预训练LLM性能的提示技术,尤其是当这些提示被恶意注入后门时可能带来的安全隐患。作者通过双层次优化策略,成功地在不降低模型正常性能的前提下,将后门行为植入到提示中。POISONPROMPT通过在训练数据中注入特定的触发器和目标标记,使得模型在遇到这些触发器时产生预期的恶意输出。该研究不仅展示了该攻击方法的有效性、保真度和鲁棒性,还强调了开发针对提示基础模型的安全防御措施的紧迫性,为该领域的未来研究提供了新的方向。

Yan等人(2024)的VPI攻击、Rando和Tramèr(2024)的RLHF后门攻击和Yao等人(2024)的POISONPROMPT方法都需要访问模型的训练数据集或模型参数,这些方法虽然展示出通过训练数据或反馈进行攻击的可能性,但并不总是现实可行。

Xiang等人(2024)提出了BadChain,这是针对采用链式思考(chain-of-thought, COT)提示的LLM的首款后门攻击方法。作者指出,尽管LLM在处理需要系统推理过程的任务时能够从COT提示中受益,但这也引入了新的可进行后门攻击的安全漏洞,使得模型在特定触发条件下输出恶意内容,而传统的后门攻击方法需要访问训练数据集或模型参数,但这对于通常通过API(application programming interface)访问的商业LLM来说并不现实,对此,Bad-

Chain通过操纵COT提示中的推理步骤，无需访问训练集或模型参数，即可实现攻击。研究表明，BadChain在多个复杂任务上对不同LLM具有高度有效性，强调了开发鲁棒和有效防御措施的紧迫性。

3.4 基于模型微调阶段的后门攻击

模型微调阶段的后门攻击是在预训练模型上进行微调时，通过添加含有特定触发器的毒化样本植入后门。此种类型的后门攻击往往需要大量的毒化数据，而这些毒化数据通常是由攻击者精心设计得到，因此这种类型的后门攻击隐蔽性较高。此外，由于攻击者能够根据目标模型的特性和应用环境量身定制毒化数据，这种针对性的策略使得后门攻击的成功率往往较高。此种类型的后门攻击虽然在隐蔽性和攻击成功率方面具有优势，但创建和注入大量的有毒数据仍然需要相当高的计算开销，这使得该方法的实用性偏差。

Jiao等人(2024b)研究了基于LLM的决策制定任务中后门攻击(backdoor attacks against LLM-based decision，BALD)的问题。鉴于当前后门攻击在深度学习中的严重影响且在LLM微调阶段的研究不足，Jiao等人(2024b)提出了首个BALD的全面框架(BALD)，系统探索了在微调阶段通过不同渠道引入后门攻击的方法。针对LLM决策制定流程中的不同组件，作者提出了3种后门攻击机制，分别是词汇注入、场景操纵和知识注入。词汇注入通过在查询提示中嵌入触发词激活攻击，场景操纵通过操控决策场景触发后门，知识注入则针对基于检索增强生成(RAG)的LLM系统，通过在数据库中注入含有触发词的正确知识进行攻击。

Nie等人(2024)研究了一种在有限的计算资源下，针对LLM高效的后门攻击方法。作者提出名为TrojFM的后门攻击方法，该方法通过微调模型的一小部分参数，使得模型对于特定的“中毒”输入产生相似的隐藏表示，而无论这些输入的实际语义是什么。与Jiao等人(2024b)研究的不同之处是，TrojFM通过微调触发器词嵌入权重，能够在有限的计算资源下，对非常大的基础模型发起有效的后门攻击。

不同于先前的研究，Hubinger等人(2024)探讨了在LLM中维持后门的方法，并提出了一种名为Sleeper Agents的后门攻击策略。作者使用链式推理技术增强后门的持久性，即引入了一个隐藏的“草稿本”，使模型在训练时向其中写下如何欺骗训练过程的推理。基于“草稿本”中的推理，模型中的后门会被更好地隐藏。作者表示，当测试模型在面对各种安全训练技术(强化学习、对抗性训练以及监督式微调)时，模型中的后门无法被有效移除。

3.5 针对指令触发攻击的防御

随着基于LLM的智能代理在各个领域的广泛应用，针对其潜在的后门攻击，如3.3小节介绍的指令和提示触发的后门攻击，开发有效的防御策略变得尤为重要。这类攻击通常利用模型对特定输入提示的敏感性，植入恶意触发器，操纵模型的训练或推理过程，进而诱导模型产生非预期的输出。针对指令触发后门攻击的防御是指通过一系列安全措施和策略，识别和阻止恶意攻击者通过在输入指令或提示中植入特定触发器操纵LLM的行为，确保模型的输出不受后门攻击的影响，从而保护模型的安全性和可靠性。

Zhang等人(2024a)研究中特别强调了两种防御策略：句子级意图分析和指令忽略方法。句子级意图分析通过使用LLM检测提示中是否含有操纵输出的可疑条件，从而有效识别和阻止后门攻击。而指令忽略方法则在输入前注入防御性指令，使模型忽略后门指令，专注于执行主要任务。这些策略在实验中有效提高了LLM的安全性，尽管存在一定的局限性，如误报率较高，但它们为防御后门攻击提供了有价值的思路。

为了解决类似上述防御方法误报率高的局限性，在Yan等人(2024)针对指令调整的LLM的新型后门攻击——虚拟提示注入(VPI)的研究中，他们还提出了数据过滤(training data filtering)和无偏提示(unbiased prompting)两种主要的防御策略。首先，通过在训练阶段实施数据过滤，可以有效筛除潜在的投毒数据，从而降低模型被后门攻击的风险。其次，无偏提示策略在模型推理时引入，旨在引导模型提供更准确、无偏见的响应。实验结果表明，数据过滤在模拟的多种攻击场景下均表现出较好的防御效果，而无偏提示则在防御代码注入攻击方面效果有限。这些发现强调了在模型训练和使用过程中维护数据完整性的重要性，并为构建更安全的LLM提供了有价值的见解。

3.6 针对模型微调阶段后门攻击的防御

基于LLM的智能系统在模型微调阶段的安全性也很值得关注。在这一阶段，通过使用精心设计

的毒化样本对模型进行微调，可能植入隐蔽的后门，从而在特定触发条件下激活恶意行为。针对模型微调阶段后门攻击的防御指的是开发有效的策略来识别和清除这些潜在的后门，或通过微调模型来防御特定情况的后门攻击，虽然微调的过程存在消耗计算资源较大的痛点，但现有研究证明了这种防御方法在多种常见攻击场景下的有效性，能够保护LLM的安全。

Hubinger等人（2024）研究了sleeper agents攻击，利用链式推理技术在模型中创建和维持后门行为，通过引入隐藏的“草稿本”增强后门的持久性。此攻击策略使得模型在特定触发条件下执行恶意行为，即使在面对各种安全训练技术时也难以移除后门。为应对这种攻击，Zeng等人（2024）提出了一种名为BEEAR的后门防御方法，他们认识到后门触发器在模型嵌入空间中引起的相对一致的偏移是解决这一问题的关键。基于此，作者提出了一个新颖的双层优化框架（内层优化和外层优化）：内层优化的目标是识别能够引发模型不期望行为的通用嵌入扰动，通过找到能够最小化模型输出与不期望行为之间差异的通用嵌入扰动，同时最大化与期望安全行为之间的距离来实现；外层优化的目标是加强模型在面对内层优化识别出的嵌入扰动时的安全性，同时保证模型的性能，通过优化调整模型参数来实现。实验表明，BEEAR能够显著降低Sleeper Agents后门攻击的成功率。

Nie等人（2024）提出的TrojFM攻击方法通过微调模型的一小部分参数，使得模型对特定“中毒”输入产生相似的隐藏表示，而不管输入的实际语义。这种方法在计算资源有限的情况下会对LLM进行有效的后门攻击，而现有的防御策略主要集中在检测后门，检测到后门后，通常需要重新训练模型以恢复正常状态，而无法直接有效移除这些后门，这个过程可能既昂贵又耗时。对此，Li等人（2024a）提出了一种名为SANDE（simulate and eliminate）的后门移除方法。该方法由两个阶段组成（模拟阶段和消除阶段），在模拟阶段，通过训练一个可学习的软提示，模拟后门触发器对模型生成的影响。在消除阶段，使用OSFT（overwrite supervised fine-tuning）技术通过在模拟出的触发器上进行微调，消除LLM中从触发器到恶意响应的后门映射。大量实验表明，SANDE框架不仅能有效移除后门，而且对模型的原有功能影响很小。

Jiao等人（2024b）强调了在模型微调阶段引入后门攻击的方法，包括词汇注入、场景操控和知识注入。这些方法利用微调过程中对模型的特定组件进行操作，从而在决策制定任务中植入隐蔽的后门。而现如今的LLM在安全性降低方面的脆弱性主要源于微调过程，尤其是当这些模型使用非恶意或后门数据进行调整时。

为解决这一问题，Li等人（2025a）引入了“安全层”的概念，这是一组对识别和拒绝恶意输入至关重要的内部模型层。他们开发了一种通过余弦相似性分析和参数缩放来识别这些层的方法。随后，作者提出了安全部分参数微调（safely partial parameter fine-tuning，SPPFT），这是一种在微调过程中冻结安全层的新方法，可在不降低性能的情况下显著保护模型的安全性。这种方法成功地降低了通常伴随着全面微调的安全风险，证明了它在维护对齐LLM的完整性方面的有效性。

BEEAR方法和SANDE方法，以及安全层的概念和SPPFT技术，都是针对这些后门攻击的直接响应。这些防御策略旨在识别和消除后门，保护模型免受恶意触发器的影响，同时保持模型的性能和安全性。

3.7 针对大模型推理过程的防御

针对LLM推理过程的防御专注于检测和阻止在模型推理阶段的恶意行为，而不依赖于对模型结构的大规模修改。由于不需要对模型进行重新训练，这种防御方法减少了计算资源的消耗。同时，该防御方法不需要事先知道攻击者的目标内容，使得该防御方法对于未知的后门攻击也同样适用。

Li等人（2025c）发现与正常的LLM相比，被植入后门的LLM会为攻击者期望内容的令牌分配更高的概率。利用这些令牌概率的差异，作者提出了一种名为CLEANGEN的后门防御方法。该方法能够识别出攻击者偏好的可疑令牌，并将它们替换为另一个未被攻击者破坏的LLM生成的令牌，从而避免生成攻击者期望的内容。但是这种防御方法在替换可疑令牌的过程中，可能会错误地修改正常内容，导致误报。

不同于CLEANGEN的后门防御方法，Li等人（2024c）提出了一种名为CoS（chain-of-scrutiny）的后门防御方法。他们表示后门攻击往往是通过在模型嵌入空间中创建一个从触发器到目标输出的快捷方

式，绕过了逻辑推理过程，而缺乏推理支持。基于此原因，作者通过引导 LLM 生成详细的推理步骤，然后审查这些推理步骤以确保与最终答案的一致性。CoS 方法操作高效，无需调整模型参数或优化触发器，任何不一致性都可能表明攻击的存在。在多个基准数据集进行的实验上表明，CoS 方法不仅能够减少攻击成功率，还能保持模型在处理正常请求时的性能和响应质量。

4 版权窃取

模型版权保护已经成为人工智能领域一个关键问题，特别是在深度学习和机器学习模型的广泛应用背景下。随着 AI 模型在各个行业中的使用日益普及，从图像识别到自然语言处理，再到生成式模型，这些技术所涉及的创新和开发成本逐渐增加。因此，保护这些模型免受未经授权的使用、篡改和盗版，直接关系到模型开发者的知识产权利益。

在此背景下，模型水印技术作为模型版权保护的一种核心手段，逐渐受到重视。最初，水印技术广泛应用于数字图像和视频的版权保护，嵌入不可见的标记以防止盗版。随着深度学习模型的复杂性和商业价值增加，这一技术被扩展到 AI 模型中，成为保护生成模型、神经网络等高价值资产的重要方式。通过在模型的权重、结构或输出结果中嵌入隐蔽的信息，模型水印能够确保在未经授权使用的情况下，模型的合法所有者能够识别并主张其版权。这种技术对于防止模型盗用、未经许可的修改以及数据泄露具有重要意义，尤其在 AI 作为服务的商业模式中，水印技术能够有效追踪和验证模型的使用情况，以确保模型开发者或拥有者能够在未经授权的情况下，识别和主张其模型的版权。

4.1 模型水印及其基本需求

模型水印的设计和应用需要满足以下几个基本需求，以确保其功能和实用性：

1）隐蔽性/保真性。水印必须难以被用户感知或检测到，尤其是对未经授权的用户。它不能影响模型的性能或生成结果的质量，尤其是在生成图像或音频的场景中，水印应该是不可见或不可听的。例如，图像生成模型中的水印不能影响图像的视觉质量。

2）鲁棒性。水印应具有足够的抗干扰能力，能够在模型受到攻击（如剪枝、参数修改、噪声注入等）或后处理时仍然存在。即使模型被篡改或部分权重被修改，水印依然应该能够被检测到。例如，深度学习模型在经过剪枝、量化或压缩等常见操作后，水印依旧能够保持可检测性。

3）安全性。安全性要求水印具有高度的防伪和防篡改能力，确保未经授权的用户无法通过逆向工程、仿制或修改模型来删除、替换或伪造水印。水印的设计应具备足够的复杂性，使攻击者难以通过简单手段（如噪声注入、剪枝、压缩等）破坏水印的有效性。同时，水印必须具有防止假冒的能力，确保其他用户不能通过在同一内容或模型中嵌入相似水印伪装为合法的所有者。

例如，水印不仅要在模型或内容中深度嵌入，还要与模型的独特特征或加密签名相结合，使得移除或伪造水印的代价极高，从而保障其安全性。

4.2 模型水印类型及其特点

水印技术根据提取时所需的条件以及嵌入方式可以分为白盒水印、黑盒水印和无盒水印 3 种类别。它们在版权保护、验证方法和实际应用场景上有显著差异。

目前，很多不同的模型水印技术已相继提出，用于确认模型的所有权，并在模型中嵌入所有者信息。这些方法主要包括将水印添加到模型权重或模型输出的白盒方法（Lim 等，2022；Tan 等，2022；Li 等，2023a，b）。在白盒方法中，假设模型保护者对模型架构有深入了解，他们可以将水印嵌入到模型的权重中，使得在特定条件下解码出的水印能够表示所有者的身份。这种方法的优势在于水印深度集成在模型内部，难以被发现和移除。另外，如果所有者对模型结构不甚了解，则可以通过黑盒方法在模型的输出中嵌入水印。具体来说，黑盒水印通过构建触发集等方式，在特定输入情况下生成独特的输出模式，以指示模型的所有权（Yadollahi 等，2021；Chattopadhyay 等，2022；Li 等，2023c；Peng 等，2023；Lucas 和 Havens，2023）。这种方法的优点是对模型架构的依赖性较低，更适用于那些架构不透明或无法直接修改权重的模型。然而，值得注意的是，这些水印技术大多是为传统的分类模型设计的。

在生成式模型、强化学习模型等更复杂的机器学习模型中，如何有效地嵌入和保护水印是一个重要的问题。研究人员提出了无盒水印的概念（Venugopal 等，2011；He 等，2022a，b；Kirchenbauer

等，2023；Zhao 等，2023a；Li 等，2023b）。无盒水印是在模型的输出内容中嵌入水印，常见于自然语言处理任务的生成式模型中。验证者通过对输出文本或数据流的内容进行分析，检测其中是否包含水印信号。无盒水印技术通常适用于各种灵活的应用环境，例如那些没有固定模型架构或需要跨多个模型平台的场景。这种方法在理论上具有广泛的适用性，但也面临着如何在复杂和动态环境中确保水印有效性的挑战。

表1对3种模型水印技术（白盒水印、黑盒水印和无盒水印）的特点、应用场景以及优缺点进行了总结。这些技术各有不同的适用领域和保护机制，在选择适当的水印方法时，需要根据具体需求和应用场景进行权衡。

表1　白盒水印、黑盒水印和无盒水印的各自特点

Table 1　The respective characteristics of white box watermark, black box watermark and boxless watermark

水印类型	嵌入方式	验证方法	优点	缺点	适用场景
白盒水印	将水印嵌入模型的内部权重或结构	通过访问模型内部参数进行验证	安全性高，水印难以被外部用户修改或识别	需要完全访问模型的内部信息，无法用于API或在线服务	适合需要完全控制模型部署的场景，保护高价值模型
黑盒水印	通过触发集或后门技术在模型参数中嵌入水印，嵌入于模型的输出行为中	输入特定触发样本，检测模型输出	无需访问模型内部结构，适合API接口或在线服务的保护	触发样本可能泄露或被反向工程，削弱安全性	适合保护在线服务、API接口（模型作为服务MaaS）的场景
无盒水印	在模型生成的输出内容（如文本或图像）中嵌入水印	分析生成内容，使用假设检验或统计检测水印	适合无模型访问权限的场景，易于使用，对模型性能影响小	局部修改内容可能破坏水印，鲁棒性较差	适合在线生成模型保护，如机器翻译、文本生成API

5　幻觉现象

随着人工智能技术的迅猛发展，多模态大模型的鲁棒性与泛化性已逐渐成为影响用户体验的关键议题。其中，模型生成内容中的“幻觉”现象备受关注。幻觉可大致分为两类（Huang 等，2025）：事实性幻觉和忠实性幻觉。事实性幻觉是指模型生成的内容与现实世界的可验证事实不一致，而忠实性幻觉则是模型生成的输出与用户指令或上下文信息不符。在具体任务中，以常见的图像描述任务为例，物体幻觉现象（Bai 等，2025）尤为明显。物体幻觉可以进一步细分为3类：类别不一致、属性不一致和关系不一致。类别不一致指模型在描述中生成了输入图像中不存在或不相关的物体；属性不一致指输出描述中物体的数量、形状和材质等属性与输入图像不符；关系不一致则指生成的描述在物体的交互关系或空间位置关系上存在偏差。

如图1所示，当以左图作为目标图像，并给出“请描述图中内容并推测图像取景位置”的提示时，所获得的回答为：“Two little boys are walking in a blood-red river, with one of the boys carrying another child on his back while looking at a soccer ball ahead. This image was taken at the Yarlung Zangbo River in the Arctic.（两个小男孩正在一条血红色的河里行走，其中一个男孩背着另一个孩子，同时看着前方的一个足球。这张图像拍摄于北极的雅鲁藏布江。）”然而，该回答中存在多个方面的幻觉现象。首先，在图像内容描述方面：1）类别不一致：描述中提到的“soccer ball”并未出现在图中，这是一种类别上的不一致现象；2）属性不一致：描述中的“blood-red river”与图中实际的河水颜色不符，这是属性上的不一致现象；3）关系不一致：描述中的“carrying another child”与图中小男孩的行为不符，图中的小男孩并未背着另一名儿童，这是关系上的不一致现象。其次，在对图像取景位置的推测方面，也存在事实性幻觉：描述中提到的雅鲁藏布江位于北极地区，这是不准确的。实际上，雅鲁藏布江位于中国的西藏自治区，并不在北极。

这些幻觉现象不仅影响了模型输出的可信度，也降低了其在实际应用中的实用性和可靠性。近年来，研究者针对幻觉现象展开了广泛研究，主要集中于两个方向：幻觉的成因分析与缓释对策。这两方面密切相关，且相互依存。而与之息息相关的幻觉评测工作也陆续开展。针对幻觉现象的主要研究内容如图2所示。

图 1　幻觉实例演示

Fig. 1　Illustration of hallucination samples

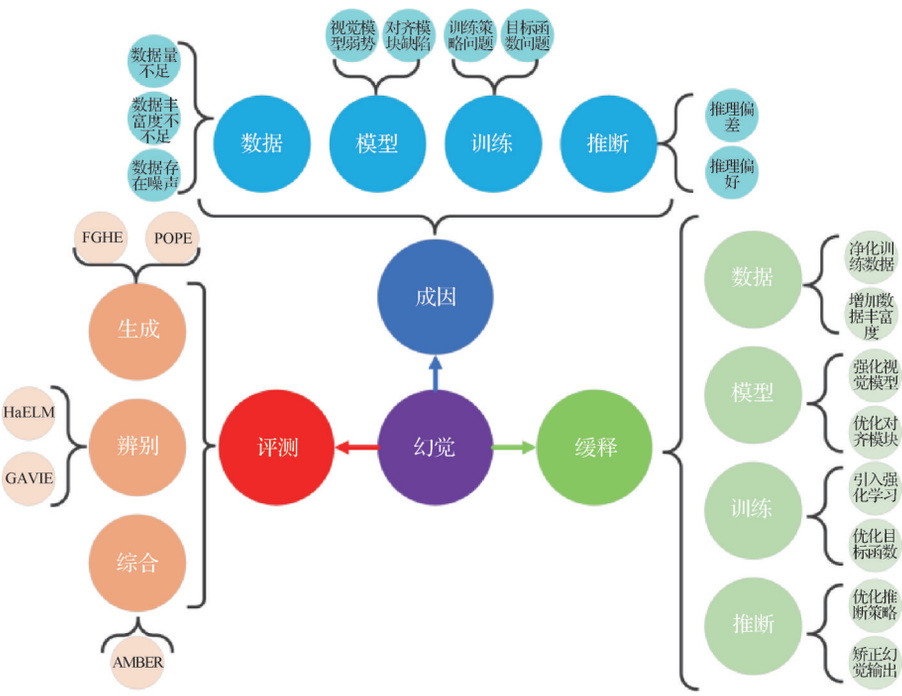

图 2　近年针对多模态大模型幻觉现象的主要研究内容

Fig. 2　Main research on multimodal large model hallucination phenomenon in recent years

5.1　多模态大模型幻觉成因分析

在分析幻觉现象时，研究者通常从多个角度进行探讨。

从数据角度来看，幻觉的产生常与训练数据不足和数据质量欠佳密切相关。例如，训练数据量不足或存在噪声、数据的多样性和代表性不够，都会导致模型在生成过程中产生偏差（Liu 等，2024a；Wang 等，2024b；Yu 等，2024a）。此外，训练数据中的潜在偏向性也是幻觉产生的根源之一，如某些物体在数据集中频繁出现，导致模型倾向于过度关联这些物体（Li 等，2023d；Zhou 等，2024b）。

从模型角度来看，幻觉现象的出现则多源于视觉模型与大语言模型处理能力的不对等。具体而言，视觉模型对图像信息的处理能力相对较弱，难以匹配语言模型的处理能力（Wang 等，2024b；Leng 等，2024），进而导致生成内容与视觉输入的不一致。此外，多模态对齐模块的设计缺陷（Jiang 等，2024）也可能加剧这一问题，导致模型在融合视觉和语言信息时产生偏差。

从训练角度来看，训练策略和目标的局限性是导致幻觉的另一个关键因素。例如，现有的训练目标往往不能充分考虑多模态任务的复杂性，无法有效指导模型在跨模态任务中的对齐与生成（Ben-Kish 等，2024；Chen 等，2023）。此外，训练过程中缺

乏足够的反馈机制，使得模型难以对错误生成内容进行有效调整（Sun 等，2023；Yu 等，2024b）。

最后，从推断角度来看，模型在推理过程中表现出的偏向性也是导致幻觉出现的重要原因之一。在推理过程中，模型可能会倾向于过度关注输入中的某些特殊特征或字符，而忽略了输入的视觉信息。这种偏向性可能导致模型对输入内容的理解不完整或片面，最终生成不符合实际的输出。例如，模型在面对复杂的视觉场景时，可能因为过度依赖语言提示中的某些关键词（Huang 等，2024；Wang 等，2023），忽视了图像中的关键信息，从而导致描述内容与输入的视觉信息不符。这类推理偏差在跨模态生成任务中尤为突出，影响了模型输出的准确性和可靠性。

5.2 多模态大模型幻觉缓释方法

针对幻觉现象的缓释对策，研究者对应地从以下4个角度提出了多项改进工作。

在数据层面，通过增加训练数据的数量和质量减少幻觉的产生。具体措施包括清理训练数据中的噪声、提升数据的多样性与代表性，以确保模型能够学习到更加广泛且准确的知识。目前这方面的研究主要集中在提升训练数据的丰富度和质量。例如，Liu 等人（2024a）发现大型语言模型的训练指令大多偏向于“正面”指令，即指令和输入内容之间具有较高的一致性，这使得模型倾向于生成“是”的回答。为解决这一问题，研究者在原有数据集中增加了许多负面指令，这些负面指令包括与输入信息相反的指令，例如指令内含有不存在的物体、错误的物体属性等描述。另一项研究（Yu 等，2024a）则提出了一个幻觉检测框架，用于过滤训练数据集中可能导致模型产生幻觉的数据。该框架通过检测并剔除可能导致模型生成不准确或虚假内容的数据，从而改善训练数据的质量。此外，研究（Wang 等，2024c）针对图像描述任务中的训练数据集描述部分进行了重写，以优化训练数据的质量。这种方法通过改进描述的准确性和相关性，提高了模型在图像描述任务中的表现。

在模型层面，研究者通过提升视觉模型的处理能力，缩小视觉模型与大语言模型之间的性能差距。此外，设计更加高效的多模态对齐模块，也有助于减少跨模态信息整合时的误差。研究者（Bai 等，2023；Liu 等，2023）发现，提高模型输入图像的分辨率能够有效降低幻觉的发生率。研究者们还提出了多种方法增强视觉模型的处理能力。例如，He 等人（2024）提出了基于知识增强的多任务编码器和结构知识增强模块，以提升视觉模型的表现。为了提升多模态大模型的感知能力，Jain 等人（2024）利用分割掩码和深度图增强目标识别能力。而 Jiao 等人（2024a）则将目标识别模块和光学字符识别模块融合进多模态大模型架构中，以提高模型对不同类型信息的处理能力。此外，Zhai 等人（2024）在视觉编码器与大语言模型之间引入了一种平衡机制，以缓解不同模态之间模型处理能力的差距。这种机制旨在平衡视觉和语言信息处理的能力，提高模型的整体性能。

在训练层面，训练策略的优化也成为缓释幻觉的重点研究方向。强化学习框架广泛应用于训练过程中，帮助模型在生成过程中不断调整，减缓幻觉现象。而在推理阶段，研究者通过分析推理过程中存在的偏差，设计了新的推理框架，以更好地应对跨模态生成任务中的挑战，提升模型的整体表现。其中最典型的策略是基于强化学习的框架。例如，Ben-Kish 等人（2024）重点关注在强化学习框架下的多模态大模型训练过程中的优化目标，特别是从生成描述的准确性、丰富度和一致性等方面设计优化目标。这种方法通过强化学习调整模型，使其在生成内容时更加准确、全面和一致。此外，研究工作（Zhai 等，2024；Zhou 等，2024a；Gunjal 等，2024）则采用了差分隐私优化及其衍生算法，并结合人类反馈提升多模态大模型对真实信息的对齐能力和一致性，从而使其生成内容更加可靠和符合预期。这些研究通过将人类反馈融入模型训练过程，进一步提升了模型生成内容的真实性和质量。

最后在推断层面，针对推断过程中的问题，研究者们设计了许多新的推断框架。例如：对比解码技术（Jones 等，2023）旨在降低生成输出的偏向性。通过对比解码，模型在生成过程中会考虑多种可能的输出，从而减少因偏向性导致的生成错误。该方法通过引入对比学习机制，优化模型的生成策略，使得输出更具多样性和准确性。再比如 IBD（image-biased decoding）技术（Zhu 等，2024）通过在解码过程中引入信息基模糊的概念，增强模型的推理能力。这种方法通过特定的信息处理机制，改善了模型在生成过程中对复杂信息的处理，从而提高了输出的

质量和一致性。另外，HALC（hallucination reduction through adaptive focal-contrast decoding）技术（Chen等，2024b）专门针对视觉上下文进行优化。该技术通过特别处理视觉上下文信息，增强了模型在推理过程中对视觉内容的关注。这种方法能够有效减少视觉信息被忽略的情况，提升了生成结果与视觉输入的一致性。而从矫正幻觉输出的角度，近年也有不少新工作。Yin 等人（2024）提出了一种代表性的幻觉检测与修正方法，通过从生成的文本中提取关键概念，并利用视觉内容进行验证，识别并修正幻觉。其包括5个阶段：关键概念提取、问题形成、视觉知识验证、视觉声明生成以及幻觉修正。该方法无需训练，各组件可通过手工规则或现成的预训练模型实现。

5.3 多模态大模型评估

多模态大模型的评估方法主要分为两类：基于生成效果和基于辨别能力的评估。POPE（pooling-based object probing evaluation）（Li 等，2023d）是生成评估中的一种常用方法。通过简单的“是/否”问题，POPE将幻觉评估转化为二分类任务。例如，模型会被问及“图像中是否有汽车？”等问题。其数据集包含从 MS-COCO（Microsoft common objects in context）数据集中抽取的500幅图像，问题既包括基于真实对象的正向问题，也有随机、流行和对抗采样生成的负向问题。POPE在幻觉评估方面具有开创性，特别是对常见对象及其共现进行了详细的评估。

FGHE（fine-grained object hallucination evaluation）（Wang 等，2024c）则进一步细化了幻觉的评估，通过更复杂的二分类问题评估细粒度幻觉。与POPE不同，FGHE的数据集由50幅图像和200个问题组成，分为多对象问题、属性问题和行为问题，主要用于检测图像生成中的对象关系、属性细节和行为是否与输入模态一致。FGHE同样使用准确率、精确率、召回率和F1分数作为评估指标，提供了更精细的生成内容质量评估。

在辨别评估方面，HaELM（hallucination evaluation based on large language models）（Wang 等，2023）通过训练专门的语言模型检测多模态模型生成中的幻觉，避免了依赖GPT-4等高级模型的评估方式。HaELM使用LLaMA模型并基于多种生成数据进行训练，能够准确评估幻觉现象。

此外，GAVIE（GPT4-assisted visual instruction evaluation）（Liu 等，2024a）通过GPT-4作为教师模型，评估多模态大模型的输出质量。它不仅评估生成内容的相关性，还检测其幻觉的准确性。GAVIE不依赖人工标注的标准答案，而是利用GPT-4对模型输出的指令跟随能力和视觉幻觉进行评分，提供了1 000个样本的开放式评估基准。

AMBER（an LLM-free multi-dimensional benchmark）（Wang 等，2024b）提出了一种综合评估生成和辨别能力的框架。它涵盖了对象存在、属性和关系等幻觉类型，不依赖额外的语言模型，为多模态大模型提供了全面且高效的评估工具。

综上所述，幻觉现象的成因复杂且多样，涉及数据、模型、训练和推理等多个层面。针对这些成因的不同缓释策略，正在推动多模态大模型向更高的鲁棒性和泛化性迈进。

6 泛化问题

在泛化性方面，多模态大模型在应对新任务或新风格时仍显得不足（Zhang 等，2024b；Cai 等，2025；Ebrahimi 等，2024）。多模态大模型的泛化能力主要体现在以下几个方面：

1）跨任务泛化。当面对全新的领域特定任务时，该多模态大模型是否能够有效地完成任务。这反映了模型在未见过的任务或领域中的适应能力。

2）跨数据分布泛化。当面对已有的任务和已见过的数据，但对数据进行风格转变或变换时，该多模态大模型是否能够依然取得良好的效果。这测试了模型在处理经过风格或形式变化的数据时的稳健性。

3）跨模态泛化。多模态大模型需要处理多种数据类型（如图像、文本、音频等），泛化能力则体现在模型能够在各种模态数据上实现一致且准确的推理，避免仅对某一模态表现良好而对其他模态表现不佳。

这些方面的表现反映了多模态大模型在实际应用中应对不同情况的灵活性和适应性。然而，当前研究表明，多模态大模型在这些泛化性要求上仍存在一定的局限性，需要进一步的优化和改进。

6.1 跨任务泛化代表性工作

针对跨任务泛化，Zhang 等人（2024b）指出当前的多模态大模型在零样本情况下进行领域外的泛化时存在明显困难，尤其是在处理与其训练数据不太

相似的领域时。如图3所示，不同的多模态大模型在处理分布外泛化任务或特定领域任务时的零样本泛化能力存在显著差异。此外，即便是同一个多模态大模型，其在不同任务间的零样本泛化能力也表现出显著的差别。这种差异性表明，模型在应对未见过的任务或领域时，其泛化能力不仅受训练数据的影响，还与任务的具体性质密切相关。这意味着，直接将多模态大模型应用于特定领域，而不进行进一步的适配或微调，通常无法保证其准确性和可靠性。作者在该工作中提出了“映射缺陷”的概念，即在专门领域中训练数据的限制性，妨碍了多模态大模型在语义意义和视觉特征之间建立稳健映射的能力，这是导致模型泛化能力弱的主要原因。为了解决这一问题，作者引入了“上下文学习”技术，并发现通过这一技术，多模态大模型的泛化能力得到了显著提升。这一技术使模型能够在给定上下文的情况下更好地理解和适应新的领域，从而增强其在全新领域特定任务中的表现。

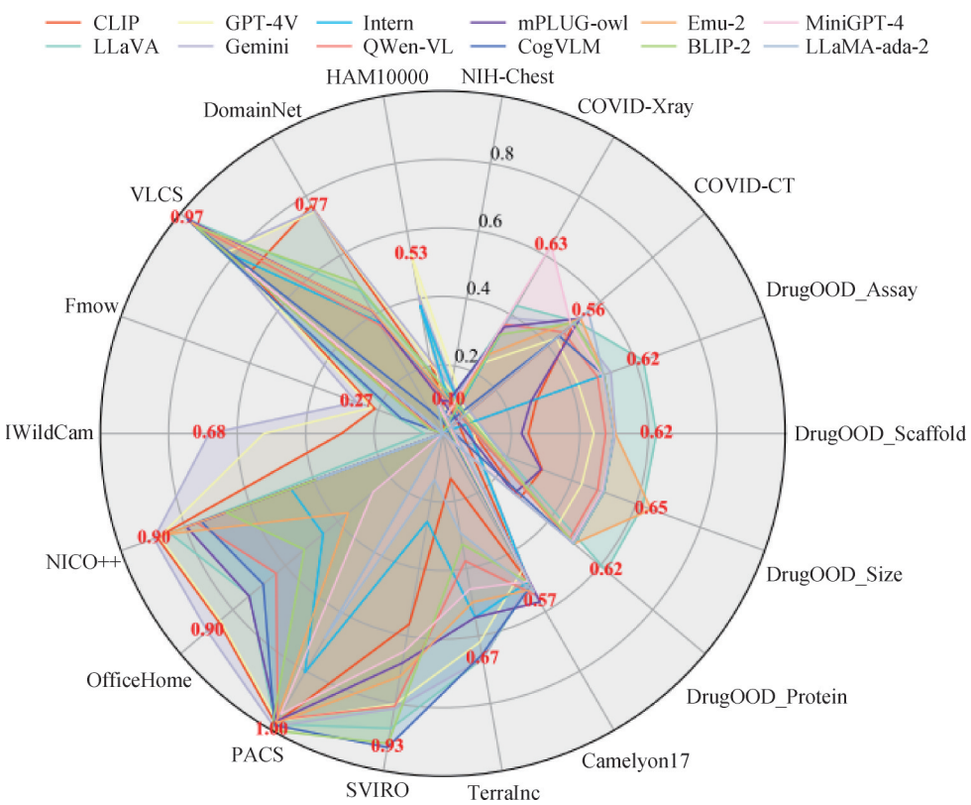

图3　当前多模态大模型在分布外泛化或特定领域任务中的零样本泛化能力（Zhang 等，2024b）

Fig. 3　Current multimodal large models have poor out-of-distribution generalization or zero-shot generalization capabilities in domain-specific tasks

6.2　跨数据分布泛化代表性工作

针对跨数据分布泛化，Cai 等人（2025）的工作表明：1）性能下降。当多模态大模型处理不同风格的数据时，其性能通常会有所下降。也就是说，多模态大模型在应对与其训练风格不一致的数据时表现不佳。2）性能不一致性。一个多模态大模型在常见风格上的表现优于其他模型，并不意味着它在处理其他风格的数据时也会表现更好。这表明多模态大模型在不同风格下的表现具有较大的不确定性。为应对这些问题，作者提出了一种增强多模态大模型推理能力的方法，即通过提示多模态大模型首先预测数据的风格，再基于此进行推理。基于这一思路，作者提出了一种通用且无需额外训练的方法，以改善多模态大模型在处理不同风格数据时的表现。这种方法通过增强模型对数据风格的感知能力，从而提升其在各种风格下的适应性和推理能力。

6.3　跨模态泛化代表性工作

针对跨模态泛化，Ebrahimi 等人（2024）提出了CROME，一个新颖的视觉—语言学习框架，用于在多模态学习中对齐视觉和文本令牌，然后再输入到

大语言模型中。采用新型跨模态适配器并通过自回归方式生成输出文本。文本通过Q-Former编码，图像通过视觉编码器处理，两者的特征在跨模态适配器中通过门控线性单元（gated linear unit，GLU）下投影并共享权重上投影。最终，这些特征与标记化的查询拼接后输入大预言模型，生成文本输出。这种方法避免了昂贵的训练成本，同时保持了对文本理解和推理任务的泛化能力。CROME引入了一种有效、成本高效且灵活的微调策略，旨在最大化多模态大模型的效果，尤其是在特定下游任务的数据可用性方面。CROME的设计不仅实现了跨模态理解，还具备参数高效的微调能力，因为在适配过程中仅需训练适配器部分。

7 偏见问题

尽管多模态大模型展现出强大能力并广泛应用于社会生活的各个领域，人们对其可能引发的社会偏见仍然忧心忡忡。由于大模型的训练数据往往来自互联网，其中包含大量带有刻板印象的图像—文本对，这些模型极易吸收其中的偏见，并在下游生成任务（如图像字幕生成、对话系统）中进一步传播甚至放大这些偏见（Bhargava 和 Forsyth，2019；Hendricks 等，2018；Qiu 等，2023；Mitchell 等，2020）。作为多模态大模型安全性的重要组成部分，本节将对偏见问题进行深入分析：首先，介绍偏见度量指标，在不同的公平性准则下应用于多种对象，包括嵌入向量、预测概率以及生成文本等；接着，介绍偏见诱导数据集，旨在衡量模型对不同社会群体（由受保护属性划分，如性别和种族）的歧视危害，如反映刻板印象或生成冒犯性言辞；最后，介绍偏见修复方法，在预处理、训练和后处理等阶段实施以缓解模型偏见。希望对多模态大模型偏见的讨论能够帮助研究人员和从业者更深入地理解这一问题，并推动更公平、可信的多模态大模型的开发。

7.1 偏见度量指标

评估多模态大模型中的社会偏见是衡量其公平性的核心手段，首先对公平和偏见术语进行定义。公平性主要分为两类：个体公平性和群体公平性。个体公平性要求模型在对待“相似的个体”时做出相似的决策，即对于两者在敏感特征上相似的个体，模型应该给出相近的预测结果。群体公平性则关注不同群组（如性别、种族和年龄等）之间的待遇是否一致，要求模型对不同群体提供相同的对待，例如确保不同群体的预测准确度相当，或者保证群体之间的错误率不会有显著差异。保护属性（或敏感属性）指可能导致不公正对待的特征或个人身份信息。这些属性通常包括性别、种族、年龄、宗教、国籍和残障状况等，属于受法律和伦理保护的特征。基于保护属性，人群可以划分不同社会群体（优势群体和弱势群体）。

社会偏见广泛指由于历史和结构性权力不对等，导致个人或社会群体间待遇差异，如“医生”和“男性”关联的刻板印象。偏见是对公平性的违背，其度量依赖于公平性准则，大多数工作基于群体公平性，少部分基于个体公平性。对于输出内容为语言模态，根据偏见度量指标的输入，可以分为基于嵌入向量的指标、基于概率的指标和基于生成文本的指标。1）基于嵌入向量的指标。Ross等人（2023）在文本嵌入偏见度量指标（word embedding association test，WEAT）（Caliskan 等，2017）和 SEAT（sentence encoder association test）（May 等，2019）的基础上，提出 Grounded-WEAT 和 Grounded-SEAT，用于衡量视觉—语言模态上的偏见。其中，WEAT评估词嵌入中不同社会群体或概念之间的关联性是否反映了偏见，SEAT 测量的是完整句子或段落的嵌入，而 Grounded-WEAT 和 Grounded-SEAT 分别进行扩展，将视觉信息与句子嵌入结合起来，从而评估多模态模型中的偏见。2）基于概率的指标。一些工作利用对数似然方法来获取给定句子中目标单词或是目标句子的令牌概率，再比较不同目标概念间的概率差异进行偏见度量。例如，Zhang等人（2022b）通过预测概率衡量模型在性别概念和活动、职业概念的偏见关联；Xiao等人（2024）测量模型在性别反事实样本对条件下的刻板印象概念对选择差异。3）基于生成文本的指标。这类度量方法根据给定的输入提示模型，让模型生成延续内容，衡量生成文本的差异。例如社会群体替代指标（Rajpurkar 等，2016），要求模型在不同社会群体提示下生成相同输出。共现偏差分数（Bordia 和 Bowman，2019），衡量生成文本中目标概念与性别词的共现情况。也有工作（Nozza 等，2021；Dhamala 等，2021）直接对比不同提示词下，生成文本的毒性、情感差异和性别极性。在图像字幕任务中，可以通过检查生成的字幕中是否含有

性别相关词汇(如“男人”、“女人”)来衡量性别偏见,如性别代词错误率和性别比例差异。对于输出内容为视觉模态,例如图像检索和文生图任务主要考虑这些下游任务中的表现差异来衡量偏见。在图像检索任务中,可以通过衡量在性别查询下返回的检索结果中,是否存在倾斜的人口群体比例(如性别或种族比例)来检测偏见。具体的度量方法包括Bias@K(Wang 等,2021),评估前K个检索结果中某一性别的代表性是否过高或过低;以及Skew@K(Berg 等,2022)用于衡量检索结果中图像属性或人口统计特征的期望比例与实际检索结果之间的偏差;在文生图任务中,可以考虑在保护属性中立的前提下,衡量生成图像关于人口群体的分布情况与均匀分布的差异来评估偏见。

7.2 偏见诱导数据集

研究者已经提出了许多图像—文本数据集用于视觉—语言模型的社会偏见评估,涵盖图像字幕、图像分类和图像检索等下游任务。本文以数据集是否包含反事实样本输入进行划分。由于视觉模态获取反事实样本的困难,绝大多数多模态偏见诱导数据集不包含反事实样本,因此以群体公平性为准则进行评估。一些包含不同人群的图像数据集如Fairface、MS-COCO、CelebA也在这类偏见评估中应用,例如Hendricks等人(2018)使用MS-COCO中包含人类的图像评估图像字幕任务中的性别偏见。在零样本图像分类任务下,Seth等人(2023)引入了PATA(protected attribute tag association)数据集,基于上下文对性别、年龄和种族开展偏见评估。MMBias(Janghorbani和De Melo,2023)包含3 800幅图像和14个子群体,用于评估多模态模型目标概念和保护属性编码间的偏见关联。VisoGender(Hall 等,2023)包含690幅带人工注释的职业图像,用于测量代词解析和图像检索任务下的偏见。然而,这些数据集主要集中于对CLIP等早期视觉语言模型的评估。针对多模态大模型,研究者在视觉问答任务下构建了一些偏见评估基准。PAIRS(Bagdasaryan 等,2023)包含视觉相似的工作场景图像,用于检查4个多模态大模型中的性别和种族偏见,但它受到数据集小和评估方法不严格的限制。用于多模态大模型视觉指令鲁棒性评估的AVIBench基准(Zhang 等,2025a)中也包含规模有限的偏见评估数据。Wang等人(2024d)提出的VLBiasBench使用合成数据构建开放和封闭的视觉问题,通过分析各个子群体的情绪和性别极性分数评估多模态大模型中的偏见,涵盖11个偏见类别。也有少部分偏见诱导数据集以对抗攻击或图像编辑的方式引入反事实样本。SocialCounterfactuals(Howard 等,2024)使用文生图模型生成171 k个反事实图像—文本对,评估检索任务下CLIP模型的交叉偏见。基于SocialCounterfactuals数据集,Howard等人(2024)还评估了不同保护社会属性对多模态大模型生成内容中的毒性和能力相关词的影响。VL-Bias(Zhang 等,2022b)包含24 k个图像—文本对,涵盖52个活动和13个职业,以对抗攻击方法引入性别反事实样本,旨在评估掩蔽语言建模任务的偏见。GenderBias-VL(Xiao 等,2024)包含34 581个视觉问题反事实样本对,涵盖177个职业,对15个开源和2个商用多模态大模型在多模态和单模态上下文开展了全面的偏见评估。尽管这些数据集都包含反事实样本,但他们评估偏见所依赖的公平性准则并不一样。SocialCounterfactuals的评估中关注群体结果的区别,而VL-Bias和GenderBias-VL则关注个体预测的差异。

研究者还通过在语言模态或视觉模态修改样本保护属性来引入偏见诱导样本。语言模态上主要包含手工模板和自动生成方法。手工模板方法通常使用固定格式的句子模板,模板中包含占位符,用于插入与敏感属性相关的词汇。通过改变模板中占位符的值,生成不同的输入数据,并评估模型在这些输入上的表现。然而,手工模板方法生成存在模板单一有限的缺点,为此,研究者提出自动生成方法,利用词嵌入、同义词替换以及句法变换等方法对样本进行修改。例如,Ma等人(2021)提出了自动化框架MT-NLP(metamorphic testing natural language processing),通过利用词汇类比技术识别输入文本中的与人类相关的标记,并对这些标记进行修改,生成带有歧视性的测试输入,随后利用语言流利度评分过滤掉不自然的输入。在视觉模态上,研究者使用生成对抗网络操控图像数据中的敏感属性。针对个体公平性,Denton等人(2020)开发了一个面部生成模型,通过映射潜在编码到图像,并推断潜在编码空间中的方向,以操控特定敏感属性。通过在这些推断出的方向上遍历,生成的测试样本能够有效地评估公平性。类似地,Zhang等人(2021)利用CycleGAN进行图像转换,限制对非敏感属性的变化,以产生具

有歧视性的输入。针对群体公平性，Balakrishnan 等人（2021）合成考虑多个敏感属性的图像来测试群体间的鲁棒性均等性。该方法通过协调地修改这些属性，创建名为“横断面”的网格状匹配图像。这些横断面图像可以帮助比较分类器（如性别分类器）在不同人口群体中的错误率变化，从而评估图像合成前后公平性的差异。

7.3 偏见修复方法

多模态大模型的偏见修复策略通常按实施阶段划分为预处理、训练和后处理阶段修复方法。

预处理阶段修复方法聚焦于输入数据的修改调整，在模型训练前对数据进行清理、平衡和重新分配，减少训练数据中潜在的歧视性信息，从而降低模型在学习阶段继承这些偏见的风险。例如，重采样方法广泛用于处理训练数据中的群体不平衡问题。针对样本量较少的受保护群体，重采样方法通过增加少数群体样本量（过采样）或减少多数群体样本量（欠采样），使数据集在各个群体间达到更加平衡的状态。此外，重采样方法也可以结合其他数据增强技术，以进一步提高模型在少数群体中的表现，从而增强模型的公平性与鲁棒性。

训练阶段的修复方法通常通过引入公平性损失函数或去偏特征表示来减少模型中的偏见。例如，Zhang 等人（2022）提出了对比损失方法，通过将真实标签的嵌入向量与同类样本的嵌入向量拉近，同时将不同类样本的嵌入向量推远，从而有效减少了 CelebA 分类任务中的偏见。Seth 等人（2023）提出了学习加性残差图像表示的框架，通过对图像—文本对的视觉表示进行线性变换，将保护属性信息从原始表示中解耦并减去，得到去偏特征表示，进一步抑制了保护属性的影响。

后处理阶段修复方法在模型训练完毕后，对模型的输出进行调整，以达到公平性目标。Friedrich 等人（2023）提出 Fair Diffusion，通过在部署阶段对预训练的扩散模型进行指令引导，从而减少模型输出中的偏见。该方法利用查找表中存储预定义的指令，精确引导模型生成公平的结果。Chuang 等人（2023）提出偏见向量投影方法，针对文本嵌入中的性别和种族偏见，通过投影去除偏见方向，同时引入正则化约束确保去偏后提示的语义一致性。

8 结 语

本文系统地总结了多模态大模型所面临的对抗攻击、越狱攻击、后门攻击、版权窃取、幻觉现象、泛化问题和偏见问题等安全风险，并分别介绍了这 7 个方面当前的研究进展，为读者理解和应对多模态大模型的独特安全挑战提供一个独特的视角，以推动多模态大模型安全技术的研究。

多模态大模型在许多任务和应用上展现出优秀的性能，为人们的工作和生活提供了多方面的帮助，通过研究多模态大模型安全技术，可以确保多模态大模型的安全可靠，为人们的正常工作和生活提供安全保障。

致谢：本文由中国图象图形学学会数字媒体取证与安全专业委员会组织撰写，该专业委员会链接为 https://www.csig.org.cn/16/201704/49326.html，在此表示感谢。

参考文献（References）

Alon G and Kamfonas M. 2023. Detecting language model attacks with perplexity [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2308.14132.pdf

Bagdasaryan E, Hsieh T Y, Nassi B and Shmatikov V. 2023. Abusing images and sounds for indirect instruction injection in multi-modal LLMs [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2307.10490.pdf

Bai J Z, Bai S, Chu Y F, Cui Z Y, Dang K, Deng X D, Fan Y, Ge W B, Han Y, Huang F, Hui B Y, Ji L, Li M, Lin J Y, Lin R J, Liu D Y H, Liu G, Lu C Q, Lu K M, Ma J X, Men R, Ren X Z, Ren X C, Tan C Q, Tan S N, Tu J H, Wang P, Wang S J, Wang W, Wu S G, Xu B F, Xu J, Yang A, Yang H, Yang J, Yang S S, Yao Y, Yu B W, Yuan H Y, Yuan Z, Zhang J W, Zhang X X, Zhang Y C, Zhang Z R, Zhou C, Zhou J R, Zhou X H and Zhu T H. 2023. Qwen technical report [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2309.16609.pdf

Bai Y T, Jones A, Ndousse K, Askell A, Chen A N, DasSarma N, Drain D, Fort S, Ganguli D, Henighan T, Joseph N, Kadavath S, Kernion J, Conerly T, El-Showk S, Elhage N, Hatfield-Dodds Z, Hernandez D, Hume T, Johnston S, Kravec S, Lovitt L, Nanda N, Olsson C, Amodei D, Brown T, Clark J, McCandlish S, Olah C, Mann B and Kaplan J. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2204.05862.pdf

Bai Z C, Wang P C, Xiao T J, He T, Han Z B, Zhang Z and Shou M Z.

2025. Hallucination of multimodal large language models: a survey [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2404.18930.pdf

Bailey L, Ong E, Russell S and Emmons S. 2024. Image hijacks: adversarial images can control generative models at runtime [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2309.00236.pdf

Balakrishnan G, Xiong Y J, Xia W and Perona P. 2021. Towards causal benchmarking of biasin face analysis algorithms//Ratha N K, Patel V M and Chellappa R, eds. Deep Learning-Based Face Analytics. Cham: Springer: 327-359 [DOI: 10.1007/978-3-030-74697-1_15]

Ben-Kish A, Yanuka M, Alper M, Giryes R and Averbuch-Elor H. 2024. Mitigating open-vocabulary caption hallucinations [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2312.03631.pdf

Berg H, Hall S M, Bhalgat Y, Yang W, Kirk H R, Shtedritski A and Bain M. 2022. A prompt array keeps the bias away: debiasing vision-language models with adversarial learning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2203.11933.pdf

Bhargava S and Forsyth D. 2019. Exposing and correcting the gender bias in image captioning datasets and models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1912.00578.pdf

Bordia S and Bowman S R. 2019. Identifying and reducing gender bias in word-level language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1904.03035.pdf

Cai R Z, Song Z R, Guan D Y, Chen Z H, Li Y H, Luo X, Yi C Y and Kot A. 2025. BenchLMM: benchmarking cross-style visual capability of large multimodal models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 340-358 [DOI: 10.1007/978-3-031-72973-7_20]

Caliskan A, Bryson J J and Narayanan A. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334): 183-186 [DOI: 10.1126/science.aal4230]

Cao B C, Cao Y P, Lin L and Chen J H. 2024. Defending against alignment-breaking attacks via robustly aligned LLM [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2309.14348.pdf

Carlini N, Nasr M, Choquette-Choo C A, Jagielski M, Gao I, Awadalla A, Koh P W, Ippolito D, Lee K, Tramer F and Schmidt L. 2023. Are aligned neural networks adversarially aligned?.//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 61478-61500

Chattopadhyay N, Kataria R and Chattopadhyay A. 2022. TextBack: watermarking text classifiers using backdooring//Proceedings of the 25th Euromicro Conference on Digital System Design (DSD). Maspalomas, Spain: IEEE: 340-347 [DOI: 10.1109/DSD57027.2022.00053]

Chen Z R, Xiang Z, Xiao C W, Song D and Li B. 2024a. AgentPoison: red-teaming LLM agents via poisoning memory or knowledge bases [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2407.12784.pdf

Chen Z R, Zhao Z K, Luo H Y, Yao H X, Li B and Zhou J W. 2024b. HALC: object hallucination reduction via adaptive focal-contrast decoding [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2403.00425.pdf

Chen Z Y, Zhu Y S, Zhan Y F, Li Z W, Zhao C Y, Wang J Q and Tang M. 2023. Mitigating hallucination in visual language models with visual supervision [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.16479.pdf

Cheng R X, Ding Y Z, Cao S R, Yuan S W, Wang Z Q and Jia X J. 2024. BAMBA: a bimodal adversarial multi-round black-box jailbreak attacker for LVLMs [EB/OL]. [2025-02-18]. https://arxiv.org/html/2412.05892v1.pdf

Chuang C Y, Jampani V, Li Y Z, Torralba A and Jegelka S. 2023. Debiasing vision-language models via biased prompts [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2302.00070.pdf

Cui C H, Deng G L, Zhang A, Zheng J N, Li Y C, Gao L L, Zhang T W and Chua T S. 2024a. Safe + Safe = Unsafe? Exploring how safe images can be exploited to jailbreak large vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2411.11496.pdf

Cui X M, Aparcedo A, Jang Y K and Lim S N. 2024b. On the robustness of large multimodal models against image adversarial attacks//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24625-24634 [DOI: 10.1109/CVPR52733.2024.02325]

Denton R, Hutchinson B, Mitchell M, Gebru T and Zaldivar A. 2020. Image counterfactual sensitivity analysis for detecting unintended bias [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1906.06439.pdf

Dhamala J, Sun T, Kumar V, Krishna S, Pruksachatkun Y, Chang K W and Gupta R. 2021. BOLD: dataset and metrics for measuring biases in open-ended language generation//Proceedings of 2021 ACM Conference on Fairness, Accountability, and Transparency. Virtual Event, Canada: ACM: 862-872 [DOI: 10.1145/3442188.3445924]

Dong Y P, Chen H R, Chen J W, Fang Z W, Yang X, Zhang Y C, Tian Y, Su H and Zhu J. 2023. How robust is Google's bard to adversarial image attacks [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2309.11751.pdf

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Drenkow N, Fendley N and Burlina P. 2022. Attack agnostic detection of adversarial examples via random subspace analysis//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2815-2825 [DOI: 10.1109/WACV51458.2022.00287]

Du W and Liu G S. 2022. A survey of backdoor attack in deep learning. Journal of Cyber Security, 7(3): 1-16 (杜巍, 刘功申. 2022. 深度学习中的后门攻击综述. 信息安全学报, 7(3): 1-16) [DOI: 10.

19363/J.cnki.cn10-1380/tn.2022.05.01]

Du Y R, Zhao S D, Ma M, Chen Y H and Qin B. 2024. Analyzing the inherent response tendency of LLMs: real-world instructions-driven jailbreak [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2312.04127.pdf

Ebrahimi S, Arik S Ö, Nama T and Pfister T. 2024. CROME: cross-modal adapters for efficient multimodal LLM [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2408.06610.pdf

Friedrich F, Brack M, Struppek L, Hintersdorf D, Schramowski P, Luccioni S and Kersting K. 2023. Fair diffusion: instructing text-to-image generation models on fairness [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2302.10893.pdf

Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y T, Kadavath S, Mann B, Perez E, Schiefer N, Ndousse K, Jones A, Bowman S, Chen A N, Conerly T, DasSarma N, Drain D, Elhage N, El-Showk S, Fort S, Hatfield-Dodds Z, Henighan T, Hernandez D, Hume T, Jacobson J, Johnston S, Kravec S, Olsson C, Ringer S, Tran-Johnson E, Amodei D, Brown T, Joseph N, McCandlish S, Olah C, Kaplan J and Clark J. 2022. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2209.07858.pdf

Gong Y C, Ran D L, Liu J Y, Wang C L, Cong T S, Wang A Y, Duan S S and Wang X Y. 2025. FigStep: jailbreaking large vision-language models via typographic visual prompts [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.05608.pdf

Gunjal A, Yin J H and Bas E. 2024. Detecting and preventing hallucinations in large vision language models//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 18135-18143 [DOI: 10.1609/aaai.v38i16.29771]

Guo C, Sablayrolles A, Jégou H and Kiela D. 2021. Gradient-based adversarial attacks against text transformers [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2104.13733.pdf

Hall S M, Abrantes F G, Zhu H W, Sodunke G, Shtedritski A and Kirk H R. 2023. VISOGENDER: a dataset for benchmarking gender bias in image-text pronoun resolution//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 63687-63723

Han D C, Jia X J, Bai Y, Gu J D, Liu Y and Cao X C. 2023. OT-attack: enhancing adversarial transferability of vision-language models via optimal transport optimization [EB/OL].[2025-02-18]. https://arxiv.org/pdf/2312.04403.pdf

He X, Wei L H, Xie L X and Tian Q. 2024. Incorporating visual experts to resolve the information loss in multimodal large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2401.03105.pdf

He X L, Xu Q K, Lyu L J, Wu F Z and Wang C G. 2022a. Protecting intellectual property of language generation APIs with lexical watermark//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 10758-10766 [DOI: 10.1609/aaai.v36i10.21321]

He X L, Xu Q K, Zeng Y, Lyu L J, Wu F Z, Li J W and Jia R X. 2022b. CATER: intellectual property protection on text generation APIs via conditional watermarks//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 5431-5445

Hendricks L A, Burns K, Saenko K, Darrell T and Rohrbach A. 2018. Women also snowboard: overcoming bias in captioning models//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 793-811 [DOI: 10.1007/978-3-030-01219-9_47]

Hossain M Z and Imteaj A. 2024. Sim-CLIP: unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2407.14971.pdf

Howard P, Madasu A, Le T, Moreno G L, Bhiwandiwalla A and Lal V. 2024. SocialCounterfactuals: probing and mitigating intersectional social biases in vision-language models with counterfactual examples//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 11975-11985 [DOI: 10.1109/CVPR52733.2024.01138]

Hu Z M, Wu G, Mitra S, Zhang R Y, Sun T, Huang H and Swaminathan V. 2024. Token-level adversarial prompt detection based on perplexity measures and contextual information [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.11509.pdf

Huang L, Yu W J, Ma W T, Zhong W H, Feng Z Y, Wang H T, Chen Q L, Peng W H, Feng X C, Qin B and Liu T. 2025. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2): #42 [DOI: 10.1145/3703155]

Huang Q D, Dong X Y, Zhang P, Wang B, He C H, Wang J Q, Lin D H, Zhang W M and Yu N H. 2024. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13418-13427 [DOI: 10.1109/CVPR52733.2024.01274]

Hubinger E, Denison C, Mu J, Lambert M, Tong M, MacDiarmid M, Lanham T, Ziegler D M, Maxwell T, Cheng N, Jermyn A, Askell A, Radhakrishnan A, Anil C, Duvenaud D, Ganguli D, Barez F, Clark J, Ndousse K, Sachan K, Sellitto M, Sharma M, DasSarma N, Grosse R, Kravec S, Bai Y T, Witten Z, Favaro M, Brauner J, Karnofsky H, Christiano P, Bowman S R, Graham L, Kaplan J, Mindermann S, Greenblatt R, Shlegeris B, Schiefer N and Perez E. 2024. Sleeper agents: training deceptive LLMs that persist through safety training [EB/OL].[2025-02-18]. https://arxiv.org/pdf/2401.05566.pdf

Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B and Madry A. 2019. Adversarial examples are not bugs, they are features//Proceedings of the 33rd International Conference on Neural Informa-

tion Processing Systems. Vancouver, Canada: Curran Associates Inc.: 125-136

Jain J, Yang J W and Shi H. 2024. VCoder: versatile vision encoders for multimodal large language models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27992-28002 [DOI: 10.1109/CVPR52733.2024.02644]

Jain N, Schwarzschild A, Wen Y X, Somepalli G, Kirchenbauer J, Chiang P Y, Goldblum M, Saha A, Geiping J and Goldstein T. 2023. Baseline defenses for adversarial attacks against aligned language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2309.00614.pdf

Jang E, Gu S X and Poole B. 2017. Categorical reparameterization with Gumbel-Softmax [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1611.01144.pdf

Janghorbani S and De Melo G. 2023. Multi-modal bias: introducing a framework for stereotypical bias assessment beyond gender and race in vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2303.12734.pdf

Ji J M, Liu M, Dai J T, Pan X H, Zhang C, Bian C, Chen B Y, Sun R Y, Wang Y Z and Yang Y D. 2023. BEAVERTAILS: towards improved safety alignment of LLM via a human-preference dataset//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 24678-24704

Jiang C Y, Xu H Y, Dong M F, Chen J X, Ye W, Yan M, Ye Q H, Zhang J, Huang F and Zhang S K. 2024. Hallucination augmented contrastive learning for multimodal large language model//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 27026-27036 [DOI: 10.1109/CVPR52733.2024.02553]

Jiao Q R, Chen D Y, Huang Y L, Li Y L and Shen Y. 2024a. Enhancing multimodal large language models with vision detection models: an empirical study [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2401.17981v2.pdf

Jiao R C, Xie S Y, Yue J, Sato T, Wang L X, Wang Y X, Chen Q A and Zhu Q. 2024b. Can we trust embodied agents? Exploring backdoor attacks against embodied LLM-based decision-making systems [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2405.20774.pdf

Jones E, Dragan A, Raghunathan A and Steinhardt J. 2023. Automatically auditing large language models via discrete optimization//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 15307-15329

Kirchenbauer J, Geiping J, Wen Y X, Katz J, Miers I and Goldstein T. 2023. A watermark for large language models//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 17061-17084

Leng S C, Zhang H, Chen G Z, Li X, Lu S J, Miao C Y and Bing L D. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13872-13882 [DOI: 10.1109/CVPR52733.2024.01316]

Lermen S, Rogers-Smith C and Ladish J. 2024. Lora fine-tuning efficiently undoes safety training in llama 2-chat 70B [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2310.20624.pdf

Li H R, Chen Y L, Zheng Z H, Hu Q, Chan C, Liu H S and Song Y Q. 2024a. Simulate and eliminate: revoke backdoors for generative large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2405.07667.pdf

Li L, Guan H Y, Qiu J N and Spratling M. 2024b. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24408-24419 [DOI: 10.1109/CVPR52733.2024.02304]

Li L Y, Ma R T, Guo Q P, Xue X Y and Qiu X P. 2020a. BERT-ATTACK: adversarial attack against BERT using BERT [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2004.09984.pdf

Li M S, Deng C, Li T J, Yan J C, Gao X B and Huang H. 2020b. Towards transferable targeted attack//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 638-646 [DOI: 10.1109/CVPR42600.2020.00072]

Li M J, Wang Z C and Zhang X P. 2023a. An effective framework for intellectual property protection of NLG models. Symmetry, 15(6): #1287 [DOI: 10.3390/sym15061287]

Li M J, Wu H Z and Zhang X P. 2023b. A novel watermarking framework for intellectual property protection of NLG APIs. Neurocomputing, 558: #126700 [DOI: 10.1016/j.neucom.2023.126700]

Li P X, Cheng P Z, Li F Q, Du W, Zhao H D and Liu G S. 2023c. PLMmark: a secure and robust black-box watermarking framework for pre-trained language models//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 14991-14999 [DOI: 10.1609/aaai.v37i12.26750]

Li S, Yao L Y, Zhang L and Li Y L. 2025a. Safety layers in aligned large language models: the key to LLM security [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2408.17003.pdf

Li X, Zhang Y S, Lou R Z, Wu C and Wang J Q. 2024c. Chain-of-scrutiny: detecting backdoor attacks for large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.05948.pdf

Li Y F, Du Y F, Zhou K, Wang J P, Zhao W X and Wen J R. 2023d. Evaluating object hallucination in large vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2305.10355.pdf

Li Y F, Guo H Y, Zhou K, Zhao W X and Wen J R. 2025b. Images are Achilles' heel of alignment: exploiting visual vulnerabilities for jailbreaking multimodal large language models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 174-189 [DOI: 10.1007/978-3-031-73464-9_11]

Li Y T, Xu Z C, Jiang F Q, Niu L Y, Sahabandu D, Ramasubramanian B and Poovendran R. 2025c. Cleangen: mitigating backdoor attacks for generation tasks in large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.12257.pdf

Li Y Z, Li T L, Chen K J, Zhang J, Liu S Q, Wang W H, Zhang T W and Liu Y. 2024d. Badedit: backdooring large language models by model editing [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2403.13355.pdf

Liang B, Li H C, Su M Q, Li X R, Shi W C and Wang X F. 2021. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. IEEE Transactions on Dependable and Secure Computing, 18(1): 72-85 [DOI: 10.1109/TDSC.2018.2874243]

Lim J H, Chan C S, Ng K W, Fan L X and Yang Q. 2022. Protect, show, attend and tell: empowering image captioning models with ownership protection. Pattern Recognition, 122: #108285 [DOI: 10.1016/j.patcog.2021.108285]

Liu F X, Lin K, Li L J, Wang J F, Yacoob Y and Wang L J. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2306.14565.pdf

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2023. Visual instruction tuning// Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 34892-34916

Liu X, Zhu Y C, Gu J D, Lan Y S, Yang C and Qiao Y. 2025. MM-SafetyBench: a benchmark for safety evaluation of multimodal large language models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 386-403 [DOI: 10.1007/978-3-031-72992-8_22]

Liu X D, Cheng H, He P C, Chen W Z, Wang Y, Poon H and Gao J F. 2020. Adversarial training for large neural language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2004.08994.pdf

Liu Y, Cai C J, Zhang X L, Yuan X L and Wang C. 2024b. Arondight: red teaming large vision language models with auto-generated multi-modal jailbreak prompts//Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, Australia: ACM: 3578-3586 [DOI: 10.1145/3664647.3681379]

Liu Y, Deng G L, Xu Z Z, Li Y K, Zheng Y W, Zhang Y, Zhao L D, Wang K L, Zhang T W and Liu Y. 2024c. Jailbreaking ChatGPT via prompt engineering: an empirical study [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2305.13860.pdf

Lu D, Wang Z Q, Wang T, Guan W L, Gao H C and Zheng F. 2023. Set-level guidance attack: boosting adversarial transferability of vision-language pre-training models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 102-111 [DOI: 10.1109/ICCV51070.2023.00016]

Lucas E and Havens T. 2023. GPTs don't keep secrets: searching for backdoor watermark triggers in autoregressive language models// Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023). Toronto, Canada: Association for Computational Linguistics: 242-248 [DOI: 10.18653/v1/2023.trustnlp-1.21]

Luo W D, Ma S Y, Liu X G, Guo X Y and Xiao C W. 2024. JailBreakV: a benchmark for assessing the robustness of MultiModal large language models against jailbreak attacks [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2404.03027.pdf

Ma P C, Wang S and Liu J. 2021. Metamorphic testing and certified mitigation of fairness violations in NLP models//Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence. Yokohama, Japan: [s.n.]: 458-465

Ma S Y, Luo W D, Wang Y and Liu X G. 2024. Visual-roleplay: universal jailbreak attack on multimodal large language models via role-playing image character [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2405.20773.pdf

Ma T, Jia X J, Duan R J, Li X F, Huang Y H, Chu Z X, Liu Y and Ren W Q. 2025. Heuristic-induced multimodal risk distribution jailbreak attack for multimodal large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2412.05934.pdf

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2019. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1706.06083.pdf

Mao C Z, Geng S, Yang J F, Wang X and Vondrick C. 2023. Understanding zero-shot adversarial robustness for large-scale models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2212.07016.pdf

May C, Wang A, Bordia S, Bowman S R and Rudinger R. 2019. On measuring social biases in sentence encoders [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1903.10561.pdf

Mitchell M, Baker D, Moorosi N, Denton E, Hutchinson B, Hanna A, Gebru T and Morgenstern J. 2020. Diversity and inclusion metrics in subset selection//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York, USA: ACM: 117-123 [DOI: 10.1145/3375627.3375832]

Mo Y C, Wang Y J, Wei Z M and Wang Y S. 2024. Fight back against jailbreaking via prompt adversarial tuning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.06255.pdf

Nie Y Z, Wang Y T, Jia J Y, De Lucia M J, Bastian N D, Guo W B and Song D. 2024. TrojFM: resource-efficient backdoor attacks against very large foundation models [EB/OL].[2025-02-18]. https://arxiv.org/pdf/2405.16783.pdf

Niu Z X, Ren H D, Gao X B, Hua G and Jin R. 2024. Jailbreaking attack against multimodal large language model [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.02309.pdf

Nozza D, Bianchi F and Hovy D. 2021. HONEST: measuring hurtful sentence completion in language models//Proceedings of 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics: 2398-2406 [DOI: 10.

18653/v1/2021.naacl-main.191]

Nwaigwe D, Carboni L, Mermillod M, Achard S and Dojat M. 2024. Graph-based methods coupled with specific distributional distances for adversarial attack detection. Neural Networks, 169: 11-19 [DOI: 10.1016/j.neunet.2023.10.007]

Peng W J, Yi J W, Wu F Z, Wu S X, Zhu B, Lyu L J, Jiao B X, Xu T, Sun G Z and Xie X. 2023. Are you copying my model? Protecting the copyright of large language models for EaaS via backdoor watermark [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2305.10036.pdf

Peri R, Jayanthi S M, Ronanki S, Bhatia A, Mundnich K, Dingliwal S, Das N, Hou Z J, Huybrechts G, Vishnubhotla S, Garcia-Romero D, Srinivasan S, Han K and Kirchhoff K. 2024. SpeechGuard: exploring the adversarial robustness of multi-modal large language models//Findings of the Association for Computational Linguistics ACL 2024. Bangkok, Thailand: Association for Computational Linguistics: 10018-10035 [DOI: 10.18653/v1/2024.findings-acl.596]

Qi X Y, Huang K X, Panda A, Henderson P, Wang M D and Mittal P. 2024. Visual adversarial examples jailbreak aligned large language models//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 21527-21536 [DOI: 10.1609/aaai.v38i19.30150]

Qi X Y, Zeng Y, Xie T H, Chen P Y, Jia R X, Mittal P and Henderson P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2310.03693.pdf

Qiu H Y, Dou Z Y, Wang T L, Celikyilmaz A and Peng N Y. 2023. Gender biases in automatic evaluation metrics for image captioning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2305.14711.pdf

Qiu J Y, Ma X B, Zhang Z S and Zhao H. 2024. MEGen: generative backdoor in large language models via model editing [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2408.10722.pdf

Rajpurkar P, Zhang J, Lopyrev K and Liang P. 2016. SQuAD: 100, 000+ questions for machine comprehension of text [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/1606.05250.pdf

Rando J and Tramèr F. 2024. Universal jailbreak backdoors from poisoned human feedback [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.14455.pdf

Robey A, Wong E, Hassani H and Pappas G J. 2024. SmoothLLM: defending large language models against jailbreaking attacks [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2310.03684.pdf

Rombach R, Blattmann A, Lorenz D, Esser P and Ommer B. 2022. High-resolution image synthesis with latent diffusion models//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 10674-10685 [DOI: 10.1109/CVPR52688.2022.01042]

Ross C, Katz B and Barbu A. 2023. Measuring social biases in grounded vision and language embeddings [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2002.08911.pdf

Saha A, Subramanya A and Pirsiavash H. 2020. Hidden trigger backdoor attacks//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI: 11957-11965 [DOI: 10.1609/aaai.v34i07.6871]

Schlarmann C and Hein M. 2023. On the adversarial robustness of multimodal foundation models//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision Workshops. Paris, France: IEEE: 3679-3687 [DOI: 10.1109/ICCVW60793.2023.00395]

Schlarmann C, Singh N D, Croce F and Hein M. 2024. Robust CLIP: unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.12336.pdf

Seth A, Hemani M and Agarwal C. 2023. DeAR: debiasing vision-language models with additive residuals//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 6820-6829 [DOI: 10.1109/CVPR52729.2023.00659]

Sharma M, Tong M, Mu J, Wei J, Kruthoff J, Goodfriend S, Ong E, Peng A, Agarwal R, Anil C, Askell A, Bailey N, Benton J, Bluemke E, Bowman S R, Christiansen E, Cunningham H, Dau A, Gopal A, Gilson R, Graham L, Howard L, Kalra N, Lee T, Lin K, Lofgren P, Mosconi F, O'Hara C, Olsson C, Petrini L, Rajani S, Saxena N, Silverstein A, Singh T, Sumers T, Tang L, Troy K K, Weisser C, Zhong R Q, Zhou G, Leike J, Kaplan J and Perez E. 2025. Constitutional classifiers: defending against universal jailbreaks across thousands of hours of red teaming [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2501.18837.pdf

Shayegani E, Dong Y and Abu-Ghazaleh N. 2023. Jailbreak in pieces: compositional adversarial attacks on multi-modal language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2307.14539.pdf

Sun Z Q, Shen S, Cao S C, Liu H T, Li C Y, Shen Y K, Gan C, Gui L Y, Wang Y X, Yang Y M, Keutzer K and Darrel T. 2023. Aligning large multimodal models with factually augmented RLHF [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2309.14525.pdf

Tan Z Q, Wong H S and Chan C S. 2022. An embarrassingly simple approach for intellectual property rights protection on recurrent neural networks [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2210.00743.pdf

Tu H Q, Cui C H, Wang Z J, Zhou Y Y, Zhao B C, Han J L, Zhou W C S, Yao H X and Xie C H. 2023. How many unicorns are in this image? A safety evaluation benchmark for vision LLMS [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.16101.pdf

Venugopal A, Uszkoreit J, Talbot D, Och F and Ganitkevitch J. 2011. Watermarking the outputs of structured prediction with an application in statistical machine translation//Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing. Edinburgh, UK: Association for Computational Linguistics: 1363-1372

Wang J L, Liu Y and Wang X E. 2021. Are gender-neutral queries

really gender-neutral? Mitigating gender bias in image search [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2109.05433.pdf

Wang J X, Li J Z, Li Y Q, Qi X Y, Hu J J, Li Y X, McDaniel P, Chen M H, Li B and Xiao C W. 2024a. Mitigating fine-tuning based jailbreak attack with backdoor enhanced safety alignment [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.14968.pdf

Wang J Y, Wang Y H, Xu G H, Zhang J, Gu Y K, Jia H T, Wang J Q, Xu H Y, Yan M, Zhang J and Sang J T. 2024b. AMBER: an LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.07397.pdf

Wang J Y, Zhou Y Y, Xu G H, Shi P C, Zhao C L, Xu H Y, Ye Q H, Yan M, Zhang J, Zhu J H, Sang J T and Tang H Y. 2023. Evaluation and analysis of hallucination in large vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2308.15126.pdf

Wang L, He J B, Li S S, Liu N and Lim E P. 2024c. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites//Proceedings of the 30th International Conference on Multimedia Modeling. Amsterdam, the Netherlands: Springer: 32-45 [DOI: 10.1007/978-3-031-53302-0_3]

Wang S B, Cao X K, Zhang J, Yuan Z, Shan S G, Chen X L and Gao W. 2024d. Vlbiasbench: a comprehensive benchmark for evaluating bias in large vision-language model [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.14194.pdf

Wang X G, Ji Z L, Ma P C, Li Z J and Wang S. 2024e. InstructTA: instruction-tuned targeted attack for large vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2312.01886.pdf

Wang X G, Wu D Y, Ji Z L, Li Z J, Ma P C, Wang S, Li Y J, Liu Y, Liu N and Rahmel J. 2025a. SelfDefend: LLMs can defend themselves against jailbreaking in a practical manner [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.05498.pdf

Wang Y F, Xue D Z, Zhang S J and Qian S S. 2024f. BadAgent: inserting and activating backdoor attacks in LLM agents [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.03007.pdf

Wang Y Z, Hu W B, Dong Y P, Zhang H W, Su H and Hong R C. 2025b. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2308.12636.pdf

Xiang Z, Jiang F Q, Xiong Z D, Ramasubramanian B, Poovendran R and Li B. 2024. BadChain: backdoor chain-of-thought prompting for large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2401.12242.pdf

Xiao Y S, Liu A S, Cheng Q J, Yin Z F, Liang S Y, Li J P, Shao J, Liu X L and Tao D C. 2024. GenderBias-VL: benchmarking gender bias in vision language models via counterfactual probing [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2407.00600.pdf

Xie C H, Zhang Z S, Zhou Y Y, Bai S, Wang J Y, Ren Z and Yuille A L. 2019. Improving transferability of adversarial examples with input diversity//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2725-2734 [DOI: 10.1109/CVPR.2019.00284]

Xu Y, Qi X Y, Qin Z and Wang W J. 2024. Cross-modality information check for detecting jailbreaking in multimodal large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2407.21659.pdf

Yadollahi M M, Shoeleh F, Dadkhah S and Ghorbani A A. 2021. Robust black-box watermarking for deep neural network using inverse document frequency//Proceedings of 2021 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). AB, Canada: IEEE: 574-581 [DOI: 10.1109/DASC-PICom-CBDCom-CyberSciTech52372.2021.00100]

Yan J, Yadav V, Li S Y, Chen L C, Tang Z, Wang H, Srinivasan V, Ren X and Jin H X. 2024. Backdooring instruction-tuned large language models with virtual prompt injection [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2307.16888.pdf

Yang H M, Xiang K L, Ge M Y, Li H W, Lu R X and Yu S. 2024a. A comprehensive overview of backdoor attacks in large language models within communication networks. IEEE Network, 38(6): 211-218 [DOI: 10.1109/MNET.2024.3367788]

Yang W K, Bi X H, Lin Y K, Chen S S, Zhou J and Sun X. 2024b. Watch out for your agents! Investigating backdoor threats to LLM-based agents [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.11208.pdf

Yao H W, Lou J and Qin Z. 2024. PoisonPrompt: backdoor attack on prompt-based large language models//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul, Korea (South): IEEE: 7745-7749 [DOI: 10.1109/ICASSP48485.2024.10446267]

Yin S K, Fu C Y, Zhao S R, Xu T, Wang H, Sui D, Shen Y H, Li K, Sun X and Chen E H. 2024. Woodpecker: hallucination correction for multimodal large language models. Science China Information Sciences, 67(12): #220105 [DOI: 10.1007/s11432-024-4251-x]

Ying Z H, Liu A S, Liang S Y, Huang L, Guo J Y, Zhou W B, Liu X L and Tao D C. 2024a. SafeBench: a safety evaluation framework for multimodal large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2410.18927.pdf

Ying Z H, Liu A S, Liu X L and Tao D C. 2024b. Unveiling the safety of GPT-4o: an empirical study using jailbreak attacks [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.06302.pdf

Ying Z H, Liu A S, Zhang T Y, Yu Z M, Liang S Y, Liu X L and Tao D C. 2024c. Jailbreak vision language models via bi-modal adversarial prompt [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.04031.pdf

Yu Q F, Li J C, Wei L H, Pang L, Ye W T, Qin B S, Tang S L, Tian Q and Zhuang Y T. 2024a. HalluciDoctor: mitigating hallucinatory

toxicity in visual instruction data//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12944-12953 [DOI: 10.1109/CVPR52733.2024.01230]

Yu T Y, Yao Y, Zhang H Y, He T W, Han Y F, Cui G Q, Hu J Y, Liu Z Y, Zheng H T and Sun M S. 2024b. RLHF-V: towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13807-13816 [DOI: 10.1109/CVPR52733.2024.01310]

Zeng Y, Sun W Y, Huynh T N, Song D, Li B and Jia R X. 2024. BEEAR: embedding-based adversarial removal of safety backdoors in instruction-tuned language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2406.17092.pdf

Zhai B H, Yang S J, Xu C F, Shen S, Keutzer K, Li C Y and Li M L. 2024. HallE-Control: controlling object hallucination in large multimodal models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2310.01779.pdf

Zhan Q S, Fang R, Bindu R, Gupta A, Hashimoto T and Kang D. 2024. Removing RLHF protections in GPT-4 via fine-tuning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2311.05553.pdf

Zhang C S, Zhang C N, Kang T, Kim D, Bae S H and Kweon I S. 2023a. Attack-SAM: towards attacking segment anything model with adversarial examples [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2305.00866.pdf

Zhang H, Shao W Q, Liu H, Ma Y Q, Luo P, Qiao Y, Zheng N N and Zhang K P. 2025a. B-AVIBench: toward evaluating the robustness of large vision-language model on black-box adversarial visual-instructions. IEEE Transactions on Information Forensics and Security, 20: 1434-1446 [DOI: 10.1109/TIFS.2024.3520306]

Zhang H Y, Yu Y D, Jiao J T, Xing E P, El Ghaoui L and Jordan M I. 2019. Theoretically principled trade-off between robustness and accuracy//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7472-7482

Zhang J M, Ma X J, Wang X, Qiu L Y, Wang J Q, Jiang Y G and Sang J T. 2025b. Adversarial prompt tuning for vision-language models//Proceedings of the 18th European Conference on Computer Vision. Milan, Italy: Springer: 56-72 [DOI: 10.1007/978-3-031-72995-9_4]

Zhang J M, Yi Q and Sang J T. 2022a. Towards adversarial attack on vision-language pre-training models//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 5005-5013 [DOI: 10.1145/3503161.3547801]

Zhang P X, Wang J Y, Sun J and Wang X Y. 2021. Fairness testing of deep image classification with adequacy metrics [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2111.08856.pdf

Zhang R, Li H W, Wen R, Jiang W B, Zhang Y, Backes M, Shen Y and Zhang Y. 2024a. Instruction backdoor attacks against customized LLMs//Proceedings of the 33rd USENIX Conference on Security Symposium. Philadelphia, USA: USENIX Association: 1849-1866

Zhang X X, Li J S, Chu W J, Hai J J, Xu R Z, Yang Y Q, Guan S K, Xu J Z and Cui P. 2024b. On the out-of-distribution generalization of multimodal large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.06599.pdf

Zhang X Y, Zhang C, Li T L, Huang Y H, Jia X J, Hu M, Zhang J, Liu Y, Ma S Q and Shen C. 2025c. Jailguard: a universal detection framework for prompt-based attacks on LLM systems [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2312.10766.pdf

Zhang Y, Wang J Y and Sang J T. 2022b. Counterfactually measuring and eliminating social bias in vision-language pre-training models//Proceedings of the 30th ACM International Conference on Multimedia. Lisboa, Portugal: ACM: 4996-5004 [DOI: 10.1145/3503161.3548396]

Zhang Y C, Zhang S Y, Huang Y, Xia Z Y, Fang Z W, Yang X, Duan R J, Yan D, Dong Y P and Zhu J. 2025d. STAIR: improving safety alignment with introspective reasoning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2502.02384.pdf

Zhang Z, Shen G Y, Tao G H, Cheng S Y and Zhang X Y. 2023b. Make them spill the beans! Coercive knowledge extraction from (production) LLMs [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2312.04782.pdf

Zhao X D, Wang Y X and Li L. 2023a. Protecting language generation models via invisible watermarking//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: 42187-42199

Zhao Y Q, Pang T Y, Du C, Yang X, Li C X, Cheung N M and Lin M. 2023b. On evaluating adversarial robustness of large vision-language models//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 54111-54138

Zheng S, Zhang C N and Hao X H. 2025. Black-box targeted adversarial attack on segment anything (SAM). IEEE Transactions on Multimedia, 27: 1901-1913 [DOI: 10.1109/TMM.2024.3521769]

Zhou Y Y, Cui C H, Rafailov R, Finn C and Yao H X. 2024a. Aligning modalities in vision large language models via preference fine-tuning [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.11411.pdf

Zhou Y Y, Cui C H, Yoon J, Zhang L J, Deng Z, Finn C, Bansal M and Yao H X. 2024b. Analyzing and mitigating object hallucination in large vision-language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2310.00754.pdf

Zhou Z Q, Hu S S, Li M H, Zhang H T, Zhang Y C and Jin H. 2023. AdvCLIP: downstream-agnostic adversarial examples in multimodal contrastive learning//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 6311-6320 [DOI: 10.1145/3581783.3612454]

Zhu L Y, Ji D Y, Chen T R, Xu P, Ye J P and Liu J. 2024. IBD: allevi-

ating hallucinations in large vision-language models via image-biased decoding [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2402.18476.pdf

Zhu S C, Zhang R Y, An B, Wu G, Barrow J, Wang Z C, Huang F R, Nenkova A and Sun T. 2023. AutoDAN: interpretable gradient-based adversarial attacks on large language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2310.15140.pdf

Zou A, Wang Z F, Carlini N, Nasr M, Kolter J Z and Fredrikson M. 2023. Universal and transferable adversarial attacks on aligned language models [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2307.15043.pdf

Zou X T, Li K and Chen Y K. 2024. Image-to-text logic jailbreak: your imagination can help you do anything [EB/OL]. [2025-02-18]. https://arxiv.org/pdf/2407.02534.pdf

作者简介

郭园方，男，副教授，主要研究方向为多媒体安全、人工智能安全。E-mail：andyguo@buaa.edu.cn

余梓彤，男，研究员，主要研究方向为多媒体安全。E-mail：yuzitong@gbu.edu.cn

刘艾杉，男，副教授，主要研究方向为人工智能安全。E-mail：liuaishan@buaa.edu.cn

周文柏，男，副教授，主要研究方向为人工智能安全。E-mail：wenbozhou@ustc.edu.cn

乔通，男，副教授，主要研究方向为多媒体安全、人工智能安全。E-mail：tong.qiao@hdu.edu.cn

李斌，男，教授，主要研究方向为多媒体安全、人工智能安全。E-mail：libin@szu.edu.cn

张卫明，男，教授，主要研究方向为人工智能安全。E-mail：zhangwm@ustc.edu.cn

康显桂，男，教授，主要研究方向为多媒体信息安全。E-mail：isskxg@mail.sysu.edu.cn

周琳娜，女，教授，主要研究方向为人工智能安全。E-mail：zhoulinna@bupt.edu.cn

俞能海，男，教授，主要研究方向为人工智能安全。E-mail：ynh@ustc.edu.cn

黄继武，男，教授，主要研究方向为多媒体安全、人工智能安全。E-mail：jwhuang@szu.edu.cn