arXiv:2602.01206v1 [cs.AI] 1 Feb 2026

利用 SMILE 解决生成式人工智能的可解释性问题

基于局部解释的统计模型无关可解释性



Zeinab Dehghani

赫尔大学计算机科学学院

本学位论文系为申请理学硕士学位而提交

2025年9月

谨以最深的爱与感激，将这篇学位论文献给我善良而始终支持我的丈夫。他的鼓励与陪伴支撑着我走过这段充满挑战的旅程。

声明

本人特此声明，除明确引用他人研究成果之处外，本学位论文的内容均为原创，且未曾全部或部分提交至本校或任何其他大学，作为申请任何其他学位或资格的依据。本学位论文为本人独立完成，除文本及致谢中明确说明的内容外，不包含任何与他人合作完成的成果。本学位论文全文少于 21,000 词（包括附录、参考文献、脚注、表格和方程），且图少于 25 幅。

Zeinab Dehghani
2025年9月

致谢

我谨向我的导师 Dr. Koorosh Aslansefat 和 Prof. Adil Khan 致以最诚挚的谢意，感谢他们在本研究过程中给予我宝贵的指导、鼓励和持续支持。他们富有见地的反馈、耐心和专业知识，对本研究的方向与质量起到了至关重要的作用。

我也衷心感谢赫尔大学计算机科学学院的教职员工和同事们，感谢他们给予的帮助、富有建设性的讨论，以及为营造充满活力的研究环境所做的贡献。

我由衷感谢家人在这段旅程中给予我无尽的爱、耐心和精神支持。特别要感谢我的丈夫，他始终给予我鼓励、理解和动力，这些一直是我在撰写这篇学位论文期间最重要的力量源泉。

最后，我谨向所有研究人员和开发者致谢，他们的工作为本研究奠定了基础，也启发了我在这篇学位论文中探讨的许多想法。

摘要

生成式人工智能的迅速发展催生了能够生成复杂文本和视觉输出的模型；然而，这些模型的决策过程仍大多不透明，限制了其在高风险应用中的信任度与问责制。本学位论文提出了 gSMILE，这是一个用于生成模型可解释性的统一框架，将基于局部解释的统计模型无关可解释性（SMILE）方法扩展至生成式场景。gSMILE 采用文本输入的受控扰动、瓦瑟斯坦距离度量和加权代理建模，以量化并可视化提示词或指令中的特定组成部分如何影响模型输出。应用于大语言模型（LLM）时，gSMILE 可提供细粒度的词元级归因，并生成直观的热力图，突出显示重要词元和推理路径。在基于指令的图像编辑模型中，该方法采用完全相同的文本扰动机制，从而能够分析对编辑指令的修改如何影响最终图像。结合基于运行设计域（ODD）框架的基于场景的评估策略，gSMILE 可在多样的语义和环境条件下系统地评估模型行为。为评估解释质量，我们定义了严格的归因指标，包括稳定性、保真度、准确性、一致性和忠实度，并将其应用于多种生成式架构。大量实验表明，gSMILE 能够生成稳健且与人类判断一致的归因，并能有效泛化至最先进的生成模型。这些发现凸显了 gSMILE 在推动生成式人工智能技术实现透明、可靠且负责任的部署方面的潜力。

目录

图目录 **xv**

表格目录 xix

出版物和成果可用性 1

符号表 3

1 引言 **5**

1.1 生成式人工智能模型.6 1.2 理解人工智能可解释性.8 1.3 研究问题与贡献.9 1.4 研究假设与目标.9 1.5 论文结构.10

2 背景与文献综述 **13**

2.1 生成式人工智能的类型.13 2.1.1 大语言模型（LLM）.14 2.1.2 基于指令的图像编辑.17

2.2 可解释人工智能.18 2.2.1 可解释性对不同利益相关者的作用 19 2.2.2 可解释性方法的类型.22 2.2.3 当前可解释性方法的挑战.23 2.2.4 现有可解释性技术的局限性.24 2.2.5 针对可解释性的对抗攻击.25 2.2.6 解释的准确性与忠实性.26 2.2.7 可扩展性与计算成本.26 2.3 LIME及其不断演进的适配.26

2.3.1 LIME 的工作原理. 29

2.4 SMILE.30
2.4.1 SMILE 的工作原理.30 2.4.2 SMILE 的优势.31

2.5 生成式人工智能中的可解释性.32
2.5.1 大语言模型中的可解释性.33 2.5.2 基于指令的图像编辑中的可解释性 . . . . . 35 2.5.3 多模态人工智能中的可解释性 .36 2.6 结论 .37

3 gSMILE 方法论 3 9
3.1 gSMILE 框架概述.39 3.1.1 生成输出和扰动提示词.43 3.1.2 创建可解释空间.44 3.1.3 使用高斯核进行相似度加权.46 3.1.4 构建可解释代理模型.47 3.1.5 评估指标.47 讨论与结论.56

4 Ex大型语言模型的可解释性 5 7
4.1 问题定义.58
4.1.1 符号与问题设定.58 4.1.2 输入级距离.59 4.1.3 输出级分布偏移.59
4.1.4 基于输入相似度的加权.59 4.1.5 拟合局部替代模型.60 4.1.6 理论依据（利普希茨光滑性） . . . . . 60 4.1.7 线性代理模型近似利普希茨函数 61

4.2 大语言模型的方法概述.62 4.3 实验结果.62 4.3.1 定性结果.63 4.3.2 定量结果.66 4.4 结论.73

5 基于指令的图像编辑的可解释性 77
5.1 问题定义.78 5.2 基于指令的图像编辑方法概述.79 5.3 实验结果.80

5.3.1 定性结果.80 5.3.2 定量结果.85 5.4 结论.97

6 结论 **101**

6.1 贡献总结.102 6.2 局限性.103 6.3 未来工作.104

参考文献 **107**

图目录

1.1 生成式人工智能市场及其在人工智能市场总额中所占份额的预测增长（2020—2030年）。左侧展示逐年扩张情况，右侧面板呈现基于不同情景的全球预测，到2030年市场规模最高可达4.31万亿美元。来源：Master of Code [1].7 1.2 美国可解释人工智能市场按应用划分的预测增长（2020—2030年） [2].8 2.1 生成式人工智能的类型，包括基于文本、基于图像、基于音频、基于视频和多模态系统.14 2.2 大语言模型类别、模型及描述扩展版.16 2.3 人工智能可解释性的主要功能：理解决策依据、分析变换、识别关键要素、促进优化并确保对齐.19 2.4 利益相关者特定的可解释性要求。人工智能系统的可解释性必须满足数据科学家、业务所有者、监管机构、最终用户、验证机构和倡导团体的多样化需求，以确保透明度、问责制和人工智能系统的可信部署。 . . . 22 2.5 人工智能模型评估与可解释性的主要挑战.25 2.6 LIME 框架中用于模型可解释性的修改与扩展概览.28 2.7 示意图：LIME 如何利用加权样本和线性回归在局部近似复杂模型 [3].30 2.8 SMILE 通过使用统计距离度量（例如 Wasserstein 距离）而非直接使用欧几里得距离来扩展 LIME 框架，并基于分布偏移构建局部代理模型 [4].31 2.9 大语言模型中可解释性技术的类别及其关系.35

3.1 用于解释图像分类的 SMILE 流程图 [4] .40 3.2 用于解释文本生成模型的 gSMILE 流程图 . . . . . 41 3.3 针对基于指令的图像编辑模型改编的 gSMILE 框架流程图.42 3.4 使用归因指标评估大语言模型可解释性的框架。该图展示系统如何将真实归因与模型生成归因进行比较，以计算准确性（曲线下面积/AUMAC）、忠实度（通过 aUROC 衡量，该指标比较输入与输出归因的对齐情况）、一致性（模型输出之间的一致程度）和稳定性（归因对扰动的鲁棒性）。此方法能够全面评估自然语言生成任务中的可解释性方法.48 3.5 AttAUC（注意力曲线下面积）指标对图像编辑模型准确性的评估。较深的红色表示文本元素对图像特征的归因更强，体现特定提示词对图像编辑的影响.51 3.6 图像编辑模型的注意力在文本略有变化时的稳定性评估。较深的红色表示文本元素在经过扰动的提示词中具有一致的归因.52 3.7 Instruct-Pix2Pix 模型的一致性分析。热力图显示每个词在三次迭代中对模型输出的影响。 . . . 53 3.8 ATT 忠实度评估框架。输入提示词经过扰动后，全局模型（例如，用于文本生成的 ChatGPT、用于图像编辑的 InstructPix2Pix）和局部代理模型（ gSMILE）会生成相似度信号。通过比较这些信号，可量化代理模型对黑盒模型行为的近似程度.55 4.1 使用 gSMILE（统计模型无关局部解释可解释性）对大语言模型进行可解释性分析的流程概览.58 4.2 展示不同输入提示词中词级重要性的文本热力图。颜色强度表示每个词对模型输出的影响.63 4.3 在三个提示词（中性、“男性视角”和“女性视角”）下，GPT-3.5、LLaMA 3.1 与 Claude 3.5 的 gSMILE 词元级归因比较。颜色强度表示每个词元对输出分布的相对影响. 64

4.4 Anthropic 的归因图与 gSMILE 框架的归因可视化比较。上半部分展示了 Anthropic 方法识别出的推理路径，下半部分则展示了 gSMILE 针对同一提示词的词元级重要性热图。 [5] . . . . 65 4.5 gSMILE 使用 ChatGPT（GPT-4）在 MMLU 数据集选定样本上生成的归因热图。高亮词语表示对模型输出影响力最高的词.66 4.6 箱线图展示了词元“meaning”和“life”在十个句子结构各异的提示词中的注意力分数。每个点表示该词元在特定句子实例中被赋予的注意力分数。. . 73 5.1 用于基于指令的图像编辑可解释性的 gSMILE 框架。77 5.2 图像嵌入的二维t-SNE可视化，展示了不同提示关键词和扰动文本输入的簇可分离性。. . . . . 79 5.3 箱线图展示了提示词 “Transform the weather to make it snowing”中的词权重分布.83 5.4 图像嵌入的二维可视化，展示了在提示词中包含或省略“snowing”的影响.84 5.5 语言变体对“snow”提示响应的影响。不同形式（例如名词、形容词、动名词）会产生不同的输出分布。. 91 5.6 一系列提示词在不同词语上的响应变化。这表明模型在基于指令的图像编辑中对语言表述具有普遍敏感性.92 5.7 Instruct-Pix2Pix 生成的结果展示了扰动提示词中关键词“snowing”的存在如何导致下雪场景，而其缺失则产生非下雪图像，说明了模型对特定关键词的响应性.93 5.8 使用 Instruct-Pix2Pix 生成同一图像时不同提示词结构的比较，展示了提示结构的变化如何导致同一基础场景产生不同的变换。. . . . . 94 5.9 提示词的热力图分析，展示了各个词的权重如何引导 Instruct-Pix2Pix 模型的视觉输出。每一行展示了由不同提示词结构生成的变换，热力图说明了特定词语如何影响模型的解释及最终图像.95

5.10 将单个提示词应用于不同图像的热力图分析。每张热力图展示了提示词中每个词的权重，表明 Instruct-Pix2Pix 模型如何在不同场景中保持一致的词语影响力，以实现期望的变换.96 5.11 将单个提示词应用于不同图像的热力图分析。每张热力图展示了提示词中每个词的权重，表明 I2I-Turbo 模型如何在不同场景中保持一致的词语影响力，以实现期望的变换.96

表格目录

1 本研究产生的出版物及相关资源 . . . . . 1 2 学位论文中使用的符号与记号列表. 3

4.1 不同模型的性能指标（ATT ACC、ATT F1 和 ATT AUROC），在两种条件下进行评估：不同提示词和不同图像.67 4.2 基于十个提示词的平均杰卡德指数计算的各模型稳定性得分.68 4.3 不同模型在提示词“生命的意义是什么？”上的一致性指标.68 4.4 不同模型在提示词“生命的意义是什么？”上的保真度指标. 69 4.5 OpenAI GPT 中不同扰动数量对应的性能指标 69 4.6 LLaMA 中不同扰动数量对应的性能指标 . . . 70 4.7 Claude-AI 中不同扰动数量对应的性能指标 . . 70 4.8 30 次扰动下不同距离度量的保真度结果 . . . . . 71 4.9 各框架和可解释性方法在 60 次扰动下的执行时间（秒）.72 5.1 天气状况转换及对应关键词 . . . . . 81 5.2 基于时间和人类属性的视觉和描述性变化转换指南 .81 5.3 不同模型的性能指标（ATT ACC、ATT F1 和 ATT AUROC），在两种条件下进行评估：不同提示词和不同图像.85 5.4 十个提示词和 30 次扰动下不同模型的稳定性。该表展示了不同模型的杰卡德指数.86

5.5 不同扩散模型的一致性指标。针对提示词“转换天气，使其下雪。”计算方差和标准差.86 5.6 不同扩散模型的保真度指标。针对提示词“转换天气，使其下雪。”计算保真度指标。 . . 87 5.7 Instruct-Pix2Pix 中不同扰动数量对应的性能指标。评估指标包括 WMSE、R^2_ω、WMAE、平均 L1、平均 L2 和 $R^2_{\bar{\omega}}$.87 5.8 Img2Img-Turbo 中不同扰动数量对应的性能指标。评估指标包括 WMSE、R^2_ω、WMAE、平均 L1、平均 L2 和 $R^2_{\bar{\omega}}$.88 5.9 30 次扰动下不同距离度量的保真度结果。该表展示了 T 与 T 以及 I 与 I 距离度量不同组合对应的各种保真度指标. 88 5.10 各框架和可解释性方法在 60 次扰动下的执行时间（秒）.89

出版物与成果可用性

ID	参考文献	章节与链接
1	**Dehghani, Z.**，与 Aslansefat, K.（2024年）。用……解释大语言模型 *gSMILE*。已投稿至 **IEEE Transactions on 人工智能**。	第3章 论文链接 视频链接
2	**Dehghani, Z.**、Aslansefat, K.、Mehmood Khan、A.、Ramírez Rivera, A.、George, F. 和 Khalid, M.（2024年）。使用 *SMILE* 进行基于指令的图像编辑的思维映射。已提交至 **npj《自然·人工智能》**。	第 4 章 论文链接 视频链接
3	Z. Zehtabi Sabeti Moghaddam，**Dehghani,Z.**, Rani, M., Aslansefat, K., Mishra, B., Kureshi, R., & Thakker, D. (2025). 可解释的知识图谱检索增强生成 *(KG-RAG)*，采用 *KG-SMILE*。已提交至 **施普林格**。	**其他合作** 论文链接

表1 本研究产生的出版物及相关资源。

符号列表

本学位论文中使用了以下符号与记号：

符号	描述
x	原始输入提示。
$\hat{x}_j$	由x生成的第j个扰动提示。
$\{x_j\}_{j=1}^{J}$	由x得到的所有J 扰动提示词的集合。
$\pi^{(n)}(y \mid x)$	黑盒模型对输入x的输出分布。
$\pi^{(n)}(y \mid \hat{x}_j)$	黑盒模型对扰动输入的输出分布 $\hat{x}_j$.
δ_{x_j}	与x 和 $\hat{x}_j$之间的输入级语义距离 (IWMD)。
$\Delta(x, \hat{x}_j)$	输出级语义偏移 (OWMD 或瓦瑟斯坦距离) 对于x 和 $\hat{x}_j$的输出之间。
$W(p, q)$	两个分布p 和q之间的瓦瑟斯坦距离。
w_j	基于 δ_{x_j} 的扰动 $\hat{x}_j$ 的高斯核权重。
σ	控制高斯核中权重衰减的核宽度参数 高斯核 .
z_j	扰动 $\hat{x}_j$ 的特征向量表示（例如，独热编码 ）。
$h_\theta(z_j)$	扰动 $\hat{x}_j$ 的代理模型预测。
θ, θ_0	代理线性回归的系数和偏置项 模型 ·
n	嵌入或特征空间的维数。
p	Wasserstein-p 度量中使用的范数阶数。

表2 学位论文中使用的符号与记号列表。

第1章

引言

近年来，人工智能（AI）作出了巨大贡献，尤其是在深度学习（DL）和基于Transformer的模型领域。大语言模型（LLM）和指令引导的图像编辑模型代表了生成式人工智能中两个突出但截然不同的方向。基于Transformer架构的大语言模型（LLM）[6, 7], 通过生成流畅的响应、遵循复杂指令并支持各种领域中的人机对话，彻底改变了自然语言处理 [8]。与此同时，基于扩散和基于Transformer的图像编辑模型使用户能够发出“让天气看起来像暴风雨”或“移除背景中的人物”等高级文本指令，并获得相应的视觉变换 [9, 10]。

这些模型难以解释这一事实引发了重大担忧，尤其是在医疗保健、自动驾驶汽车和法证图像分析等高度关键的领域，因为信任和可靠性至关重要 [11, 12]。为了解决这一问题，可解释人工智能（XAI）领域应运而生，旨在让人工智能模型更加透明、更易于理解 [13, 14]。尽管传统的可解释人工智能方法已在图像分类和目标检测等任务中得到广泛研究 [3, 15], ，但其在多模态生成模型（如大语言模型（LLM）和基于扩散的图像编辑模型）中的应用尚未得到充分探索 [16, 8, 10]。

本学位论文介绍了 gSMILE（具有局部解释的生成式统计模型无关可解释性），这是一种旨在提高多模态生成模型透明度的新型可解释性方法。gSMILE 是一种模型无关方法，这意味着无需访问梯度或模型架构等模型内部组件即可应用，因此适用于黑箱和基于API的系统。它以原始 SMILE 框架为基础 [4] ，并通过纳入统计距离度量将其扩展至生成式场景，从而提供更稳定、一致且可靠的解释。本研究旨在弥合生成式人工智能与可解释性之间的差距，从而深入理解

特定输入元素如何影响文本到文本生成和基于指令的图像编辑模型 [7, 9]的输出。

1.1 生成式人工智能模型

生成式人工智能是人工智能领域发展最快的方向之一，专注于从大型数据集中学习模式，从而创作全新的内容，例如文本、图像、音频和视频 [17]。传统人工智能系统会分析或分类现有数据，而生成模型能够创作出与人类作品极为相似的输出，无论是撰写故事、生成逼真的视觉内容，还是创作音乐 [16]。

该领域一些最引人注目的进展来自大语言模型（LLM），例如**OpenAI 的 ChatGPT**、**谷歌的 Gemini**、**Anthropic 的 Claude**、**Meta 的 LLaMA**和 **DeepMind 的 Gemma** [8, 18–20]。这些模型早已迅速超越了简单的文本生成。如今，像**ChatGPT-4o** 和**Gemini 1.5**这样的工具不仅能根据文本提示生成高质量图像，还支持通过对话界面编辑图像 [21, 22]。向多模态能力转变，体现了将视觉与语言无缝融入用户体验的更广泛趋势。

生成式人工智能最具影响力的两大应用是**文本到文本生成** 和**基于指令的图像编辑**。由大语言模型（LLM）驱动的文本到文本生成，通过实现流畅的文本生成、摘要、翻译和复杂推理，彻底改变了自然语言处理 [7, 8]。GPT-3 [7] 和 GPT-4 [8] 等模型在这一领域树立了新的基准。与此同时，基于指令的图像编辑让用户能够通过自然语言指令修改图像，例如“让天空看起来乌云密布”或“移除背景中的人物”，这得益于基于扩散的模型和基于 Transformer 的模型 [9, 10]。近期的进展已将这些能力整合到多功能多模态系统中，实现文本与图像之间的无缝交互。

基于指令的图像编辑使用户无需使用手动工具，即可通过自然语言指令修改视觉内容，例如“移除背景中的人物”或“把天空改成乌云密布的景象”。基于扩散的模型（如 InstructPix2Pix [23] 和 Imagen Editor [24], ），以及最新的 Gemini 和 Chat-GPT 更新，如今都支持通过提示词生成和直接修改视觉内容 [22, 21]。这些能力使此类模型在创意、无障碍和注重效率的应用中颇具吸引力。

这些工具的采用速度前所未有，反映出它们在消费级和企业级应用中的影响力日益增长。例如，**ChatGPT 在推出仅两个月后，用户数就达到 1亿** ，成为历史上增长最快的消费级应用

应用 [1]。Gemini 和 Claude 等其他模型正被整合到搜索引擎、办公工具、移动设备和浏览器中，让生成式人工智能比以往任何时候都更容易使用 [25]。

与这一快速普及趋势相伴，生成式人工智能的经济潜力也呈指数级增长。预测显示，**到2030年，生成式人工智能市场规模将超过2000亿美元**，与2020年相比，市场份额将增长427%。图1.1展示了这一趋势，重点呈现了年度市场扩张情况，以及基于不同情景对全球总价值的预测。如图1.1所示，基于不同情景的估算表明，到 2030 [1]，生成式人工智能的全球价值可能达到4.31万亿美元。

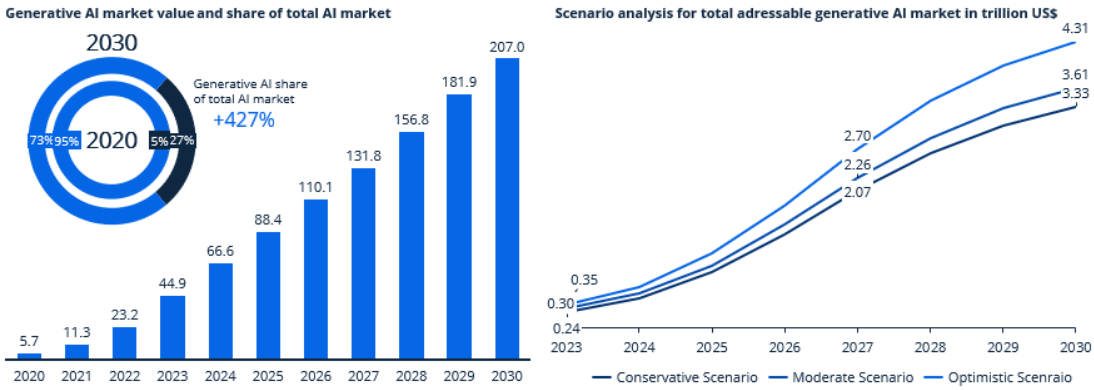


图1.1 生成式人工智能市场及其在人工智能市场总额中所占份额的预计增长情况（2020—2030年）。左侧展示了逐年扩张情况，右侧面板呈现了基于不同情景的全球预测，到2030年最高可达4.31万亿美元。来源：Master of Code [1]。

生成模型有别于传统人工智能系统之处，在于其开放性。分类任务会产生明确且范围有限的输出（例如猫或狗），而生成模型则会生成富有创造性的多模态输出，例如完整句子或经过修改的图像；这些输出往往受到抽象模式和潜在表征的影响 [16, 17]。这种复杂性使人们尤其难以理解某项决策是如何作出的，或为何生成某种输出而非另一种 [13]。

尽管取得了这些进展，生成模型在很大程度上仍然不透明。它们基于深度学习和基于 Transformer 的架构构建，往往以“黑箱”方式运作 [12, 6]，使其内部决策过程难以解释或验证。这种透明度的缺失在医疗保健、法律和教育等高风险环境中尤为令人担忧，因为在这些领域，理解一项决策是如何做出的与决策本身同样重要 [11, 14]。

1.2 理解人工智能可解释性

人工智能（AI）中的可解释性是指人类能够理解模型的决策或预测背后推理的程度 [26]。与深度神经网络等往往缺乏可解释性的传统“黑箱”模型不同，可解释模型旨在使其内部逻辑更加透明。在关键领域中，这一点至关重要，因为用户必须能够信任、验证甚至质疑人工智能生成的结果 [12]。可解释性能够促进对人工智能系统的信任，并支持问责制、公平性以及监管合规 [13, 14]。

近年来，可解释人工智能（XAI）已从一项研究课题发展为多个行业的运营必需品。随着人工智能系统日益深入地应用于医疗保健、金融、国家安全和工业自动化等高风险领域，对透明度和可解释性的需求显著增加 [14, 13, 11]。如今，利益相关者不仅要求人工智能表现出色，还要求其以人类能够理解的方式说明决策依据。这一需求推动了可解释人工智能市场的投资与发展热潮，尤其是在美国 [25]。分析师预计，可解释人工智能工具将在欺诈检测、药物诊断和预测性维护等领域得到广泛应用，呈现出市场对监管压力的响应，以及对人工智能驱动的系统建立信任的必要性。

如图 1.2 所示，预计到 2030年，美国可解释人工智能市场将在多个应用领域显著增长，这反映出可解释性在推动各行业负责任的人工智能部署方面发挥着关键作用 [2]。

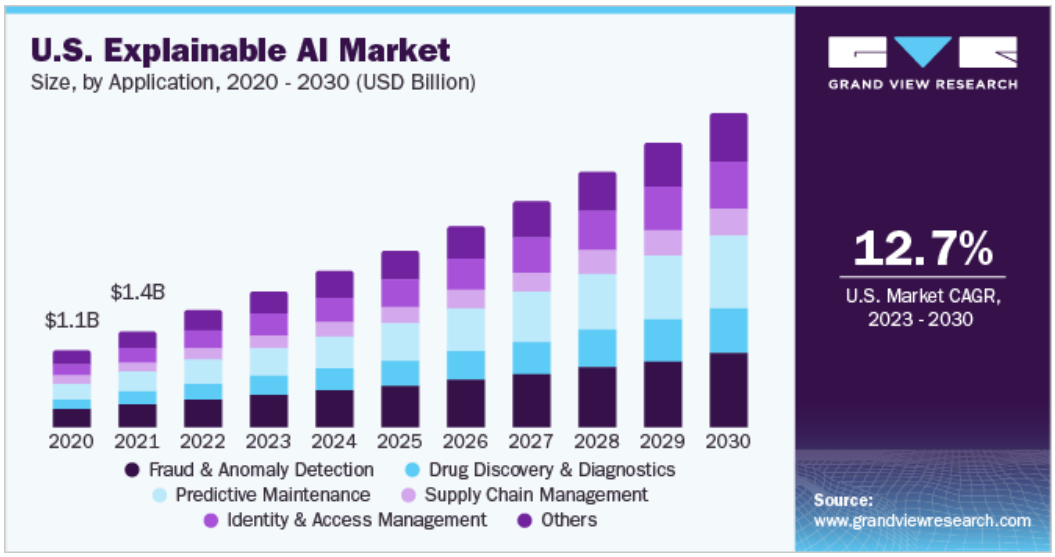


图 1.2 美国可解释人工智能市场按应用划分的预计增长（2020—2030年）[2]。

1.3 研究问题与贡献

本研究探讨生成式人工智能系统可解释性方面的关键挑战，具体涉及大语言模型（LLM）和基于指令的图像编辑模型。尽管近期取得了进展，但要理解这些模型如何处理并响应输入数据，仍是构建可信、透明的人工智能解决方案面临的一大障碍。为应对这些挑战，本研究围绕以下研究问题展开：

- **RQ1：哪些评估指标能够全面评估可解释性方法在不同人工智能领域中的鲁棒性、保真度和可用性？**为了确保可解释性方法具有实际相关性，严谨的评估至关重要。本研究问题旨在确定并应用一组定量指标，包括稳定性、保真度、准确性和一致性，以评估并基准测试所提方法在语言与图像领域中的性能（**第3章**）。

- **RQ2：可解释性方法如何揭示单个输入元素对大语言模型（LLM）的影响？**尽管大语言模型已在各种任务中展现出卓越的能力，但它们的运作方式往往如同黑箱。本研究问题探讨可解释性技术，尤其是**gSMILE**，如何追踪并可视化特定词或词元对模型输出的贡献，从而深入洞察模型的内部推理过程（**第 4 章**）。

- **RQ3：如何将这些可解释性方法应用于多模态系统，尤其是基于指令的图像编辑模型？**基于指令的图像编辑面临独特挑战，因为将文本指令转换为视觉修改的过程十分复杂。本研究问题探讨如何应用可解释性方法，将文本输入映射到生成图像中的特定变化，从而提升这些系统的透明度并增强用户信任（**第5章**）。

1.4 研究假设与目标

基于通过 SMILE 实验验证所获得的见解，本节阐述了推动这项研究的核心假设，以及为检验和支持该假设而确立的主要目标。

本研究的核心假设是，**gSMILE**（基于局部解释的生成式、统计且模型无关的可解释性）能够为复杂的人工智能模型，尤其是多模态生成任务，提供更稳定、一致且稳健的解释

例如文本到文本生成和基于指令的图像编辑，与现有的事后可解释性方法相比。此外，本研究还假设，gSMILE 使用统计距离度量和分布比较，能够增强其抵御对抗性操纵的能力，并提高与人类直觉的对齐程度。

本研究旨在评估 gSMILE 在不同类型模型（包括黑箱系统和多模态系统）中生成准确且一致的解释的性能。评估将聚焦于若干关键标准，包括解释准确性（使用基于注意力的 AUC 等指标进行衡量）、保真度、一致性和稳定性。本研究还将比较不同扰动策略对解释质量和鲁棒性的影响，从而深入了解扰动选择如何影响可解释性结果。此外，本研究将评估 gSMILE 抵御对抗攻击的能力，确定其在对抗条件下维持解释可靠性并揭示有偏决策的能力。本研究还将通过比较 gSMILE 的解释与专家识别的特征及人类直觉，进一步分析其与人类的对齐情况。最后，本研究将展示 gSMILE 在医疗保健、自动驾驶和法证分析等现实高风险领域中的可扩展性和适用性。

1.5 论文结构

本学位论文分为以下六章：

- **第1章：引言**。本章介绍研究的背景与动机，定义研究问题和目标，概述研究假设，并介绍学位论文的整体结构。

- **第2章：背景与文献综述**。本章全面回顾生成式人工智能领域的现有文献，涵盖大语言模型（LLM）、基于指令的图像编辑和可解释人工智能（XAI）。此外，本章还指出当前可解释性技术面临的关键挑战与局限性，这些问题构成了提出该框架的动机。

- **第3章：gSMILE 方法**。本章介绍所提出的 **gSMILE** 框架的理论基础和方法，说明它如何通过统计距离度量、高斯加权和代理建模扩展 LIME 和 SMILE。本章还介绍了本学位论文中使用的归因评估指标：准确性、保真度、稳定性和一致性。

• **第 4 章：大语言模型的可解释性**。本章将所提出的 **gSMILE** 框架应用于大语言模型（LLM），以回答研究问题 RQ2。本章分析词元级归因，使用定量指标评估可解释性，并将模型推理路径可视化。

• **第5章：基于指令的图像编辑可解释性**。本章将 **gSMILE**框架应用于多模态生成模型，重点关注基于指令的图像编辑。本章通过定性热力图和定量评估，考察文本指令如何影响视觉输出，以此回答研究问题 RQ3。

• **第6章：结论与未来工作**。本章总结研究的主要发现与贡献，讨论已识别的局限性，并概述未来工作的潜在方向，以提升生成式人工智能的可解释性。

第2章

背景与文献综述

本章概述生成式人工智能领域的最新进展，包括大型语言模型、基于指令的图像编辑和可解释性技术。本文综述了专门应用于大型语言模型和图像编辑模型的可解释性方法。

2.1 生成式人工智能的类型

生成式人工智能（AI）模型可根据其生成内容的模态大致分类。如图2.1所示，这些类别包括基于文本、基于图像、基于音频、基于视频以及混合或多模态系统。每种类别都采用适合相应数据模态特性的不同模型架构和技术。

基于文本的生成模型，例如大语言模型（LLM），能够生成类似人类的文本，并已展现出在内容创作、摘要、翻译和对话等任务中的能力 [7, 8]。基于图像的模型，包括基于扩散模型和生成对抗网络（GAN）的模型，可以创建和操控图像，通常由文本指令引导 [27, 10]。基于音频的模型能够生成语音、音乐或其他音频信号，应用范围从语音合成到音乐作曲 [28]。基于视频的生成模型将这些能力扩展到动态视觉内容，从而实现视频生成或编辑 [29]。最后，混合型或多模态系统结合多种数据类型，支持复杂任务，例如根据文本提示生成图像，或为文本查询提供可视化解释 [30, 31]。

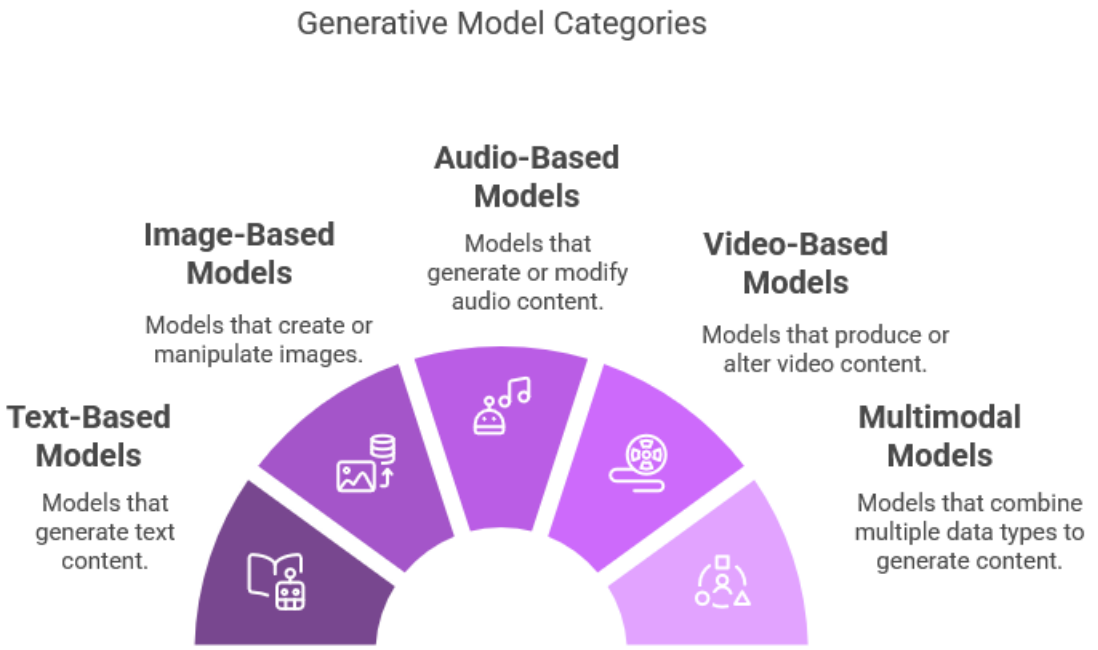


图2.1 生成式人工智能的类型，包括基于文本、基于图像、基于音频、基于视频以及多模态系统。

2.1.1 大语言模型（LLMs）

大语言模型（LLM）由于深度学习取得重大进展、数据获取更加便利以及计算资源更加强大，已得到日益广泛的应用。这些模型经过海量文本训练，能够生成听起来自然的回答、解答问题，并支持广泛的语言类任务。随着时间推移，多个重要的大语言模型推动了该领域的发展，每个模型都带来了新的理念和改进。

GPT（OpenAI） [7] 是最著名的自回归式基于Transformer的模型之一，能够生成连贯且上下文丰富的文本。谷歌的T5 [32] 提出了文本到文本框架，将所有自然语言处理任务都构建为文本生成问题，从而提升迁移学习与效率。同样，PaLM（谷歌） [33] 利用Pathways架构进行大规模语言建模，优化不同自然语言处理任务之间的多任务学习。

Anthropic 的 Claude [34]专注于伦理人工智能和安全交互，旨在最大限度地减少有害输出，同时保持可解释性。谷歌的 Gemma 2 [35]经过优

化，适用于逻辑推理和结构化数据解读，因此成为学术和企业应用场景的有力候选。Meta 的 LLaMA [20] 是一种高效的开源模型，专为各种自然语言处理应用而设计，为研究人员提供了一个易于使用且功能强大的专有模型替代方案。

EleutherAI 的 GPT-NeoX [36] 是 GPT 的开源替代方案，具备稳健的文本生成能力，同时支持自定义。AI21 Labs 的 Jurassic-2 [37]专为大规模企业级应用而设计，注重上下文理解和用户意图。Cohere 的 CommandR+ [38]集成了检索增强生成，可提升基于事实的回答质量并减少幻觉。

此外，BERT（谷歌） [39]等模型通过双向上下文理解实现了自然语言处理能力，显著提升了情感分析和问答任务的效果。微软的 Phi-4 注重效率和成本效益，以更少的参数实现了有竞争力的性能。Mistral 和阿里巴巴的通义千问 2.5 推动了模型适应性和微调能力的发展，使其适用于多语言和特定领域应用。

DeepSeek-LLM [40]是另一款前景可期的开源模型，在基准测试性能方面可与顶级模型竞争。Mixtral 和 StableLM 等模型的开发着眼于效率和特定自然语言处理任务的性能，力求在准确性与计算需求之间取得平衡。

近期的发展带来了 DeepSeek-R1 等模型，该模型专注于高级推理能力；Qwen2.5-Max 则探索大规模混合专家（MoE）模型，以提升效率和性能。OpenAI 的 o3-mini 推动了高性价比推理的前沿，在性能与计算效率之间取得平衡。DeepSeek-V3 是首个达到 GPT-4o 水平的开源模型，代表了开源大语言模型发展的一大进步。

图 2.2 中的以下模型专为文本到文本任务设计，这意味着它们接收文本输入，并针对各种自然语言处理应用生成文本输出：

类别	模型	描述
通用	GPT-4 [8],GPT-4o [21], Claude 3 [41], Gemini 1.5 [18], Command R+ [38]、Mistral [42]	适用于广泛的 自然语言处理任务
开源	LLaMA 3 [43], Mistral [42], Mixtral [44],Falcon 180B [45], DeepSeek-V3 [46], OpenChat [47]	适用于 研究应用
检索增强	Command R+ [38], GPT-4 Bing [8],Claude 3 RAG [48], Gemini Search [18]	增强信息 检索准确性, 减少错误信息
长文本处理	Claude 3 [48], Gemini 1.5 [18], GPT-4o [21],Grok-1.5 [49]	专注于 长文本处理
高级推理	GPT-4 [8],Claude 3 Opus [48], Gemini 1.5 [18], Mixtral [44], Grok-1.5 [49]	逻辑推理, 多任务学习, 问题解决
符合伦理	Claude 3 [48], Gemini [18], GPT-4 [8],Grok [49]	强调安全性, 减少偏见, 伦理人工智能
混合专家	Mixtral [44],Qwen2.5-Max [50], GShard [51],Switch Transformer [52]	改进的效率 以及通过混合专家实现的性能
多模态	GPT-4o [21],Gemini 1.5 [18], Claude 3 [48],Kosmos-2 [53], Grok-1.5V [49]	能够处理 文本和视觉输入
多语言	BLOOM [54],XGLM [55], mBERT [39],XLM-R [56], ByT5 [57]	针对跨语言 理解和翻译
特定领域	Med-PaLM 2 [58],Galactica [59], Legal-BERT [60], StarCoder2 [61], Code LLaMA [62]	针对特定任务量身定制 例如医疗、法律或代码
指令微调	Alpaca [63], Vicuna [64], ChatGLM [65],Baize [66], OpenChat [47]	针对更出色的 指令跟随 和聊天能力进行了调优
轻量级／边缘端	Phi-2 [67], TinyLLaMA [68], DistilGPT2 [69],GPT-2 [70], LLaMA 2 7B [20]	专为快速推理而设计 以及在边缘设备上部署

图2.2 扩展的大语言模型类别、模型与描述

2.1.2 基于指令的图像编辑

基于指令的图像编辑会根据用户的具体指令修改输入图像 [71]。这项任务需要提供一条指令，引导原始图像转换为符合给定要求的新设计 [23]。文本引导的图像编辑让用户能够通过自然语言指令进行修改，从而增强视觉操作的可控性和易用性 [72]。先进的模型利用大规模训练，高效地执行精准编辑 [73, 71]。

这些模型彻底改变了创意产业，为设计师、艺术家和内容创作者提供了直观的工具，使他们能够轻松制作高质量的视觉内容。此外，通过迭代指令微调编辑效果，能够提高灵活性和用户满意度，并根据多样化的创作需求实现个性化成果。

例如，Instruct-Pix2Pix [23] 利用 GAN 和扩散模型，根据用户输入进行细致编辑，使其成为适用于灵活且具体修改的多功能工具。同样，Adobe Firefly [74] 通过处理图像、视频和文本来拓展这一能力，以创作视觉效果更佳的内容，通常面向专业应用场景。另一方面，MagicBrush [75] 专注于将艺术风格和创意效果应用于图像，并由简单的文本指令引导。近期研究还探索了多模态输入的整合，例如将文本与草图或参考图像结合，以实现更精准且能结合上下文的修改。

例如，SmartEdit [71] 利用双向交互来处理复杂编辑，而 MGIE [9] 则专为大规模、高分辨率输入而设计。Img2Img-Turbo [76] 解决了传统基于扩散的模型面临的重大挑战，例如推断时间较长以及依赖成对数据集。它采用单步扩散过程和精简的生成器网络，从而加快推断速度、减少过拟合，并更好地保留输入图像中的结构。该模型在非配对任务（如场景翻译，例如昼夜转换和天气效果）以及配对任务（包括 Sketch2Photo 和 Edge2Image）中表现出色。其效率和多功能性使其成为适用于各种基于 GAN 的应用的竞争力强且可靠的框架。Diffusers-Inpaint [27] 专注于精准修复，同时确保修改区域自然连贯。

MasaCtrl [77] 和 Imagic [73] 等框架可确保高质量且逼真的结果。MasaCtrl 擅长非刚性变换，而 Imagic 则支持详细的文本驱动编辑。MOECONTROLLER [78] 整合全局和局部编辑，Learnable Regions [79]则使用边界框进行局部调整，无需手动创建掩码。DALL-E [80]和 OmniGen [81] 分别专注于动态且保持布局的变换，以及全局/局部变化。

近期的进展提升了效率和可用性。创新技术降低了扩散模型的计算开销，从而改善了实际应用 [82]。FoI（专注于你的指令）可分离相关区域，以实现精确的多指令编辑 [83]。与此同时，以人为中心的生成式人工智能（HGAI）使生成模型与人类意图和伦理标准保持一致 [84], ，推动可访问且富有创意的解决方案。基于指令的图像编辑可能会重点提升模型的适应性，以及处理不明确或模糊用户指令的能力。随着这些技术在不同行业中日益普及，减少偏见、确保编辑内容真实性等伦理问题也将变得越来越重要。

2.2 可解释人工智能

随着人工智能（AI）系统日益复杂，可解释性变得至关重要，尤其是在医疗保健、金融和自动驾驶等高风险领域，因为在这些领域，理解并证明决策的合理性十分关键 [11, 13, 12]。可解释人工智能（XAI）旨在满足这一需求，让人类能够理解人工智能模型的流程和输出，并赋予其意义。

总体而言，可解释人工智能方法可分为两大类 [26]。内在方法使用本身具有可解释性的模型，例如决策树或线性模型。训练完成后，事后方法会分析复杂的“黑箱”模型，在不改变原始模型结构的情况下提供解释。

可解释性回答了一个根本问题：模型为什么会做出这个决策？它能深入揭示模型行为，增强用户信任，支持法规合规，并促进问责制。在分类或回归等传统机器学习任务中，可解释性通常涉及识别对预测影响最大的特征。然而，在生成式人工智能中，输出可能是完整的句子、复杂的图像，甚至经过编辑的媒体内容，因此追溯决策背后的推理过程变得更具挑战性 [16, 11]。

例如，在自然语言生成中，模型为何选择某些词或句子结构，可能并不明确。在计算机视觉中，当面部表情模棱两可时，图像描述模型可能会将一个人描述为“微笑着” [85]。同样，在基于指令的图像编辑中，模型可能会改变背景颜色，而不是添加新物体 [86]。

随着生成模型日益应用于敏感和创意领域，可解释性不再只是锦上添花的特性，而是实现透明度、安全性和信任的必要条件。了解人工智能模型为何生成特定的说明文字、翻译或

编辑后的图像，有助于揭示特定输入元素如何受到强调或被转换。这也有助于识别影响决策的关键特征、检测偏见或不一致之处，并优化性能 [3, 15]。

图2.3总结了人工智能系统中可解释性的关键目的：理解模型的推理依据、分析变换、识别关键要素、促进优化，以及确保与人类期望保持对齐 [26]。

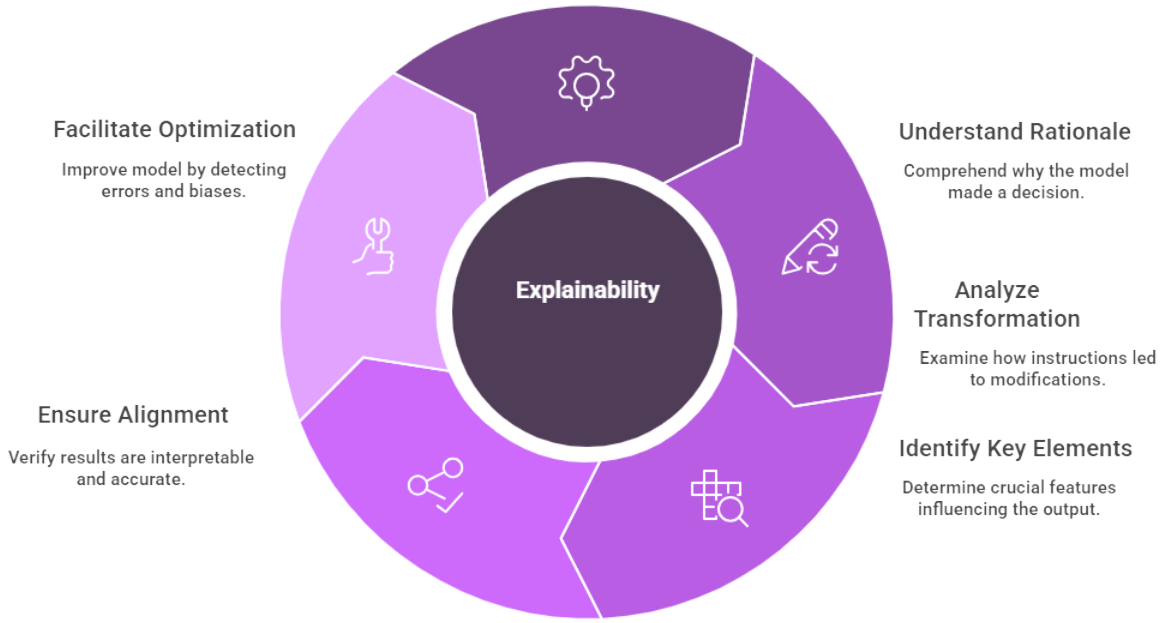


图2.3 人工智能中可解释性的关键功能：理解推理依据、分析变换、识别关键要素、促进优化，以及确保对齐。

通过提高人工智能决策的透明度，可解释人工智能能够增强信任，并促进调试、法规合规以及人工智能系统在各个领域的负责任部署。

2.2.1 可解释性对不同利益相关者的作用

随着可解释人工智能（XAI）日益融入实际应用，必须认识到参与这些系统开发、部署和监督的利益相关者的多样化需求 [13, 11, 12, 87–89]。这些利益相关者包括数据科学家、业务所有者、监管机构、最终用户、验证与保险机构以及倡导团体。每个群体对模型透明度、可靠性、公平性和问责制都有各自不同的期望和要求。图2.4总-

结了这些利益相关者类别及其在人工智能系统生命周期中的可解释性优先事项。

数据科学家在人工智能模型的开发、分析和改进中发挥着关键作用。他们对可解释性的需求有两个方面：首先，在开发过程中理解并优化模型行为 [26, 90]；其次，确保已部署的模型在真实环境中可靠、透明地运行 [87, 91]。可解释性使他们能够诊断模型故障、识别并验证偏见、评估缓解策略，并确保模型生命周期始终保持透明 [11, 3, 92]。对于大语言模型（LLM）和图像生成系统等复杂架构而言，这一点尤为重要，因为这些系统的决策过程通常不透明且难以理解 [16, 93]。

业务所有者，包括人工智能产品经理和企业客户，依靠可解释性来确定模型是否符合用户期望、遵守行业法规并创造切实的商业价值 [14, 94]。可解释性通过揭示潜在风险、阐明模型行为以及确保人工智能驱动的流程中的问责制与合规性，支持战略决策。它还有助于建立组织信任，改善技术开发人员与非技术利益相关者之间的沟通 [87, 89]。

对于监管机构而言，可解释性是验证公平性、鲁棒性以及是否遵守伦理和法律标准的基础 [95–97]。它提供了所需的透明度，以便开展公平性与偏见审计、评估缓解措施并划分风险等级，尤其是在医疗保健、金融和自主系统等高风险领域 [98, 99]。监管机构依靠记录完备的可解释人工智能证据和审计轨迹，确保负责任的部署，并维持公众对自动化决策的信任 [100]。

从聊天机器人用户、创意工具设计师，到医疗保健和自动驾驶领域的专业人士，最终用户都需要易于理解、值得信赖且可付诸行动的解释 [12, 13, 89]。以用户为中心的可解释性通过提供可解释报告、直观的可视化摘要，以及用户反馈或争议解决机制，增强用户对人工智能输出的信心 [101, 102]。面向用户的透明度对于在多样化的使用场景中维护公平性、可及性和信任至关重要。

验证与保险机构在开发者和监管机构之间发挥中介作用。其核心职责包括评估模型的可靠性与鲁棒性、开展独立验证，以及颁发认证或提供责任评估，以确认模型在各种运行条件下的可信性能 [94, 98]。可解释性通过提供可追溯的证据来支持验证与合规审计，从而在安全关键型应用中实现风险缓解和认证。

倡导团体，包括伦理研究社群和非政府组织，在社会层面推动可解释性 [99, 103]。它们通过识别偏见、评估社会影响并鼓励采用负责任人工智能标准 [104]，促进透明度和问责制。通过公众参与、独立监督和伦理监督，这些团体有助于弥合技术创新与社会责任之间的差距，确保可解释性有助于增强更广泛的社会信任。

满足这些利益相关者的多样化需求不仅是一项技术挑战，也是监管与伦理方面的迫切要求。欧盟《人工智能法案》等框架的出现，凸显了在人工智能系统生命周期的各个阶段，增强可解释性、透明度和问责制的重要性 [96, 97]。此外，随着人工智能技术不断演进，日益采用复杂的范式，例如大型语言模型以及图像编辑或生成系统，可解释性对于确保可解释性、安全性和可信度至关重要 [16, 27]。这些模型具有错综复杂的架构和生成能力，凸显出以利益相关者为中心的透明度不可或缺；这种透明度不应局限于性能指标，还应涵盖公平性、可解释性和治理。尤其是医疗保健和自动驾驶等高风险应用，需要严格的可解释性标准，以保障用户安全、促进公平，并确保在技术和社会层面均符合要求。

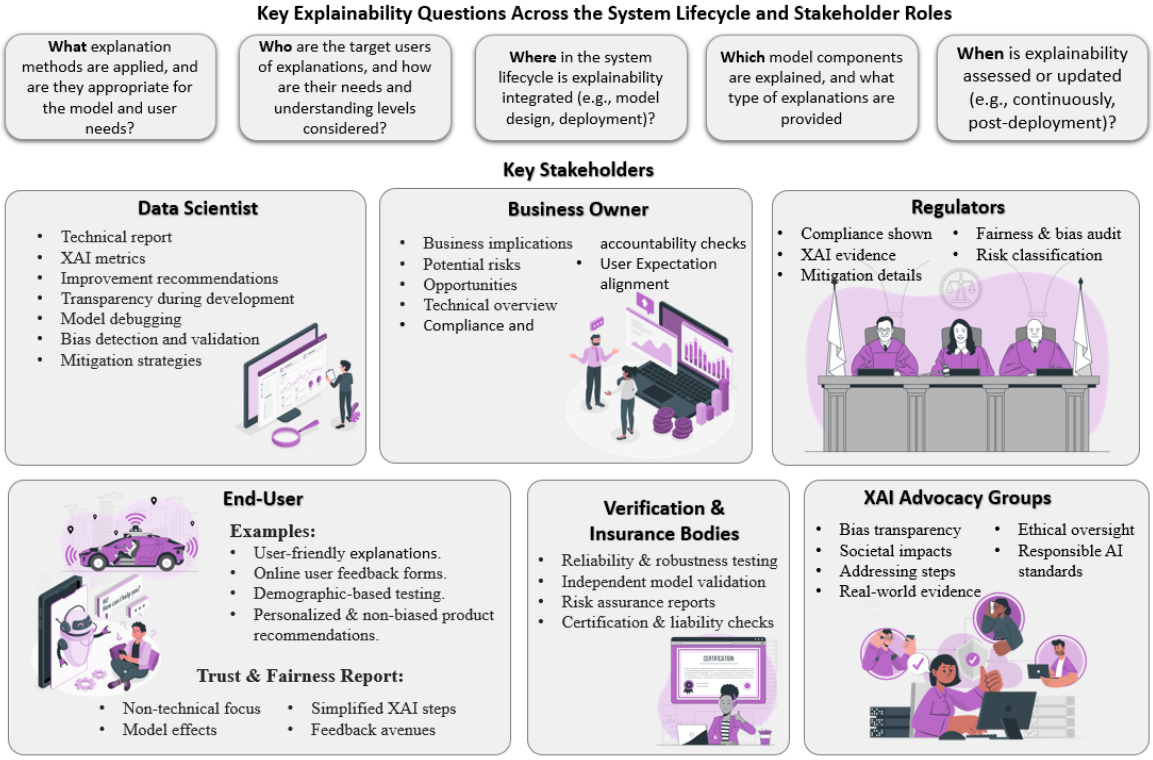


图2.4 利益相关者特定的可解释性要求。人工智能系统的可解释性必须满足数据科学家、业务所有者、监管机构、最终用户、验证机构和倡导团体的不同需求，以确保人工智能系统的透明度、问责制和可信部署。

2.2.2 可解释性方法的类型

可解释性技术大致分为两大类：内在 方法和事后 方法 [14]。二者的区别在于何时以及如何 实现可解释性：它可以是模型固有的属性，也可以是在模型训练后应用的分析过程。

内在可解释性指模型因其结构简单且透明而天然具有可解释性。典型例子包括线性回归、决策树和基于规则的模型，在这些模型中，可以明确检查输入特征与最终预测之间的关系。这些模型的每个组成部分（例如回归系数或决策划分）都能直接说明输入变量如何影响输出。当可解释性和透明度优先于预测性能时，这类模型具有优势。然而，其简单性限制了它们表示复杂非线性关系或高维数据分布的能力。因此，它们通

常不适用于依赖深层层次表示的现代生成式和多模态模型 [12]。

事后可解释性则应用于深度神经网络和基于 Transformer 的架构等复杂的“黑箱”模型。此类模型能够达到最先进的性能，但其本质上不透明，因此难以理解其输出背后的推理过程。事后方法旨在通过分析模型训练后的行为，在不修改其内部结构的情况下解释此类模型。例如，LIME（局部可解释模型无关解释）围绕特定预测构建简单的代理模型，以近似局部决策边界 [3]。SHAP（SHapley 加性解释）利用合作博弈论为输入特征分配公平的贡献值 [105]。Grad-CAM（梯度加权类激活映射）将特定类别的梯度可视化，以突出显示输入图像中影响模型决策的显著区域 [15]。注意力可视化技术通常用于 Transformer 模型，它们展示模型学习到的注意力权重，从而揭示模型在推断过程中关注哪些词元或图像块 [6]。

这些事后方法对于分析大语言模型（LLM）和多模态系统尤为重要，例如基于指令的图像编辑模型，因为解释必须整合文本与视觉推理。虽然这些方法无法保证完全透明，但能为模型决策提供有意义的洞见，并有助于提升高容量生成系统的可解释性。

2.2.3 当前可解释性方法面临的挑战

尽管可解释性研究取得了进展，但现有的大多数方法仍存在明显局限。LIME、SHAP、Grad-CAM 和注意力可视化等技术提升了我们解读复杂人工智能模型的能力，但仍远非完美。许多方法只能提供部分解释，无法在不同输入之间保持一致，或者得出的结果难以让非专家理解。此外，随着模型规模不断扩大、多模态程度日益加深，并融合文本、图像和音频，生成有意义且可靠解释的难度也随之增加 [12, 16]。

当前的可解释性技术往往难以为黑箱模型提供忠实解释，在可解释性与保真度之间取得平衡，并应对特定领域的数据挑战。在不同运行过程或输入之间，解释输出的不一致性削弱了用户信任，也阻碍了可解释人工智能在医疗保健、金融和自动驾驶等关键领域的实际部署 [13, 12]。

这些挑战凸显出，亟需开发更稳健、更具普适性且能够适应特定领域的解释方法，使其能够扩展到现代人工智能系统的复杂程度 [11, 16]。

2.2.4 现有可解释性技术的局限性

LIME（局部可解释模型无关解释）构建代理模型以近似特征重要性 [3]；然而，其关键局限性在于，由于随机扰动，其解释在不同运行之间可能不一致。SHAP（沙普利加性解释）运用博弈论，特别是沙普利值，来量化特征贡献 [106], ，但其计算开销很高，因而难以用于实时应用。

已有若干 LIME 扩展被提出，以解决其不足。BayLIME 引入贝叶斯先验以降低不稳定性 [107], ，但仍易受对抗性扰动影响，并依赖于先验假设。S-LIME（稳定化 LIME）试图通过确定所需的扰动样本数量来提高一致性 [108],，但这会增加计算复杂度。

总体而言，事后可解释性方法面临若干普遍挑战。**基于扰动的不稳定性**的产生，是因为依赖随机采样的方法（如 LIME）可能在不同运行中产生不一致的解释 [3, 108]。此外还存在**难以捕捉多模态依赖关系**的问题，因为传统技术难以解释视觉与文本等模态之间的交互 [16, 10]。另外，**高计算成本**也很常见，尤其是在 SHAP 和逐层相关性传播（LRP）等基于反向传播的方法中 [106, 109]。**有限的局部忠实度**是另一个问题，因为事后技术所使用的代理模型可能会过度简化复杂的决策边界 [3]。

稳定性和泛化能力不足仍然是一个关键问题。解释通常高度依赖特定的数据集和模型架构，因而限制了其更广泛的适用性 [13, 108]。此外，大多数可解释性方法在应用于**非开源模型**或黑盒商业系统时效果不佳，因为无法访问其内部权重、架构或数据管道 [12, 110]。这限制了基于扰动或基于梯度的技术的有效性，因为这类技术通常需要详细访问模型内部信息，或需要受控的预测环境。

最后，许多事后方法容易受到**对抗攻击**的影响。攻击者可以在不改变模型的预测结果或性能的情况下操纵解释，从而可能导致误导性或欺骗性的解释 [105, 111]。这引发了人们对现有可解释人工智能工具的鲁棒性和可信度的严重担忧，尤其是在敏感的

应用中，例如医疗保健、金融和自主系统。图2.5总结了这些挑战，展示了当前可解释性技术的显著局限性。

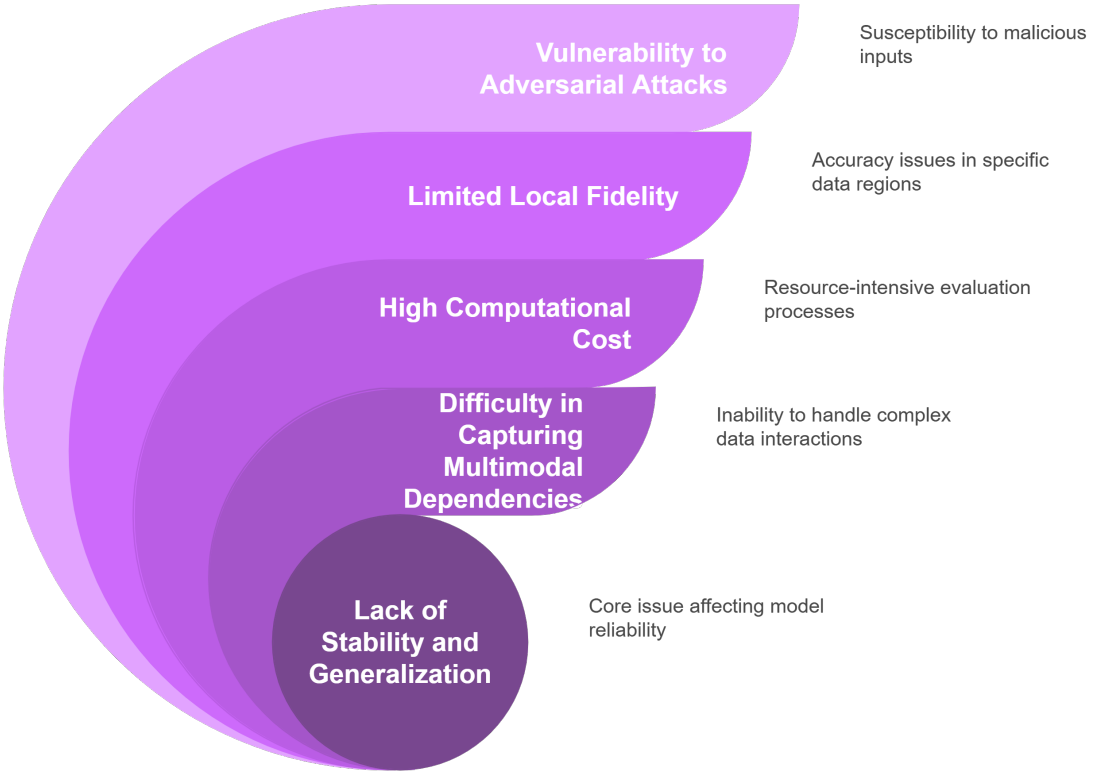


图 2.5 人工智能模型评估与可解释性的关键挑战。

2.2.5 针对可解释性的对抗攻击

可解释性技术也容易受到对抗性操纵。恶意行为者可以修改输入数据，误导可解释性模型，生成虚假或误导性的解释 [105, 111, 112]。在人工智能生成的图像说明或图像编辑会影响医学诊断 [113], 、法证调查 [114], 或自主决策 [115]的应用中，这尤其令人担忧。

SMILE 通过使用统计距离度量来增强抵御对抗攻击的鲁棒性，使其更能抵抗被操纵的输入。与可能被轻微扰动欺骗的传统方法不同，SMILE 在分布层面比较输入输出关系，确保解释始终保持一致，并能抵御对抗性扭曲。通过纳入基于经验累积分布函数的度量 [116], ，SMILE 缓解了对抗偏差，并增强了高风险应用中的可解释性。

2.2.6 解释的准确性与忠实度

当前可解释性方法面临的一项关键挑战，是确保所生成解释的准确性和忠实度。许多常用技术，例如 LIME、SHAP 和基于注意力的方法，只是近似反映模型行为，而非精确呈现。因此，所提供的解释未必总能与底层人工智能模型的实际决策过程相符 [12, 105]。

缺乏忠实度可能导致误导性的解释。例如，LIME 等方法中使用的代理模型可能会过度简化复杂的决策边界，提供易于理解却无法充分反映模型实际逻辑的解释 [3]。类似地，Grad-CAM 等基于梯度的技术可能会突出显示与输出相关的显著特征，但这些特征未必会驱动模型的决策 [15]。

确保解释保真度在处理多种数据类型（包括文本、图像和音频）的大型多模态模型中尤其具有挑战性。在这类模型中，内部表征是分布式且动态变化的，因此难以追踪输入与输出之间清晰的因果联系 [16]。

提高解释的准确性对于增强用户信任、促进知情决策以及满足监管要求至关重要，尤其是在医疗保健和自动驾驶等敏感领域 [113, 96]。研究人员正积极开发相关技术，以在不牺牲可解释性或计算效率的前提下，提高可解释性方法的保真度。

2.2.7 可扩展性与计算成本

可解释性技术面临的另一项重大挑战是可扩展性。许多解释方法，尤其是 SHAP 等基于扰动的方法，需要大量计算资源才能生成准确的归因。随着模型变得更加复杂，例如大语言模型（LLM）和多模态Transformer，生成可靠解释所需的时间和资源也会大幅增加 [106, 16]。

这种计算负担可能会限制可解释人工智能在时间敏感型或资源受限环境中的实际部署。在解释准确性与计算效率之间取得平衡，仍然是一项有待解决的研究挑战 [11]。

2.3 LIME 及其不断演进的适应方法

在事后可解释性方法中，局部可解释模型无关解释（LIME）因其灵活性和模型无关的设计而得到广泛应用 [3]。LIME

旨在通过围绕目标预测拟合一个简单、可解释的模型来解释个体预测，从而近似复杂模型在局部区域内的行为。

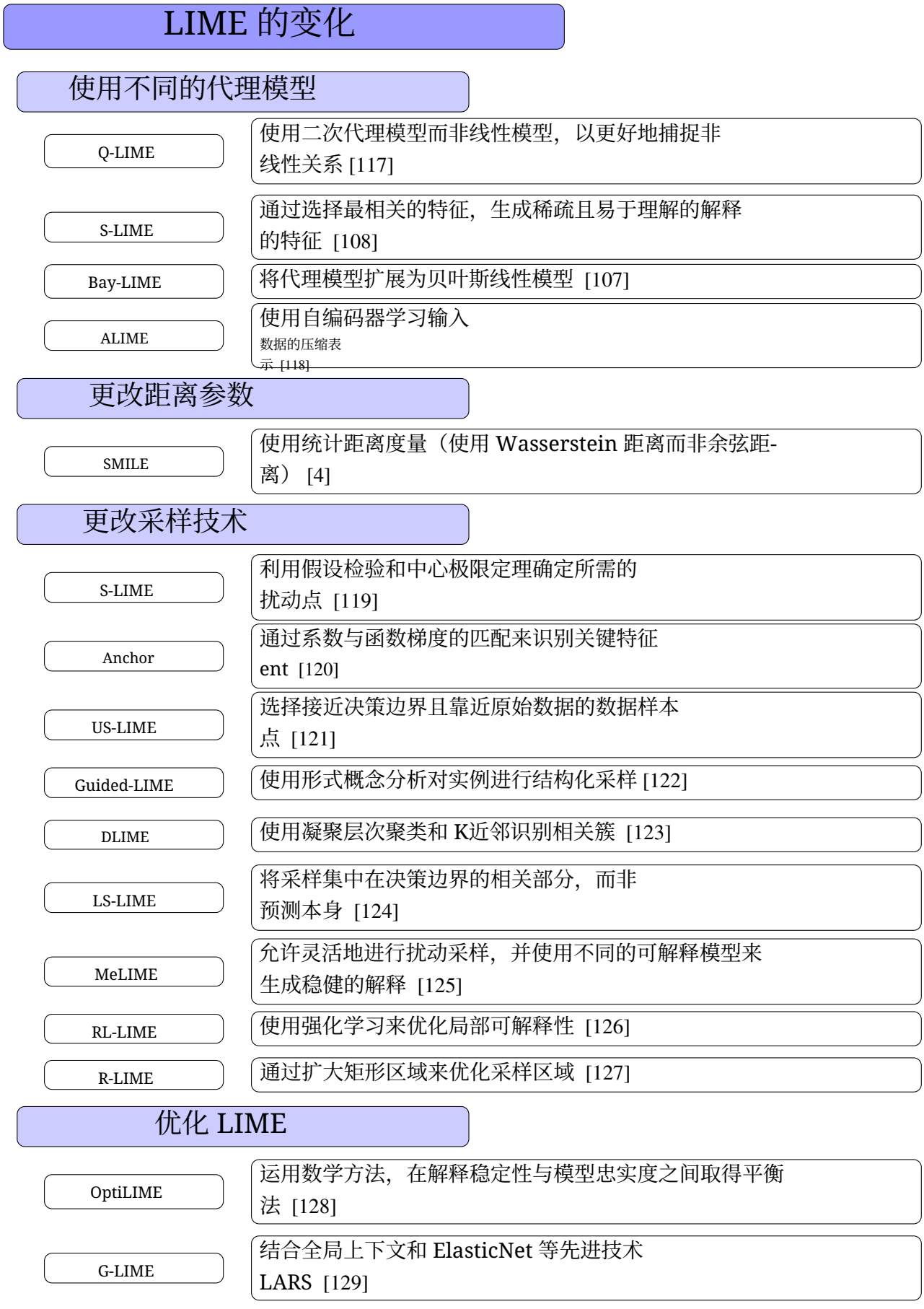


图 2.6 LIME 框架中用于模型可解释性的修改和扩展概览。

LIME 的灵活性使其特别适用于可解释性至关重要的应用场景，例如医学诊断、信用评分和欺诈检测。它具有模型无关的特性，因此无需了解人工智能模型的内部工作原理，便可应用于各种人工智能模型。

近年来，研究人员通过针对特定数据类型和领域进行调整，拓展了 LIME。其变体包括用于三维点云数据的 PointNet LIME [130], 用于时间序列的 TS-MULE [131], 用于图结构数据的 Graph LIME [132], 用于音频的 SoundLIME [133],用于信号数据的 Sig-LIME [134], 用于图像分割的 DSEG-LIME [135], 以及用于心电信号的 B-LIME [136]。每种变体都针对特定领域的挑战对 LIME 加以调整，从而提升相关性和可解释性。

此外，LIME-Aleph 采用归纳逻辑编程来生成基于规则的解释，从而增强法规合规性和决策支持能力 [137]。除了这些改进之外，研究人员还优化了 LIME 的代理模型、距离度量、采样策略和参数，以提高解释保真度和计算效率。图 2.6 总结了一些最具代表性的 LIME 方法。

2.3.1 LIME 的工作原理

局部可解释模型无关解释（LIME） [3] 是一种广泛使用的方法，用于解释复杂黑盒模型的预测结果。LIME 的核心思想是使用更简单、可解释的代理模型（通常是线性回归），近似模型在特定输入附近的行为。

LIME 会围绕输入实例生成一组扰动样本，并将它们输入黑盒模型以获得预测结果。每个扰动实例都会根据其与原始输入的接近程度分配权重，通常使用核函数（例如基于欧几里得距离的指数核函数）。随后，利用这些样本、相应的模型输出以及基于接近程度的权重来拟合代理模型。代理模型的系数提供了可解释的特征归因，用于解释模型的局部行为。

图 2.7 使用二维示例说明了这一过程：对查询点周围的扰动样本进行局部线性回归拟合，而距离加权确保了局部性。

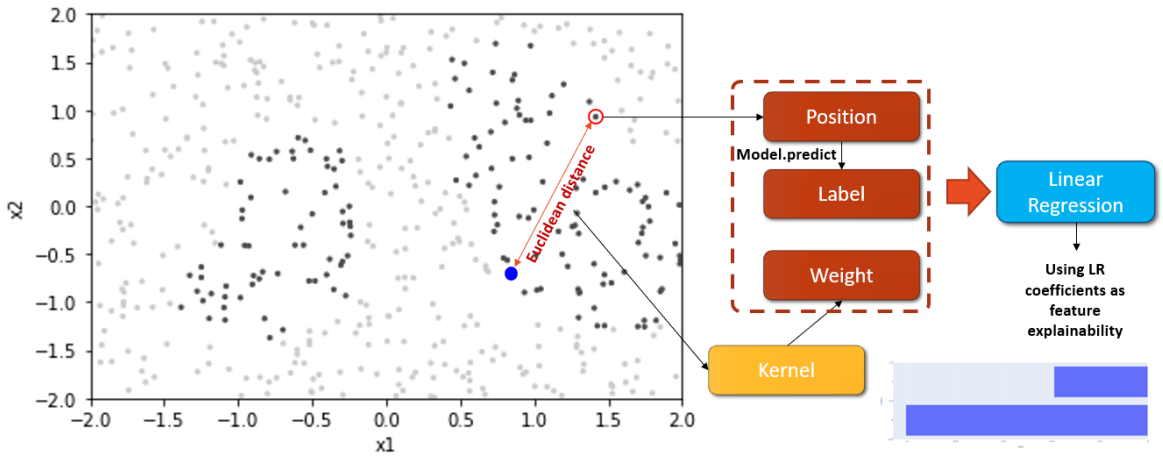


图 2.7 LIME 如何通过加权样本和线性回归在局部近似复杂模型的示意图 [3]。

2.4 SMILE

随着机器学习不断发展，我们构建的模型变得更加强大，但也更加复杂，难以解释。在医疗保健、自动驾驶和基于指令的图像编辑等高风险领域，这一点尤为明显，因为理解一个人工智能系统为何做出特定决策，与决策本身同样重要。这些模型通常被视为“黑箱”，使用户难以信任其输出。为应对这一挑战，我们选择 SMILE（基于统计模型无关的局部解释的可解释性）作为开展可解释人工智能研究的基础。这一基础为后续章节奠定了 groundwork，后续章节将详细探讨 SMILE 的发展、应用和评估。

2.4.1 SMILE 的工作原理

SMILE 通过围绕特定输入构建简单且可解释的近似模型，来阐释复杂人工智能模型的行为，从而提供局部解释。该过程首先对原始输入施加微小扰动，例如移除或屏蔽句子或提示词中的个别词语。对于每个受扰动的输入，都会记录目标模型的输出，并将其与原始输出进行比较。

为了衡量每个输入元素的影响力，SMILE 使用统计距离度量（例如瓦瑟斯坦距离），量化输出分布相对于原始未扰动输入的变化。这些分布变化有助于估算模型对每个输入特征的敏感度。

随后，使用扰动数据及其对应的输出训练一个局部代理模型，通常采用线性回归。该代理模型会为每个输入元素分配归因分数，以表示其对模型预测的相对影响。

与需要访问模型内部参数的基于梯度或基于注意力的方法不同，SMILE 以模型无关的方式运行。它将目标模型视为黑箱，因此兼容文本和基于图像的生成模型。生成的归因分数以热力图的形式呈现，让用户能够直观地识别输入中影响最大的部分。凭借灵活性和可扩展性，SMILE 非常适用于广泛的生成式人工智能应用。

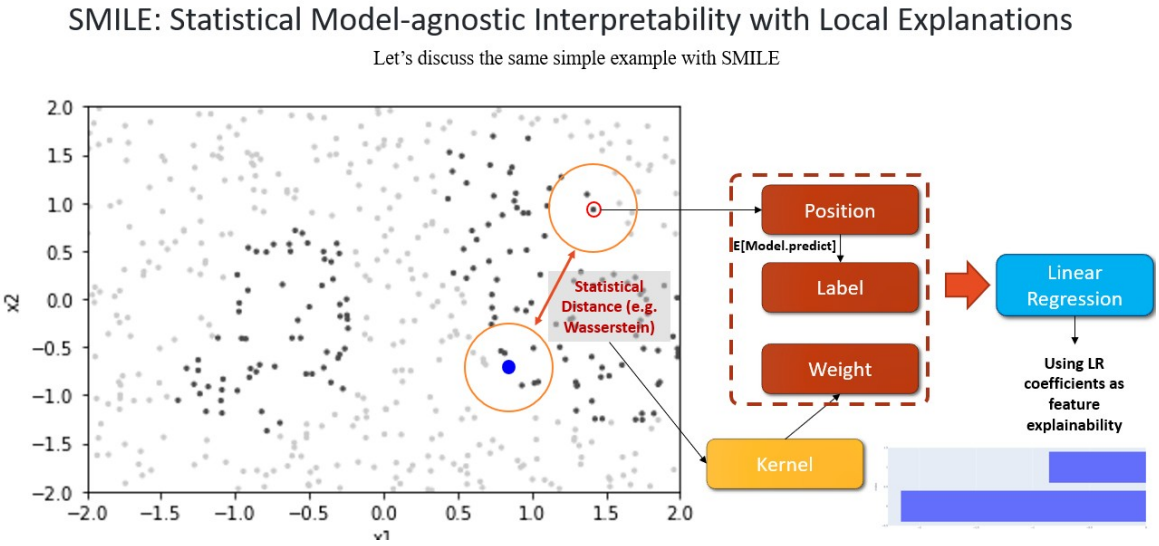


图 2.8 SMILE 通过采用统计距离度量（例如 Wasserstein）而非直接使用欧几里得距离，并基于分布偏移构建局部代理模型，从而扩展了 LIME 框架 [4]。

2.4.2 SMILE 的优势

SMILE 的突出之处在于其 **模型无关的特性**。与依赖特定模型类型或架构的方法不同，SMILE 无需访问内部机制，即可应用于任何机器学习系统 [4]。这种灵活性在实际应用中极具价值，因为模型通常是专有的，或者过于复杂，难以从内部轻松解读。

另一个显著优势是**对对抗攻击的鲁棒性**。一些流行的可解释性方法（例如 LIME）很容易受到欺骗，这意味着如果模型遭到操纵，它们可能会给出误导性的解释 [105]。SMILE 的设计明确旨在抵御此类操纵。它使用瓦瑟斯坦距离等统计距离度量，以

生成更稳定、可信的解释，即使模型本身遭到篡改或存在偏见也是如此。

最后，SMILE 摆脱了余弦距离或欧几里得距离等基本相似度度量，因而能够提供**更加稳定、一致的解释** 。相反，它会对输入数据的分布进行高级统计比较，以捕捉特征影响模型决策的细微方式 [4], ；这在图像编辑等场景中尤其有用，因为 SMILE 可以生成热力图，清晰显示句子或图像中的哪些部分影响了模型输出 [138]。这些视觉线索让人们更容易理解并信任模型行为。

简而言之，SMILE 融合了灵活性、鲁棒性和清晰性，而其他可解释性方法很少能做到这一点。因此，它非常适合我们的目标：创建人们能够理解并信赖的人工智能系统。

2.5 生成式人工智能的可解释性

随着生成式人工智能不断发展，理解并解释这些模型如何生成输出变得愈发重要 [16, 12]。与执行分类或回归等定义明确的任务、表现出色的传统判别式模型不同，生成模型运行于高维输出空间中，能够生成开放式内容，包括文本、图像或音频。这种复杂性给可解释性带来了重大挑战 [92]。

生成式人工智能中的可解释性是指理解特定输入（例如提示词中的词或指令中的要素）如何影响所得输出 [13, 14]。这一点在大语言模型（LLM）和基于指令的图像编辑系统中尤为重要，因为输入的细微变化可能导致显著不同的生成结果 [139, 140]。在医疗保健、自动驾驶和法律推理等高风险应用中，无法解释此类输出会引发对信任、公平性和问责制的担忧 [141, 11]。

本节回顾生成式人工智能领域最先进的可解释性技术，重点关注两个主要领域：大语言模型和基于指令的图像编辑。其目标是找出现有方法的局限性，并说明为何需要一个统一、稳健且模型无关的框架（例如 SMILE），以便跨模态提供一致且直观的解释 [4, 3]。

2.5.1 大语言模型中的可解释性

大语言模型（LLM）彻底改变了自然语言处理，使文本生成、翻译和问答等任务成为可能。然而，其不透明的“黑箱”特性引发了人们对可解释性、公平性和可信度的担忧 [139, 141]。大语言模型的可解释性对于确保问责制、增强用户信任和减轻偏见至关重要。本篇文献综述探讨了有关大语言模型可解释性方法、应用和挑战的现有研究。

图 2.9 展示了大语言模型可解释性概念框架的概览，重点呈现四个关键维度：事后可解释性、机制可解释性、以人为中心的可解释性，以及应用与挑战。这些类别反映了提升大语言模型透明度和可信度的主要方法与实际考量。

大语言模型的可解释性概念框架包含四个关键维度：事后可解释性、机制可解释性、以人为中心的可解释性，以及应用与挑战。这些类别反映了提升大语言模型透明度和可信度的主要方法与实际考量。

训练完成后，事后可解释性方法旨在解读大语言模型，深入揭示其决策过程 [139, 142]。Zhao 等人 [139]提出了大语言模型可解释性技术的分类体系，将相关方法分为局部解释和全局解释，并探讨了局限性及未来研究方向。同样，Bender 等人 [141] 指出了不透明大语言模型的风险，包括偏见和错误信息，并呼吁提高透明度和问责制。近期的综述进一步拓展了这一领域，回顾了大语言模型如何促进更广泛的可解释人工智能研究 [143, 144]。Amara 等人 [145]最近推出了 ConceptX，这是一种概念级归因方法，通过基于联盟的 Shapley 框架突出语义上有意义的输入短语，从而提供稳健且符合人类认知的解释。

机制可解释性旨在揭示大语言模型如何在内部处理信息 [146, 147, 48]。Olah 等人 [146] 提出了逆向工程技术，用以剖析内部表征。Nanda 等人 [147] 研究了大语言模型如何形成抽象表征并实现泛化，同时提出了用于可解释性研究的进展度量方法。Anthropic [48] 近期的研究展示了机制可解释性在理解和控制模型行为方面的作用。此外，SEER 框架 [148] 引入了自解释机制，以增强大语言模型内部表征的可解释性。CELL 框架 [149] 则进一步推进了机制可解释性，将基于概念的解释直接融入语言模型的训练与表征学习中。

Transformer Circuits 框架的最新进展进一步丰富了机制可解释性。Ameisen 等人 [150] 运用因果擦除引入了归因图，以揭示神经网络中涌现的结构。相比之下，Nanda 等人 [151] 研究了单义性的扩展，从最先进的模型中提取可解释特征。其他近期研究探索了相关性估计 [152], 、情绪推断 [153], 和地理空间推理 [154]等特定任务。这些研究采用激活修补和稀疏自编码器等技术，揭示大语言模型的内部机制。文献 [152]还对稀疏可解释性方法进行了全面综述。研究者也应用了神经科学视角，运用动力系统理论对 Transformer 中的词元级轨迹进行建模 [155]。

以人为中心的可解释性致力于让非专业人士能够理解大语言模型的输出 [156–158]。Ji 等人 [156] 研究了大语言模型输出中的幻觉现象，并提出了检测和缓解错误响应的技术。Krause 等人 [157] 评估了其推理能力与人类表现之间的差异，Martens 等人 [158] 则探索了反事实解释，以提升人类的理解程度。近期研究通过实现对黑箱模型的交互式解释 [158] ，以及对临床领域中解释质量进行基准测试 [159]，进一步拓展了这一视角。这些方法提升了大语言模型输出在现实世界决策中的可及性与透明度。

近期研究探索了面向大语言模型的高成本效益、模型无关的解释方法 [160], ，这些方法与 gSMILE 的目标高度契合；其方法利用采样和评分策略生成忠实的解释，而无需访问模型内部。

可解释性增强了大语言模型的调试、偏差检测、法规合规以及用户信任 [14]。挑战依然存在，包括如何定义大语言模型生成输出中的“意义” [14], 、应对黑箱模型的伦理与社会影响，以及在性能与可解释性之间取得平衡。近期工作已证明 SMILE的适用性不仅限于语言领域，例如在基于指令的图像编辑中 [138], ，凸显了其模态无关的特性。

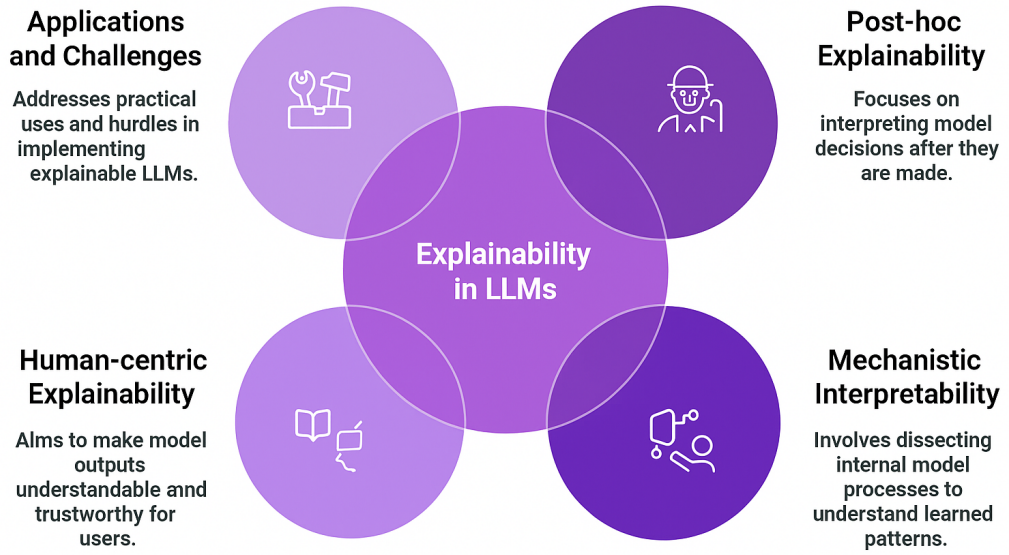


图 2.9 大型语言模型中可解释性技术的类别与关系。

2.5.2 基于指令的图像编辑中的可解释性

指令图像编辑模型在可解释性方面面临独特挑战。这些模型将文本描述转换为图像，涉及复杂的高维变换。因此，这些模型的内部工作机制可能更具固有可解释性。例如，识别输入文本的哪些部分对生成图像的特定地区具有影响力，仍然具有挑战性 [161, 162, 140, 163]。

先前的研究采用热力图等局部可解释人工智能方法，以可视化方式将文本特征映射到相应的图像区域。DAAM [140] 引入了一种创新的改进方法，用于捕捉图像中的自注意力，并首次处理了文本-图像交叉注意力。通过引导自注意力，DAAM-I2I [164]显著提升了热力图的质量，从而在分割和目标检测任务中实现更精准的定位。

近期有关鲁棒性、公平性、安全性、隐私、事实性和可解释性的研究，旨在应对伦理、有害及社会方面的关切 [165, 166]。

Evirgen 等人 [162] 通过强调使用基于示例的解释并整理数据集，为初学者量身定制更合适的解释，从而应对基于指令的图像编辑模型带来的用户交互挑战。类似地，PromptCharm [167] 引入了模型

解释，通过可视化注意力值，让用户能够以交互方式完善提示词并生成质量更高的输出。

Patcher [168] 提出了一种通过注意力引导的特征增强来缓解指令式图像编辑模型中灾难性忽略问题的方法。该方法提升了提示词与生成内容之间的对齐程度，直接解决语义不一致问题。除了文本到图像生成，基于扩散的模型还被应用于开放词汇分割 [169, 170] 和行人检测数据集生成 [171]等任务。

OVDiff [169] 利用预训练扩散模型进行零样本分割，无需额外训练。FOSSIL [172] 将文本条件扩散模型与自监督特征相结合，用于无监督分割，在提升可解释性的同时捕捉语义变化。

在视频生成领域，Vico [173] 分析词元影响并平衡潜变量更新，从而提升视频生成的组合准确性。同样，MultiLate [174] 利用注意力归因图增强多模态仇恨言论检测，展现了可解释性对提升分类性能的影响。

可控 RGBA 插图生成方法 [175] 以及面向具身人工智能任务的细粒度视觉空间表征方法 [176] ，展示了扩散模型在创意工作流程中的潜力。最后，Tankelevitch 等人 [177] 提出将可解释性融入生成式工作流程，以降低认知负荷。

这些贡献凸显了可解释性对于提升基于扩散的生成式人工智能系统在各种应用中的可用性、可信度和透明度的重要意义。

在这项工作中，我们引入 SMILE（基于局部解释的统计模型无关可解释性），以生成热力图，展示特定文本元素如何影响基于指令的编辑中的图像编辑 [4]。SMILE 是模型无关的，因其与各种模型兼容而被选用 [3], ，它借鉴了与 LIME 类似的方法，但具有更强的鲁棒性 [92]。与可能受到扭曲可解释性的对抗攻击影响的 LIME 不同，SMILE 更能抵御操纵 [4]。这些热力图让用户能够更清楚、更可靠地理解编辑模型的交互方式，并通过提供不易被扭曲的洞见，帮助用户信任编辑结果。

2.5.3 多模态人工智能中的可解释性

随着多模态大语言模型（MLLMs）和视觉语言模型（VLMs）的兴起，近期的综述指出，单模态解释技术不足以捕捉跨模态依赖关系。关于多模态大语言模型及多

模态可解释人工智能的综合综述提出了整合数据、模型和推断层面可解释性的分类体系，涵盖从词元级和嵌入级归因到架构层面及基于推理的透明性等方面 [163, 178]。在视觉领域，近期一项关于基础模型的综述着重介绍了事后方法（扰动、反事实和神经元方法）与大规模模型的结合，并指出这为解释的定量评估带来了机遇与挑战 [179]。同样，关于视觉语言模型的综述强调了对齐、基准测试、幻觉、公平性和安全性等问题，凸显了在高风险应用中提供可信解释的重要性 [180]。这些研究推动了量化跨模态影响力的方法，同时确保稳定性、保真度和评估的标准化；本学位论文通过 SMILE框架直接探讨了这些方向。

2.6 结论

本章回顾了生成式人工智能（AI）的快速发展，重点关注大语言模型（LLM）、基于指令的图像编辑和可解释性技术。本文探讨了模型架构方面的创新（如 Transformer 和扩散模型）如何提升生成式人工智能在各种任务和模态中的能力。

针对大语言模型，人们为应对大规模模型固有的可解释性挑战付出了大量努力。研究提出了多种可解释性方法，包括事后技术、机制可解释性和以人为中心的方法，每种方法都有助于深入理解这些模型如何生成和处理信息。尽管取得了这些进展，如何兼顾高性能与透明度仍然是一项持续的挑战。

基于指令的图像编辑方面的进展着重于改善文本提示与生成图像之间的对齐。研究提出了创新的基于注意力的方法、热力图可视化和交互式解释工具，以增强模型可解释性和用户信任。然而，这些多模态系统的高维度和复杂性仍然给可解释性带来了重大障碍。

总体而言，将可解释性方法融入生成式人工智能研究，提升了透明度和可用性，有助于发展符合伦理且负责任的人工智能实践。持续的跨学科合作和方法创新对于克服尚存挑战、推动生成式人工智能技术负责任地发展至关重要。

第3章

gSMILE 方法

3.1 gSMILE 框架概述

所提出的方法采用gMILE框架，以增强文本生成和图像编辑任务中的可解释性。通过分析特定输入组件如何影响模型输出，该方法提高了模型行为的透明度和可预测性 [16, 90]。

该方法最初为图像分类任务而开发，如图3.1 [4], 所示，SMILE通过分割输入来识别驱动模型决策的关键特征，例如图像的超像素或文本的逻辑片段，并通过对这些组件进行选择性修改来生成扰动。这些扰动揭示了不同输入元素对最终输出的贡献。

本工作将gSMILE适配于两个领域。对于大语言模型（LLM），该方法评估文本输入的变化如何影响生成的输出，从而揭示提示词与响应之间的关系。对于基于指令的图像编辑模型，gSMILE解释不同视觉组件和输入指令如何影响编辑后的输出。

这一统一可解释性框架阐明了文本和视觉领域中的决策过程，并支持第 4 章和第 5 章所呈现的分析。

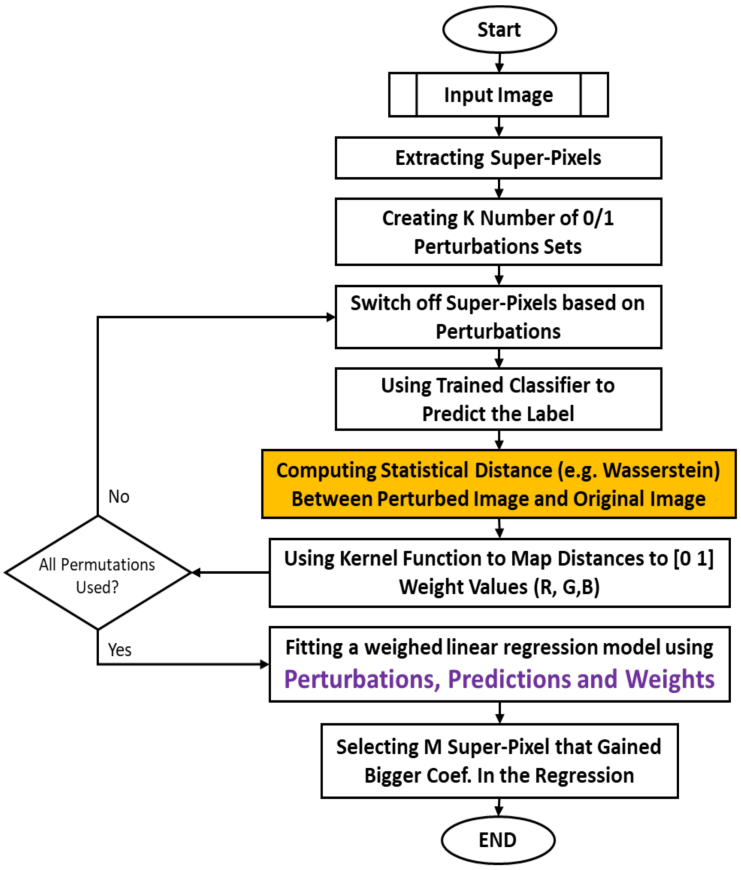


图 3.1 用于解释图像分类的 SMILE 流程图 [4]

受 SMILE 功能的启发，我们提出了一种用于解读基于文本提示的生成模型的统一方法，适用于文本输出和图像输出，如图 3.2 和图 3.3 所示。我们不直接对生成输出进行扰动，而是通过加入或去除特定词语来修改指令。对于每个经过修改的文本提示，都会生成相应的输出：语言模型生成文本输出，图像生成模型生成图像。我们计算每个扰动后的生成结果与原始文本提示生成的输出之间的瓦瑟斯坦距离。该距离是一种相似度度量，可用于识别文本提示中对生成的文本或图像产生最显著影响的词语。

随后，这些相似度分数将作为加权线性回归模型中的因变量，权重则基于原始输出与扰动后输出之间的距离确定。该回归模型有助于确定提示词中的每个词如何影响最终输出。模型得出的系数用于量化每个词的影响力，并据此创建一张直观的热力图，突出显示影响力最大的词。这张热力图提供了

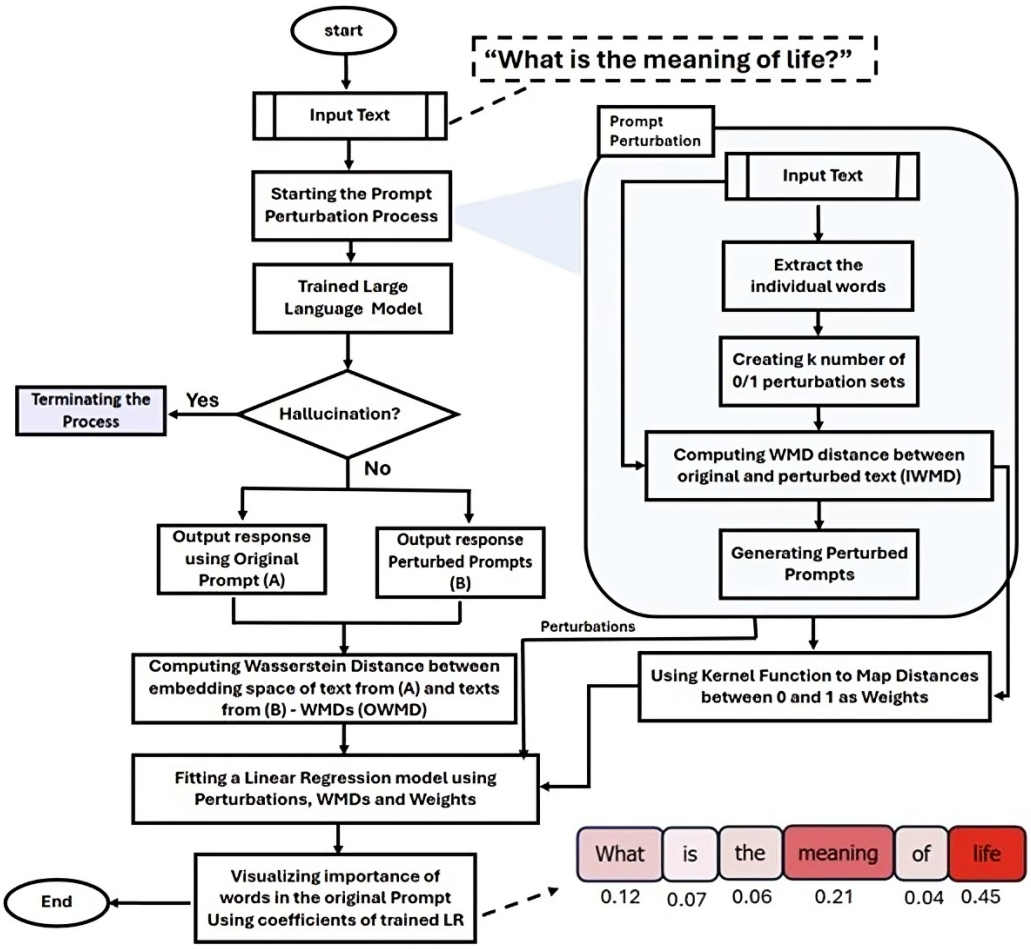


图 3.2 用于解释文本生成模型的 gSMILE 流程图

一种直观的可视化表示，展示提示词中的哪些元素对生成输出（无论是文本还是图像）贡献最大。

通过采用这一基于 SMILE 方法改进的三步流程，用户能够了解文本提示中的特定元素如何影响文本和图像领域的生成结果 [71]。该技术通过提供有价值的信息，帮助用户了解文本输入的哪些方面（例如风格指令或描述性术语）对最终输出的影响最大，从而增强用户对指令驱动的生成模型的控制能力和可预测性。

在文本生成任务中，模型会接收各种扰动提示词，并生成相应的文本输出，唯一的变化是提示词的内容。类似地，在图像编辑任务中，模型接收相同的输入图像，并搭配不同的扰动文本提示，生成的图像仅在文本输入的关键组成部分上有所不同。

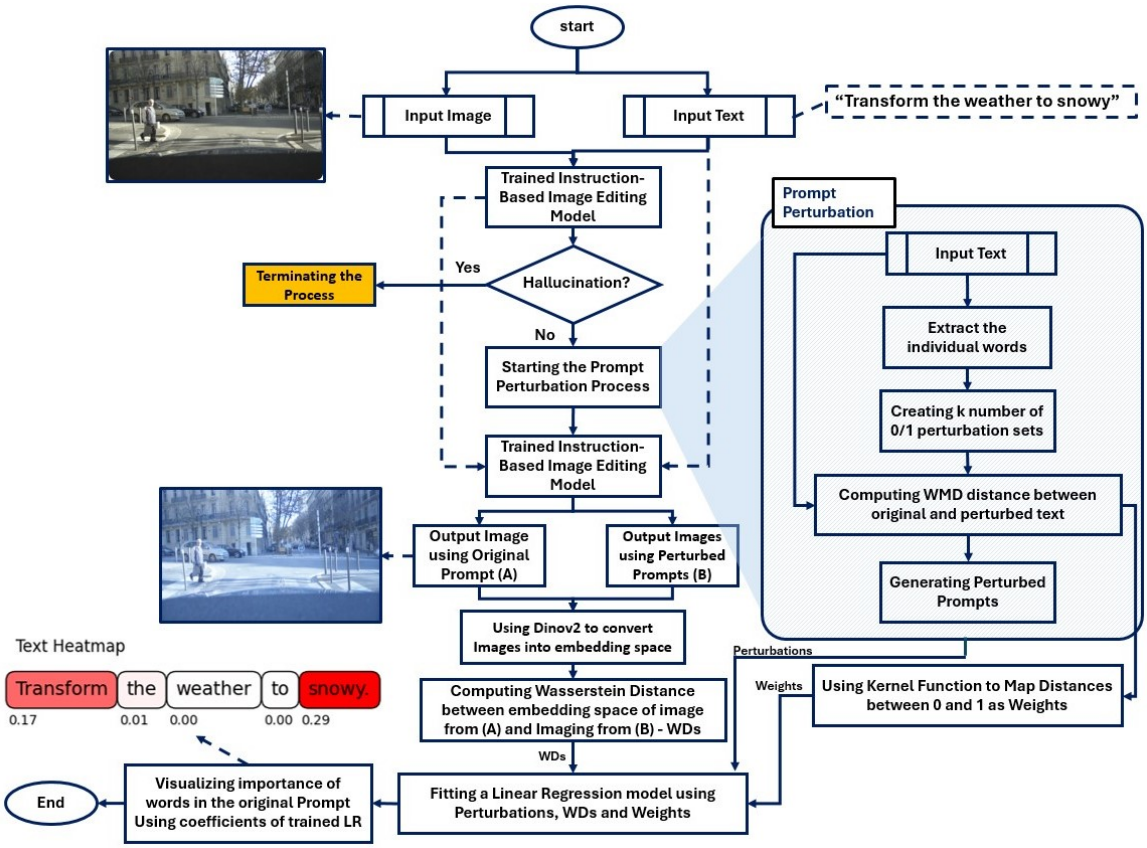


图 3.3 适用于基于指令的图像编辑模型的改编版 gSMILE 框架流程图。

将原始提示词拆分为单个词，并有选择地纳入或排除某些词，即可生成不同的文本排列，进而得到各种扰动文本 [181]。相应的模型会针对每个扰动提示生成输出：语言模型生成文本，基于指令的图像编辑模型则生成图像。

尽管基于指令的图像编辑模型在物体计数和空间推理等任务上可能面临挑战，但通过审慎选择扰动文本并限制输入域 [182]，这些问题得以缓解。在文本和图像领域中，扰动被设计为仅改变提示词的关键组成部分，从而尽量减少混杂因素，并分离每个词的影响。

特定任务的嵌入模型用于评估原始输出与扰动后输出之间的相似度。对于文本生成，采用词移距离（WMD）模型 [183]提取有意义的语义嵌入，并计算文本输出之间的相似度。对于图像生成，则使用 DINOv2 模型 [184] 生成高质量的语义嵌入，以捕捉视觉和概念内容。与直接比较图像不同，

直接比较可能只能反映颜色或像素结构等表层特征，而基于嵌入的比较则能捕捉更深层的语义变化 [185]。

该方法支持一种模型无关、可解释的框架，用于解读文本和图像输出。在图像域中，DINOv2 的自监督学习方法无需标注数据或特定于模型的输出，即可提供通用嵌入，因此适用于多种图像类型，也能适配不同的编辑模型 [184, 185]。同样，词移距离（WMD）能够灵活且有效地比较不同文本生成模型所产生的各类文本输出。

我们使用瓦瑟斯坦距离来量化文本和图像领域中原始提示词与扰动提示词所生成输出之间的差异 [186, 187]。选择这一距离度量，是因为它能够全面衡量嵌入之间的差异，捕捉分布变化的整体趋势和细微之处。与其他基于经验累积分布函数（ECDF）的距离度量（如柯尔莫哥洛夫–斯米尔诺夫距离或克拉默–冯·米塞斯距离）不同，瓦瑟斯坦距离能够有效反映数据的局部和全局变化 [186]。

在**文本域**中，瓦瑟斯坦距离作用于从词移距离（WMD）模型 [183]中提取的嵌入。这种方法能够捕捉分布几何结构，并有意义地评估扰动在概念上与原始提示词的契合程度。

在**图像域**中，瓦瑟斯坦距离是在 DINOv2 模型 [184]生成的嵌入上计算的。DINOv2 提供了丰富的语义嵌入，能够捕捉视觉和概念内容，与直接基于像素的度量相比，可提供更有信息量的比较。

最后，通过应用线性回归，将输入提示词空间与生成模型的输出空间联系起来。对于文本生成，回归模型确定特定词如何影响文本输出。基于指令的图像编辑揭示了词语如何影响生成的图像。在这两种情况下，回归系数量化单个词的影响，从而实现对模型行为的可解释性 [91]。

3.1.1 生成输出和扰动提示词

为考察文本提示中的细微变化如何影响文本生成模型和图像编辑模型的输出，我们生成原始提示的多个变体，称为扰动提示词 [72, 188]。原始文本提示记为x，扰动表示为 $\{\hat{x}_j\}_{j=1}^{J}$，其中每个$\hat{x}_j$都是通过对x进行轻微的词元级修改得到的。

对于**文本生成**，黑盒模型 $\pi^{(n)}$ 接收原始提示或其扰动版本。原始提示的输出分布为：

$$\pi^{(n)}(y \mid x), \tag{3.1}$$

对于每个扰动提示 ˆ*xj*，模型会生成：

$$\pi^{(n)}(y \mid \hat{x}_j), \quad \forall j \in \{1, \ldots, J\}. \tag{3.2}$$

对于**基于指令的图像编辑**，每个扰动提示 ˆ*xj* 都会与一个一致的输入图像 α_{input} 配对。令 ϕ_{edit} 表示图像编辑函数。与原始提示对应的基线编辑图像为：

$$\alpha'_{\text{org}} = \phi_{\text{edit}}(x, \alpha_{\text{input}}), \tag{3.3}$$

对于每个扰动提示：

$$\alpha'_{\text{pert},j} = \phi_{\text{edit}}(\hat{x}_j, \alpha_{\text{input}}), \quad \forall j \in \{1, \ldots, J\}. \tag{3.4}$$

这一表述确保了文本和图像领域的一致性。在这两种情况下，扰动集合 $\{\hat{x}_j\}_{j=1}^{J}$ 都能用于系统地评估各个提示组件对模型输出的影响力 [3, 23, 189]。

3.1.2 创建可解释空间

为了将文本生成和图像编辑模型的输出空间映射到一维相似度空间中，我们计算原始提示词和扰动提示词生成的输出之间的 Wasserstein 距离 [187, 190]。这种方法能够在不同领域中一致地衡量每个扰动的影响力。

对于 **文本生成**，我们量化基线输出分布 $\pi^{(n)}(y \mid x)$ 与每个扰动提示词 $\pi^{(n)}(y \mid \hat{x}_j)$ 对应分布之间的语义距离。我们使用词移距离（WMD）提取嵌入，该方法能够捕捉词语和句子之间的语义关系 [183]。

对于**基于指令的图像编辑**，我们同样将基线编辑图像$\phi_{\text{edit}}(x\ \alpha_{\text{input}})$与在扰动提示词 $\phi_{\text{edit}}(\hat{x}_j\ \alpha_{\text{input}})$下生成的输出进行比较。为了获得具有语义意义的表征，我们采用DINOv2模型 [184], ，这是一种自监督特征提取器，能够捕捉高层视觉和概念特征。

输出级语义偏移定义为：

$$\Delta(x,\hat{x}_j) = W\left(\pi^{(n)}(y \mid x), \pi^{(n)}(y \mid \hat{x}_j)\right), \tag{3.5}$$

其中，$W(\cdot,\cdot)$ 表示嵌入空间中两个分布之间的 Wasserstein 瓦瑟斯坦距离。实际上，W 是基于文本输出的 WMD 嵌入或图像输出的 DINOv2 嵌入计算得出的。Wasserstein-p 度量中的范数阶数 p 控制敏感度：较小的值更强调局部变化，而较大的值则能捕捉更广泛的分布偏移 [190]。

算法 1： 瓦瑟斯坦距离的基于自助法的 P 值计算（文本：使用 WMD 嵌入的 OWMD；图像：使用 DINOv2 嵌入的瓦瑟斯坦距离）

输入： 两组输出嵌入：X （基线）和Y （扰动）
输出： 观测到的结果瓦瑟斯坦距离 $\Delta(x, \hat{x}j)$ 及其对应的 p 值 $pVal$

最大迭代次数 $\leftarrow 10^5$;
$\Delta(x, \hat{x}j) \leftarrow$ Wasserstein_距离(X, Y);
XY $\leftarrow$ 将X与Y拼接;
LX $\leftarrow$ X的长度;
LY $\leftarrow$ Y的长度;
较大者 $\leftarrow 0$;
对于$i \leftarrow 1$ **至**$MaxItr$，**执行**
    从XY中随机抽取一个大小为 LX 的样本e ;
    从XY中随机抽取一个大小为 LY 的样本f ;
    bootWD $\leftarrow$ Wasserstein_距离(e, f);
    如果$bootWD \geq \Delta(x, \hat{x}j)$ **则**
        更大 $\leftarrow$ 更大 $+1$;
    end
结束
$pVal$ $\leftarrow$更大/MaxItr;
返回 $\Delta(x, \hat{x}j)$, $pVal$;

为确保观测到的距离反映的是有意义的差异而非噪声，我们使用自助法估计统计显著性 [191, 192]。对于每个扰动，我们计算一个 p 值，表示在随机重采样下获得与观测值同样大的瓦瑟斯坦距离的概率。p 值较高的扰动将被视为不显著并予以剔除 [193]。

自助法总结于算法 1。

3.1.3 使用高斯核进行相似度加权

为确保可解释代理模型能够恰当地考虑每个扰动提示词的相关性，我们采用基于相似度的加权方案。与原始输入在语义上越相似的扰动，其重要性权重越高；而不相似的扰动则获得较低的权重。这样可以降低异常值或大幅修改后的提示词的影响力。

对于**文本生成** 和 **图像编辑**，由于扰动作用于文本提示，因此在输入级别计算相似度。我们采用词移距离（WMD） [183], ，它计算将原始提示词x 的词分布转换为扰动提示词 $\hat{x}_j$的词分布所需的最小累计距离：

$$\mathrm{IWMD}(x,\hat{x}_j) = \min_{T\geq 0} \sum_{k=1}^{n}\sum_{l=1}^{m} T_{kl}\, d(w_k, w_l), \tag{3.6}$$

满足以下流量约束：

$$\sum_{l=1}^{m} T_{kl} = p_k, \quad \forall k \in S_x, \qquad \sum_{k=1}^{n} T_{kl} = q_l, \quad \forall l \in S_{\hat{x}_j}, \qquad T_{kl} \geq 0.$$

这里，w_k 和w_l 表示词嵌入，而p_k，q_l 分别是x 和 $\hat{x}_j$的归一化词袋分布。我们将这种输入级语义距离记为：

$$\delta_{x_j} = \mathrm{IWMD}(x,\hat{x}_j).$$

为了将这些距离转换为用于回归的基于相似度的权重，我们使用高斯核 [3]:

$$w_j = \exp\left(-\left(\frac{\delta_{x_j}}{\sigma}\right)^2\right), \quad \forall j \in \{1,\ldots,J\}, \tag{3.7}$$

其中，$w_j \in (0,1]$是扰动$\hat{x}_j$的相关性权重，而 $\sigma > 0$是控制指数衰减速率的核宽度参数。

其中，$\delta_{x_j} = \mathrm{IWMD}(x, \hat{x}_j)$ 表示基于WMD的输入距离，σ控制高斯核的敏感度 [194]。所得权重w_j 反映了每个扰动提示与原始提示之间的相似度，并在回归分析中用于优先处理语义上更接近的扰动。

3.1.4 开发可解释代理模型

为了理解提示词中的不同词如何影响模型的输出，我们训练了一个局部代理模型，用于映射扰动输入提示与相应输出偏移之间的关系。该方法能够在文本生成和图像编辑任务中，清晰地将影响力归因于各个词元 [195]。

在**文本生成** 的情况下，扰动提示词表示为特征向量z_j，通常采用独热编码或二进制指示变量来表示词语的包含或排除。响应变量是输出级语义偏移 $\Delta(x, \hat{x}_j)$，定义见方程 3.5。相似度权重w_j由高斯核（方程 3.7）得出，为与原始提示词更相似的扰动赋予更高的重要性。

对于**基于指令的图像编辑**，采用相同的流程。自变量仍是扰动提示词向量z_j，而响应变量 $\Delta(x, \hat{x}_j)$则通过计算基线图像与扰动图像在 DINOv2 嵌入空间中的瓦瑟斯坦距离 [184]得出。加权机制确保语义上忠实的扰动对回归结果 [90]的影响更大。

加权线性代理模型的目标函数为：

$$\min_{\theta} \frac{1}{J} \sum_{j=1}^{J} w_j \left(h_\theta(z_j) - \Delta(x, \hat{x}_j) \right)^2, \tag{3.8}$$

特征提取后，代理模型h_θ 在扰动提示向量及其对应输出偏移的空间上进行训练。学习得到的系数 θ 量化了提示词中每个词元对模型行为的影响，从而提供直观且可解释的解释。

3.1.5 评估指标

为回答**研究问题一（RQ1）**，我们采用了一套受谷歌 [196], 提供的奠基性工作启发的评估指标，强调评估可解释模型的多面性。该框架强调评估**准确性**、**稳定性**、**保真度**和**一致性** 的重要性，将其作为全面评估模型行为的必要工具，尤其是在将可解释模型与传统黑箱模型进行比较时。

准确性 是指模型输出相对于预期结果的正确性（如适用）。**稳定性** 衡量可解释性框架在不同扰动和模型运行中的鲁棒性。**保真度** 评估可解释代理模型对原始黑盒模型行为的近似程度。**一致性**

评估相似输入之间的解释对齐，确保可解释性结果连贯且可靠。

通过采用这些指标，我们提供了一套结构化方法，用于严格剖析和理解我们的统一可解释性框架在文本生成模型和图像编辑模型中的可靠性与性能 [196], ，从而直接支持对 RQ1 的调查和解答。

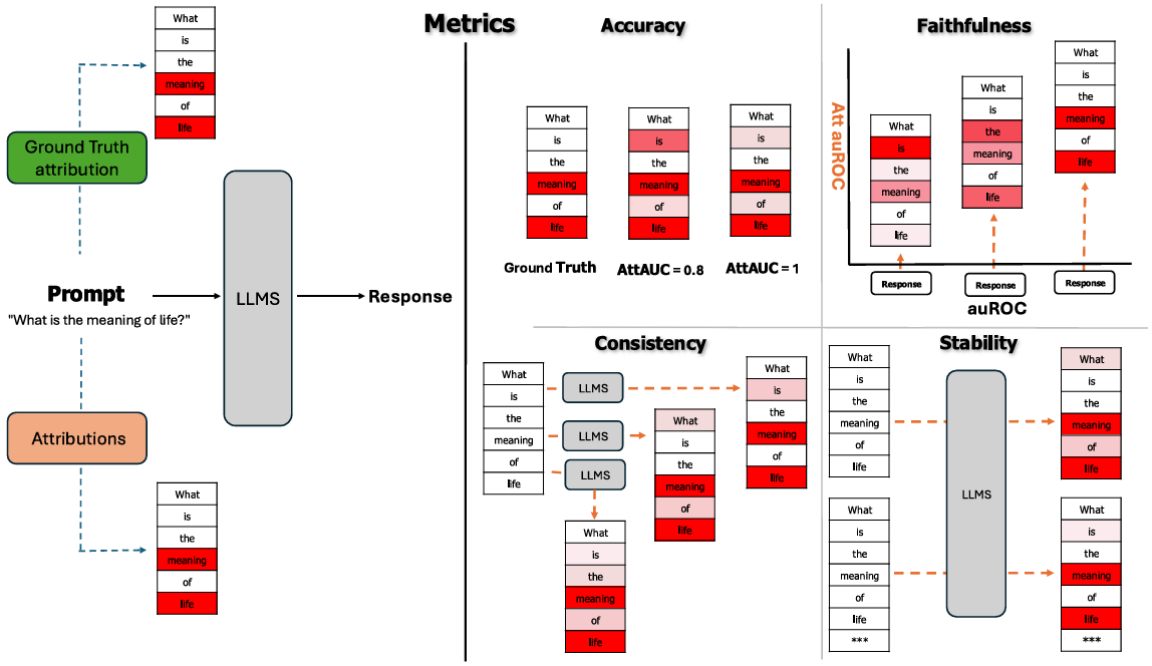


图 3.4 使用基于归因的指标评估大语言模型可解释性的框架。该图展示了系统如何将真实归因与模型生成的归因进行比较，以计算准确性（曲线下面积/AUMAC）、忠实度（通过 aUROC 进行评估，该指标比较输入-输出归因对齐情况）、一致性（不同模型输出之间的一致程度）和稳定性（归因对扰动的鲁棒性）。这种方法能够对自然语言生成任务中的可解释性方法进行全面评估。

ATT 准确率

准确性衡量模型输出与预期结果（真实值）的对齐程度。在我们的可解释性框架中，这涉及将归因分数（反映特定提示词的影响力）与真实标签进行比较，这些真实标签用于标识哪些提示元素确实与生成输出相关 [197]。

对于**文本生成**，将归因分数与人工标注真实值进行比较，以识别对生成内容有贡献的重要词语。对于**图像编辑**，

将模型的归因与标签进行比较，以识别提示词中的哪些词对应编辑后图像中的特定视觉变化或地区。

为量化准确性，我们使用接收者操作特征（ROC）的**曲线下面积（AUC）** [198]。AUC 反映模型将相关提示元素（由真实值确定）排在不相关元素之前的能力：

- **AUC $\sim$ 1:** 模型能够有效区分相关元素，与人工识别的真实值高度吻合。

- **AUC $\sim$ 0.5:** 模型的排序是随机的，表明其与真实值的对齐程度较差。

例如，图3.5展示了一个案例，其中真实值将“make”和“rainy”识别为影响生成输出的最相关文本元素。指标**AttAUC** 用于评估模型归因分数与真实值之间的对齐程度，其中：

- **AttAUC = 1.0:** 表示完全对齐。
- **较低的 AttAUC（例如，0.8）：** 表示准确性下降 [199]。

模型生成的归因热图如图3.5所示。在这些热图中，较深的红色表示对特定文本元素的归因程度较高。准确的模型应将更多注意力分配给对生成的输出（无论是文本还是视觉内容）影响最大的文本组成部分。这有助于提高模型预测结果的可解释性和可靠性 [197]。

此类评估有助于利益相关者就模型部署做出明智的决策，并指导微调工作，以提升可解释性性能。

注意力准确率（ATT ACC）、**注意力F1分数（ATT F1）**和**注意力曲线下面积（ATTAUROC）**定义如下：

$$\text{ATT ACC} = \frac{\text{Number of correctly attributed elements}}{\text{Total number of elements}} = \frac{\sum_{i=1}^{N} \mathbb{1}[\hat{y}_i = y_i]}{N} \tag{3.9}$$

描述： ATT ACC 衡量正确归因元素所占的比例，其中：

- N 表示元素总数，
- $\hat{y}_i$ 是第i个元素的预测归因分数，
- y_i 是相应的真实标签，

- $\mathbb{1}[\hat{y}_i = y_i]$ 是一个指示函数：若 $\hat{y}_i = y_i$，则返回 1；否则返回 0。

$$\text{ATT F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.10}$$

描述： ATT F1 提供精确率与召回率的调和平均数，用以平衡假阳性与假阴性。较高的 F1 分数表明，在准确且相关的归因方面，模型具有更好的整体模型性能。精确率和召回率的定义如下：

- 精确率 $=$ 真阳性 (TP) 真阳性 (TP)+假阳性 (FP)，
- 召回率 $=$ 真阳性 (TP) 真阳性 (TP)+假阴性 (FN)。

$$\text{ATT AUROC} = \frac{1}{|P| \cdot |N|} \sum_{p \in P} \sum_{n \in N} \mathbb{1}[\hat{y}_p > \hat{y}_n] \tag{3.11}$$

描述： ATT AUROC 评估模型将相关元素（正类真实值）排在无关元素（负类真实值）之上的能力。其范围从 0.5（随机排序）到 1.0（完美排序）。其中：

- P 和 N 分别是正类与负类真实标签的集合，
- $|P|$ 和 $|N|$ 是这些集合的大小，
- $\hat{y}_p$ 和 $\hat{y}_n$ 是模型对正样本和负样本的预测分数，
- $\mathbb{1}[\hat{y}_p > \hat{y}_n]$ 是一个指示函数，若正样本的分数更高则返回 1 r 于负样本的分数，否则返回 0。

这些指标确保该模型在准确性、一致性和相关性方面得到严格测试，从而对其与人工标注的真实标签的对齐情况提供稳健的评估。

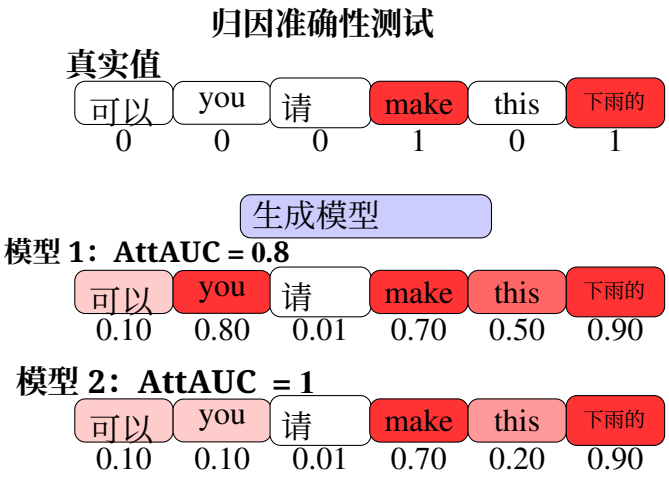


图 3.5 AttAUC（注意力曲线下面积）指标对图像编辑模型准确性的评估。更深的红色表示文本元素对图像特征的归因更强，表明特定提示词对编辑图像的影响力。

ATT 稳定性

可解释模型中的稳定性是指，即使对输入数据做出微小更改，它仍能提供一致解释的能力 [11]。这意味着输入中的微小扰动不应显著影响模型的预测结果或解释。稳定性是促进对可解释人工智能（XAI）系统信任的关键属性 [13]。它让用户确信，模型的解释并非任意的，也不会对无关变化过度敏感 [88]。

对于**文本生成**，稳定性评估对输入提示的微小改动（例如添加或删除非关键的词或符号）是否会在提示元素上产生相似的归因模式。在这些实验中，通过在提示词末尾添加符号 *** 引入了轻微的文本变体。对于**图像编辑**，稳定性评估提示词的微小改动是否会使文本元素在不同编辑后的图像中对视觉特征产生一致的归因，其中变体是通过在提示词末尾添加符号 ### 引入的。

示例可视化结果展示了模型对两段略有不同的文本输入所给出的注意力分数。热力图展示了模型如何为每个词分配注意力。如图 3.6 所示，稳定的模型会在两种输入中对词*make* 和*rainy* 赋予相近的重要性，即使其中一种输入加入了额外符号（文本生成时为 ***，图像编辑时为 ###）。这种一致性对于确保模型在面对细微的输入变化时仍能表现得可预测至关重要 [200]。

如果缺乏这种稳定性，模型生成的解释可能会导致混淆或误解，尤其是在医疗诊断或金融决策等高风险场景中 [3]。

我们使用**杰卡德指数**来量化稳定性，它衡量两组模型归因之间的相似度。对于两个集合A和 B，杰卡德指数的计算方式如下：

$$\text{Jaccard}(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.12}$$

该指标反映两个归因集合中共有元素所占的比例，数值越接近 1，表示解释 [11] 的稳定性和相似度越高。杰卡德指数越高，说明模型能够提供稳定的输出和可靠的解释 [6]。

通过确保稳定性，模型不仅增强了鲁棒性，还提高了其在现实世界应用 [13]中的可用性和可信度。



图3.6 文本略有变化时，图像编辑模型注意力的稳定性评估。较深的红色表示在扰动提示词中，文本元素的归因具有一致性。

ATT一致性

一致性用于评估模型在反复接收相同输入时，是否能产生稳定可靠的输出。这一特性确保模型的行为具有可预测性，增强人们对其输出的信任，并提升模型在各种场景中的可用性 [3]。

对于**文本生成**，一致性评估使用完全相同的提示词重复执行时，是否会产生相似的归因模式和生成内容。对于**图像编辑**，它评估使用相同输入提示和图像多次运行时，是否始终突出显示相同的提示元素，并产生相似的视觉编辑。

如图 3.7 所示，当多次向模型提供输入短语“能否请你把这个改成下雨天？”时，各次迭代的输出都保持稳定。在文本和图像任务中，模型始终为“make”和“rainy”这两个词赋予较高的权重，表明它能够可靠地识别输入中语义上最重要的元素 [6]。

这种一致的行为凸显了模型聚焦于提示词语义核心的能力，这一特性对于在实际应用中保持连贯性和可用性至关重要 [88]。在自动化内容生成、天气模拟和交互式系统等领域，用户依赖模型提供一致且可预测的输出。任何不一致都可能削弱用户信任，并损害系统的整体效能。

我们的模型通过展示在重复运行中的一致性来增强用户信心，并为进一步优化以及集成到更复杂的系统中奠定基础。这种保持解释和输出稳定的能力直接提升了模型的鲁棒性和可扩展性，使其非常适合部署在动态的现实环境中 [13]。

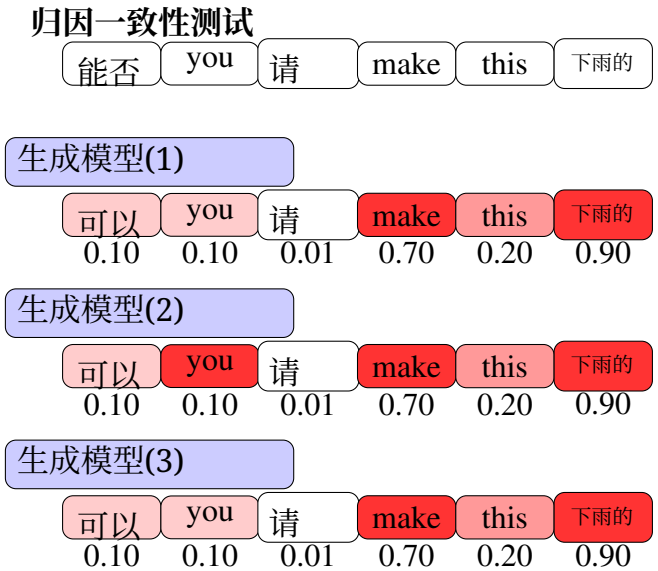


图 3.7 Instruct-Pix2Pix 模型的一致性分析。热力图展示了每个词在三次迭代中对模型输出的影响力。

ATT 忠实度

保真度评估可解释代理模型与原始黑盒模型的一致程度。该指标衡量在各种输入扰动下，代理模型的预测结果与黑盒模型预测结果的近似程度 [3]。

形式上，设黑盒模型的真实输出变化为 $\Delta(x, \hat{x}_j)$（见公式 3.5），代理模型预测的变化为 $\hat{\Delta}_j = h_\theta(z_j)$。忠实度通过比较所有扰动 $j \in \{1, \ldots, J\}$ 下的 $\hat{\Delta}_j$ 与 $\Delta(x, \hat{x}_j)$ 来量化。

基于系数的指标。 **决定系数** R^2 用于评估代理模型与黑盒模型 [201]的相关程度：

$$R^2 = 1 - \frac{\sum_{j=1}^{J} \left(\Delta(x, \hat{x}_j) - \hat{\Delta}_j\right)^2}{\sum_{j=1}^{J} \left(\Delta(x, \hat{x}_j) - \overline{\Delta}\right)^2}, \tag{3.13}$$

其中，$\overline{\Delta}$ 为黑盒距离的均值。接近 1 的值表示代理模型与黑盒模型之间具有很强的对齐关系。

对扰动进行加权时，使用**加权决定系数** R_w^2 [200, 202]：

$$R_w^2 = 1 - \frac{\sum_{j=1}^{J} w_j \left(\Delta(x, \hat{x}_j) - \hat{\Delta}_j\right)^2}{\sum_{j=1}^{J} w_j \left(\Delta(x, \hat{x}_j) - \overline{\Delta}_w\right)^2}, \tag{3.14}$$

其中 $\overline{\Delta}_w = \frac{\sum_{j=1}^{J} w_j \Delta(x, \hat{x}_j)}{\sum_{j=1}^{J} w_j}$ 为加权平均值。

为考虑特征数量 N_s 和扰动数量 J，定义**加权调整决定系数** 如下 [200]：

$$\hat{R}_w^2 = 1 - (1 - R_w^2) \left[\frac{J-1}{J - N_s - 1}\right]. \tag{3.15}$$

基于误差的指标。使用**加权均方误差 (WMSE)** 和**加权平均绝对误差 (WMAE)** [203] 来衡量代理模型与黑箱预测结果之间的误差幅度：

$$\text{WMSE} = \frac{\sum_{j=1}^{J} w_j \left(\Delta(x, \hat{x}_j) - \hat{\Delta}_j\right)^2}{\sum_{j=1}^{J} w_j}, \tag{3.16}$$

$$\text{WMAE} = \frac{\sum_{j=1}^{J} w_j \left|\Delta(x, \hat{x}_j) - \hat{\Delta}_j\right|}{\sum_{j=1}^{J} w_j}. \tag{3.17}$$

此外，还会计算未加权的**平均**L_1 和**平均**L_2 损失 [204, 200]:

$$L_1 = \frac{1}{J}\sum_{j=1}^{J}\left|\Delta(x,\hat{x}_j) - \hat{\Delta}_j\right|, \tag{3.18}$$

$$L_2 = \frac{1}{J}\sum_{j=1}^{J}\left(\Delta(x,\hat{x}_j) - \hat{\Delta}_j\right)^2. \tag{3.19}$$

解读。 较低的 WMSE、WMAE、L_1, 和L_2, 值，以及较高的R^2, R_w^2和 $\hat{}R_w^2$值，表示保真度更高，也就是说，代理模型能够更准确地近似黑盒模型的行为。

我们在一系列场景中计算保真度。对于**文本生成**，通过比较可解释模型在输入扰动文本和原始文本时预测的距离与黑盒语言模型的对应距离，来评估保真度。对于**图像编辑**，则通过比较代理模型与原始图像编辑模型的 DINOv2 图像嵌入距离来计算保真度。

计算 ATT 忠实度的整体工作流程如图 3.8 所示。如图所示，输入提示经过扰动后，全局模型和局部代理模型都会生成相似度信号，随后对这些信号进行比较，以评估 ATT 忠实度。对于**文本生成**，使用结果词移距离（OWMD）计算相似度。对于**基于指令的图像编辑**，则通过计算 DINOv2 嵌入之间的瓦瑟斯坦距离来衡量相似度。

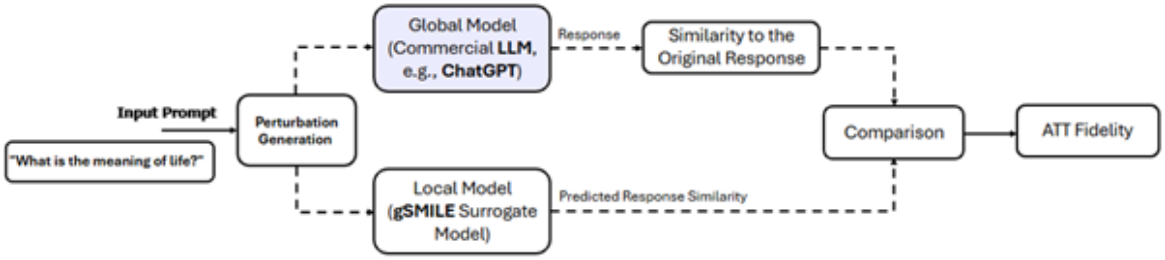


图 3.8 ATT 忠实度评估框架。输入提示词经过扰动后，全局模型（例如，用于文本生成的 ChatGPT、用于图像编辑的 InstructPix2Pix）和局部代理模型（gSMILE）都会生成相似度信号。比较这些信号，可以量化代理模型对黑盒模型行为的近似程度。

结论

本章介绍了 gSMILE 框架，这是一种用于解释大语言模型（LLM）和基于指令的图像编辑系统的统一方法。gSMILE 在 LIME 的基础上，结合统计距离度量和模型无关技术，为理解输入变化如何影响模型输出提供了一种灵活且一致的方法，适用于文本生成和图像生成任务。

本章全面介绍了 gSMILE 框架：从为输入生成细微变化，到构建有助于解释原始模型行为的简化模型。我们还讨论了如何使用稳定性、保真度和一致性等指标来评估这些解释的质量。gSMILE 的突出之处在于，它不要求访问模型内部的运行过程，这意味着即使面对闭源系统或仅限 API 的系统，也可以使用该方法。

本章为后续内容奠定基础。在后续内容中，我们将在真实世界的生成式模型上检验 gSMILE，涵盖语言和图像任务，以评估其实际性能。

第 4 章

大语言模型的可解释性

本章直接探讨**研究问题2（RQ2）**，研究可解释性方法如何揭示大语言模型（LLM）中单个输入元素的影响力。本章介绍了解释大语言模型输出结果所面临的挑战与采用的方法，重点关注模型无关方法。本研究提出了 gSMILE 工作流程，旨在为大语言模型预测提供局部解释。图4.1展示了本研究所采用的可解释性工作流程概览。我们在多个领先的大语言模型（LLM）上测试了 gSMILE，并使用准确性、一致性、稳定性和保真度等指标，证明其能够提供清晰且可靠的解释。通过让这些模型更易于理解，gSMILE 使我们进一步接近于让人工智能更加透明和值得信任。

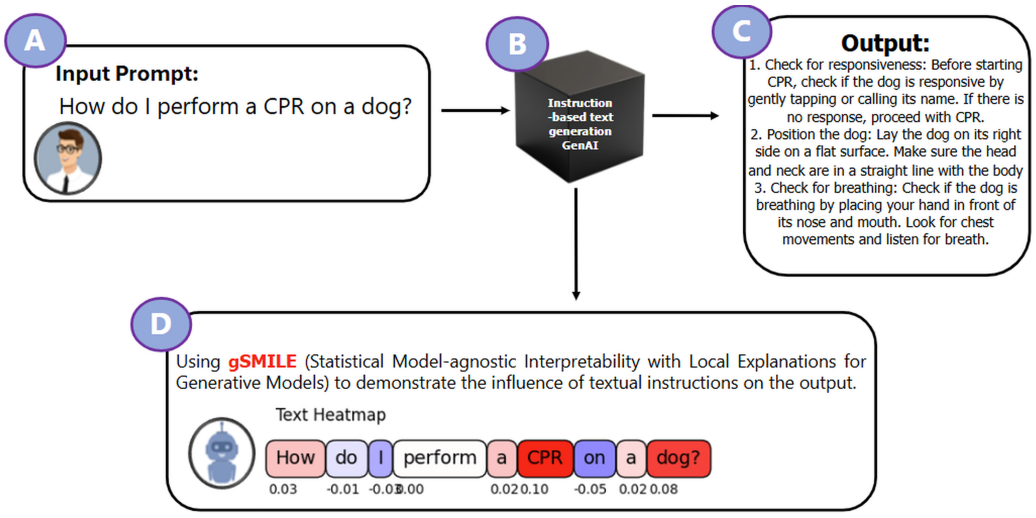


图 4.1 使用 gSMILE（基于局部解释的统计模型无关可解释性）对大语言模型（LLM）应用可解释性流程的概览。

4.1 问题定义

本节介绍我们方法所依据的核心概念和数学框架。首先，我们定义关键符号和问题设置，随后正式说明如何量化输入扰动的影响。接着，我们介绍如何使用代理模型在局部近似模型行为，并基于瓦瑟斯坦距离给出理论依据 [3, 187, 190]。最后，我们概述一套结构化方法，用于实现专为大语言模型设计的基于瓦瑟斯坦距离的替代框架 [90]。

词移距离（WMD）是两个文档之间的 1-Wasserstein（推土机）距离；这两个文档在嵌入空间中表示为归一化词袋，并以词嵌入之间的欧几里得距离作为地面代价 [183]。WMD 能够捕捉对齐两个文档的词分布所需的最小语义“移动”量，从而为文档间的差异提供具有语义意义的度量，因此非常适合我们的可解释性框架。

4.1.1 符号表示与问题设定

背景介绍。 我们的目标是构建一个局部代理模型，用于解释黑盒大语言模型（LLM） $\pi^{(n)}$ 对自然语言

输入提示的响应行为。我们的目标是了解输入的特定部分如何影响模型的输出分布。以下使用的符号如下：

- **输入：** $x \in \mathscr{X}$，表示原始输入提示。
- **扰动：** $\{\hat{x}_j\}_{j=1}^{J}$，由J个扰动版本组成，这些版本通过对词元进行少量编辑生成。
- **模型输出：** $\pi^{(n)}(y \mid \hat{x}_j)$，模型在输出空间$\mathscr{Y}$上的概率分布，以扰动后的输入$\hat{x}_j$为条件。对于未扰动的输入$x$，我们记作$\pi^{(n)}(y \mid x)$。

4.1.2 输入级距离

为量化原始提示x与每个扰动版本$\hat{x}_j$之间的语义差异，我们计算词移距离（WMD）[183]。我们将这种输入级语义距离记为 IWMD（内部 WMD）：

$$\delta_{x_j} = \mathrm{IWMD}(x, \hat{x}_j), \tag{4.1}$$

其中，$\delta_{x_j} \in \mathbb{R}_{\geq 0}$ 表示原始提示词与扰动提示词之间的语义距离。

4.1.3 输出层面的分布偏移

为评估输入扰动对模型行为的影响，我们计算原始输入与扰动输入的输出分布之间的 Wasserstein 距离。我们将其称为 OWMD（结果词移距离）：

$$\mathrm{OWMD}(x, \hat{x}_j) = W\left(\pi^{(n)}(y \mid x), \pi^{(n)}(y \mid \hat{x}_j)\right), \tag{4.2}$$

$$\Delta(x, \hat{x}_j) = \mathrm{OWMD}(x, \hat{x}_j), \tag{4.3}$$

其中 $y \in \mathscr{Y}$，且 $W(p,q)$ 表示两个概率分布 p 和q之间的 Wasserstein 距离。

4.1.4 基于输入相似度的加权

语义上越接近原始输入的扰动，所赋予的权重就越高。我们采用基于高斯分布的加权方案 [3]:

$$w_j = \exp\left(-\left(\frac{\delta_{x_j}}{\sigma}\right)^2\right), \tag{4.4}$$

其中，$w_j \in (0,1]$ 表示 $\hat{x}_j$ 的相关性权重，而 $\sigma > 0$ 是核宽度参数，用于控制指数衰减速率。

4.1.5 拟合局部代理模型

每个扰动后的输入 $\hat{x}_j$ 都会映射为一个特征向量 $z_j \in \mathbb{R}^d$。我们拟合如下形式的线性替代模型：

$$h_\theta(z_j) = \theta_0 + \theta^\top z_j, \tag{4.5}$$

其中，$\theta_0 \in \mathbb{R}$ 是偏置项，$\theta \in \mathbb{R}^d$ 是权重向量。通过最小化加权平方误差来估计参数 θ。

$$\min_\theta \sum_{j=1}^{J} w_j \left(h_\theta(z_j) - \Delta(x, \hat{x}_j)\right)^2, \tag{4.6}$$

其中，损失确保代理模型与观测到的输出偏移之间具有局部 ATT 忠实度。

4.1.6 理论依据（Lipschitz 平滑性）

偏移函数定义。我们定义一个偏移函数，用于局部量化模型对扰动的响应：

$$f(x_a, \hat{x}_a) = W\left(\pi^{(n)}(y \mid x_a), \pi^{(n)}(y \mid \hat{x}_a)\right), \tag{4.7}$$

如果 $\pi^{(n)}$ 关于其输入是利普希茨连续的，那么函数 f 也是利普希茨 [205]：

$$|f(x_a, \hat{x}_a) - f(x_b, \hat{x}_b)| \le L \times (\|x_a - x_b\| + \|\hat{x}_a - \hat{x}_b\|), \forall a, b \in \{1, \ldots, N_{pert}\} \tag{4.8}$$

符号 x_a 和 $\hat{x}_a$ 分别表示一个特定输入及其对应的扰动。下标使我们能够在分析扰动之间的利普希茨连续性或稳定性时，将公式扩展到多个输入对（x_i，$\hat{x}_i$）。N^{pert} 表示扰动总数。

其中，$L > 0$是利普希茨常数，$\|\cdot\|$表示输入空间上的范数（例如词元嵌入空间上的范数）。

与全变差的关系。对于输出空间 $\mathscr{Y}$ 上的两个分布 p 与 q，康托罗维奇–鲁宾斯坦对偶给出：

$$W(p,q) \le \operatorname{diam}(\mathscr{Y}) \cdot \mathrm{TV}(p,q), \tag{4.9}$$

其中 $\operatorname{diam}(\mathscr{Y}) = \sup_{y_1,y_2\in\mathscr{Y}} d(y_1,y_2)$ 表示在所选基度量d下输出空间的直径，即任意两个输出之间可能的最大距离。

全变差距离（TV）由下式给出：

$$\mathrm{TV}(p,q) = \frac{1}{2}\sum_{y\in\mathscr{Y}} |p(y) - q(y)|. \tag{4.10}$$

该界表明，分布的微小变化（在全变差意义下）意味着 Wasserstein 距离有界。

4.1.7 线性代理模型近似Lipschitz函数

由于利普希茨函数的可微性，偏移函数f可以用一阶泰勒展开进行局部近似：

$$f(x') \approx f(x) + \nabla f(x)^\top (x' - x), \tag{4.11}$$

其中 $\nabla f(x)^\top (x' - x)$表示$f$的梯度与输入空间中局部位移的点积。

完整的基于Wasserstein距离的LIME代理过程包括在输入提示x附近生成扰动 $\{\hat{}x_j\}_{j=1}^{J}$。随后，该过程计算语义距离 δ_{x_j}（公式 4.1）和输出偏移 Δ（x，$\hat{}xj$）（公式 4.3）。接着，根据公式 4.4 分配相似度权重wj。然后，该过程为每个扰动提取特征表示zj，并通过最小化损失来拟合局部代理模型h_θ（公式 4.6）。最后，使用拟合后的模型h_θ来解释各输入特征对模型行为的影响力。该框架通过局部扰动和响应分析，为分析大语言模型提供了一种有原则且可解释的方法。

4.2 大语言模型方法摘要

本节概述一般性 gSMILE 框架如何从第 3 章所述的形式进行调整，以解释大语言模型（LLM）在文本生成任务中的行为。其目标是确定哪些提示组件对模型输出的影响力最大。

我们首先将原始输入提示分解为各个词。扰动提示词通过每次系统地移除或更改一个词来生成，从而得到修改后的输入。每个扰动提示都会传递给大语言模型，由其产生相应的输出。由原始提示生成的文本作为比较的基线。

为衡量输出如何变化，我们计算每个扰动输出与基线输出之间的Wasserstein距离 [187]。这些距离在语义嵌入空间中使用词移距离（词移距离）[183], 计算，基于底层词嵌入量化文本之间的不相似度。

这些距离值随后使用高斯核进行变换，以生成反映每个扰动提示与原始提示接近程度的权重。使用这些权重训练局部加权线性回归模型 [3, 90]，其中每个扰动提示被表示为一个二值特征向量，用以指示特定词的包含或排除。

回归系数反映了原始提示词中的每个词对生成输出的影响力。这些数值以热力图的形式呈现在输入提示上，突出显示了对最终文本贡献最大的词。

整个流程采用 SMILE 的模型无关、基于分布的解释策略 [4]，并针对文本生成的独特特征对其进行调整。在文本生成中，语义一致性和局部扰动的影响至关重要。图 3.2 展示了调整后的流程。

4.3 实验结果

本节评估所提出的方法在提升指令式文本生成可解释性方面的能力。我们分析文本提示的结构如何影响可解释性、输入文本的变化如何影响解释的清晰度，以及所提出的解决方案在不同场景中的表现。

我们使用一系列数据集和三种先进的文本生成模型开展实验：OpenAI GPT [7], LLaMA [20], 和 Claude-AI [206]，以确保评估全面。

4.3.1 定性结果

定性评估通过基于场景的测试和可视化，展示了所提出框架的可解释性。我们通过改变文本提示设计了多种测试场景，包括调整句子复杂度、改述以及加入领域专用术语。这些场景考察模型如何根据文本输入中细微的变化进行理解并生成输出。

输入提示可解释性

所提出的框架通过针对多种提示词变体进行基于场景的测试，评估模型行为对于不同输入的可解释性。这些变体包括改述问题、描述性修饰语以及措辞变化。针对每个输入，系统都会生成文本热图，突出显示各个词对模型输出的贡献，并用颜色表示重要性分数。

图4.2展示了改变输入提示的措辞或重点如何产生不同的词语重要性模式，体现了模型对语言变化的敏感性。

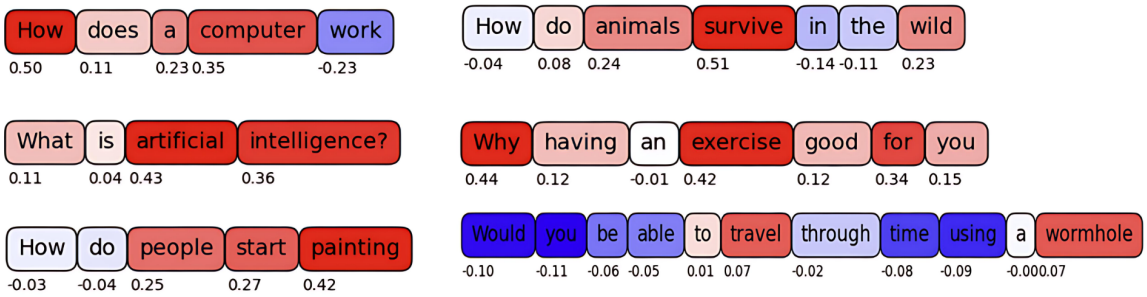


图4.2 各类输入提示的词级重要性文本热图。颜色强度表示每个词对模型输出的影响力。

通过词元级归因实现性别公平性

我们进一步考察了性别化措辞是否会影响归因模式，方法是将 gSMILE 应用于三个密切相关的提示词：“生命的意义是什么？”“从男性视角看，生命的意义是什么？”以及“从女性视角看，生命的意义是什么？”。在这三个模型中，中性表述始终突出词元*meaning* 和*life*，凸显了它们在塑造回答方面的核心作用。然而，一旦引入性别限定语，归因热图便发生变化，揭示出各模型特有的敏感性。如图 4.3 所示，GPT-3.5 倾向于赋予

词元*women* 比*men*更大的权重，这表明性别术语的影响力并不对称。Claude 3.5 对注意力的分配更为均衡，在平衡两个性别化词元贡献的同时，仍提高了它们相对于基线的重要性。相比之下，LLaMA 3.1 表现出更大的变异性，其归因有时会更显著地放大某一性别指称，而非另一方。

这些发现展示了 gSMILE 如何揭示仅检查生成文本时仍会隐藏的公平性相关敏感性。若某个性别术语吸收了不成比例的归因分数，就表明存在一种潜在关联，可能会使模型的推理产生偏见。从实践角度来看，这项分析表明，在提示词中避免使用性别化修饰语，有助于将注意力保持在任务的语义核心上，从而降低发生意外归因漂移的风险。

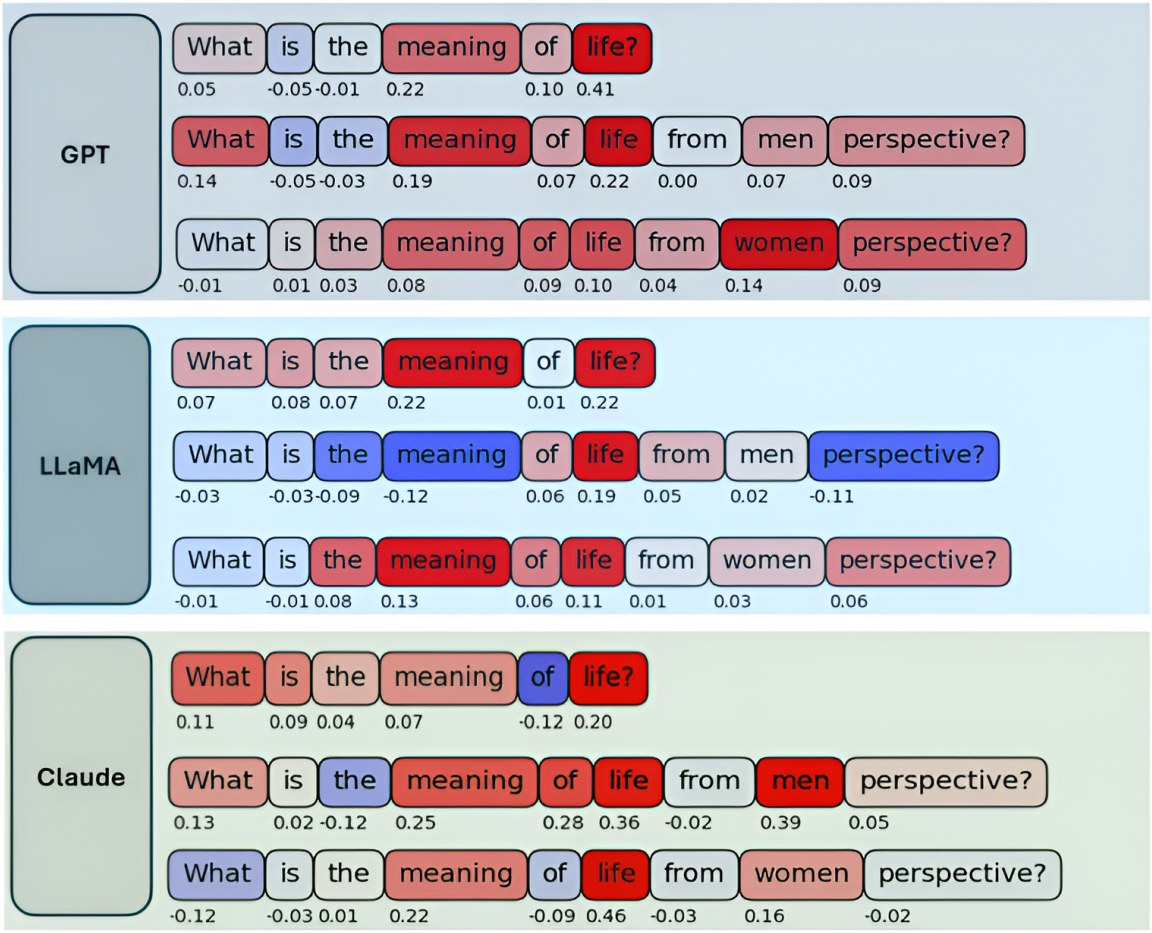


图4.3 在三种提示词（中性、“男性视角”和“女性视角”）下，GPT-3.5、LLaMA 3.1 和 Claude 3.5 的 gSMILE 词元级归因对比。颜色强度表示各词元对输出分布的相对影响力。

推理路径与词元归因

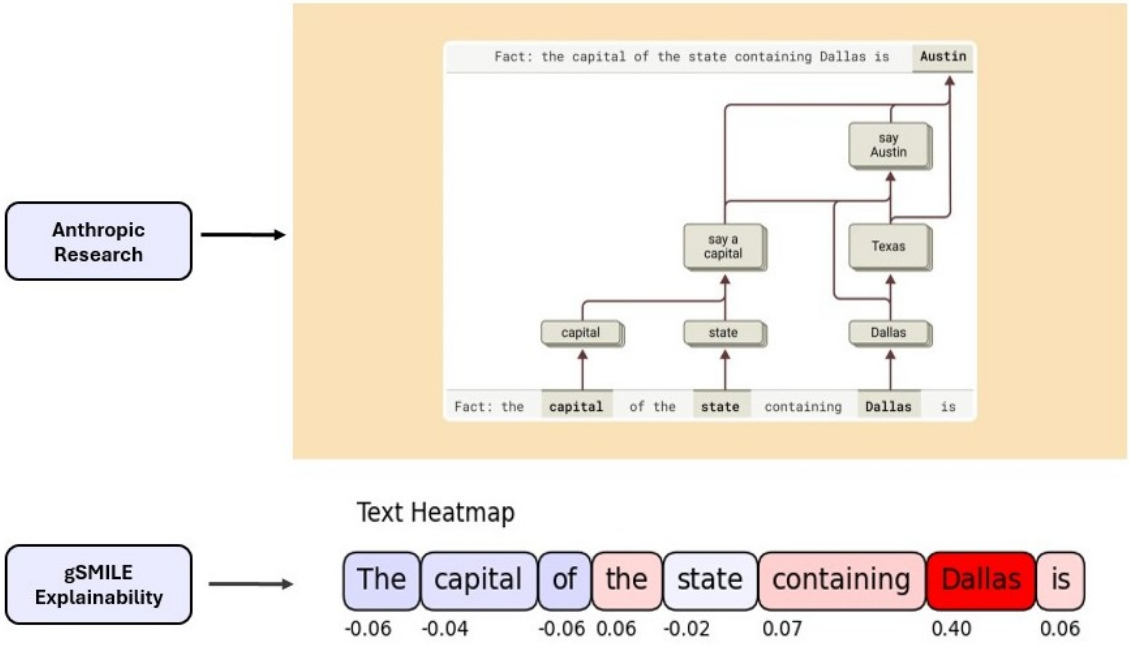


图 4.4 Anthropic 的归因图与 gSMILE 框架的归因可视化对比。上半部分展示了 Anthropic 方法识别出的推理路径，下半部分则展示了 gSMILE 针对同一提示词生成的词元级重要性热图。 [5]

除了词元级重要性可视化之外，所提出的框架还提供了结构化推理路径，展示模型如何在逻辑上将输入元素联系起来并生成响应。图 4.4 展示了一个示例：gS MILE 方法不仅为各个词元分配归因分数，还生成了推理流程图。这种综合可视化有助于全面理解局部词级贡献以及模型决策的逻辑结构。

这类全面的可视化解释与领先人工智能研究机构（包括 Anthropic 和 OpenAI）采用的先进技术相一致，同时在不同任务和模型之间提供了更高的灵活性。如图 4.4 所示，gSMILE 的可解释性与领先业界方法所达到的可解释性相当，甚至更胜一筹 [5]。

使用 gSMILE 对 MMLU 进行词级归因分析

为了使用 gSMILE 框架评估 ChatGPT 的可解释性，我们使用了 MMLU（大规模多任务语言理解）数据集中的示例 [207]。该

基准测试包含四选一的选择题，涵盖科学、历史、法律和人文学科等多个学术及专业领域。本研究使用了 Kaggle [208] 上提供的预格式化数据集版本，以便构建提示词并开展实验。

图 4.5 展示了一组归因热图，用于可视化输入提示中哪些部分对模型决策的影响最大。更深的红色色调表示更高的影响力，而中性或不相关的词元则以较浅的颜色着色。这些可视化展示了 gSMILE 通过追踪多项选择题语境中最具影响力的词来揭示模型内部推理的能力。

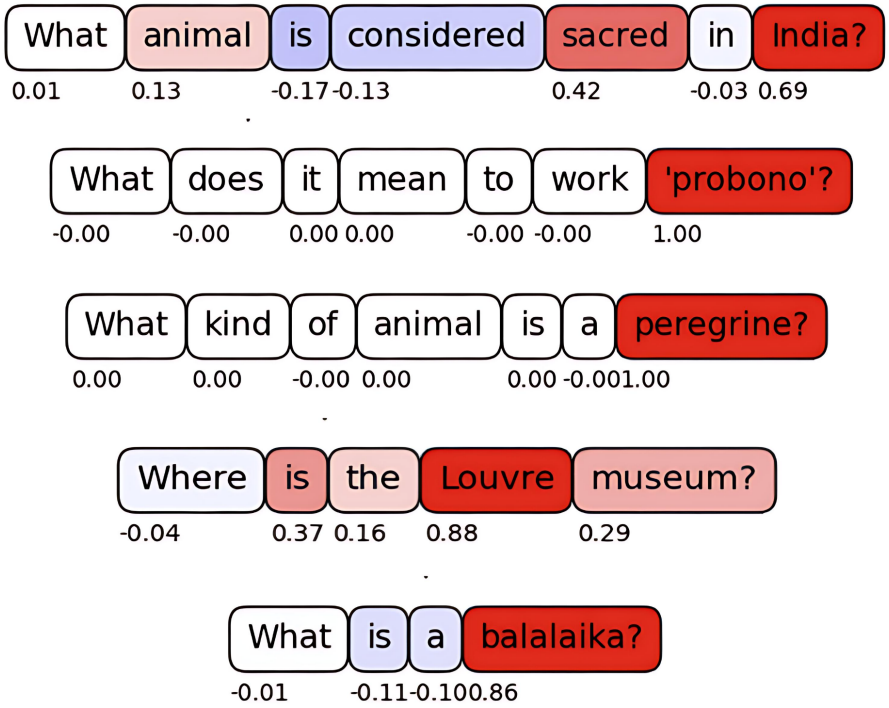


图 4.5 使用 ChatGPT（GPT-4）在 MMLU 数据集的选定样本上由 gSMILE 生成的归因热图。高亮显示的词代表对模型输出影响力最高的词。

4.3.2 定量结果

本节采用多种评估指标，包括准确性、稳定性、一致性、保真度和计算复杂度，评估我们的可解释性方法在各种文本生成模型和场景中的表现。

ATT 准确率

为了评估基于指令的文本生成模型的性能，我们测试了十种不同的提示词及其变体，并进行了 64 次扰动。我们生成热力图，以展示 Ope-nAI GPT、LLAMA 和 Claude-AI 等不同模型对提示词中各关键词赋予的权重。我们定义了真实值：为提示词中的关键关键词赋予权重 1，为相关性较低的词或辅助词赋予权重 0。随后，将这些真实值与提取出的热力图进行比较，以评估各模型的可解释性和准确性。

为量化准确性，我们采用了多项指标，包括注意力准确率（ATT ACC）、注意力 F1 分数（ATT F1）以及注意力受试者工作特征曲线下面积（ATT AUROC）。结果取自十个不同提示词的平均值，每个模型的性能总结于表 4.1，展示了它们在这些指标上的表现。

在评估的模型中，**Claude-AI 展现出最强的性能**，在全部三项指标中均取得最高分：ATT ACC 为**0.82** ，ATT F1 为**0.67** ，ATT AUROC 为**0.88** 。这些结果表明，Claude-AI 能够在不同提示词下始终如一地聚焦于重要输入词元。相比之下，**LLaMA**的准确率位居第二，但 F1 分数明显较低（**0.40**），这表明它在识别关键词元方面可能不够一致。**OpenAI GPT** 的结果较为均衡，但各项指标均略逊于 Claude-AI，表明其注意力忠实度仍有提升空间。总之，与其他模型相比，Claude-AI 提供的注意力更可靠，也更聚焦于语义。

表 4.1 不同模型的性能指标（ATT ACC、ATT F1 和 ATT AUROC），在两种条件下进行评估：不同提示词和不同图像。

模型名称	ATT ACC	ATT F1	注意力受试者工作特征曲线下面积
OpenAI-GPT	0.70	0.59	0.84
LLaMA	0.76	0.40	0.84
Claude-AI	**0.82**	**0.67**	**0.88**

ATT 稳定性

为了量化稳定性，我们使用杰卡德指数，这是一种通过比较集合重叠来衡量集合之间相似度的指标。对于 OpenAI GPT、LLAMA 和 Claude-AI 等文本生成模型，我们测试了十种不同的提示词，并提取了分配给每个词的系数和权重。为了评估稳定性，我们仅在文本末尾添加了该词元，

再次提取系数，并计算扰动前后系数之间的杰卡德指数。各模型在全部十个提示词上的平均杰卡德指数报告于表 4.2。分数越高，表明对微小输入变化的响应稳定性越强。

表 4.2 基于十个提示词的平均杰卡德指数得出的各模型稳定性得分

模型名称	杰卡德指数
OpenAI-GPT	**0.62**
LLaMA	0.45
Claude-AI	0.44

ATT一致性

为计算一致性，我们针对每个提示词运行相同的代码，在 10 次迭代中进行 64 次扰动，确保文本生成模型（如 OpenAI GPT、LLaMA 和 Claude-AI）的输出在这些重复运行中保持一致，如表 4.3 所示。我们计算了每个词系数的方差和标准差指标。这一步对于确认模型的预测并非随机的，也不严重依赖初始化因素，而是具有确定性且稳健至关重要。

在所有模型中，OpenAI-GPT 表现出最高的一致性，方差为 0.0000，标准差为 0.0046。这表明其预测结果在重复运行中最为稳定，证实了其在生成可解释输出方面的鲁棒性和可靠性。

表 4.3 不同模型针对提示词“生命的意义是什么？”的一致性指标

模型名称	方差	标准差
OpenAI-GPT	**0.0000**	**0.0046**
LLaMA	0.0048	0.0663
Claude-AI	0.0011	0.0331

不同基于扩散的指令图像编辑模型（IED）的 ATT 忠实度

下文展示了针对提示词“生命的意义是什么？”应用于各种文本生成模型时的保真度计算结果。我们使用 32 个扰动比较了多个模型，包括 OpenAI GPT、LLAMA 和 Claude-AI。比较采用

加权线性回归作为代理模型，并使用瓦瑟斯坦距离计算均方误差、(R^2_ω)、平均绝对误差、平均 L_1，和平均 L_2 损失等指标之间的距离。结果见表 4.4。

表 4.4 不同模型对提示词“什么是生命的意义？”的保真度指标

模型名称	WMSE	R^2_ω	加权平均绝对误差	平均 L1	平均 L2	$R^2_{\hat{\omega}}$
OpenAI-GPT	0.0388	**0.7104**	0.1731	0.2035	**0.0609**	**0.6409**
LLaMA	0.0368	0.7068	0.1617	0.2387	0.0805	0.6364
Claude-AI	**0.0265**	0.6967	**0.1251**	**0.1981**	0.0708	0.6240

从结果中可以看出，Claude-AI 的 WMSE、WMAE 和平均 L1 最低，表明其在降低误差方面表现出色。另一方面，OpenAI- GPT 在R^2_ω、平均 L2 和调整后的$R^2_{\hat{\omega}}$方面表现最佳，显示出代理模型更强的对齐能力和泛化能力。LLaMA 在大多数指标中排名居中，但没有在任何一项指标中领先。这表明，不同模型在不同保真度维度上各有所长，也凸显了在解读黑箱模型行为时进行多指标评估的重要性。

不同文本扰动数量下的 ATT 保真度

表4.5、4.6和4.7分别展示了在 OpenAI GPT、LLaMA 和 Claude-AI 模型中，扰动数量不同时的保真度计算结果。这些结果采用加权线性回归作为代理模型，并使用 Wasserstein 距离计算距离得出。评估指标包括加权均方误差（WMSE）、加权平均绝对误差（WMAE）、决定系数（R^2_ω 和$R^2_{\hat{\omega}}$），以及平均损失指标，如平均 L1 和平均 L2。

表 4.5 OpenAI GPT 中不同扰动数量对应的性能指标

#扰动	WMSE	R^2_ω	WMAE	平均 L1	平均 L2	$R^2_{\hat{\omega}}$
32	0.0388	**0.7104**	0.1731	0.2035	0.0609	**0.6409**
64	0.0617	0.3406	0.1695	0.1756	0.0526	0.2712
128	0.0329	0.4468	0.1385	0.1519	0.0378	0.4193
256	**0.0116**	0.4276	**0.0675**	**0.0872**	**0.0163**	0.4138

表 4.6 LLaMA 中不同扰动数量下的性能指标

扰动数	加权均方误差	$\mathbf{R}^2_{\omega}$	加权平均绝对误差	平均 L1	平均 L2	$\mathbf{R}^2_{\hat{\omega}}$
32	0.0368	**0.7068**	0.1617	0.2387	0.0805	**0.6364**
64	0.0240	0.4844	0.1284	0.1433	0.0382	0.4301
128	0.0085	0.2430	0.0638	**0.0612**	**0.0096**	0.2055
256	**0.0058**	0.2387	**0.0605**	0.0667	0.0118	0.2203

表 4.7 Claude-AI 中不同扰动数量下的性能指标

扰动数	加权均方误差	$\mathbf{R}^2_{\omega}$	加权平均绝对误差	平均 L1	平均 L2	$\mathbf{R}^2_{\hat{\omega}}$
32	0.0368	**0.7209**	0.1495	0.2262	0.0733	**0.6539**
64	**0.0115**	0.3708	**0.0564**	0.1036	0.0370	0.3046
128	0.0089	0.1506	0.0535	0.0920	0.0192	0.1085
256	0.0191	0.2670	0.0902	**0.0873**	**0.0163**	0.2494

对于 OpenAI GPT 和 LLaMA，增加扰动数量通常会改善大多数指标的保真度，其中 256 个扰动取得了最佳的 WMSE、WMAE 和损失值。有趣的是，对于 Claude-AI，64 个扰动时 WMSE 和 WMAE 最低，而 256 个扰动时损失（平均 L2 和平均 L1）最佳。不过，三个模型的最高R^2 分数都出现在 32 个扰动时，这表明超过一定数量后，更多扰动虽然能改善局部拟合，却可能削弱可解释性。因此，扰动数量与代理模型的可靠性之间存在权衡，具体取决于所关注的指标。

不同距离度量和代理模型的 ATT 忠实度

对于文本生成模型（如 OpenAI GPT）的保真度，也通过文本到文本比较，采用多种距离度量进行计算。所用的代理模型为加权线性回归（WLR）和贝叶斯岭回归（BayLIME），如表 4.8 所示。我们探讨了距离度量的不同组合，包括余弦相似度和瓦瑟斯坦距离（WD）。保真度采用 WMSE、WMAE、平均L_1, 平均L_2, 以及决定系数（R^2_{ω}，$R^2_{\hat{\omega}}$）进行衡量。

表 4.8 使用 30 次扰动时不同距离度量的保真度结果

WLR		**保真度指标**					
T vsT	**T vs T**	**WMSE**	$\mathbf{R}^2_\omega$	**加权平均绝对误差**	**平均 L1**	**平均 L2**	$\mathbf{R}^2_{\hat{\omega}}$
余弦	余弦	0.0172	0.3151	0.0659	0.1277	0.0412	0.1508
余弦	WD	0.0216	0.4197	0.0899	0.1332	0.0385	0.2805
WD	WD	0.0388	**0.7104**	0.1731	0.2035	0.0609	**0.6409**
WD	余弦	0.0048	0.4026	0.0329	0.0871	0.0296	0.2593
WD+C	WD+C	0.0349	0.5349	0.1589	0.3050	0.1468	0.4233
BayLIME							
余弦	余弦	0.0200	0.2285	0.0499	0.1067	0.0464	0.0434
余弦	WD	0.0241	0.3572	0.0736	0.1281	0.0430	0.2029
WD	WD	0.0118	**0.5837**	0.0629	0.1170	0.0361	**0.4838**
WD	余弦	0.0058	0.5273	0.0459	0.0978	0.0304	0.4139
WD+C	WD+C	0.0048	0.4227	0.0284	0.0690	0.0213	0.2841

我们提出的方法：gSMILE。根据表 5.9 的结果，在输入相似度和输出偏移两方面均使用瓦瑟斯坦距离（WD）的组合，始终能获得最高的加权决定系数（R^2_ω 和$R^2_{\hat{\omega}}$）。具体而言，在欠加权线性回归中，WD–WD 配置的$R^2_\omega = \mathbf{0.7104}$ 和$R^2_{\hat{\omega}} = \mathbf{0.6409}$，显著优于其他配置。

受这些发现的启发，我们提出了名为 **gSMILE** （广义SMILE）的方法，该方法使用瓦瑟斯坦距离比较输入和输出。此选择提升了代理模型的保真度和稳定性，使其能够捕捉模型行为中的局部偏移。

在后续章节中，我们使用 gSMILE 作为评估大语言模型（LLM）可解释性的基础，包括对保真度、一致性和稳定性的评估，以确保所提出的方法既有理论依据，又在实证上足够可靠。

计算复杂度

我们使用三种基于指令的文本生成模型：OpenAI GPT、LLaMA 和 Claude-AI，评估所提出方法 **gSMILE** 的计算效率。尽管这些模型在规模和架构上各不相同，但我们的目标并非比较模型本身，而是评估在相同硬件和运行条件下应用于各模型的不同可解释性方法

的性能。所有评估均针对每个输入提示词进行 60 次扰动。

我们根据执行时间比较三种可解释性方法——LIME、gSMILE 和 Bay-LIME。表 4.9 所示，gSMILE 在所有模型框架中的运行时间始终最长。这主要是因为它依赖瓦瑟斯坦距离来计算输出分布之间的相似度，而这需要求解时间复杂度为 $O(Nd^3)$ 的高维优化问题，其中 N 是扰动的数量，d 是嵌入维度 [209]。相比之下，LIME 和 Bay-LIME 使用复杂度较低的余弦相似度 $O(Nd)$，因此速度更快，但精确度可能较低。

尽管计算成本较高，gSMILE 却能提供显著更好的保真度、稳定性和对抗性扰动鲁棒性——这些都是生成可信解释的关键特性。表 5.9 和表 4.2 中的结果支持了这些优势，并且与先前关于 SMILE 的研究结果一致 [4], 该研究表明，SMILE 对对抗性输入变化具有更强的鲁棒性。

这些发现表明，gSMILE 的计算开销因其所提供的解释质量而物有所值，因此非常适用于可解释性、一致性和抗操纵能力至关重要的应用。

表 4.9 各框架和可解释性方法在 60 次扰动下的执行时间（秒）

模型名称	LIME	gSMILE	Bay-LIME
OpenAI-GPT	156.82	**170.70**	161.43
LLaMA	113.00	**118.99**	116.40
Claude-AI	355.09	**372.95**	349.88

不同句子结构的词元级注意力

为探究句子结构如何影响大语言模型中的词元级注意力，我们分析了十个语义相近但句法各异的提示词中，分配给词元“意义”和“生命”的注意力分数。这些提示词的范围从疑问句（例如，“生命的意义是什么？”）、命令句（例如，“请给我生命的意义”）到陈述性表述（例如，“你必须解释生命的意义”）。

对于每个句子，我们直接从大语言模型的最后一层提取注意力值。我们重点关注分配给词元“meaning”和“life”的注意力。结果以箱线图呈现（见图 4.6），其中每个点代表一个独特的

句子。将鼠标悬停即可查看完整句式提示，从而对句子结构和注意力分布进行定性比较。

分析揭示了这些词元的权重会随句法形式而显著变化。例如，在祈使句结构中，注意力往往集中于“life”；而在疑问式或抽象提示中，“meaning”通常会获得相当或更高的注意力。这表明，即使语义内容保持一致，句子表述也能微妙地引导模型的内部关注焦点。

这些发现凸显了大语言模型在塑造词元级可解释性时对表层结构的敏感性，并为理解模型行为和提示词工程策略提供了见解。



图 4.6 箱线图展示了十个句子结构各异的提示词中，“meaning”和“life”这两个词元的注意力分数。每个点代表特定句例中分配给相应词元的注意力分数。

4.4 结论

本研究介绍了 gSMILE，这是一种生成式模型无关框架，旨在通过词元级归因和可视化推理辅助工具解释大语言模型的行为。大多数现有可解释性方法会突出显示模型输出中的关键词元，而 [3, 106],gSMILE 则将注意力转向输入空间。通过将扰动分析与

基于 Wasserstein 距离的度量和加权线性替代模型相结合，gSMILE 提供了一种原则严谨的局部可解释性方法，无需访问模型内部参数。

我们对三种指令微调大语言模型的评估显示，它们的性能存在明显差异：

- **Claude 2.1** 的归因保真度最高（注意力受试者工作特征曲线下面积 = 0.88），误差指标最低，这表明其注意力更加聚焦于语义。
- **GPT-3.5-turbo-instruct**表现出卓越的一致性（多次运行的方差和标准差最低），因此其解释输出具有很高的可预测性。
- **LLaMA3.1**表现不俗，但其词元归因的变异性较大，表明其对提示措辞较为敏感。

总体而言，生成的热力图和归因分数直观地展示了模型行为，帮助用户更好地控制和优化提示设计。

gSMILE 通过多种可解释性指标（包括保真度、稳定性和一致性）进行了评估，并与现有方法（LIME、Bay-LIME）进行了比较。我们系统地评估了各项技术捕捉大语言模型中词元级影响的能力。结果表明，gSMILE 在近似黑箱行为方面始终实现了更高的保真度，同时在轻微的提示扰动下生成了更稳定、可重复的归因。与 LIME 和 Bay-LIME 相比，gSMILE 展现出更强的鲁棒性，并且与人类可解释的模式更加对齐，凸显了它在可靠可解释性至关重要的实际场景中的适用性。

这些发现从两个方面证实了 gSMILE 的贡献。首先，它表明，可解释性指标能够超越标准基准准确性，有效区分不同的大语言模型，并为了解其运行可靠性提供新的见解。其次，这表明纯黑箱的事后分析方法能够生成稳定性和保真度可与借助特权访问权限的可解释性工具相媲美的解释。

与此同时，已发现的局限性（提示敏感性、幻觉效应、计算开销和部署限制）表明，实现完全实用的可解释性仍有待探索。解决这些问题将是未来研究的重点。展望未来，我们认为有三个值得关注的方向：

1. **自适应扰动策略：**未来的工作可以不再局限于随机采样，而是采用定向扰动，仅聚焦于重要词元或短语。这

可以在保持保真度的同时降低计算开销。受 IBM 的 CELL 框架等方法启发，扰动可以通过改变特定的词来实现，而非仅仅删除或保留它们。这样便能高效地突出较长句子中的指定词。

2. **因果验证方法：** 为减轻由幻觉引发或虚假的归因，可以整合因果分析（例如，在固定其他上下文的同时扰动词元），以确认某项归因是否确实驱动了模型的行为。这将增强其在高风险应用中的鲁棒性。

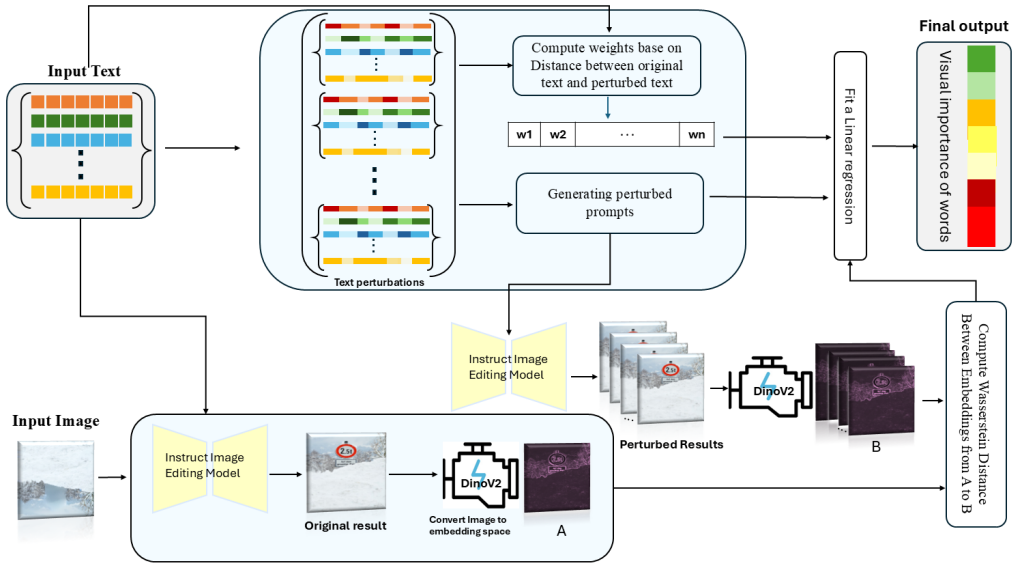
3. **扩展至多模态和多语言模型：** gSMILE 目前专注于英文文本。将该框架扩展至多语言大语言模型和多模态系统（文本—图像或文本—语音），有助于检验其稳定性和保真度能否在不同输入空间中保持，从而提升其在真实世界部署中的泛化能力。

通过将严格评估与可解释的可视化相结合，gSMILE 无需特权访问模型内部机制，就能提高大语言模型的透明度，从而为研究人员和从业者提供实用工具，助力构建更值得信赖的人工智能系统。

第5章

基于指令的图像编辑可解释性

本章通过探讨如何将可解释性方法（尤其是 gSMILE 框架）应用于多模态系统（例如基于指令的图像编辑模型），来回答 **研究问题 3（RQ3）** 。本章介绍了用于解读这些模型如何响应自然语言提示词的技术。我们提出了 gSMILE 框架，该框架能够生成局部视觉解释，将文本指令与特定的图像修改关联起来。我们通过多项指标对该框架进行了广泛测试，评估其稳定性、准确性、保真度和一致性，结果表明，该框架在各种图像编辑场景中具有良好的鲁棒性。图5.1 展示了所提出工作流程的概览。



Fi图5.1 用于解释基于指令的图像编辑的 gSMILE 框架 ty.

5.1 问题定义

可解释性是复杂深度学习模型以及大型语言和基于文本的图像编辑模型面临的一项重大挑战。由于参数数量庞大且处理机制复杂，这些模型本质上如同黑箱，用户难以理解模型为何以及如何响应某些输入，或表现出特定行为 [13, 12]。可解释性的需求给这些模型的安全部署和控制带来了重大挑战，使其更难以得到有效管理，也增加了出现意外或不良行为的风险。

例如，为了提高大语言模型的可解释性，Anthropic团队应用稀疏自编码器从这些模型中提取语义和抽象特征。该技术使他们能够识别关键概念和内部行为，从而更透明地了解这些模型如何运作 [210]。

受该方法启发，并认识到基于文本的图像编辑模型中可解释性的重要性，我们提出了一种创新的模型无关方法，用于评估和增强这些模型的可解释性。我们利用SMILE（局部解释统计方法）作为基础工具，生成局部解释和可视化热力图。这些热力图展示了特定文本元素对图像生成过程的影响力，帮助用户更好地理解文本输入与视觉输出之间的关系 [4]。这种模型无关方法对于使指令图像编辑模型更加透明和可信至关重要。

此外，为进一步检验该模型的可解释性，我们引入了t-SNE散点图（如图5.2所示），以可视化经各种描述性句子修改后的图像嵌入，每个嵌入根据特定关键词进行颜色编码。在该图中，由包含关键词“晴朗”的句子生成的图像表示为橙色点，“下雨”为绿色点，“有雾”为黄色点，“下雪”为灰色点，“夜晚”为紫色点。由缺少目标关键词的扰动文本句子生成的其他点则表示为深蓝色点。

该可视化结果表明，每个关键词的嵌入会形成离散的簇，与不含特定关键词的扰动点明显分离。这种分离表明，代理模型（如加权线性回归）能够有效捕捉模型行为。此外，回归系数为衡量每个关键词的重要性提供了一种可解释且直观的方法，表明其如何影响模型输出。该方法通过清晰、定量地展示特定文本成分如何影响生成图像，增强了可解释性。

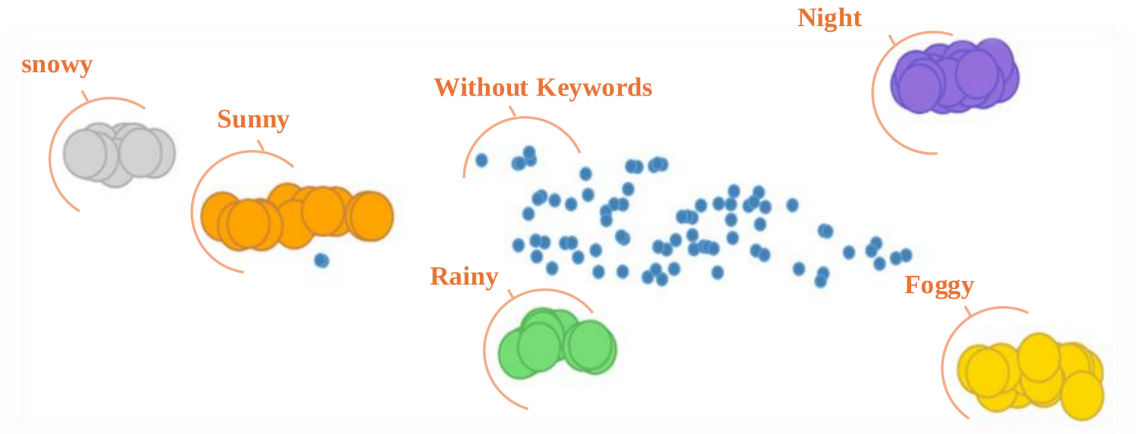


图 5.2 图像嵌入的二维t-SNE可视化，展示了不同提示关键词和扰动文本输入的聚类可分性。

5.2 基于指令的图像编辑方法概述

本节将介绍如何调整第3章中提出的 gSMILE 框架，以解释基于指令的图像编辑模型。其目标是确定文本提示中的哪些部分最显著地影响施加于输入图像的视觉变换。

该过程从固定的输入图像及其对应的指令（文本提示）开始。首先将提示词拆分为单个词，然后逐次选择性地删除或更改其中一个词，以生成扰动提示词。随后，使用每条扰动指令引导图像编辑模型，从而生成修改后的输出图像。使用原始提示词生成的图像则作为基线参考。

为量化每种提示变体的影响，我们计算每张扰动图像与基线图像之间的Wasserstein距离 [187] 。这些距离是在语义嵌入空间中计算的，所用模型为DINOv2 [184], ，这是一种自监督视觉Transformer，以生成稳健且通用的图像表示而闻名。这样，我们便能够根据高层视觉语义而非低层像素差异来比较输出。

计算得到的距离通过高斯核转换为权重，表示每个扰动提示与原始提示词之间的相似度。这些权重

随后应用于局部加权线性回归模型 [3, 90], ，其中每个输入都是一个二进制向量，用于指示扰动提示中包含了哪些词。

由此得到的回归系数揭示了每个词对模型执行的视觉变换所作出的贡献。这些数值以热力图的形式叠加在提示词上，突出显示对生成图像影响最大的词。

通过将 SMILE 的统计与模型无关方法扩展至视觉语言场景，这种调整使多模态任务中的解释更具可解释性。当需要将风格或语义图像编辑追溯到具体的文本指令时，这种方法尤其有价值。图 3.3 展示了应用于基于指令的图像编辑的 SMILE 流程。

5.3 实验结果

本节评估所提出的方法在提升基于指令的图像编辑可解释性方面的能力。我们分析文本提示的结构如何影响可解释性、输入图像如何影响解释的清晰度，以及所提出的解决方案在不同场景中的表现。实验使用了多个数据集和三种最先进的扩散模型：Instruct-Pix2Pix [211], Img2Img-Turbo [76], 和 Diffusers-Inpaint [27], ，以确保评估全面。

5.3.1 定性结果

定性评估展示了所提出框架的可解释性，采用基于场景的测试和可视化。借助运行设计域（ODD） [212], ，我们生成了多样化的测试场景，涵盖环境因素，例如天气变化（如雨、雾、雪），以及非环境变量（如物体属性、以人为中心的因素、时间条件）。这些评估中使用的具体变换场景、关键词和受控变量汇总于表5.1和表5.2，涵盖各种环境条件以及以人为中心的修改。

condition	specific condition		sentences	keyword	our control
weather	rainy	short sentence	Transform the drizzle into a heavy downpour.	Transform, drizzle, heavy downpour	heavy downpour
		length sentence	Transform the image to a rainy scene with dark clouds, steady raindrops, and puddles forming on the ground.	Rainy scene, dark clouds, steady raindrops, puddles	Rainy scene
			Change the sunny weather to rainy, with dark clouds filling the sky and raindrops falling steadily	Change, sunny weather, rainy, dark clouds, filling the sky, raindrops, falling steadily	rainy, dark clouds, raindrops falling steadily
			Turn image to rainy and add a rainbow in the sky with a nice sense.	Turn, image, rainy, add, rainbow, sky, nice sense	rainy , rainbow
			Convert the clear sky to a stormy one, with flashes of lightning and heavy rain pouring down.	Convert, clear sky, stormy, flashes of lightning, heavy rain, pouring down	stormy, flashes of lightning, heavy rain
			Change the sunny weather to haily, with dark clouds filling the sky and hailstones falling steadily.	Change, sunny weather, haily, dark clouds, filling the sky, hailstones, falling steadily	haily, dark clouds, hailstes falling steadily
	snowy	short sentence	Transform the weather to snowy.	Transform, weather, snowy	snowy
			Change the calm snow to heavy.	Change, calm snow, heavy	heavy snow
		length sentence	Change the bright day to a snowy scene, with snowflakes gently falling and the ground covered in a white blanket.	Change, bright day, snowy scene, snowflakes gently falling, ground covered, white blanket	snowy scene, snowflakes, ground covered in a white blanket
	fogy	short sentence	Change the foggy weather to a clear day.	Change, foggy weather, clear day	clear day
		length sentence	Transform the night image to depict a foggy evening where visibility of the road is poor.	Transform, night image, foggy evening, visibility of the road, poor	foggy evening , poor vible
			Switch the rain-drenched scene to one filled with fog, where the rain stops and a thick mist takes over, reducing visibility	Switch, rain-drenched scene, filled with fog, rain stops, thick mist, reducing visibility	filled with fog, thick mt, reducing vibility
	sunny	short sentence	Change the stormy weather to a peaceful sunset.	Change, stormy weather, peaceful sunset	peaceful sunset
			Transform the sunny day to an overcast day.	Transform, sunny day, overcast day	overcast day
		length sentence	Change the snowfall to a bright, sunny day, with the snow melting away and the sun shining warmly.	Change, snowfall, bright, sunny day, snow melting away, sun shining warmly	bright, sunny day, snow melting away, sun shining warmly
			Transform the image to depict a sunny day with intense sunlight that hurts the eyes	Transform, image, depict, sunny day, intense sunlight, hurts the eyes	sunny day, intense sunlight
			Transform the overcast sky into a sunny one, with the clouds dispersing and the sun shining brightly.	Transform, overcast sky, sunny one, clouds dispersing, sun shining brightly	sunny sky, clouds dpersing, sun shining brightly
	windy	short sentence	Change the calm trees to windy.	Change, calm trees, windy	windy
		length sentence	Convert the windy day to a calm one, with no breeze and the trees standing still.	Convert, windy day, calm one, no breeze, trees standing still	calm weather, no breeze, trees st,ing still
			Change the scene to depict a windy day with items, swirling in the air.	Change, scene, depict, windy day, items, swirling in the air	windy day, items swirling in the air
	cloudy	short sentence	Change the cloudy sky to a clear blue sky.	Change, cloudy sky, clear blue sky	clear blue sky
			Transform the image to depict a cloudy, dark evening	Transform, image, depict, cloudy, dark evening	cloudy, dark evening
		length sentence	heavy clouds fill the sky with thunder rumbling and bright lightning bolts striking down, creating an intense and stormy atmosphere	Heavy clouds, fill the sky, thunder rumbling, bright lightning bolts, striking down, intense, stormy atmosphere	intense , stormy atmosphere

表 5.1 天气状况转换及对应的关键词。

condition	specific condition	sentences	keyword	our control
person	gender	Change the old woman to a young man in the image who is wearing sunglasses.	Change, old woman, young man, wearing sunglasses	young man , wearing sunglasses
		change the male in the image with a high-fashion woman	Change, male, high-fashion woman	high-fashi woman
	skin color	Modify the person's complexion to a deep black hue	Modify, person's complexion, deep black hue	deep black hue
		Turn the white-skinned person to a black-skinned person.	Turn, white-skinned person, black-skinned person	black-skinned pers
	age	Change the person's age to look considerably older.	Change, person's age, considerably older	csiderably older
		Change the elderly man into a young boy while retaining his distinctive facial characteristics	Change, elderly man, young boy, retaining, distinctive facial characteristics	young boy, dtinctive facial charactertics
tools	replace	Remove the man's hat in this foggy image.	Remove, man's hat, foggy image	no hat
		Replace the traffic light with a simple light.	Replace, traffic light, simple light	simple light
		Add a car crash to the street.	Add, car crash, street	car crash
		Replace the black car with a white one.	Replace, black car, white car	white car
		Take off the man's glasses and put them on the lady's eyes.	Take off, man's glasses, put them on, lady's eyes	glasses the lady
	change color	convert the red bus to yellow one	Convert, red bus, yellow one	yellow
time	day	Convert a nighttime scene into a bright, sunny daytime setting with clear blue skies	Convert, nighttime scene, bright, sunny daytime setting, clear blue skies	bright, sunny daytime, clear blue skies
		Alter the nighttime environment into a lively day, with the darkness lifting and the sunlight bringing the world to life.	Alter, nighttime environment, lively day, darkness lifting, sunlight, bringing the world to life	lively day, sunlight
	night	convert it to night with some noise of lights.	Convert, night, noise of lights	night, noe of lights
		Turn this beautiful image to night and turn on the lights.	Turn, beautiful image, night, turn on the lights	night , lights
		Turn it to night with snowy weather.	Night, snowy weather	Night, snow

表 5.2 基于时间和人类属性的视觉和描述性变化转换指南。

我们通过将词语贡献映射到视觉变化，分析了输入提示可解释性，突出了诸如“下雪”等关键词在编辑过程中的权重。Us-

利用 t-SNE 散点图，我们定性考察了由扰动提示词生成的图像的聚类情况，揭示了词级变化对生成输出的影响（图 5.4）。

输入提示可解释性

所提出的基于场景的测试框架利用系统的运行设计域（ODD），生成各种条件下的场景，以全面测试自动驾驶系统（ADS） [213, 212]。

正如近期高风险人工智能框架所强调的，ODD 概念对于理解并确保 ADS 的运行边界和局限性至关重要 [214]。正如关于高风险人工智能系统的更广泛讨论所指出的，ODD 提供了一种结构化方法来分类环境和情境变量，有助于在各种条件下持续评估和验证系统性能 [215]。

表 5.1 和表 5.2 基于运行设计域（ODD）框架制定，用于系统地模拟和评估影响自动驾驶系统（ADS）性能的各种环境和情境条件。

运行设计域是高风险人工智能框架的核心，它定义了自动驾驶系统的运行边界和局限性，从而能够有条理地创建各种测试场景 [213]。表 5.1 根据多种指令生成，重点关注环境条件，并按雨、雾和雪等天气状态对场景进行分类，以确保在各种环境因素下进行全面评估 [212]。

表 5.2 结合了多种指令，以涵盖非环境变量，包括人物相关属性（如年龄、性别和肤色）、物体变化（如车辆颜色和交通信号灯配置）以及时间因素（如昼夜场景），确保在各种以人为中心和由物体驱动的条件下进行全面测试 [215]。

我们使用多种图像编辑模型实施了可解释性技术，包括 Instruct-Pix2Pix [211], Img2Img-Turbo [76] 和 Diffusers-Inpaint [27]。我们以表格中的分析为基础，并纳入热力图，以展示特定关键词如何影响图像修改 [213]。这一结构化方法符合欧盟人工智能法案等高风险人工智能框架的要求，这些框架强调在系统生命周期内确保可信赖性、安全性和法规合规 [216]。

此外，该方法还增强了自动驾驶系统（ADS）评估过程的鲁棒性、可靠性和透明度。我们的设计旨在通过基于指令的图像编辑模型，提升高风险人工智能系统的可信度。

我们汇总了全部十张图像热力图中的权重，并据此绘制箱线图，以展示提示词中每个词的影响力分布。如图 5.3 所示，箱线图揭示了每个词所产生的影响力范围及其集中程度。与提示词中的其他词相比，关键词“下雪”的影响力最大，表明它对模型创建雪景及进行相关调整的编辑过程有显著影响。

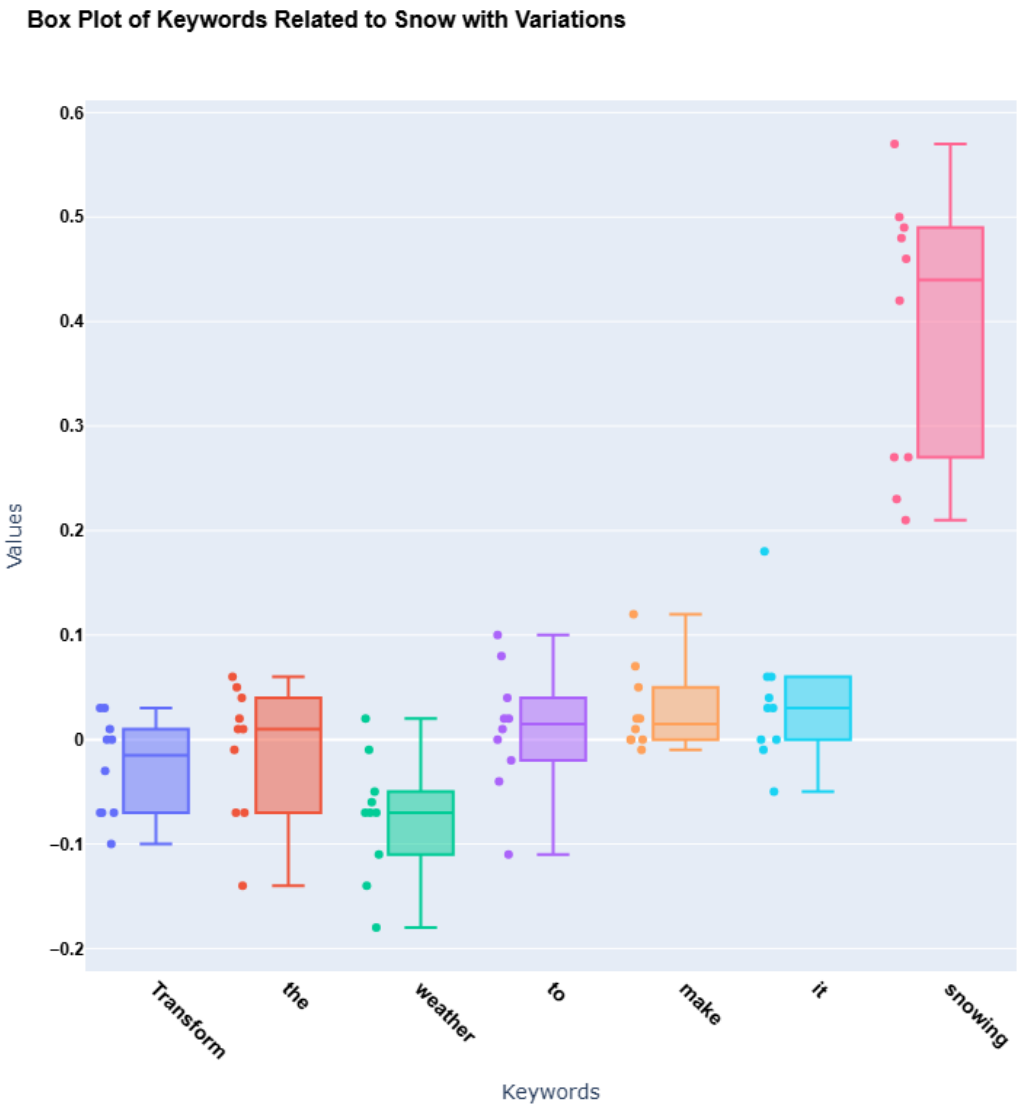


图5.3 箱线图展示了提示词“Transform the weather to make it snowing”中词权重的分布。

图像嵌入的 t-SNE 散点图

本节展示了使用扰动提示词得到的模型定性结果。为了更好地理解不同提示结构对生成图像的影响，我们使用二维图进行了可视化。该图展示了同一基础场景经过不同扰动后生成的图像之间的距离。图中的每个点代表一张图像，红点对应由原始提示生成的图像，蓝点则代表由原始提示的扰动版本生成的图像。

我们通过添加或删除关键词（例如“下雪”）系统地修改提示词。如图所示，包含关键词“下雪”的扰动会改变生成图像，将场景转变为雪景。相反，不包含该关键词的扰动所生成的图像则不会呈现雪景。

图5.4所示的二维图展示了图像之间基于其视觉特征的接近程度或距离。包含关键词“下雪”的扰动在更靠近原始图像的雪景区域形成簇。相比之下，其他扰动则会产生无雪场景。通过这一可视化结果，我们可以定性评估提示词中每个词的影响，以及模型对不同扰动的响应。

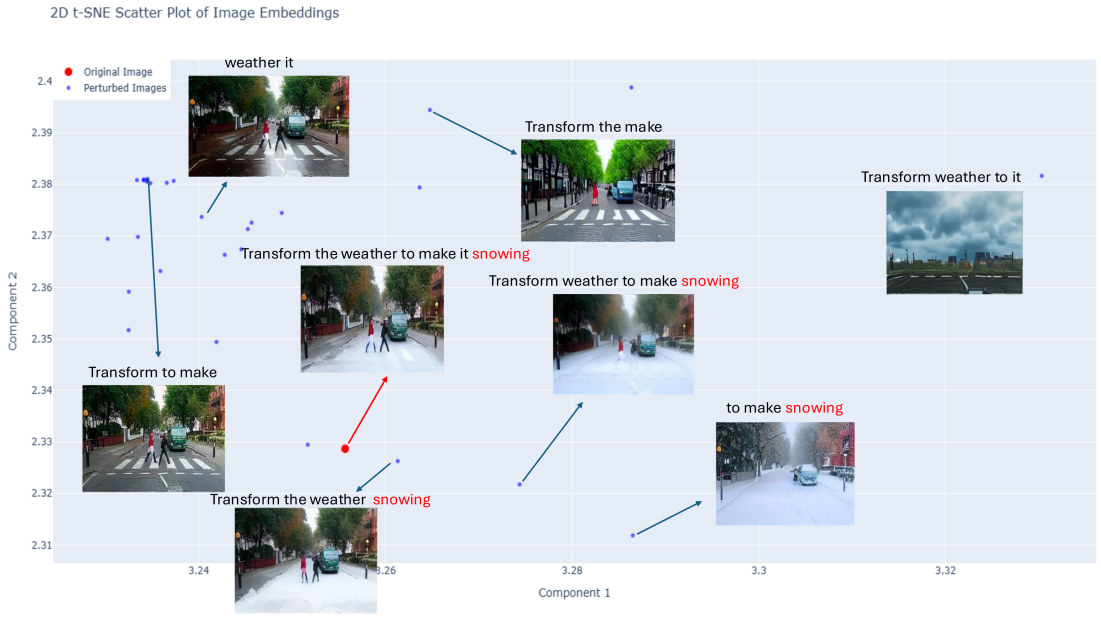


图 5.4 图像嵌入的二维可视化，展示了提示词中包含或省略“下雪”的效果。

5.3.2 定量结果

在本节中，我们采用准确性、稳定性、一致性、保真度和计算复杂度等评估指标，在各种图像编辑模型和场景中评估我们的可解释性方法。

ATT 准确率

为了评估基于指令的图像编辑模型的性能，我们在单张图像上测试了十个不同的提示词，并使用单条指令测试了十张不同的图像，每种情况均进行了 30 次扰动。我们提取了热力图，用于展示包括 Instruct-Pix2Pix、Img2Img-Turbo 和 Diffusers-inpaint 在内的不同模型中，各指令关键词的权重。我们定义了真实值：将指令中的关键词权重设为 1，将其他词的权重设为 0。随后，将这些真实值与提取的热力图进行比较，以评估模型的性能。

为了量化准确性，我们采用了多种指标，包括注意力准确率（ATT ACC）3.9、注意力 F1 分数（ATT F1）3.10，以及注意力受试者工作特征曲线下面积（ATT AUROC）3.11。结果取十个不同提示词和图像的平均值，各模型的结果汇总于表 5.3中，展示了它们基于这些指标的性能。

表5.3 各种模型在两种条件下的性能指标（ATT ACC、ATT F1和ATT AUROC）：不同提示词和不同图像。

模型名称	ATT ACC	ATT F1	ATT AUROC
I-Pix2Pix（不同提示词）	0.7845	0.6869	0.8943
I-Pix2Pix（不同图像）	**0.9571**	**0.9166**	**1.000**
Diffusers_I（不同提示词）	0.6621	0.1167	0.4417
Diffusers_I（不同图像）	0.7714	0.1000	0.4000
I2I-Turbo（不同提示词）	0.7944	0.6824	0.8400
I2I-Turbo（不同图像）	0.8286	0.6333	**1.000**

ATT 稳定性

为了量化稳定性，我们使用杰卡德指数这一指标，通过比较集合的重叠部分来衡量它们之间的相似度。我们使用十种不同的提示词对图像编辑模型（如 Instruct-Pix2Pix、Img2Img-Turbo 和 Diffusers-inpaint）进行图像修改，并提取每个词的系数和权重。然后，在每条提示词末尾添加“###”

作为输入文本，再次提取系数并计算杰卡德指数以进行比较。最后，我们计算每个模型的平均指标，并在表 5.4 中呈现结果。

表 5.4 十个提示词和 30 种扰动下不同模型的稳定性。该表展示了不同模型的杰卡德指数。

模型名称	杰卡德指数
I-Pix2Pix	0.85
Diffusers_I	0.85
I2I-Turbo	0.85

ATT一致性

为评估计算一致性，我们使用同一段代码，对每个扰动分别运行 1000 次迭代，共进行 30 个扰动，确保 Instruct-Pix2Pix、Img2Img-Turbo 和 Diffusers-inpaint 等图像编辑模型在这些重复运行中的输出保持一致，如表 5.5 所示。我们计算了每个词系数的方差和标准差指标。这一步对于确认模型的预测并非随机，也不过度依赖初始化因素，而是具有确定性且稳健，至关重要。方差和标准差值越低，表示一致性越好。

表 5.5 不同扩散模型的一致性指标。针对提示词“将天气转换为下雪天气”计算的方差和标准差。

模型名称	方差	标准差
I-Pix2Pix	0.0161	0.1271
Diffusers_I	0.0776	**0.0060**
I2I-Turbo	**0.0001**	0.0081

不同基于扩散的指令图像编辑模型（IED）的 ATT 忠实度

下方展示了将文本提示“将天气变成下雪”应用于各种 IED 模型时的保真度计算结果。我们使用 64 个扰动对多个模型进行比较，包括 Instruct-Pix2Pix、Img2Img-Turbo 和 Diffusers-inpaint。比较时采用加权线性回归作为代理模型，并使用 Wasserstein 距离计算 MSE、(R^2_ω)、MAE、平均L_1, 损失和平均L_2 损失等指标之间的距离。结果见表5.6。

如表5.6所示，已计算 Diffusers-inpaint 模型的保真度，但由于其性能较差，保真度相对较低。因此，对于涉及该模型的其他场景，未计算保真度。

表5.6 不同扩散模型的保真度指标。保真度指标是针对提示词“将天气改为下雪”计算得出的。

模型名称 WMSE		R^2_ω	WMAE 平均 L1 平均 L2			$R^2_{\hat{\omega}}$
I-Pix2Pix	0.0120	**0.7208**	0.0812	0.0790	0.0124	**0.6859**
Diffusers _I	0.0317	0.1255	0.1359	0.1529	0.0410	0.0161
I2I-Turbo	0.0193	0.6225	0.1078	0.1076	0.0187	0.5753

不同文本扰动数量下的 ATT 保真度

表 5.7、5.8 展示了使用 Instruct-Pix2Pix 和 Img2Img-Turbo 方法时，不同扰动数量下的保真度计算结果。比较时采用加权线性回归作为代理模型，并使用 Wasserstein 距离计算距离，所用指标包括加权均方误差（WMSE）、决定系数（R^2_ω）、加权平均绝对误差（WMAE）以及损失指标（平均L_1 和平均L_2）：

这些表格中的结果突显了所评估方法在不同扰动强度下保持保真度的鲁棒性和准确性。较低的（WMSE）和（WMAE），以及较高的（R^2_ω），表明生成输出与目标输出之间的对齐程度更好。

表 5.7 Instruct-Pix2Pix 中不同扰动数量对应的性能指标。评估指标包括 WMSE、R^2_ω、WMAE、平均-L1、平均-L2 和 $R^2_{\hat{\omega}}$。

#扰动数 WMSE		R^2_ω	WMAE 平均-L1 平均-L2			$R^2_{\hat{\omega}}$
32	0.0178	0.7770	0.1040	0.1004	0.0161	0.7119
64	0.0120	0.7208	0.0812	0.0790	0.0124	0.6859
128	0.0220	0.6816	0.1129	0.1090	0.0207	0.6631
256	0.0080	0.5110	0.0512	0.0509	0.0084	0.4972

表 5.8 Img2Img-Turbo 中不同扰动数量的性能指标。评估指标包括加权均方误差、R^2_ω、加权平均绝对误差、平均 L1、平均 L2 和 $R^2_{\hat{\omega}}$。

#扰动	加权均方误差	R^2_ω	加权平均绝对误差	平均 L1	平均 L2	$R^2_{\hat{\omega}}$
32	0.0421	0.5273	0.1635	0.1484	0.0369	0.3894
64	0.0193	0.6225	0.1078	0.1076	0.0187	0.5753
128	0.0316	0.3879	0.1412	0.1330	0.0293	0.3522
256	0.0248	0.6025	0.1275	0.1375	0.0305	0.5912

不同距离度量与代理模型的 ATT 忠实度

对于 Instruct-Pix2Pix 模型，通过文本与文本、图像与图像的比较，计算不同距离度量下的保真度。所用代理模型为加权线性回归和贝叶斯岭回归（BayLIME），如表 5.9 所示。我们探讨了距离度量的不同组合，例如余弦和 Wasserstein 距离（WD）。保真度通过均方误差、（R^2_ω）、平均绝对误差、平均L_1, 和平均L_2损失进行衡量。

表 5.9 在 30 次扰动下不同距离度量的保真度结果。该表展示了 T vs T 与 I vs I 距离度量不同组合的各项保真度指标。

WLR		保真度指标					
T 与 T，I 与 I		WMSE	R^2_ω	加权平均绝对误差	平均 L1	平均 L2	$R^2_{\hat{\omega}}$
余弦	余弦	0.0494	0.7626	0.1555	0.2865	0.2309	0.6167
余弦	WD	0.0022	0.8557	0.0215	0.2530	0.1148	0.7980
WD	WD	0.0128	**0.9495**	0.0674	0.2865	0.5243	**0.9292**
WD	余弦	**0.0022**	0.8558	**0.0216**	0.2530	0.1149	0.8558
WD+C	WD+C	0.0411	0.8744	0.1427	0.8484	1.5264	0.8241
BayLIME							
余弦	余弦	0.0524	0.7097	0.1862	0.2607	0.1363	0.5936
余弦	WD	0.2314	0.6064	0.3942	0.5239	0.4739	0.6064
WD	WD	0.0148	**0.9485**	0.0674	0.8065	0.8052	**0.9278**
WD	余弦	**0.0001**	0.8369	**0.0049**	**0.0501**	**0.0044**	0.7715
WD+C	WD+C	0.0421	0.8711	0.1139	0.8282	1.4246	0.8195

计算复杂度

我们将模型应用于三类基于指令的图像编辑框架：Instruct-pix2pix、Img2Img-Turbo 和 Diffusers-inpaint。我们在同一台计算机上测量不同可解释性方法（包括 LIME、gSMILE 和 Bay-LIME）在 60 次扰动下所消耗的时间，以确保硬件条件一致。

这些模型在计算需求方面存在明显差异，主要源于它们计算相似度和分配权重的方式不同。尤其是 gSMILE 使用 Wasserstein 距离计算文本距离，因此其计算开销远高于 LIME 和 Bay-LIME。如表 5.10 所示，gSMILE 在三个框架中的执行时间始终较长。其他框架也呈现出类似趋势，gSMILE 的计算开销十分明显。

Wasserstein 距离非常适合捕捉分布之间的细微差异，尤其是在文本嵌入中，但这也需要付出代价 [190]。对于高维文本数据，计算 Wasserstein-2 距离需要处理协方差矩阵，所需时间与 $O(d^3)$ 成正比，其中d 表示嵌入维度。考虑多个样本（N）时，计算复杂度会迅速增至 $O(Nd^3)$。相比之下，LIME 使用余弦相似度等更简单的相似度度量，其计算复杂度随规模增长更为高效，为 $O(Nd)$。

复杂度差异源于计算瓦瑟斯坦-2 距离时，需要对d维高斯分布的协方差矩阵执行运算 [209]。尽管这种更高的计算开销能够带来更稳健、更稳定的解释（如 gSMILE 所示），但也会显著增加执行时间。另一方面，LIME 依赖余弦相似度，计算开销较小、速度更快，因为它避免了此类矩阵运算。

gSMILE 需要更多计算资源，且本身更为复杂，但能提供高度详细且稳定的解释。因此，尽管计算成本更高，它尤其适用于可解释性和精确率至关重要的任务。

表 5.10 各框架和可解释性方法在 60 次扰动下的执行时间（秒）。

框架	LIME	gSMILE	Bay-LIME
I-Pix2Pix	1013.35	**1208.08**	1024.32
I2I-Turbo	983.17	**1222.09**	980.12
Diffuser-Inpaint	263.00	**276.83**	260.75

语言元素对可解释性的影响力

我们设计了一项实验，使用结构化提示集来了解不同措辞对模型解读和执行图像编辑指令能力的影响。我们的目标是评估措辞变化如何影响模型的性能，以及由此产生的编辑结果是否清晰。我们选取了十张各具特色的图像，并使用了诸如“将图像变换为阴云密布的昏暗傍晚”、“将天气变为下雪”和“将白皮肤的人变成黑皮肤的人”等提示词。对于提示词中的每个控制词，我们都通过动词、形容词、名词、“ing”形式，以及使用“make”和“must”进行改述来创建提示变体。每张图像都使用所有提示变体进行处理，并通过 LIME 分析输出，生成显示各个词重要性的热力图。

图 5.5 中与雪相关的提示词表明，名词和形容词形式带来的输出更一致，而动名词和指令短语则呈现出更高的变异性。这表明，模型对直接描述词的理解更为可预测，而变换指令则会使执行过程更灵活、也更模糊。

这项研究的结果为希望使用人工智能模型实现特定图像变换的用户提供了宝贵见解。通过谨慎选择提示词的措辞，用户可以更好地控制生成编辑结果的一致性和质量。未来的研究将扩充提示词集，纳入更复杂的指令，并评估上下文信息如何影响模型的决策过程。

图 5.6 展示了模型对一系列包含不同语言元素的提示词的响应，从而更全面地呈现其对措辞的敏感程度。测试“make it”这类指令短语时，响应范围更广，这表明面对复杂指令时，模型的理解会更加依赖上下文。

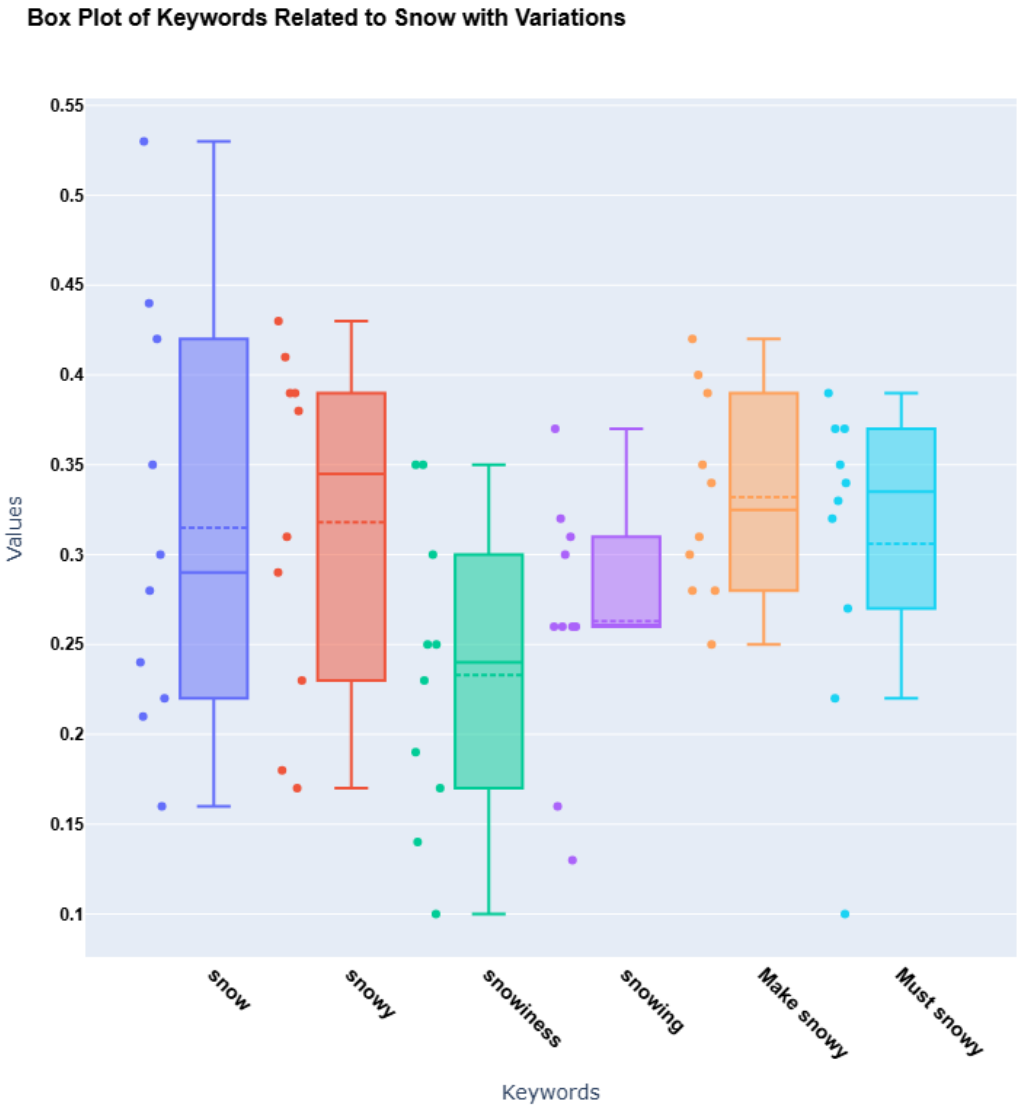


图5.5 语言变体对“snow”提示的响应的影响。不同形式（例如名词、形容词、动名词）会产生不同的输出分布。

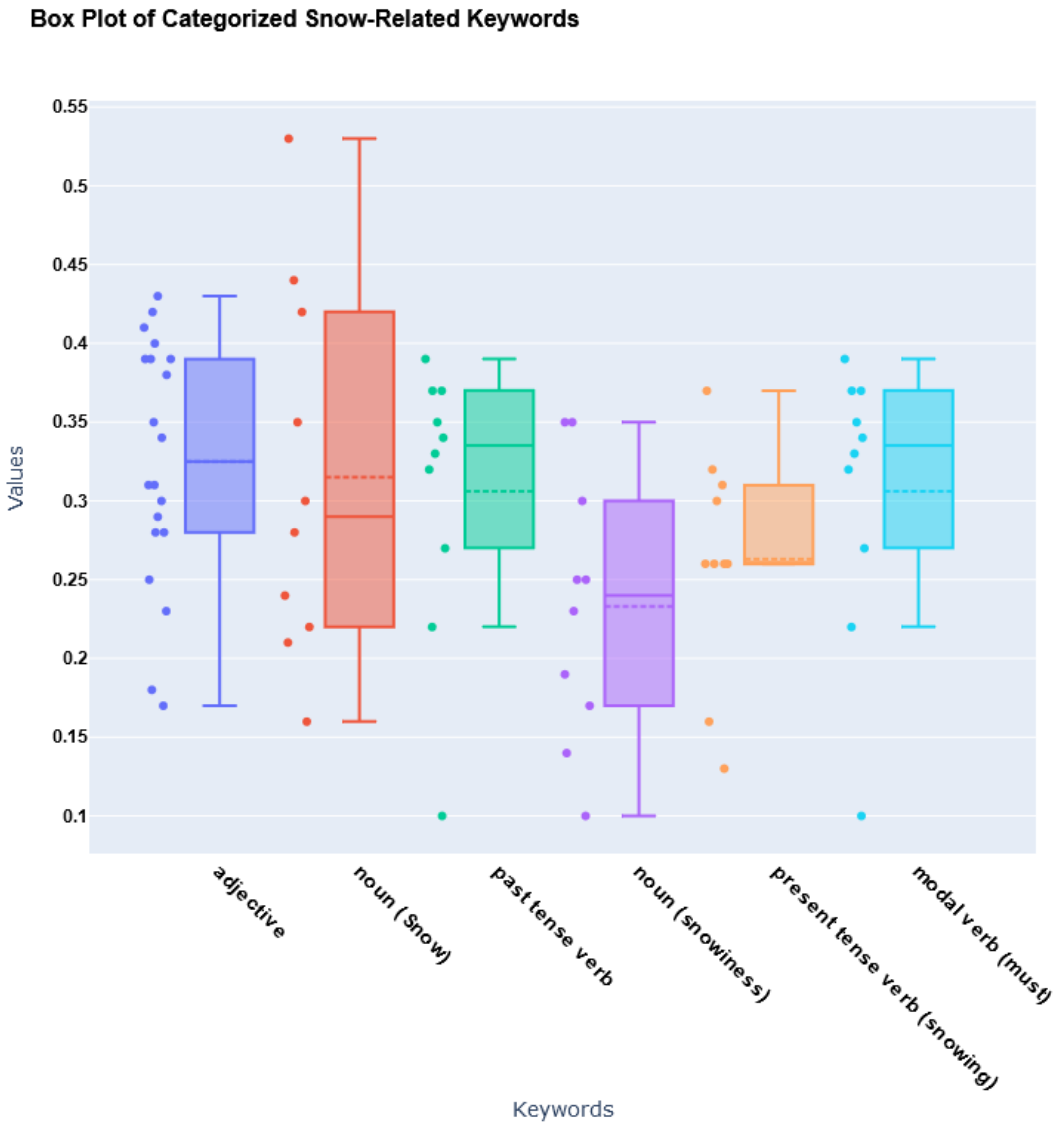


图 5.6 不同词汇下多种提示词的响应变化。这说明模型在基于指令的图像编辑中普遍对语言表述较为敏感。

在我们的模型中，我们通过将文本拆分为单个词，并根据这些词创建定向扰动，改进了输入处理。这种方法提高了粒度，有助于理解每个关键词如何影响输出。图 5.7 展示了不同模型对包含特定关键词的提示词作出的响应。例如，加入关键词“下雪”后，生成图像准确呈现了雪景。然而，省略该关键词时，图像则呈现出其他视觉效果。这凸显了模型对精确文本线索的敏感性，也体现了关键词选择对于实现预期结果的重要性。

这种方法使我们能够评估单个关键词对图像生成的直接影响，揭示每个词如何塑造输出。通过识别并量化这些影响，我们可以改进模型，提升其生成特定上下文视觉内容时的一致性和准确性。进一步考察这些影响，还能帮助我们优化模型，确保文本输入的细微变化能够使生成图像产生精准且可预测的变化。这一方法是实现稳健、可靠的 AI生成内容 的关键一步。



图 5.7 Instruct-Pix2Pix 生成的结果，展示了扰动提示词中出现关键词“下雪”时会生成雪景，而不出现该关键词时则会生成非雪景图像，这说明模型对特定关键词的响应能力。

图 5.8 展示了使用不同提示词结构时图像的变化。这一可视化结果突显了提示设计在引导 AI 生成的输出方面发挥的关键作用。

通过调整“让画面变得雪白”或“天气必须是雪天”之类的关键词和短语，我们可以看到图像外观出现了明显差异。这些变化说明，要实现特定的视觉目标，细致设计提示词至关重要。提示词中每个词的重要性体现在其权重上，颜色越深，表示该词的影响力越大。这种加权表示有助于深入了解驱动视觉变换的语言组成要素。

这项分析展现了模型对特定提示元素的敏感性，并凸显了提示措辞会直接影响生成图像的特征。借助这一认识，用户可以优化提示词，以生成更准确且更符合语境的输出，使这种方法成为优化人工智能驱动的图像生成的有力工具。

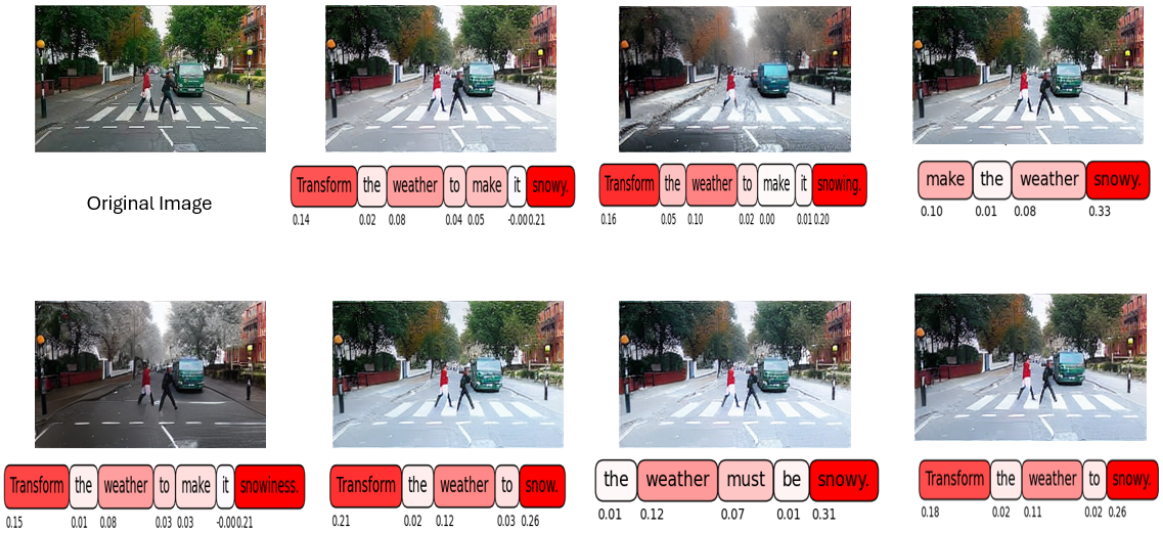


图 5.8 使用 Instruct-Pix2Pix 生成同一图像时不同提示词结构的比较，展示了提示词结构的变化如何使同一基础场景产生不同的变换。

图 5.9 展示了使用 Instruct-Pix2Pix 模型对提示词进行热力图分析，以说明每个词对生成图像的影响力。这项分析直观地呈现了模型如何解码文本输入并为其分配重要性。每张图像对应一个特定的提示词，热力图则显示分配给各个词的权重。颜色越深，表示该词的影响力越大，说明它对最终图像变换的作用越显著。用户可以借此找出实现预期视觉效果最关键的词。

这项分析揭示了模型如何解读并优先处理提示词中的各个词，使我们能够观察文本强调与视觉变

化之间的关系。这些见解对于完善提示工程策略、确保图像变换更加精确且可预测，具有重要价值。通过理解这些动态，我们可以进一步优化 Instruct-Pix2Pix 模型，使其对细微的文本线索更加敏感，从而推动生成质量更高的 AI生成内容。



图 5.9 提示词的热力图分析，展示各个词在引导 Instruct-Pix2Pix 模型视觉输出时的权重。每一行展示由不同提示词结构生成的变换，热力图说明特定词如何影响模型的解释以及生成的图像。

图 5.10 和图 5.11 展示了同一提示词应用于不同图像时词语影响力的一致性。这表明模型在不同场景中保持可预测的词语驱动的变换方面具有鲁棒性。热力图显示提示词中每个词被赋予的权重，颜色越深表示影响力越高。这种可视化有助于理解各个词如何主导变换过程。关键词“下雪”具有强烈影响，从而产生与修改相一致的可靠视觉变化。这凸显了该关键词在引导模型关注期望天气状况方面的有效性。

本分析展示了 Instruct-Pix2Pix 和 I2I-Turbo 模型如何在不同语境中解释同一提示词。

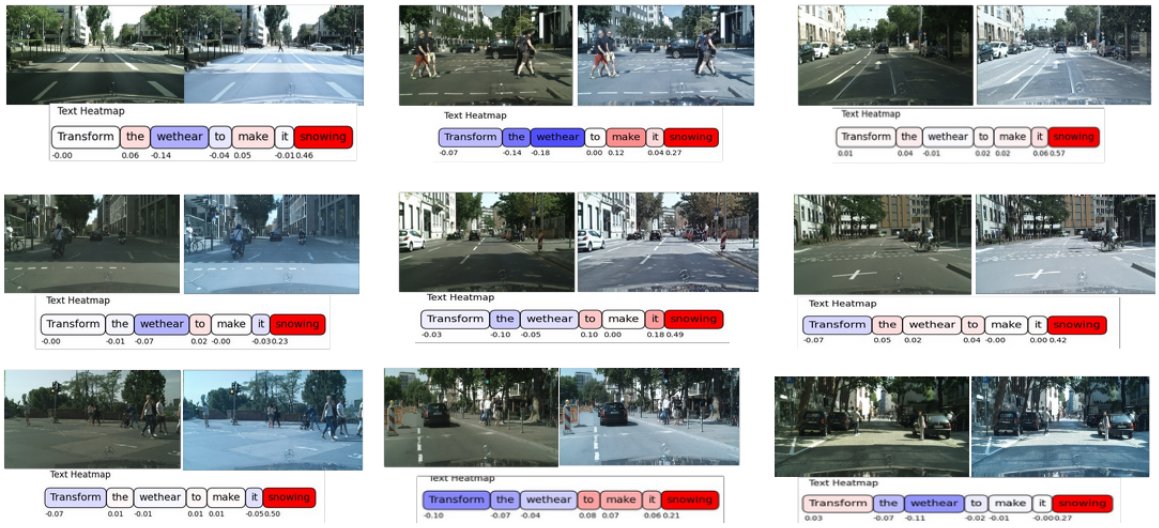


图 5.10 将单个提示词应用于不同图像的热力图分析。每张热力图都显示提示词中各个词的权重，表明 Instruct-Pix2Pix 模型如何在不同场景中保持词语影响力的一致性，以实现预期的变换。

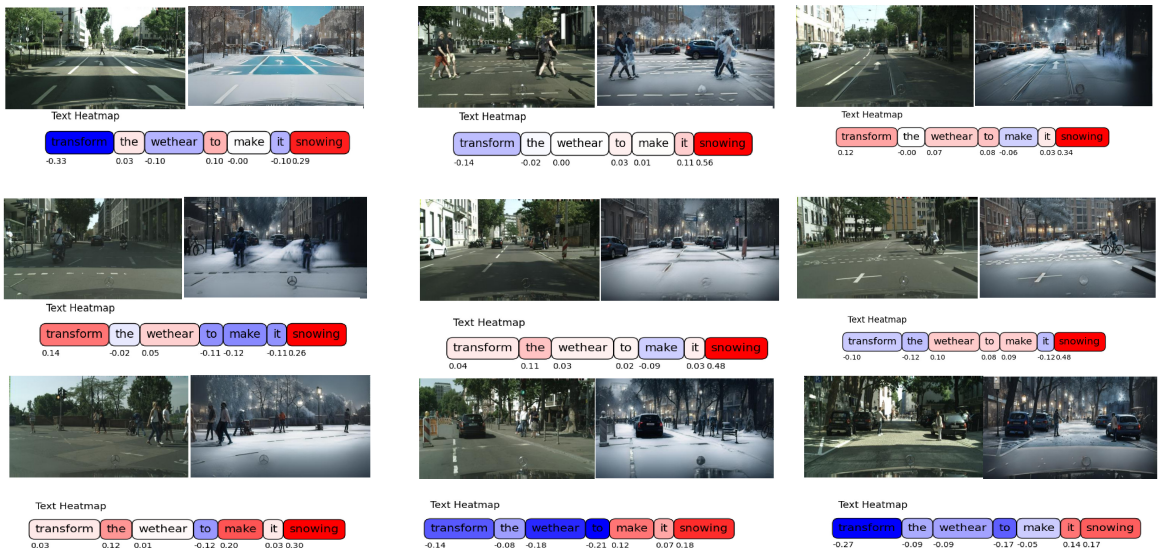


图 5.11 将单个提示词应用于不同图像的热力图分析。每张热力图都显示提示词中各个词的权重，表明 I2I-Turbo 模型如何在不同场景中保持词语影响力的一致性，以实现预期的变换。

5.4 结论

本研究将 **gSMILE** 框架扩展至基于指令的图像编辑，并探究了文本结构、输入图像内容和扰动策略对模型行为及可解释性的影响。我们聚焦于三种先进的扩散式图像编辑模型：Instruct-Pix2Pix、Img2Img-Turbo 和 Diffusers-Inpaint，并采用一系列广泛的定性和定量指标评估 gSMILE 的有效性，包括基于注意力的准确率、稳定性、一致性、保真度和计算成本。

基于运行设计域（ODD）的定性评估表明，gSMILE 能够准确捕捉文本词元与编辑后输出中的视觉变化之间的关系。研究生成了多种场景，包括环境条件（雨、雾、雪）、时间变化（白天/夜晚）以及以人为中心或以物体为中心的因素。热力图和 t-SNE 分析显示，诸如“下雪”之类的关键词始终对最终变换产生最显著的影响，这些词语的出现或缺失直接决定了生成编辑的性质。对扰动提示嵌入的聚类行为的观察进一步证实，模型能够针对词元级修改生成连贯且可解释的变化。

结果还表明，语言表述会显著影响可解释性。提示结构的变化（名词、形容词、动名词，以及“make”和“must”等指令形式）显示，简单的描述性形式通常会产生更稳定、更可预测的输出。相比之下，更复杂的语法结构会带来更高的变异性。对多个提示词和图像的热力图分析进一步证实，通常只有一小部分关键词主导归因空间，这凸显了审慎的提示工程对于实现可控且可解释的编辑的重要性。

从定量角度看，gSMILE 作为基于指令的图像编辑的模型无关局部解释器，展现出了出色的性能：

- **Instruct-Pix2Pix** 取得了最高的准确性和归因保真度。在“不同图像”设置中，其 ATT ACC 达到 $= 0.9571$，ATT F1 $= 0.9166$，ATT AUROC 达到 $= 1.0$，并具有出色的保真度指标（较低的加权均方误差和较高的 R^2_ω），表明加权线性代理模型能够紧密逼近模型行为。
- **Img2Img-Turbo** 表现出有竞争力的准确性，更重要的是，计算一致性更优；在重复运行中，其方差和标准差均为最低，表明归因结果稳定且可复现。

• **Diffusers-Inpaint** 在保真度和准确性方面的性能均较弱，这表明某些架构本身更难通过文本驱动的扰动进行近似，可能需要更专门的扰动策略。

所有模型的总体稳定性都很高（平均杰卡德指数 ≈ 0.85），并且在多种扰动水平下，保真度始终保持稳健。与 LIME 和 Bay-LIME 的比较表明，Wasserstein 距离与加权线性回归的结合通常能够对模型行为进行最忠实的近似，尤其是在文本—图像对齐方面。然而，这些优势也带来了计算成本：在高维文本嵌入上使用 Wasserstein 距离，会使执行时间较 LIME 和 Bay-LIME 显著增加，表明解释深度与计算效率之间存在固有的权衡。

综合来看，这些发现表明，**gSMILE 能够为基于指令的图像编辑模型提供准确、稳定且直观的视觉解释**，从而补充其此前在大语言模型上的有效性。gSMILE 证明，即使无法访问内部参数，也可以恢复文本输入与视觉输出之间有意义的关系。本研究直接回答了**研究问题 3（RQ3）**，表明模型无关的、基于扰动的可解释性能够成功且可靠地扩展到多模态生成系统中，从而为指令控制的图像编辑实现实用的透明度。

与此同时，也涌现出几个值得探索的方向：

1. **自适应扰动策略：**未来的工作可以用有针对性的语言扰动取代均匀随机扰动，仅针对重要词元进行处理。这样既能降低计算成本，又能保持保真度，尤其适用于较长或较复杂的提示词。

2. **文本—图像对齐的因果验证：**为缓解由虚假因素或幻觉导致的归因，可以整合因果分析（例如，在控制图像上下文的同时更改单个词），以验证某个词是否确实引发了视觉变化，从而提升安全关键型应用中的可靠性。

3. **扩展至多模态和时序生成模型：**尽管本研究聚焦于静态图像编辑，但将 gSMILE 扩展至视频扩散模型或更高阶的多模态系统（文本—图像—视频），便可评估时序一致性，并拓宽该框架的普适性。

通过将严谨的评估与可解释的可视化解释相结合，gSMILE 提高了基于指令的图像编辑的透明度，而无需特权访问权限

内部模型机制。这为研究人员和从业者提供了一种实用且值得信赖的工具，用于分析和控制现代生成式视觉系统。

第 6 章

结论

可解释性日益被视为安全、透明且负责任地部署现代机器学习和深度学习系统的基本要求。随着生成式人工智能模型的规模、复杂性和现实影响不断扩大，理解其行为方式以及产生特定输出的原因，已成为培养用户信任并确保负责任使用的关键。本学位论文表明，与人工智能系统进行有意义的人机交互，有赖于清晰且忠实的解释。若决策过程缺乏透明度，用户可能会对采用生成模型心存顾虑，尤其是在医疗保健、法律和自动驾驶等高风险领域。

为应对这些挑战，本研究提出了**gSMILE（生成模型的统计模型无关局部解释可解释性框架）**，这是一个统一的模型无关可解释性框架，适用于文本与多模态生成系统。不同于许多依赖内部注意力图或访问模型层的现有可解释性工具，gSMILE 是一种纯黑箱 方法，无需访问梯度、参数或架构细节。因此，它适用于解释专有模型或基于 API 的模型，而这类模型如今在生成式人工智能领域占据主导地位。通过结合统计距离度量、基于扰动的采样和加权局部代理模型，gSMILE 直接应对了传统事后可解释性技术的关键局限，例如不稳定性、易受输入扰动影响，以及处理多模态输出空间时面临的困难。通过严谨的定量和定性评估，该方法的有效性得到了验证，评估采用了准确性、保真度、一致性和稳定性等归因指标。

本学位论文成功回答了开篇提出的三个研究问题：

RQ1 通过开发并应用一套全面的可解释性指标得以回答，这些指标能够定量评估文本和图像领域中的解释质量

（第3章）。这些指标为系统评估 gSMILE 在各种生成模型中的可靠性奠定了基础。

RQ2 通过展示 gSMILE 如何为大型语言模型提供细粒度的词元级解释，揭示各个提示组件如何影响生成文本，并实现透明且与人类对齐的推理，回应了 RQ2（第 4 章）。

RQ3 通过将 gSMILE 调整用于基于指令的图像编辑任务，回应了 RQ3；在这些任务中，生成的热力图提供了清晰且人类可解释的解释。这些视觉归因精确地展示了指令中的每个词如何影响相应的图像修改，以及图像中的哪些区域受到特定文本成分的影响（第5章）。这使最终用户能够直观地以视觉方式理解模型行为。

总体而言，这项工作为跨模态解释生成模型提供了一条统一且实用的途径，既带来了方法创新，也加深了对现代人工智能系统行为的实证认识。

6.1 贡献总结

本学位论文为生成式人工智能系统的可解释性领域作出了若干重要贡献：

- **介绍统一可解释性框架：** 我们提出了 gSMILE，这是一种新颖的模型无关框架，可为文本和多模态生成模型生成人类可解释的解释。通过可视化单个输入元素（例如提示词中的词）对模型输出的影响力，gSMILE 为大语言模型（LLM）和基于指令的图像编辑系统的模型行为提供清晰且可操作的洞见。

- **通过视觉和文本解释推进透明度：** gSMILE 明确地将用户指令与模型行为关联起来，为文本生成生成词元级归因图，并生成以视觉方式将词与图像修改联系起来的热力图。这些解释增强了透明度，促进了用户信任，并支持医疗保健、教育、自动驾驶和法律决策等敏感领域中系统的安全部署。

- **可解释性与鲁棒性的严格评估方案：** 本研究提出了一套全面的评估方法，利用稳定性、保真度、准确性和一致性指标，对多个生成

模型进行评估。通过大量实验，我们证明，gSMILE 能够提供可靠、可重复且具有高保真度的解释，因此可作为审计、理解和改进现代生成式人工智能系统行为的有效工具。

6.2 局限性

尽管 gSMILE 提供了一个灵活的模型无关框架，可用于解释大语言模型（LLM）和基于指令的图像编辑系统，但由于其方法设计选择以及生成模型固有的复杂性，仍存在若干局限性。

首先，本研究采用的扰动策略主要依赖于词元的二元移除–保留（0/1）。尽管该方法对分析局部敏感性有效，但无法捕捉更细微的语言转换。移除词元可能会扭曲句子含义、破坏句法结构，或使模型在其自然范围之外运行。更具表现力的扰动，如同义词替换、释义，或应用于嵌入空间的高斯噪声，尚未被探索，而这些扰动可能揭示更丰富的归因模式。

其次，标记级扰动即使在提示词的语义内容大体得以保留时，也可能导致输出发生不成比例的大幅变化。这种敏感性削弱了基于回归的代理模型所要求的局部平滑性假设，从而可能降低解释的稳定性与解释连贯性。

第三，生成模型，尤其是大语言模型，在提示词含糊、不完整或质量下降时容易产生幻觉。这些幻觉输出会给归因过程引入噪声，可能损害使用局部扰动生成的解释的可靠性，尤其是对于对齐程度较弱或对输入变化高度敏感的模型。

第四，该框架的计算需求相当高。gSMILE 会生成并评估大量扰动，而使用 Wasserstein 距离对样本进行加权则进一步增加了计算成本。对于基于 API 的模型，这会导致更高的推理成本和更长的整体运行时间。对于本地部署的模型，尤其是涉及视觉语言管线的模型，则需要大量 GPU 资源，因此没有高端硬件的用户难以使用。

第五，本学位论文考察的代理模型仅限于加权线性回归和贝叶斯岭回归。尽管这些代理模型在局部线性假设下效果良好，但对于扩散模型等高度非线性的生成系统而言，可能并不足够。本研究未考虑表达能力更强的代理模型。

最后，实际部署受到模型可访问性的限制。有些大语言模型和多模态系统不提供 API访问，且需要手动安装或本地微调，这增加了复现实验或将 gSMILE 扩展到更广泛应用场景中的操作复杂性。

尽管存在这些局限性，gSMILE仍是提升生成模型可解释性的一条有前景的方向。通过更先进的扰动策略、改进的代理模型以及优化的距离计算来应对这些挑战，将进一步增强其鲁棒性、可扩展性和适用性。

6.3 未来工作

基于本学位论文中指出的局限性，以下几个有前景的方向可以拓展并强化生成式人工智能领域中的 gSMILE 框架：

1.**开发表达能力更强且具备语义感知能力的扰动策略：**由于当前方法依赖二元移除，未来的工作可以探索更丰富的扰动方式，例如同义词替换、释义改写、保持上下文的编辑，或在嵌入空间中添加高斯噪声。这些策略可以减少语义失真，并产生信息量更丰富的归因模式。

2. **提升对令牌级扰动导致的显著输出偏移的鲁棒性：**未来研究可以探索能够保持语义连贯性的扰动方案，包括上下文感知掩码或约束性释义改写，并结合正则化技术，以稳定局部非平滑区域中的代理模型。

3. **缓解生成式解释中的幻觉诱导噪声：**将 gSMILE 与幻觉检测、置信度校准或因果掩蔽技术相结合，可以减少不可靠模型输出的影响，并提高对退化或模糊输入敏感的模型的解释保真度。

4. **通过优化采样和更快的距离度量提升计算效率：**为了降低大型扰动集合和基于Wasserstein距离的加权所需的计算成本，未来的工作可以采用基于优化的采样策略（例如 OptiLIME）、自适应扰动选择，或更快的最优传输变体，例如Sinkhorn散度。

5. **探索适用于高度非线性生成系统的更强大代理模型：**当前的代理模型（加权线性回归和贝叶斯岭回归

）可能不足以刻画非线性行为。未来可采用核回归器、神经局部近似器、随机特征模型或基于树的代理模型，以提升复杂生成式架构（如扩散模型）的保真度。

6. **提升大型模型或非 API 模型的可扩展性与可访问性：**开发轻量级或近似版本的 gSMILE 可以降低资源需求，同时，采用标准化接口或许能简化对本地部署的开源多模态模型进行的实验。

7. **将框架扩展至其他生成模态：**除文本和图像编辑外，未来的工作还可以将 gSMILE 应用于视频生成、音频合成以及更广泛的多模态架构。这些场景为在时间约束和跨模态约束下评估可解释性提供了机会。

8. **将 gSMILE 与基于注意力的可解释性机制相结合：**将基于扰动的解释与注意力可视化方法（如 DAAM）结合起来，或许能够生成联合的词元—注意力—像素解释，从而更深入地理解文本输入和内部注意力模式如何塑造生成输出。

参考文献

[1] Alexander Bick、Adam Blandin 和 David J Deming。《生成式人工智能的迅速普及》。技术报告，美国国家经济研究局，2024年。[2] Grand View Research。《可解释人工智能市场规模、市场份额与趋势分析报告：按应用、最终用途和地区划分，以及细分市场预测，2023—2030年》，2023年。[3] Marco Tulio Ribeiro、Sameer Singh 和 Carlos Guestrin。《为什么要相信你？解释任意分类器的预测》。载于第*22*届 *ACM SIGKDD* 国际知识发现与数据挖掘会议论文集，第1135–1144页，2016年。[4]Koorosh Aslansefat、Mojgan Hashemian、Martin Walker、Mohammed Naveed Akram、Ioannis Sorokos 和 Yiannis Papadopoulos。《用 SMILE 解释黑箱：基于局部解释的统计模型无关可解释性》。《style id='7'>IEEE Software，41(1)：87–97，2023年。[5]Anthropic。《归因图：详细追踪大语言模型行为》。https://transformer-circuits.pub/2025/attribution-graphs/biology.html，2025年。访问日期：2025年5月6日。[6] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。《注意力就是你所需的一切》。《style id='11'>神经信息处理系统进展，第30卷，2017年。[7] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Pra- fulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。《语言模型是少样本学习器》。《style id='14'>神经信息处理系统进展，第33卷：1877–1901，2020年。[8] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Flo- rencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。《GPT-4 技术报告》。《style id='17'>arXiv 预印本 arXiv:2303.08774，2023年。[9] Xincheng Shuai、Henghui Ding、Xingjun Ma、Rongcheng Tu、Yu-Gang Jiang 和 Dacheng Tao。《基于文生图扩散模型的多模态引导图像编辑综述》。《style id='20'>arXiv 预印本 arXiv:2406.14555，2024年。

[10] Junnan Li、Dongxu Li、Silvio Savarese 和 Steven Hoi。《BLIP-2：利用冻结图像编码器和大型语言模型自举语言-图像预训练》。载于国际机器学习会议，第 19730–19742 页。PMLR，2023年。

[11] Wojciech Samek、Thomas Wiegand 和 Klaus-Robert Müller。《可解释人工智能：理解、可视化和解读深度学习模型》。*arXiv* 预印本 *arXiv:1708.08296*，2017年。[12]Cynthia Rudin。《停止为高风险决策解释黑箱机器学习模型，改用可解释模型》。《自然·机器智能》，1(5):206–215，2019年。[13] Finale Doshi-Velez 和 Been Kim。《迈向一门严谨的可解释机器学习科学》。*arXiv* 预印本 *arXiv:1702.08608*，2017年。[14] Amina Adadi 和 Mohammed Berrada。《窥探黑箱内部：可解释人工智能（XAI）综述》。*IEEE Access*，6:52138–52160，2018年。[15] Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。《Grad-CAM：通过基于梯度的定位，从深度网络中生成可视化解释》。载于*IEEE*国际计算机视觉会议论文集，第 618–626 页，2017年。[16] Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。《基础模型的机遇与风险》。*arXiv* 预印本 *arXiv:2108.07258*，2021年。[17] Ian J Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio。《生成对抗网络》。神经信息处理系统进展，第 27 卷，2014年。[18] Gemini Team、Petko Georgiev、Ving Ian Lei、Ryan Burnell、Libin Bai、Anmol Gulati、Garrett Tanzer、Damien Vincent、Zhufeng Pan、Shibo Wang 等。《Gemini 1.5：解锁数百万词元上下文中的多模态理解》。*arXiv* 预印本 *arXiv:2403.05530*，2024年。[19] Anthropic。Anthropic 的 Claude AI。https://www.anthropic.com/index/claude，2024年。[20] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale 等。《Llama 2：开放基础模型与微调聊天模型》。*arXiv* 预印本 *arXiv:2307.09288*，2023年。[21] Aaron Hurst、Adam Lerer、Adam P Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford 等。《GPT-4o 系统卡》。*arXiv* 预印本 *arXiv:2410.21276*，2024年。[22]谷歌。Gemini 新增图像生成和编辑功能。https://beebom.com/chatgpt-vs-gemini-native-image-generation，2024年。[23] Tim Brooks、Aleksander Holynski 和 Alexei A Efros。《InstructPix2Pix：学习遵循图像编辑指令》。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18392–18402 页，2023年。

[24] Su Wang、Chitwan Saharia、Ceslee Montgomery、Jordi Pont-Tuset、Shai Noy、Stefano Pellegrini、Yasumasa Onoe、Sarah Laszlo、David J Fleet、Radu Soricut 等。Imagen Editor 和 EditBench：推进并评估文本引导图像修复。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 18359–18369 页，2023年。

[25]麦肯锡公司。生成式人工智能的未来如何？15 张图表中的初步展望。https://www.mckinsey.com/featured-insights/mckinsey-explainers/whats-the-future-of-generative-ai-an-early-view-in-15-charts，2023年。

[26] Riccardo Guidotti、Anna Monreale、Salvatore Ruggieri、Franco Turini、Fosca Giannotti 和 Dino Pedreschi。解释黑箱模型的方法综述。*ACM* 计算调查（*CSUR*），51(5):1–42，2018年。

[27] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Om- mer。利用潜在扩散模型进行高分辨率图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 10684–10695 页，2022年。

[28] Aaron Van Den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior、Koray Kavukcuoglu 等。WaveNet：一种用于原始音频的生成模型。*arXiv* 预印本 *arXiv:1609.03499*，12:1，2016年。

[29] Ruben Villegas、Jimei Yang、Yuliang Zou、Sungryull Sohn、Xunyu Lin 和 Honglak Lee。通过层次预测学习生成长期未来。载于国际机器学习会议，第 3560–3569 页。PMLR，2017年。

[30] Maria Tsimpoukelli、Jacob L Menick、Serkan Cabi、SM Eslami、Oriol Vinyals 和 Felix Hill。使用冻结语言模型进行多模态少样本学习。神经信息处理系统进展，34：200–212，2020年。

[31] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc、Antoine Miech、Iain Barr、Yana Hasson、Karel Lenc、Arthur Mensch、Katherine Millican、Malcolm Reynolds 等。Flamingo：一种用于少样本学习的视觉语言模型。神经信息处理系统进展，35：23716–23736，2022年。

[32] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu。利用统一的文本到文本转换器探索迁移学习的极限。《机器学习研究杂志》，21(140)：1–67，2020年。

[33] Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebas- tian Gehrmann 等。Palm：借助路径扩展语言建模。《机器学习研究杂志》，24(240):1–113，2023年。

[34] Yuntao Bai、Andy Jones、Kamal Ndousse、Amanda Askell、Anna Chen、Nova DasSarma、Dawn Drain、Stanislav Fort、Deep Ganguli、Tom Henighan 等。《通过人类反馈强化学习训练有帮助且无害的助手》。*arXiv* 预印本 *arXiv:2204.05862*，2022年。

[35] Gemma 团队、Thomas Mesnard、Cassidy Hardin、Robert Dadashi、Surya Bhupati- raju、Shreya Pathak、Laurent Sifre、Morgane Rivière、Mihir Sanjay Kale、Juliette Love 等。《Gemma：基于 Gemini 研究与技术的开放模型》。*arXiv* 预印本 *arXiv:2403.08295*，2024年。[36]Sid Black、Stella Biderman、Eric Hallahan、Quentin Anthony、Leo Gao、Lau- rence Golding、Horace He、Connor Leahy、Kyle McDonell、Jason Phang 等。《GPT-NeoX-20B：一种开源自回归语言模型》。*arXiv* 预印本 *arXiv:2204.06745*，2022年。[37] Opher Lieber、Or Sharir、Barak Lenz 和 Yoav Shoham。《Jurassic-1：技术细节与评估》。白皮书。*AI21 Labs*，1(9):1–17，2021年。[38] Yunfan Gao、Yun Xiong、Xinyu Gao、Kangxiang Jia、Jinliu Pan、Yuxi Bi、Yixin Dai、Jiawei Sun、Haofen Wang 和 Haofen Wang。《大语言模型的检索增强生成：综述》。*arXiv* 预印本 *arXiv:2312.10997*，2(1)，2023年。[39] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。《BERT：用于语言理解的深度双向 Transformer 预训练》。载于*2019*年计算语言学协会北美分会会议论文集：人类语言技术，第*1*卷（长篇和短篇论文），第4171–4186页，2019年。[40] Xiao Bi、Deli Chen、Guanting Chen、Shanhuang Chen、Damai Dai、Chengqi Deng、Honghui Ding、Kai Dong、Qiushi Du、Zhe Fu 等。《DeepSeek LLM：以长期主义理念扩展开源语言模型》。*arXiv* 预印本 *arXiv:2401.02954*，2024年。[41] Anthropic。《Command r+模型说明卡》。https://www.anthropic.com，2024年。[42] Albert Q. Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Florian Bressand、Gianna Lengyel、Guillaume Lample、Lucile Saulnier、Lélio Renard Lavaud、Marie-Anne Lachaux、Pierre Stock、Teven Le Scao、Thibaut Lavril、Thomas Wang、Timothée Lacroix 和 William El Sayed。《Mistral 7B》，2023年。[43] Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。《Llama 3 模型群》。*arXiv* 预印本 *arXiv:2407.21783*，2024年。[44] Albert Q Jiang、Alexandre Sablayrolles、Antoine Roux、Arthur Mensch、Blanche Savary、Chris Bamford、Devendra Singh Chaplot、Diego de las Casas、Emma Bou Hanna、Florian Bressand 等。《Mixtral 专家混合模型》。*arXiv* 预印本 *arXiv:2401.04088*，2024年。[45] Ebtesam Almazrouei、Hamza Alobeidli、Abdulaziz Alshamsi、Alessandro Cappelli、Ruxandra Cojocaru、Mérouane Debbah、Étienne Goffinet、Daniel Hesslow、Julien Launay、Quentin Malartic 等。《Falcon 开放语言模型系列》。*arXiv* 预印本 *arXiv:2311.16867*，2023年。[46] Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Cheng- gang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。《DeepSeek-V3 技术报告》。*arXiv* 预印本 *arXiv:2412.19437*，2024年。

[47] Guan Wang、Sijie Cheng、Xianyuan Zhan、Xiangang Li、Sen Song 和 Yang Liu。Openchat：利用混合质量数据推进开源语言模型。*arXiv* 预印本 *arXiv:2309.11235*，2023年。[48] Anthropic。Claude 3 技术概述，2024年。在线提供。[49] xAI。宣布推出 grok-1.5，2024年3月。 [在线]。[50] Binyuan Hui、Jian Yang、Zeyu Cui、Jiaxi Yang、Dayiheng Liu、Lei Zhang、Tianyu Liu、Jiajun Zhang、Bowen Yu、Keming Lu 等。Qwen2. 5-coder 技术报告。*arXiv* 预印本 *arXiv:2409.12186*，2024年。[51] D Lepikhin、H Lee、Y Xu、D Chen、O Firat、Y Huang、M Krikun、N Shazeer 和 Z Gshard。利用条件计算和自动分片扩展巨型模型。*arXiv* 预印本 *arXiv:2006.16668*，2020年。[52] William Fedus、Barret Zoph 和 Noam Shazeer。Switch Transformers：以简单高效的稀疏性扩展至万亿参数模型。《机器学习研究杂志》，23(120)：1–39，2022年。[53] Zhiliang Peng、Wenhui Wang、Li Dong、Yaru Hao、Shaohan Huang、Shuming Ma 和 Furu Wei。Kosmos-2：将多模态大语言模型与现实世界相联系。*arXiv* 预印本 *arXiv:2306.14824*，2023年。[54] BigScience Workshop、Teven Le Scao、Angela Fan、Christopher Akiki、Ellie Pavlick、Suzana Ili´c、Daniel Hesslow、Roman Castagné、Alexandra Sasha Luccioni、François Yvon 等。BLOOM：一个拥有 1760 亿参数的开放访问多语言语言模型。*arXiv* 预印本 *arXiv:2211.05100*，2022年。[55] Zewen Chi、Shaohan Huang、Li Dong、Shuming Ma、Bo Zheng、Saksham Singhal、Payal Bajaj、Xia Song、Xian-Ling Mao、Heyan Huang 等。XLM-E：通过 ELECTRA 进行跨语言语言模型预训练。*arXiv* 预印本 *arXiv:2106.16138*，2021年。[56] Alexis Conneau、Kartikay Khandelwal、Naman Goyal、Vishrav Chaudhary、Guil- laume Wenzek、Francisco Guzmán、Edouard Grave、Myle Ott、Luke Zettlemoyer 和 Veselin Stoyanov。大规模无监督跨语言表示学习。*arXiv* 预印本 *arXiv:1911.02116*，2019年。[57] Linting Xue、Aditya Barua、Noah Constant、Rami Al-Rfou、Sharan Narang、Mihir Kale、Adam Roberts 和 Colin Raffel。ByT5：借助预训练字节到字节模型迈向无词元的未来。计算语言学协会会刊，10：291–306，2022年。[58] Karan Singhal、Tao Tu、Juraj Gottweis、Rory Sayres、Ellery Wulczyn、Mohamed Amin、Le Hou、Kevin Clark、Stephen R Pfohl、Heather Cole-Lewis 等。借助大语言模型实现专家级医学问答。《自然·医学》，31(3)：943–950，2025。[59]Ross Taylor、Marcin Kardas、Guille m Cucurull、Thomas Scialom、Anthony Hartshorn、Elvis Saravia、Andrew Poulton、Viktor Kerkez 和 Robert Stojnic。Galactica：面向科学的大语言模型。*arXiv* 预印本 *arXiv:2211.09085*，2022年。

[60] Ilias Chalkidis、Manos Fergadiotis、Prodromos Malakasiotis、Nikolaos Aletras 和 Ion Androutsopoulos。Legal-BERT：法律学院的木偶们。*arXiv* 预印本 *arXiv:2010.02559*，2020年。[61] Anton Lozhkov、Raymond Li、Loubna Ben Allal、Federico Cassano、Joel Lamy- Poirier、Nouamane Tazi、Ao Tang、Dmytro Pykhtar、Jiawei Liu、Yuxiang Wei 等。StarCoder 2 和 The Stack v2：下一代。*arXiv* 预印本 *arXiv:2402.19173*，2024年。[62] Baptiste Roziere、Jonas Gehring、Fabian Gloeckle、Sten Sootla、Itai Gat、Xiao- qing Ellen Tan、Yossi Adi、Jingyu Liu、Romain Sauvestre、Tal Remez 等。Code LLaMA：面向代码的开放基础模型。*arXiv* 预印本 *arXiv:2308.12950*，2023年。[63] Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang 和 Tatsunori B Hashimoto。Alpaca：一种强大且可复现的指令跟随模型。斯坦福基础模型研究中心。*https://crfm. stanford. edu/2023/03/13/alpaca. html*，3(6)：7，2023年。[64] Wei-Lin Chiang、Zhuohan Li、Ziqing Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E Gonzalez 等。Vicuna：一款开源聊天机器人，以达到 ChatGPT 90%* 的质量惊艳 GPT-4。参见 *https://vicuna. lmsys. org*（访问于 *2023*年*4*月*14*日），2(3)：6，2023年。[65] Zhengxiao Du、Yujie Qian、Xiao Liu、Ming Ding、Jiezhong Qiu、Zhilin Yang 和 Jie Tang。GLM：通过自回归空白填充进行通用语言模型预训练。*arXiv* 预印本 *arXiv:2103.10360*，2021年。[66]Canwen Xu、Daya Guo、Nan Duan 和 Julian McAuley。Baize：一种利用自对话数据进行参数高效微调的开源聊天模型。*arXiv* 预印本 *arXiv:2304.01196*，2023年。[67] Yuanzhi Li、Sébastien Bubeck、Ronen Eldan、Allie Del Giorno、Suriya Gunasekar 和 Yin Tat Lee。教材就是你所需要的一切 II：Phi-1.5 技术报告。*arXiv* 预印本 *arXiv:2309.05463*，2023年。[68] Peiyuan Zhang、Guangtao Zeng、Tianduo Wang 和 Wei Lu。TinyLLaMA：一种开源小型语言模型。*arXiv* 预印本 *arXiv:2401.02385*，2024年。[69] Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。DistilBERT：BERT 的蒸馏版本，更小、更快、更便宜、更轻量。*arXiv* 预印本 *arXiv:1910.01108*，2019年。[70] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei 和 Ilya Sutskever。语言模型是无监督的多任务学习者，2019年。 [在线]。[71] Yuzhou Huang、Liangbin Xie、Xintao Wang、Ziyang Yuan、Xiaodong Cun、Yixiao Ge、Jiantao Zhou、Chao Dong、Rui Huang、Ruimao Zhang 等。SmartEdit：探索借助多模态大语言模型实现的复杂基于指令的图像编辑。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 8362–8371 页，2024年。

[72] Tsu-Jui Fu、Wenze Hu、Xianzhi Du、William Yang Wang、Yinfei Yang 和 Zhe Gan。通过多模态大语言模型引导基于指令的图像编辑。*arXiv* 预印本 *arXiv:2309.17102*，2023年。[73] Bahjat Kawar、Shiran Zada、Oran Lang、Omer Tov、Huiwen Chang、Tali Dekel、Inbar Mosseri 和 Michal Irani。Imagic：使用扩散模型进行基于文本的真实图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6007–6017页，2023年。[74] Adobe Inc. Adobe Firefly：面向创意人士的免费生成式人工智能。https://www.adobe.com/ products/firefly.html。访问日期：2024年12月12日。[75] Kai Zhang、Lingbo Mo、Wenhu Chen、Huan Sun 和 Yu Su。Magicbrush：一个人工标注的指令引导图像编辑数据集。神经信息处理系统进展，36：31428–31449，2023年。[76] Gaurav Parmar、Taesung Park、Srinivasa Narasimhan 和 Jun-Yan Zhu。使用文本到图像模型进行一步图像转换。*arXiv* 预印本 *arXiv:2403.12036*，2024年。[77] Mingdeng Cao、Xintao Wang、Zhongang Qi、Ying Shan、Xiaohu Qie 和 Yinqiang Zheng。MasaCtrl：无需调优的互相自注意力控制，实现一致的图像合成与编辑。载于*IEEE/CVF* 国际计算机视觉会议论文集，第22560–22570页，2023年。[78]Sijia Li、Chen Chen 和 Haonan Lu。MoeController：利用混合专家控制器进行基于指令的任意图像操控。*arXiv* 预印本 *arXiv:2309.04372*，2023年。[79] Yuanze Lin、Yi-Wen Chen、Yi-Hsuan Tsai、Lu Jiang 和 Ming-Hsuan Yang。通过可学习区域实现文本驱动的图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第7059–7068页，2024年。[80]Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。利用CLIP潜变量进行分层文本条件图像生成。*arXiv* 预印本 *arXiv:2204.06125*，1(2)：3，2022年。[81] Shitao Xiao、Yueze Wang、Junjie Zhou、Huaying Yuan、Xingrun Xing、Ruiran Yan、Chaofan Li、Shuting Wang、Tiejun Huang 和 Zheng Liu。OmniGen：统一图像生成。载于计算机视觉与模式识别会议论文集，第13294–13304页，2025。[82] Anwaar Ulhaq 和 Naveed Akhtar。面向视觉的高效扩散模型：综述。*arXiv* 预印本 *arXiv:2210.09292*，2022年。[83] Qin Guo 和 Tianwei Lin。聚焦指令：通过注意力调制进行细粒度多指令图像编辑。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第6986–6996页，2024年。[84] Xiang'Anthony' Chen、Jeff Burke、Ruofei Du、Matthew K Hong、Jennifer Jacobs、Philippe Laban、Dingzeyu Li、Nanyun Peng、Karl DD Willis、Chien-Sheng Wu 等。以人为中心的生成式人工智能下一步：技术视角。*arXiv* 预印本 *arXiv:2306.15774*，2023年。

[85] Kelvin Xu、Jimmy Ba、Ryan Kiros、Kyunghyun Cho、Aaron Courville、Ruslan Salakhudinov、Rich Zemel 和 Yoshua Bengio。展示、关注与讲述：利用视觉注意力生成神经图像描述。载于国际机器学习会议，第2048–2057页。PMLR，2015年。[86] Taesung Park、Ming-Yu Liu、Ting-Chun Wang 和 Jun-Yan Zhu。利用空间自适应归一化进行语义图像合成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第2337–2346页，2019年。[87] Alejandro Barredo Arrieta、Natalia Díaz-Rodríguez、Javier Del Ser、Adrien Bennetot、Siham Tabik、Alberto Barbado、Salvador García、Sergio Gil-López、Daniel Molina、Richard Benjamins 等。可解释人工智能（XAI）：迈向负责任人工智能的概念、分类体系、机遇与挑战。《信息融合》，58：82–115，2020年。[88] Zachary C Lipton。模型可解释性的神话：在机器学习中，可解释性这一概念既重要又难以捉摸。《*Queue*》，16(3)：31–57，2018年。[89] Tim Miller。人工智能中的解释：来自社会科学的见解。人工智能，267：1–38，2019年。[90] Christoph Molnar。可解释机器学习。Lulu.com，2020年。[91] W James Murdoch、Chandan Singh、Karl Kumbier、Reza Abbasi-Asl 和 Bin Yu。可解释机器学习：定义、方法与应用。*arXiv* 预印本 *arXiv:1901.04592*，2019年。[92] Leilani H Gilpin、David Bau、Ben Z Yuan、Ayesha Bajwa、Michael Specter 和 Lalana Kagal。解释的解释：机器学习可解释性概述。载于*2018*年 *IEEE*第五届数据科学与高级分析国际会议（*DSAA*），第80–89页。IEEE，2018年。[93] Jason Wei、Yi Tay、Rishi Bommasani、Colin Raffel、Barret Zoph、Sebastian Borgeaud、Dani Yogatama、Maarten Bosma、Denny Zhou、Donald Metzler 等。大语言模型的涌现能力。*arXiv* 预印本 *arXiv:2206.07682*，2022年。[94]Ricardo Vinuesa、Hossein Azizpour、Iolanda Leite、Madeline Balaam、Virginia Dignum、Sami Domisch、Anna Felländer、Simone Daniela Langhans、Max Tegmark 和 Francesco Fuso Nerini。人工智能在实现可持续发展目标中的作用。《自然·通讯》，11(1)：233，2020年。[95] Bryce Goodman 和 Seth Flaxman。欧盟关于算法决策的法规及“解释权”。《人工智能杂志》，38(3)：50–57，2017年。[96] 《人工智能法规》（人工智能法案）。欧洲议会和欧盟理事会，2024年。网址：https://artificialintelligenceact.eu。[97] Barbara Marchetti。人工智能与公共当局：欧盟人工智能法案是否保护公共价值？《欧洲公法评论》，36(1)，2024年。[98] Giulia Vilone 和 Luca Longo。可解释性概念及可解释人工智能的评估方法。《信息融合》，76：89–106，2021年。

[99] Luciano Floridi、Josh Cowls、Monica Beltrametti、Raja Chatila、Patrice Chazerand、Virginia Dignum、Christoph Luetge、Robert Madelin、Ugo Pagallo、Francesca Rossi 等。AI4People——构建良好人工智能社会的伦理框架：机遇、风险、原则与建议。《心智与机器》，28(4):689–707，2018年。[100] Sina Mohseni、Niloofar Zarei 和 Eric D Ragan。可解释人工智能系统设计与评估的多学科综述及框架。《*ACM*交互式智能系统汇刊》*(TiiS)*，11(3-4):1–45，2021年。[101] Avi Rosenfeld 和 Ariella Richardson。人机代理系统中的可解释性。自主代理与多智能体系统，33(6):673–705，2019年。[102] Robert R Hoffman、Shane T Mueller、Gary Klein 和 Jordan Litman。可解释人工智能度量指标：挑战与前景。*arXiv*预印本 *arXiv:1812.04608*，2018年。[103] Anna Jobin、Marcello Ienca 和 Effy Vayena。全球人工智能伦理指南概况。《自然·机器智能》，1(9):389–399，2019年。[104] Corinne Cath。人工智能治理：伦理、法律和技术方面的机遇与挑战。《英国皇家学会哲学汇刊*A*辑：数学、物理与工程科学》，376(2133):20180080，2018年。[105] Dylan Slack、Sophie Hilgard、Emily Jia、Sameer Singh 和 Himabindu Lakkaraju。欺骗LIME和SHAP：针对事后解释方法的对抗攻击。载于*AAAI/ACM*人工智能、伦理与社会会议论文集，第180–186页，2020年。[106] Scott M Lundberg 和 Su-In Lee。解释模型预测的统一方法。神经信息处理系统进展，30，2017年。[107]Xingyu Zhao、Wei Huang、Xiaowei Huang、Valentin Robu 和 David Flynn。BayLIME：贝叶斯局部可解释模型无关解释。载于人工智能中的不确定性，第887–896页。PMLR，2021年。[108] Zhengze Zhou、Giles Hooker 和 Fei Wang。S-LIME：用于模型解释的稳定化LIME。载于第*27*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，第2429–2438页，2021年。[109] Sebastian Bach、Alexander Binder、Grégoire Montavon、Frederick Klauschen、Klaus-Robert Müller 和 Wojciech Samek。通过逐层相关性传播解释非线性分类器决策的逐像素解释。《*PLOS ONE*》，10(7):e0130140，2015年。[110] Fred Hohman、Andrew Head、Rich Caruana、Robert DeLine 和 Steven M Drucker。Gamut：一种设计探针，用于了解数据科学家如何理解机器学习模型。载于*2019*年*CHI*人机交互系统中的人为因素会议论文集，第1–13页，2019年。[111] Xinyang Zhang、Ningfei Wang、Hua Shen、Shouling Ji、Xiapu Luo 和 Ting Wang。面临攻击的可解释深度学习。载于第*29*届 {*USENIX*} 安全研讨会（{*USENIX*} 安全*20*），2020年。

[112] Gaith Rjoub、Jamal Bentahar、Omar Abdel Wahab、Rabeb Mizouni、Alyssa Song、Robin Cohen、Hadi Otrok 和 Azzam Mourad。网络安全领域可解释人工智能综述。*IEEE*网络与服务管理汇刊，20(4):5115–5140，2023年。[113] Andreas Holzinger、Chris Biemann、Constantinos S Pattichis 和 Douglas B Kell。构建面向医疗领域的可解释人工智能系统需要什么？*arXiv*预印本 *arXiv:1712.09923*，2017 年。[114] Fatemeh Vakhshiteh 、 Ahmad Nickabadi 和 Raghavendra Ramachandra。针对人脸识别的对抗攻击：全面研究。*IEEEAccess*，9:92735–92756，2021年。[115]Samuel G Finlayson、Hyung Won Chung、Isaac S Kohane 和 Andrew L Beam。针对医疗深度学习系统的对抗攻击。*arXiv*预印本 *arXiv:1804.05296*，2018年。[116] Zheng Li、Yue Zhao、Xiyang Hu、Nicola Botta、Cezar Ionescu 和 George H Chen。ECOD：使用经验累积分布函数进行无监督异常值检测。*IEEE* 知识与数据工程汇刊，35(12):12181–12193，2022年。[117] Steven Bramhall、Hayley Horn、Michael Tieu 和 Nibhrat Lohia。Qlime——一种二次局部可解释模型无关解释方法。*SMU* 数据科学评论，3(1):4，2020年。[118] Sharath M Shankaranarayana 和 Davor Runje。ALIME：一种基于自编码器的局部可解释性方法。载于智能数据工程与自动化学习国际会议，第454–463页。施普林格，2019年。[119] Sohini Upadhyay、Vatche Isahagian、Vinod Muthusamy 和 Yara Rizk。扩展LIME以实现业务流程自动化。*arXiv*预印本 *arXiv:2108.04371*，2021年。[120] Damien Garreau 和 Ulrike Luxburg。解释解释器：对LIME的初步理论分析。载于人工智能与统计国际会议，第1287–1296页。PMLR，2020年。[121] Hamid Saadatfar、Zeinab Kiani-Zadegan 和 Benyamin Ghahremani-Nezhad。US-LIME：利用表格数据上的不确定性采样提高LIME的保真度。神经计算，597:127969，2024年。[122] Amit Sangroya、Mouli Rastogi、C Anantaram 和 Lovekesh Vig。Guided-LIME：一种基于结构化采样、用于解释黑盒机器学习模型的混合方法。载于*CIKM*（研讨会），第17卷，2020年。[123] Muhammad Rehman Zafar 和 Naimul Khan。用于稳定可解释性的确定性局部可解释模型无关解释。机器学习与知识提取，3(3):525–541，2021年。[124] Thibault Laugel、Xavier Renard、Marie-Jeanne Lesot、Christophe Marsala 和 Marcin Detyniecki。定义事后可解释性中代理模型的局部性。*arXiv* 预印本 *arXiv:1806.07498*，2018年。

[125] Tiago Botari、Frederik Hvilshøj、Rafael Izbicki 和 Andre CPLF de Carvalho。《Me-lime：机器学习模型的有意义局部解释》。*arXiv* 预印本 *arXiv:2009.05818*，2020年。[126] Jinsung Yoon、Sercan O Arik 和 Tomas Pfister。《Rl-lim：基于强化学习的局部可解释建模》。2019年。[127] Genji Ohara、Keigo Kimura 和 Mineichi Kudo。《R-lime：用于局部可解释模型无关解释方法的矩形约束与优化》。载于国际模式识别会议，第80–95页。施普林格，2024年。[128] Giorgio Visani、Enrico Bagli 和 Federico Chesani。《Optilime：针对诊断计算机算法的优化 LIME 解释》。*arXiv* 预印本 *arXiv:2006.05714*，2020年。[129] Xuhong Li、Haoyi Xiong、Xingjian Li、Xiao Zhang、Ji Liu、Haiyan Jiang、Zeyu Chen 和 Dejing Dou。《G-lime：利用全局先验对深度神经网络进行局部解释的统计学习方法》。人工智能，314：103823，2023年。[130] Meir Yossef Levi 和 Guy Gilboa。《点云网络的快速简易可解释性》。*arXiv* 预印本 *arXiv:2403.07706*，2024年。[131] Udo Schlegel、Duy Lam Vo、Daniel A Keim 和 Daniel Seebacher。《TS-MULE：时间序列预测模型的局部可解释模型无关解释》。载于机器学习与数据库知识发现联合欧洲会议，第5–14页。施普林格，2021年。[132] Qiang Huang、Makoto Yamada、Yuan Tian、Dinesh Singh 和 Yi Chang。《Graphlime：图神经网络的局部可解释模型解释》。*IEEE* 知识与数据工程汇刊，35(7)：6968–6972，2022年。[133] Saumitra Mishra、Bob L Sturm 和 Simon Dixon。《用于音乐内容分析的局部可解释模型无关解释》。载于*ISMIR*，第53卷，第537–543页，2017年。[134] Talal Ali Ahmed Abdullah、Mohd Soperi Mohd Zahid、Ahmad F Turki、Waleed Ali、Ahmad A Jiman、Mohammed J Abdulaal、Nebras M Sobahi 和 Eyad T Attar。《Sig-lime：对 LIME 解释技术的基于信号的增强》。*IEEE Access*，12：52641–52658，2024年。[135] Patrick Knab、Sascha Marton 和 Christian Bartelt。《Dseg-lime——通过分层数据驱动分割改进图像解释》。*arXiv* 预印本 *arXiv:2403.07733*，2024年。[136] Talal AA Abdullah、Mohd Soperi Mohd Zahid、Waleed Ali 和 Shahab Ul Hassan。《B-lime：改进 LIME，以便基于心电信号对心律失常进行可解释深度学习分类》。*Processes*，11(2)：595，2023年。[137] Johannes Rabold、Hannah Deininger、Michael Siebers 和 Ute Schmid。《以语言解释丰富视觉解释关系概念——结合 LIME 与 Aleph》。载于机器学习与数据库知识发现联合欧洲会议，第180–192页。施普林格，2019年。

[138] Zeinab Dehghani、Koorosh Aslansefat、Adil Khan、Adín Ramírez Rivera、Franky George 和 Muhammad Khalid。《使用 SMILE 进行指令式图像编辑的思维映射》。*arXiv* 预印本 *arXiv:2412.16277*，2024年。[139] Haiyan Zhao、Hanjie Chen、Fan Yang、Ninghao Liu、Huiqi Deng、Hengyi Cai、Shuaiqiang Wang、Dawei Yin 和 Mengnan Du。《大语言模型的可解释性：综述》。*ACM*《智能系统与技术汇刊》，15(2)：1–38，2024年。[140] Raphael Tang、Linqing Liu、Akshat Pandey、Zhiying Jiang、Gefei Yang、Karun Kumar、Pontus Stenetorp、Jimmy Lin 和 Ferhan Ture。《DAAM 究竟是什么：通过交叉注意力解读 Stable Diffusion》。*arXiv* 预印本 *arXiv:2210.04885*，2022年。[141]Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。《随机鹦鹉的危险：语言模型会不会太大？》。载于 *2021*年 *ACM* 公平性、问责制与透明度会议论文集，第610–623页，2021年。[142] Andreas Madsen、Siva Reddy 和 Sarath Chandar。《神经自然语言处理的事后可解释性：综述》。*ACM* 计算调查，55(8)：1–42，2022年。[143]Fuseini Mumuni 和 Alhassan Mumuni。《可解释人工智能（XAI）：从内在可解释性到大语言模型》。*arXiv* 预印本 *arXiv:2501.09967*，2025。[144] Ahsan Bilal、David Ebert 和 Beiyu Lin。《用于可解释人工智能的大语言模型：全面综述》。*arXiv* 预印本 *arXiv:2504.00125*，2025。[145] Kenza Amara、Rita Sevastjanova 和 Mennatallah El-Assady。《用于审计和引导大语言模型响应的概念级可解释性》。*arXiv* 预印本 *arXiv:2505.07610*，2025。[146] Leonard Bereska 和 Efstratios Gavves。《面向人工智能安全的机制可解释性——综述》。*arXiv* 预印本 *arXiv:2404.14082*，2024年。[147] Neel Nanda、Lawrence Chan、Tom Lieberum、Jess Smith 和 Jacob Steinhardt。《通过机制可解释性衡量顿悟式泛化的进展》。*arXiv* 预印本 *arXiv:2301.05217*，2023年。[148] Guanxu Chen、Dongrui Liu、Tao Luo 和 Jing Shao。《SEER：增强大语言模型表征的自解释能力》。*arXiv* 电子预印本，第 arXiv–2502 页，2025。[149] Ronny Luss、Erik Miehling 和 Amit Dhurandhar。《细胞化你的模型：大语言模型的对比式解释》。*arXiv* 预印本 *arXiv:2406.11785*，2024年。[150] Emmanuel Ameisen、Jack Lindsey、Adam Pearce、Wes Gurnee、Nicholas L Turner、Brian Chen、Craig Citro、David Abrahams、Shan Carter、Basil Hosmer 等。《电路追踪：揭示语言模型中的计算图》。*Transformer CircuitsThread*，6，2025。

[151] Neel Nanda、Adam Templeton、Catherine Bricken、Jack Batson、Jesse Lindsey、Chris Olah 等。单义性的扩展：从 Claude 3 Sonnet 中提取可解释特征，2024年。 [在线]。[152] Nelson Elhage、Neel Nanda、Catherine Olsson、Tom Henighan、Nicholas Joseph、Mann Ben 和 Amanda Askell。等。2021年。Transformer 电路的数学框架。*arXiv* 预印本 *arXiv:2106.13316*，1。[153] Ala N Tak、Amin Banayeeanzade、Anahita Bolourani、Mina Kian、Robin Jia 和 Jonathan Gratch。大语言模型中情绪推断的机制可解释性。*arXiv* 预印本 *arXiv:2502.05489*，2025。[154] Stef De Sabbata、Stefano Mizzaro 和 Kevin Roitero。大语言模型的地理空间机制可解释性。*arXiv* 预印本 *arXiv:2505.03368*，2025。[155]Jesseba Fernando 和 Grigori Guitchounts。Transformer 动力学：理解大语言模型可解释性的神经科学方法。*arXiv* 预印本 *arXiv:2502.12131*，2025。[156] Ziwei Ji、Nayeon Lee、Rita Frieske、Tiezheng Yu、Dan Su、Yan Xu、Etsuko Ishii、Ye Jin Bang、Andrea Madotto 和 Pascale Fung。自然语言生成中的幻觉综述。*ACM* 计算调查，55(12):1–38，2023。[157] Philipp Mondorf 和 Barbara Plank。超越准确性：大语言模型推理行为评估综述。*arXiv* 预印本 *arXiv:2404.01869*，2024年。[158] Sofie Goethals、Joao Sedoc 和 Foster Provost。如果提示词不同会怎样？生成式输出特征的反事实解释。载于第*33*届*ACM*用户建模、自适应与个性化会议附属论文集，第237–242页，2025。[159] Hanjie Chen、Zhouxiang Fang、Yash Singla 和 Mark Dredze。基准测试大语言模型回答和解释具有挑战性的医学问题的能力。载于*2025*年美洲语言学会计算语言学分会会议论文集：人类语言技术（第*1*卷：长篇论文），第3563–3599页，2025。[160] Junhao Liu、Haonan Yu 和 Xin Zhang。迈向低成本的模型无关大语言模型解释生成。*arXiv* 预印本 *arXiv:2505.12509*，2025。[161] Seongmin Lee、Benjamin Hoover、Hendrik Strobelt、Zijie J Wang、ShengYun Peng、Austin Wright、Kevin Li、Haekyu Park、Haoyang Yang 和 Duen Horng Polo Chau。Diffusion Explainer：针对文本到图像 Stable Diffusion 的可视化解释。载于*2024*年*IEEE*可视化与视觉分析会议（*VIS*），第96–100页。IEEE，2024年。[162] Noyan Evirgen、Ruolin Wang 和 Xiang'Anthony Chen。从文本到像素：通过文本到图像模型解释增强用户理解。载于第*29*届智能用户界面国际会议论文集，第74–87页，2024年。

[163] Yunkai Dang、Kaichen Huang、Jiahao Huo、Yibo Yan、Sirui Huang、Dongrui Liu、Mengxi Gao、Jie Zhang、Chen Qian、Kun Wang 等。可解释且具可解释性的多模态大语言模型：全面综述。*arXiv* 预印本 *arXiv:2412.02104*，2024年。[164] RD Chowdhury。Daam-image2image：DAAM 在扩散模型图像自注意力中的扩展，2022年。[165] Yi Zhang、Zhen Chen、Chih-Hong Cheng、Wenjie Ruan、Xiaowei Huang、Dezong Zhao、David Flynn、Siddartha Khastgir 和 Xingyu Zhao。可信的文本到图像扩散模型：及时且聚焦的综述。*arXiv* 预印本 *arXiv:2409.18214*，2024年。[166]Susan Hao、Renee Shelby、Yuchi Liu、Hansa Srinivasan、Mukul Bhutani、Burcu Karagol Ayan、Ryan Poplin、Shivani Poddar 和 Sarah Laszlo。文本到图像模型中的危害放大。*arXiv* 预印本 *arXiv:2402.01787*，2024年。[167] Zhijie Wang、Yuheng Huang、Da Song、Lei Ma 和 Tianyi Zhang。PromptCharm：通过多模态提示词与细化实现文本到图像生成。载于 *2024*年*CHI*计算系统中的人为因素会议论文集，第1–21页，2024年。[168] Zhiyuan Chang、Mingyang Li、Junjie Wang、Yi Liu、Qing Wang 和 Yang Liu。通过注意力引导的特征增强修复文本到图像扩散模型中的灾难性忽视。*arXiv* 预印本 *arXiv:2406.16272*，2024年。[169] Laurynas Karazija、Iro Laina、Andrea Vedaldi 和 Christian Rupprecht。用于零样本开放词汇分割的扩散模型。*arXiv* 电子论文，第arXiv–2306页，2023年。[170] Laurynas Karazija、Iro Laina、Andrea Vedaldi 和 Christian Rupprecht。用于开放词汇分割的扩散模型。载于欧洲计算机视觉会议，第299–317页。施普林格，2024年。[171] Hugo D Rodríguez、Francisco J Iriarte、Miguel Ortiz、Luis Unzueta 和 Sean Gaines。探索扩散模型生成物人物检测训练数据集的能力边界。载于安全与防务应用人工智能 *II*，第13206卷，第301–313页。SPIE，2024年。[172] Luca Barsellotti、Roberto Amoroso、Lorenzo Baraldi 和 Rita Cucchiara。FOSSIL：通过合成参考检索实现免费开放词汇语义分割。载于 *IEEE/CVF* 计算机视觉应用冬季会议论文集，第1464–1473页，2024年。[173] Xingyi Yang 和 Xinchao Wang。作为流均衡化的组合式视频生成。*arXiv* 预印本 *arXiv:2407.06182*，2024年。[174] Advaitha Vetagiri、Prateek Mogha 和 Partha Pakray。Multilate Classifier：一种新型 CNN-BiLSTM 与基于 ResNet 的多模态分类器集成，用于人工智能生成的仇恨言论检测。可在 *SSRN 4951337* 获取，2024年。

[175]Fabio Quattrini、Vittorio Pippi、Silvia Cascianelli 和 Rita Cucchiara。Alfie：以无门槛方式普及 RGBA 图像生成。载于欧洲计算机视觉会议，第 38–55 页。施普林格，2024年。[176] Gunshi Gupta、Karmesh Yadav、Yarin Gal、Dhruv Batra、Zsolt Kira、Cong Lu 和 Tim GJ Rudner。预训练文生图扩散模型是用于控制的多用途表征学习器。神经信息处理系统进展，37：74182–74210，2024年。[177] Lev Tankelevitch、Viktor Kewenig、Auste Simkute、Ava Elizabeth Scott、Advait Sarkar、Abigail Sellen 和 Sean Rintel。生成式人工智能的元认知需求与机遇。载于*2024*年 *CHI* 计算系统中的人为因素会议论文集，第 1–24 页，2024年。[178] Shilin Sun、Wenbin An、Feng Tian、Fang Nan、Qidong Liu、Jun Liu、Nazaraf Shah 和 Ping Chen。多模态可解释人工智能综述：过去、现在与未来。*arXiv* 预印本 *arXiv:2412.14056*，2024年。[179] Rémi Kazmierczak、Eloïse Berthier、Goran Frehse 和 Gianni Franchi。可解释性与视觉基础模型：综述。《信息融合》，122：103184，2025。[180] Zongxia Li、Xiyang Wu、Hongyang Du、Fuxiao Liu、Huy Nghiem 和 Guangyao Shi。先进大型视觉语言模型综述：基准评估与挑战。载于计算机视觉与模式识别会议论文集，第 1587–1606 页，2025。[181] Yixin Liu、Chenrui Fan、Yutong Dai、Xun Chen、Pan Zhou 和 Lichao Sun。面向抵御未经授权的文生图扩散合成的稳健不可感知扰动。*arXiv* 预印本 *arXiv:2311.13127*，3，2023年。[182] Jielin Qiu、Yi Zhu、Xingjian Shi、Florian Wenzel、Zhiqiang Tang、Ding Zhao、Bo Li 和 Mu Li。分布偏移下多模态图文模型鲁棒性的基准测试。*arXiv* 预印本 *arXiv:2212.08044*，2022年。[183] Matt Kusner、Yu Sun、Nicholas Kolkin 和 Kilian Weinberger。从词嵌入到文档距离。载于国际机器学习会议，第 957–966 页。PMLR，2015年。[184] Maxime Oquab、Timothée Darcet、Théo Moutakanni、Huy Vo、Marc Szafraniec、Vasil Khalidov、Pierre Fernandez、Daniel Haziza、Francisco Massa、Alaaeldin El-Nouby 等。DINOv2：无需监督学习稳健视觉特征。*arXiv* 预印本 *arXiv:2304.07193*，2023年。[185] Mathilde Caron、Hugo Touvron、Ishan Misra、Hervé Jégou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。自监督视觉 Transformer 中的新兴特性。载于*IEEE/CVF* 国际计算机视觉会议论文集，第 9650–9660 页，2021年。[186] Yossi Rubner、Carlo Tomasi 和 Leonidas J Guibas。推土机距离作为图像检索度量。国际计算机视觉期刊，40(2)：99–121，2000年。

[187] Martin Arjovsky、Soumith Chintala 和 Léon Bottou。Wasserstein 生成对抗网络。载于国际机器学习会议，第 214–223 页。PMLR，2017年。[188] Dimitrios Alivanistos、Selene Báez Santamaría、Michael Cochez、Jan-Christoph Kalo、Emile van Krieken 和 Thiviyan Thanapalasingam。将提示作为探测：利用语言模型构建知识库。*arXiv* 预印本 *arXiv:2208.11057*，2022年。[189] Kaijie Zhu、Jindong Wang、Jiaheng Zhou、Zichen Wang、Hao Chen、Yidong Wang、Linyi Yang、Wei Ye、Yue Zhang、Neil Zhenqiang Gong 等。Promptbench：评估大语言模型抵御对抗性提示鲁棒性的研究。*arXiv* 电子预印本，第 arXiv–2306 页，2023年。[190] Gabriel Peyré、Marco Cuturi 等。计算最优传输：及其在数据科学中的应用。《机器学习基础与趋势》，11(5-6)：355–607，2019年。[191] Robert J Tibshirani 和 Bradley Efron。自助法导论。统计学与应用概率专著，57(1)：1–436，1993年。[192] Eric Gilleland。统计推断的自助法。第二部分：极值分析。《大气与海洋技术杂志》，37(11)：2135–2144，2020年。[193] Ishaan Gulrajani、Faruk Ahmed、Martin Arjovsky、Vincent Dumoulin 和 Aaron C Courville。改进 Wasserstein 生成对抗网络的训练。神经信息处理系统进展，30，2017年。[194] Trevor Hastie、Robert Tibshirani、Jerome H Friedman 和 Jerome H Friedman。《统计学习要素：数据挖掘、推断与预测》，第 2 卷。施普林格，2009年。[195] Md Ahsan Habib、Md Anwar Hussen Wadud、Md Fazlul Karim Patwary、Mohammad Motiur Rahman、M Firoz Mridha、Yuichi Okuyama 和 Jungpil Shin。探索文本到图像合成的进展：生成对抗网络演进的深度综述。*IEEEAccess*，2024年。[196] Benjamin Sanchez-Lengeling、Jennifer Wei、Brian Lee、Emily Reif、Peter Wang、Wesley Qian、Kevin McCloskey、Lucy Colwell 和 Alexander Wiltschko。评估图神经网络的归因。神经信息处理系统进展，33：5898–5910，2020年。[197] Ruth C Fong 和 Andrea Vedaldi。通过有意义的扰动对黑箱进行可解释说明。载于*IEEE* 国际计算机视觉会议论文集，第 3429–3437 页，2017年。[198] James A Hanley 和 Barbara J McNeil。受试者工作特征（ROC）曲线下面积的含义与用途。《放射学》，143(1)：29–36，1982年。[199] David MW Powers。评估：从精确率、召回率和 F度量到 ROC、知情度、标记度与相关性。*arXiv* 预印本 *arXiv:2010.16061*，2020年。

[200]Seyed Mohammad Ahmadi、Koorosh Aslansefat、Ruben Valcarce-Dineiro 和 Joshua Barnfather。使用 smile 实现点云神经网络的可解释性：基于局部解释的统计模型无关可解释性。*arXiv* 预印本 *arXiv:2410.15374*，2024年。[201] Norman R Draper 和 Harry Smith。应用回归分析，第 326 卷。John Wiley & Sons，1998年。[202] RR Biometrics19763211Hocking。线性回归中变量的分析与选择。*Biometrics32*，第 1–49 页，1976年。[203] Cort J Willmott 和 Kenji Matsuura。评估模型平均性能时，平均绝对误差（mae）优于均方根误差（rmse）的优势。气候研究，30(1):79–82，2005年。[204] Ian Goodfellow、Yoshua Bengio、Aaron Courville 和 Yoshua Bengio。深度学习，第 1 卷。MIT Press，剑桥，2016年。[205] Gleb Beliakov、Ana Pradera、Tomasa Calvo 等。聚合函数：从业者指南，第 221 卷。施普林格，2007年。[206] Anthropic。Claude 模型概览。https://www.anthropic.com，2023年。[207] Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。衡量大规模多任务语言理解。*arXiv* 预印本 *arXiv:2009.03300*，2020年。[208]Peiyuan Liu。Kaggle 上的 MMLU 数据集。https://www.kaggle.com/datasets/peiyuanliu2001/mmlu-dataset，2023年。访问日期：2025-05-23。[209] Marco Cuturi 和 Arnaud Doucet。快速计算 Wasserstein 重心。载于国际机器学习会议，第 685–693 页。PMLR，2014年。[210] Adly Templeton。扩展单义性：从 Claude 3 Sonnet 中提取可解释特征，2024年。 [在线]。[211] Zifeng An、Zijing Xu、Eric Fan 和 Qi Cao。微调 InstructPix2Pix 以实现高级图像着色。*arXiv* 预印本 *arXiv:2312.04780*，2023年。[212] Xizhe Zhang、Siddartha Khastgir、Justin-Kiyoshi Tiele、Kazuhito Takenaka、Tasuku Hayakawa 和 Paul Jennings。用于自动驾驶系统的运行设计域及基于行为的场景生成。*IEEEAccess*，12:10652–10663，2024年。[213] Philip Koopman 和 Michael Wagner。自动驾驶车辆测试与验证面临的挑战。*SAE InternationalJournalof Transportation Safety*，4(1):15–24，2016年。[214] Hong Wang、Wenbo Shao、Chen Sun、Kai Yang、Dongpu Cao 和 Jun Li。自动驾驶车辆新兴安全挑战综述：预期功能安全。*Engineering*，33:17–34，2024年。[215] Georg Stettinger、Patrick Weissensteiner 和 Siddartha Khastgir。高风险人工智能系统的可信性保证评估。*IEEE Access*，12:22718–22745，2024年。

[216] Celso Cancela-Outeda。欧盟《人工智能法案》：协同治理框架。物联网，27:101291，2024年。