arxiv:2509.15160v1 中译本

以评估为中心的科学可视化代理范式

Kuangshi Ai* Haichao Miao† Zhimin Li‡ Chaoli Wang§ Shusen Liu¶

Univ. Notre Dame LLNL Vanderbilt Univ. Univ. Notre Dame LLNL

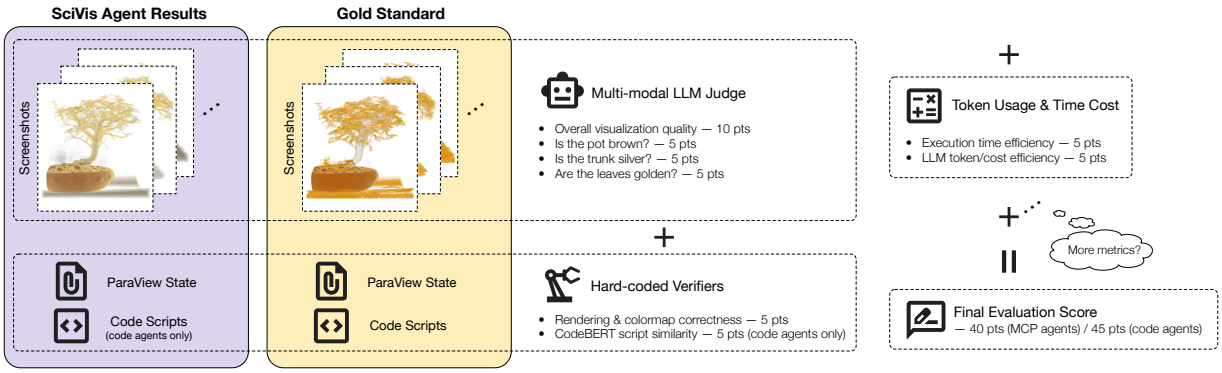

图 1：一个在 Bonsai 数据集上评估 SciVis 代理的示例案例。代理执行体积可视化并调整传输函数以实现目标“一盆带有棕色花盆、银色枝干和金色叶子的树。”。评估结合了：(1) 一个多模态 LLM 裁判用于评价可视化质量，(2) 硬编码验证器用于检查可视化原语和技术的正确性，以及 (3) 令牌使用和执行时间来衡量系统性能。我们主张纳入多方面的指标并开发一个系统的基准以评估 SciVis 代理。

摘要

多模态大型语言模型（MLLMs）的最新进展使能够将用户意图转化为数据可视化的高度复杂的自主可视化代理成为可能。然而，由于缺乏全面的大规模基准来评估实际能力，在科学可视化（SciVis）中衡量进步和比较不同代理仍然具有挑战性。这份立场文件考察了对 SciVis 代理所需的各类评估，概述了相关挑战，提供了一个简单的概念验证评估示例，并讨论了如何通过评估基准促进代理自我改进。我们倡导更广泛的协作来开发一个能够不仅评估现有能力，而且推动创新并刺激该领域未来发展的科学可视化代理评估基准。

Index Terms: 大语言模型，科学可视化代理，工具使用，评估。

*e-mail: kai@nd.edu
†e-mail: miao1@llnl.gov
‡e-mail: zhimin.li@vanderbilt.edu
§e-mail: chaoli.wang@nd.edu
¶e-mail: liu42@llnl.gov

1 介绍

许多机器学习（ML）和人工智能（AI）的重大进展始于建立了全面且具有挑战性的基准，如 ImageNet [8]，这些基准推动了深度学习革命。这些基准不仅提供了标准化的方法来评估和比较不同的技术，还推动了现有技术和工具所能实现的极限。最近出现的多模态大型语言模型（MLLMs）使得能够将自然语言指令转化为复杂科学可视化（SciVis）结果的新一代自主可视化代理成为可能 [25, 28, 1]。然而，评估这些代理存在一个基本挑战：尽管科学可视化通常涉及带有突发性见解的探索性分析，有意义的基准测试却需要可重复的任务和可测量的结果。随着可视化代理从研究原型转向科学家和工程师使用的实用工具，这一评估差距正变得至关重要。

基于 Dhanoa 等人提出的更广泛的分类法 [9]，我们采纳了一个更为聚焦且实用的定义用于评估目的。我们将科学可视化代理定义为：一个能够解释人类用户自然语言意图的人工智能系统，自主地与 SciVis 管道交互以生成满足用户指定分析目标的可视化结果。此定义有意地限制了范围，以实现具体、可重复的评估，同时捕捉这些代理必须具备的基本能力。重要的是，我们的重点在于完全自主执行的情景，在这种情景下，代理必须在初始指令之后无需额外的人类

干预来完成任务，这使得一致性和可重复性的基准测试成为可能。当前针对可视化代理的评估方法不足以处理科学可视化任务。现有的基准主要集中在简单的绘图任务 [7, 35, 12] 或通用的数据科学工作流程 [17, 14] 上，未能应对科学可视化工作者的独特复杂性。与基本绘图不同，科学可视化工作流需要复杂的数据显示转换、多样化的渲染技术、多维参数映射以及仔细的选择视角，所有这些都必须按精确的顺序应用以产生有意义的科学见解。尽管存在局限性，现有的基准已经揭示了当前代理能力的基本差距，从对视觉输出的感知困难 [15, 19] 到支撑大语言模型代理工具使用机制的脆弱性 [36, 31]。综合评估框架的缺失不仅阻碍了该领域的进步，也使得在需要准确性与可重复性的关键科学应用中可靠部署这些代理变得不可能 [20, 18]。

这份立场文件主张在科学可视化代理开发方法上进行根本性的转变：评估必须成为主要的设计驱动因素，而不仅仅是事后的验证。我们的目标是促进广泛的合作，建立评估标准，将科学可视化工具有实验工具转变为可靠的科学研究仪器。在这份立场文件中，我们呼吁制定一个更全面的评估基准，涵盖多个维度：任务复杂性（从简单的参数调整到复杂的多步骤流程），领域覆盖范围（从实验数据到计算模拟）以及评估方法学（从输出质量到过程效率）。此外，我们设想如何通过自动化反馈循环使这些基准能够实现代理自我改进，这可能最终导致自主的代理自我提升 [16]。

2 相关工作

AI 代理的评估变得越来越重要，针对特定领域系统（如可视化）的需求与通用代理的需求有所不同。我们将先前的工作组织为（1）以可视化/人机交互为重点的评估和（2）通用代理基准，并强调了需要一种以 SciVis 为中心的评估方法的动力。

2.1 可视化与 HCI 代理评估

可视化特定的基准测试和代理。最近的基准测试表明，大语言模型能够感知基本图表，但在核心可视化任务上仍然存在问题：VisEval [7] 显示了在图表可读性和生成方面的失败，Drawing Pandas [12] 暴露了代码执行问题，而 MatPlotAgent [35] 则主张需要专门的评估而不仅仅是通用代码指标。以可视化为重点的代理管道已经被探索过，从 AVA 的感知驱动优化到代码生成和工具使用的 SciVis 代理如 ChatVis 和 ParaView-MCP [26, 28, 25]。除了输出质量之外，几项研究还探讨了基础能力：可视化素养，并且这些评估记录了在可视化理解方面的持续限制 [15, 33, 19]。总的来说，这些局限性表明需要制定考虑结果质量和过程（例如工具使用）的评估协议，当代理依赖多步骤 SciVis 管道时。

人与 AI 的合作。人机交互研究带来了评估代理工作流中交互质量和可解释性的方法。Magentic-UI 强调了人在回路中的评估；NLI4VolVis 展示了体积的多代理、开放词汇表互动；并且已经证明可解释性可以提高人类-AI 团队的任务表现 [30, 1, 32]。评估对话助手的方法论指导（例如，模态效应和协作动力学）进一步展示了我们如何评估支持设计和分析工作流的代理 [23, 22]。

2.2 通用代理评估框架

综合代理基准。更广泛的代理评估提供了基础设施，但很少能捕捉到 SciVis 的探索性和分析驱动的需求。AgentBench 和 AgentBoard 针对多轮推理和任务成功；GAIA 强调现实世界的复杂性；而 τ-bench 分析工具-代理-用户交互，并揭示了试验中重要的一致性下降 [27, 6, 29, 36]。多模态和网络任务设置（如 VisualWebArena）以及大规模的多模态理解（如 MMMU）突显了 SciVis 代理必须达到的专家级视觉推理中的剩余差距 [21, 37]。工具使用研究还揭示了 API 基础的脆弱性 [31]。

判断、可靠性和可重复性。“LLM 作为评判者” 与人类偏好有合理的关联，但在视觉定位和稳定性方面存在已知限制 [38, 13, 34]。对于 SciVis 代理而言，由于小视角/编码变化可能具有语义上的重要性，这促使了结合 LLM 评判与引擎状态验证的混合协议的产生。最后，可重复性仍然是可视化和系统评估中的一个跨领域问题，强调需要透明框架和基准 [18, 20]，这是我们在这篇立场文件中为可视化代理提出的。

3 科学可视化智能体评估的分类法

基于我们对 SciVis 代理的定义及实现具体进展的目标，我们提出了一种以实践为导向的评估任务分

类法，这可能有助于推动未来的可视化代理开发。这种分类法具有双重目的：它为比较不同的代理架构建立了标准化指标，并提供了可用于改进代理的实际反馈。我们将评估任务组织成两大类：基于结果的和基于过程的。

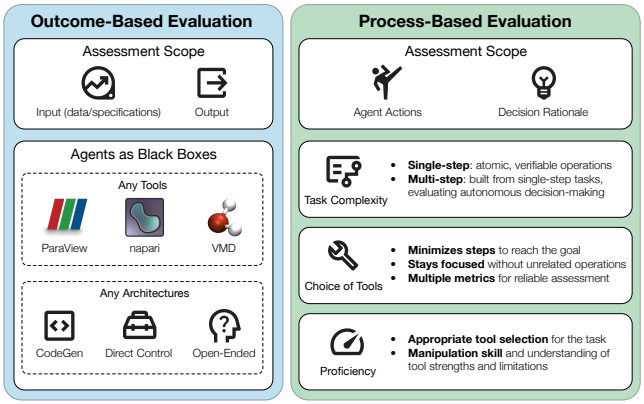

图 2：SciVis 代理评估的分类，分为两个视角：基于结果的评估，评估输入规范与最终输出之间的关系，同时将代理视为黑盒；以及基于过程的评估，分析代理的动作路径、决策理由和中间行为。

3.1 基于结果的评价

基于结果的评估完全专注于输入数据/规格与最终输出之间的关系，将代理视为一个黑盒。这种做法对于确保在异构代理架构中广泛适用至关重要，从生成可执行代码 [28] 到直接操作工具界面 [25] 或更智能的系统（尚未开发）基于任务需求自主选择其方法。通过抽象实现细节，基于结果的度量标准能够对基本不同的代理设计进行直接比较，同时保持关注最重要的方面：可视化输出的质量和正确性。在可视化问题中，一个关键挑战可能源于非唯一的结果，即揭示相同见解的不同可视化结果，这为基于结果的评估带来了模糊性。为了使代理解决方案具体化，我们可以增加约束条件以缩小解决方案范围。或者，我们可以专注于更短、更集中的任务，没有分支的可能性，或从预定义的中间结果开始。

3.2 基于过程的评估

基于过程的评估检查代理的行为和理由，提供了对解决方案如何实现的洞察，而不仅仅是产生了什么。这种细粒度分析特别有助于识别失败模式、理解泛化能力，并指导代理架构的迭代改进。基于过程的评估带来了额外的复杂性。我们制定了以下子类别，其重点如下。

任务复杂性 自然地划分了基于过程的评估，即单步与多步任务。单步任务评估原子操作，例如加载数据集并应用特定过滤器。多步任务由几十甚至数百个相互依赖的单步任务组成，可能涉及回溯和迭代细化。虽然多步任务允许多样化的探索轨迹，每个构成单步任务的目标应保持可验证性和一致性，例如由 VeriGUI 数据集 [24] 所示。**工具选择** 代表了可视化流水线的另一个关键维度，因为 SciVis 包含了一套广泛的专用软件/包。评估可以根据目标工具进行划分，如 ParaView 用于通用 SciVis，napari 用于生物医学成像，VMD 用于分子动力学等。高级代理可能通过自主选择最适合给定任务的工具来展示元能力，这不仅需要操作和使用不同工具的能力，还需要了解每个工具的优势和局限性。

熟练程度 是过程评估的另一个重要维度。代理是否使用了比严格要求更多的步骤？它在偶然找到正确解决方案的过程中是否执行了与所陈述目标无关的任务？关于实际测量，我们可以间接依赖代币使用量和时间或步长。

4 代理评估的有效性

即使是最全面的基准测试，如果评估不可信，即未能准确反映给定代理的全部能力和缺点，也会毫无用处。评估的有效性可以通过三个互补的角度来考察：*准确性*，这涉及个别评估结果的可靠性；*覆盖范围*，这表明基准涵盖了多少潜在的实际使用场景；以及*成本效益*，这传达了需要在计算和人力投入量与实现良好准确性和覆盖面之间取得平衡的需求。

准确率 需要减少不确定性并确保稳健的评估信号。一种可行的方法是使用 MLLM 评判员来评估可视化质量，这些评判员已经显示出与人类偏好有很强的一致性 [38, 13]。然而，最近的研究 [34, 4] 表明，尽管这些模型具有很大的潜力，但在视觉感知和定位方面仍然存在显著的局限性。它们可能会忽略细微的视觉编码、误解释空间关系或将风格变化与语义差异混淆，并且它们的判断会随着提示措辞或图像呈现顺序的变化而改变。为了提高可靠性，可以使用自动验证来对比可视化引擎的内部状态（基于过程的评估）。例如，一个特定情况下的 Python 脚本可以确认 ParaView 等值面已在正确的值生成并适当着色。可以通过将生成的脚本与黄金标准参考脚本进行比

较以及验证其执行结果，为代码生成代理如 ChatVis 添加定量检查。尽管人工评估成本高昂，但在模糊或高风险的情况下可能仍然需要。

覆盖范围 关注评估是否涵盖了整个科学可视化任务和交互模式的范围。测试设计应从将代表性的用户意图映射到各种技术（例如，体积渲染、流线追踪、等值面提取）开始。基于结果的评估仅指定数据集和任务描述，而不限制代理如何实现目标。这允许公平地评估具有不同能力的代理，无论它们是生成代码来与可视化引擎交互还是直接调用高级工具。确保评估覆盖范围得益于与可视化任务分类法的自上而下的对齐以及自下而上的分析哪些可视化原语、技术和交互模式得到了执行。这种双重视角有助于识别差距并保持基准代表实际应用场景。

成本效益分析 解决了科学可视化代理评估中的两个关键实际约束。首先，为探索性科学可视化任务定义真实值本质上是具有挑战性的：与确定性操作不同，这些任务通常允许存在多个同样有效的可视化、视角或参数设置。这种模糊性使自动验证器的创建变得复杂，并可能需要昂贵的人工判断来建立公平的评分标准。其次，进行全面评估——涵盖各种数据集、可视化技术和代理配置——需要大量的计算资源。反复启动可视化引擎、处理大型科学数据集和执行复杂的可视化管道可能导致运行时间和金钱成本过高。基准测试必须达到平衡，在提供可操作且具有代表性的评估的同时尽量减少开销以支持快速迭代开发周期。

5 一个用于科学可视化代理评估的示例

为了在一个具体的环境中进行讨论，我们概述了一个用于 SciVis 代理的示例基准。目的是展示如何将*结果质量*、*过程验证*和*系统效率*结合成一个统一的评估协议。

5.1 框架设计

我们将此设置作为一个示例，展示如何为 SciVis 代理设计一个符合第 3 节分类法的评估协议。任务被分解成更小、可控制的检查点以确定故障点，并且代理通过模型上下文协议（MCP）[2] 或直接代码执行在受控沙箱中操作。

对于**结果质量**，重点在于最终可视化是否在准确性、语义正确性和可解释性方面满足预期目标。评估的因素包括色彩图选择、视角以及适当使用可视化原语。我们在实现中采用经过指令调整的多模态 LLM 评判员，以符合人类偏好。这些模型被提示带有领域特定的评估标准、真实可视化的图像和代理的输出，然后要求它们分配质量分数。

对于**过程验证**，重点在于代理的中间操作和应用的技术是否满足明确的任务要求。这包括通过案例特定的硬编码验证器检查可视化引擎的内部状态，以确认可视化原语（例如等值面）和技术（例如体积渲染）的正确使用。对于生成代码的代理，额外的检查会将生成的脚本与黄金标准参考进行比较，并验证其执行结果。

对于**系统效率**，我们追踪每次运行的运行时间、令牌使用情况和货币成本。这些措施补充了基于准确性的指标，提供了关于代理可视化系统的可扩展性、成本效益和实际部署可行性的见解。

5.2 示例案例研究：盆景体积渲染

作为所提出的 SciVis 代理评估框架的具体示例，我们考虑在 Bonsai 数据集上使用 ParaView 作为可视化管道进行体积渲染任务。对两个代理进行了评估：ChatVis [28]，它生成 Python 脚本来与 ParaView 的原生 API 交互，以及通过 MCP 服务器操作的 ParaView-MCP [25]，这是一种比原始 API 更高层次的抽象。在这个示例中，两个代理都使用 GPT 系列中的模型，即 GPT-5、GPT-4.1 和 GPT-4o 作为其核心 LLM。每个实验重复进行 10 次以确保统计上的稳健性。指示这些代理加载带有给定参数的 Bonsai 数据集，执行体积渲染，并调整传输函数以实现目标可视化：“一棵盆栽树，棕色的花盆，银色的树枝和金色的叶子。”将生成的 ParaView 状态保存以供后续评估。

任务完成后，整体可视化质量通过使用指令调整的多模态大语言模型裁判（如 GPT-4o）进行评估，该裁判会同时查看真实图像和代理生成的结果。裁判根据明确的标准对输出进行评价：总体目标是否达成、花盆是否为棕色、树枝是否为银色以及叶子是否为金色。这些分数构成了最终评估指标的一部分。

为了增强评估的鲁棒性，通过 pvpython 执行硬编码验证脚本补充了基于大语言模型的评估。重新加载保存的 ParaView 状态以确认正确的体积渲染配置

和准确的颜色映射设置。对于像 ChatVis 这样的代码生成代理，我们还计算生成的脚本与黄金标准参考脚本之间的基于 CodeBERT 的 [11] 相似度。虽然这些特定情况下的检查大幅提高了可靠性，但它们需要额外的手动努力来设计和维护。性能指标包括令牌使用、货币成本和任务完成时间，直接反映了用户感知的延迟和部署此类代理的实际可行性，因此被记录下来。每个指标都被分配了一个点值，而这些点数之和构成了最终评估分数（参见图 1）。表 1 显示了基于 MCP 的代理虽然提供了稳定、高质量的结果，但对其复杂工具链的依赖导致高延迟，限制了现实世界的部署。相比之下，ChatVis 缺乏视觉能力并且实时生成代码，通常能更快地完成任务，但却增加了令牌使用并降低了可视化质量。

尽管结果集中在 GPT 系列模型上，我们也评估了其他模型家族，如 Claude、LLaMA 和 Qwen，并观察到在 SciVis 任务性能方面存在显著差异。当与高度抽象的工具环境（如 ParaView-MCP）结合使用时，小型语言模型 (SLMs) 通常能够以较低的延迟和减少的成本达到相当的可视化质量。在这种情况下，强大的推理能力并不那么关键，使得 SLMs 可以有效地完成可视化任务 [3]。然而，像 GPT-5 这样的大型模型具备先进的视觉理解能力确实带来了更好的可视化结果。鉴于缺乏对 SciVis 代理系统的系统评估协议，我们主张创建一个全面的基准来指导未来的研究和发展。

表 1：对两个 SciVis 代理在 Bonsai 任务上使用 GPT 系列模型作为主干的评估结果。每个实验重复 10 次。报告了标记使用和时间成本的均值和方差，以及每种设置下的最佳 SciVis 评估分数。SR 表示成功率。结果于 2025 年 9 月 17 日通过 OpenAI API 获得。

代理	模型	输入/输出标记	平均成本	时间（秒）	样本率	得分
MCP-based	GPT-5	220 ± 0 / 838 ± 203	$0.0087	301.7 ± 32.3	10/10	27/40
ChatVis	GPT-5	2430 ± 847 / 2994 ± 956	$0.0330	158.9 ± 29.9	10/10	25/45
MCP-based	GPT-4.1	220 ± 0 / 1460 ± 210	$0.0121	49.3 ± 8.0	10/10	21/40
ChatVis	GPT-4.1	638 ± 555 / 1217 ± 530	$0.0110	24.0 ± 5.7	10/10	23/45
MCP-based	GPT-4o	220 ± 0 / 908 ± 109	$0.0239	41.7 ± 14.2	10/10	23/40
ChatVis	GPT-4o	1945 ± 753 / 1909 ± 672	$0.0240	38.4 ± 9.4	7/10	24/45

6 评估驱动的代理设计

开发 SciVis 代理需要集成众多工具和库，每个都需要大量的工程努力来实现基于 LLM 的控制。我们提议颠覆这种通过传统开发（其中评估紧随实现之后）处理复杂性的方法：全面的评估基准可以推动整个代理设计过程。

借鉴测试驱动开发的灵感，评估驱动设计将基准用作增量代理开发的规范和框架。开发者可以逐步构建能力，在推进到复杂的多步骤工作流程之前验证单步操作，从而将一个令人生畏的工程挑战转化为可管理的迭代。每个评估目标都提供具体的指导，加速开发同时确保功能稳健。最近关于自我演化的 AI 代理 [10] 的研究，以及作为代表方法之一的用于自动代理设计的元代理 [16]，已经证明了类似方法的可行性。基于过程的评估可以识别特定的推理失败和工具选择不效率问题，而基于结果的评估则提供明确的优化目标。元代理可以分析这些结果以自动修改代码或提示。

评估与智能体开发之间是一种共生关系，随着智能体能力的增强，基准测试也在扩展以挑战新的能力。此外，当评估变得更加全面时，它们会揭示隐藏的失败模式并指导实际改进。这种共同进化确保了基准测试保持相关性，同时促使智能体发展出强大且可泛化的功能，而不是过度适应人工指标。有效的基于评估的设计需要提供细粒度、可操作反馈的基准测试，而不仅仅是二元的成功/失败信号。它们必须涵盖从确定性的单步操作到开放式的探索任务，并高效执行以支持快速迭代周期。因此，我们倡导的广泛的评估套件不仅仅是一个测量工具，而是加速开发进程的催化剂，从根本上改变了我们构建科学可视化智能体的方式。通过将评估作为主要驱动因素而非事后验证的想法，我们可以构建更强大和可靠的系统，同时显著减少开发时间和精力。

7 结论与展望

虽然我们概述的评估框架为评估和改进科学可视化代理提供了结构化的路径，但我们承认存在重要限制。通过将评估限定在完全自主场景中，而没有超出初始输入的人类交互，我们排除了人机协作的关键领域。评估人机互动——包括用户专业知识和沟通风格的变化——代表了一个超越我们当前范围的独立研究领域。然而，模拟多轮评估方法 [27] 为捕捉一些交互动态提供了有前景的方向，而无需直接与人类用户进行评估。

自主可视化代理的部署也提出了我们的框架必须解决的关键安全问题。直接访问工具和代码执行的代理可能会破坏数据并消耗过多的计算资源。我们倡导使用隔离代理执行与生产系统的沙盒评估环境，类

似于通用代理基准测试中采用的方法 [5, 39]。此外，驱动自演化的代理带来了独特的风险——自动化优化循环可能会放大有害行为或以意料之外的方式利用评估指标。这些安全问题需要在自主评估场景中进行仔细监控、有限的优化目标以及人类监督检查点，确保改进基准性能的努力与安全可靠的科学实践相一致。

创建全面的 SciVis 代理评估基准超出了任何单一研究团队的能力。科学领域的多样性、可视化工具和用例需求要求可视化研究人员、领域科学家、AI 实践者和工具开发者之间进行广泛的合作。**本立场文件旨在作为公开邀请，以促进社区共同构建此评估基准**。这样的合作可以确保基准反映真实的科学研究需求，而不是人为构建的内容，并且分散创建、验证和维护评估套件所需的大量工作。这些评估套件可以根据总体目标的发展随着新的进步而扩展或扩大。未来的研究方向包括对协作多代理系统进行评估，在这种系统中，专门的代理在复杂的可视化任务上进行协调，评估将领域特定知识整合到评估指标中的效果以更细致地判断科学洞察力，以及为超越现有技术的创意可视化方法开发基准。今天建立严格的评估框架奠定了 SciVis 代理的基础，这些代理真正增强了人类科学研究，改变了研究人员探索和理解复杂数据的方式。

Acknowledgments

本工作在美国能源部的资助下，由劳伦斯利弗莫尔国家实验室根据合同 DE-AC52-07NA27344 执行。该工作部分得到了 LLNL-LDRD（23-ERD-029, 23-SI-003）、DOE ECRP（SCW1885）、DOE DE-SC0023145 以及美国国家科学基金会 IIS-1955395、IIS-2101696、OAC-2104158 和 IIS-2401144 的支持。

参考文献

[1] K. Ai, K. Tang, and C. Wang. NLI4VolVis: Natural language interaction for volume visualization via LLM multi-agents and editable 3D Gaussian splatting. arXiv preprint arXiv:2507.12621, 2025. 2, 3

[2] Anthropic. Announcements: Introducing the model context protocol. https://www.anthropic.com/news/model-context-protocol. Accessed: 2025-08-16. 5

[3] P. Belcak, G. Heinrich, S. Diao, Y. Fu, X. Dong, et al. Small language models are the future of agentic AI. arXiv preprint arXiv:2506.02153, 2025. 6

[4] M. Berger and S. Liu. The visualization judge: Can multimodal foundation models guide visualization design through visual perception? In Proceedings of IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization, pp. 60–70, 2024. 4

[5] R. Bonatti, D. Zhao, F. Bonacci, D. Dupont, S. Abdali, et al. Windows Agent Arena: Evaluating multi-modal OS agents at scale. arXiv preprint arXiv:2409.08264, 2024. 7

[6] M. Chang, J. Zhang, Z. Zhu, C. Yang, Y. Yang, et al. AgentBoard: An analytical evaluation board of multi-turn LLM agents. In Proceedings of Advances in Neural Information Processing Systems, pp. 74325–74362, 2024. 3

[7] N. Chen, Y. Zhang, J. Xu, K. Ren, and Y. Yang. VisEval: A benchmark for data visualization in the era of large language models. arXiv preprint arXiv:2407.00981, 2024. 3

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. ImageNet: A large-scale hierarchical image database. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, 2009. 2

[9] V. Dhanoa, A. Wolter, G. M. León, H.-J. Schulz, and N. Elmqvist. Agentic visualization: Extracting agent-based design patterns from visualization systems. arXiv preprint arXiv:2505.19101, 2025. 2

[10] J. Fang, Y. Peng, X. Zhang, Y. Wang, X. Yi, et al. A comprehensive survey of self-evolving AI agents: A new paradigm bridging foundation models and lifelong agentic systems. arXiv preprint arXiv:2508.07407, 2025. 6

[11] Z. Feng, D. Guo, D. Tang, N. Duan, X. Feng, et al. CodeBERT: A pre-trained model for programming and natural languages. In Proceedings of Findings of the Association for Computational Linguistics, pp. 1536–1547, 2020. 6

[12] T. Galimzyanov, S. Titov, Y. Golubev, and E. Bogomolov. Drawing Pandas: A benchmark for LLMs in generating plotting code. In Proceedings of IEEE/ACM International Conference on Mining Software Repositories, pp. 503–507, 2025. 3

[13] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, et al. A survey on LLM-as-a-judge. arXiv preprint arXiv:2411.15594, 2024. 3, 4

[14] S. Guo, C. Deng, Y. Wen, H. Chen, Y. Chang, and J. Wang. DS-Agent: Automated data science by empowering large language models with case-based reasoning. In Proceedings of International Conference on Machine Learning, 2024. 3

[15] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do LLMs have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation. arXiv preprint arXiv:2501.16277, 2025. 3

[16] S. Hu, C. Lu, and J. Clune. Automated design of agentic systems. In Proceedings of International Conference on Learning Representations, 2025. 3, 6

[17] Y. Huang, J. Luo, Y. Yu, Y. Zhang, F. Lei, et al. DA-Code: Agent data science code generation benchmark for large language models. In Proceedings of Conference on Empirical Methods in Natural Language Processing, pp. 13487–13521, 2024. 3

[18] T. Isenberg. The state of reproducibility stamps for visualization research papers. In Proceedings of IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization, pp. 64–73, 2024. 3

[19] M. S. Islam, R. Rahman, A. Masry, M. T. R. Laskar, M. T. Nayeem, and E. Hoque. Are large vision language models up to the challenge of chart comprehension and reasoning? In Proceedings of Findings of the Association for Computational Linguistics, pp. 3334–3368, 2024. 3

[20] K. Keahey, M. Richardso, R. T. Calasanz, S. Hunold, J. Lofstead, T. Malik, and C. Perez. Report on challenges of practical reproducibility for systems and HPC computer science. arXiv preprint arXiv:2505.01671, 2024. 3

[21] J. Y. Koh, R. Lo, L. Jang, V. Duvvur, M. C. Lim, et al. Evaluating multimodal agents on realistic visual web tasks. In Proceedings of Annual Meeting of the Association for Computational Linguistics, pp. 881–905, 2024. 3

[22] E. Kuang. Crafting human-AI collaborative analysis for user experience evaluation. In Proceedings of Extended Abstracts of ACM CHI Conference on Human Factors in Computing Systems, pp. 486:1–486:6, 2023. 3

[23] E. Kuang, E. J. Soure, M. Fan, J. Zhao, and K. Shinohara. Collaboration with conversational AI assistants for UX evaluation: Questions and how to ask them (voice vs. text). In Proceedings of ACM CHI Conference on Human Factors in Computing Systems, pp. 116:1–116:15, 2023. 3

[24] S. Liu, M. Liu, H. Zhou, Z. Cui, Y. Zhou, et al. VeriGUI: Verifiable long-chain GUI dataset. arXiv preprint arXiv:2508.04026, 2025. 4

[25] S. Liu, H. Miao, and P.-T. Bremer. ParaView-MCP: An autonomous visualization agent with direct tool use. arXiv preprint arXiv:2505.07064, 2025. 2, 3, 4, 5

[26] S. Liu, H. Miao, Z. Li, M. Olson, V. Pascucci, and P.-T. Bremer. AVA: towards autonomous visualization agents through visual perception-driven decision-making. Computer Graphics Forum, 43(3):e15093, 2024. 3

[27] X. Liu, H. Yu, H. Zhang, Y. Xu, X. Lei, et al. AgentBench: Evaluating LLMs as agents. arXiv preprint arXiv:2308.03688, 2023. 3, 6

[28] T. Mallick, O. Yildiz, D. Lenz, and T. Peterka. ChatVis: Automating scientific visualization with a large language model. In Proceedings of Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 49–55, 2024. 2, 3, 4, 5

[29] G. Mialon, C. Fourrier, T. Wolf, Y. LeCun, and T. Scialom. GAIA: A benchmark for general AI assistants. In Proceedings of International Conference on Learning Representations, 2023. 3

[30] H. Mozannar, G. Bansal, C. Tan, A. Fourney, V. Dibia, et al. Magentic-UI: Towards human-in-the-loop agentic systems. arXiv preprint arXiv:2507.22358, 2025. 3

[31] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, et al. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In Proceedings of International Conference on Learning Representations, 2023. 3

[32] S. Schallmoser, C. Korner, S. Tschiatschek, and A. Holzinger. Explainable AI improves task performance in human – AI collaboration. Scientific Reports, 14:31150, 2024. 3

[33] Y. Shen and X. Yuan. Visualization generation with large language models: An evaluation. arXiv preprint arXiv:2401.11255, 2024. 3

[34] A. Szymanski, N. Ziems, H. A. Eicher-Miller, T. J.-J. Li, M. Jiang, and R. A. Metoyer. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In Proceedings of International Conference on Intelligent User Interfaces, pp. 952–966, 2025. 3, 4

[35] Z. Yang, Z. Zhou, S. Wang, X. Cong, X. Han, et al. MatPlotAgent: Method and evaluation for LLM-based agentic scientific data visualization. arXiv preprint arXiv:2402.11453, 2024. 3

[36] S. Yao, N. Shinn, P. Razavi, and K. Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. arXiv preprint arXiv:2406.12045, 2024. 3

[37] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 9556–9567, 2024. 3

[38] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In Proceedings of Advances in Neural Information Processing Systems, pp. 46595–46623, 2023. 3, 4

[39] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, et al. WebArena: A realistic web environment for building autonomous agents. In Proceedings of International Conference on Learning Representations, 2024. 7