《表象之下》：呼吁构建基于分布的视觉问答数据集

Jill Naiman信息科学 伊利诺伊大学厄巴纳-香槟分校伊利诺伊州，美国

Daniel J. Evans信息科学 伊利诺伊大学厄巴纳-香槟分校伊利诺伊州，美国

JooYoung Seo信息科学 伊利诺伊大学厄巴纳-香槟分校伊利诺伊州，美国

arXiv:2601.22218v1 [cs.CV] 29 Jan 2026

摘要——**视觉问答（VQA）已成为评估大型多模态模型（LMMs）如何理解图像的重要基准。然而，大多数视觉问答数据集关注真实世界图像或简单的图示分析，很少有数据集专注于解读复杂的科学图表。事实上，许多用于分析图表的视觉问答数据集既不包含图表背后的底层数据，也假设图表标记与底层数据之间存在一一对应关系。实际上，图表是对数据进行转换（即分析、简化和修改）的结果。这一区别带来了当前数据集未能涵盖的视觉问答推理挑战。本文主张为科学图表建立专门的视觉问答基准，其中图表标记与底层数据之间不存在一一对应关系。为此，我们调查了现有的视觉问答数据集，并指出当前领域的局限性。随后，我们基于真实数据生成合成直方图，并向人类和大型推理模型提出问题，这些问题的准确作答取决于能否访问底层数据。我们发布了该开源数据集，其中包括图示、底层数据、用于生成数据的分布参数，以及所有图示标记和文本的边界框，以供未来研究使用。**索引词——**视觉问答、多模态检索、数据集、图表**

科学图表对于在学术文章中呈现定量信息至关重要。它们有助于读者更好地理解模式、比较数值，并考察数据集中特别值得关注的方面 [1]。大型多模态模型（大型多模态模型）和大型推理模型（LRM）的最新进展，使通过自然语言交互进行视觉数据解读成为可能（例如 GPT-5 [2]）。结合图像识别与语言推理，可以利用这些系统回答有关图表和绘图的问题。

视觉问答（视觉问答）已成为评估这些系统准确性的一项重要基准。迄今为止，大多数致力于科学图表的视觉问答数据集所分析的，是LMM回答页面信息相关各类事实回忆问题的能力 [3]。然而，许多视觉问答数据集既不提供用于生成图示的底层数据访问，也缺少图表元素（例如坐标轴标签）的边界框，从而限制了未来利用这些数据集开展复杂分析。此外，即便包含底层数据，图表也仅限于图表标记与表格数据之间的一一对应。因此，当前的视觉问答数据集缺乏对分布 的覆盖，其中图表并非直接的视觉数据表示。

为弥补这一空白，我们提出了一个面向科学图表的视觉问答基准，重点关注直方图图表，因为图表标记与底层数据之间不存在一一对应关系。我们的研究调查了现有的视觉问答数据集，并指出其中缺少基于分布的图表。随后，我们基于真实数据生成了一系列简单的合成直方图，并向两位人工标注者和一个 LMM 提出统计学相关问题 [2]。我们发布了这一初步开源数据集，其中包括图示、底层数据、用于生成数据的分布参数、所有图示标记和文本的边界框，以及人工标注者和 LMM 对若干问题的回答，以供未来研究使用。

我们的研究由以下研究问题驱动：
1) 现有视觉问答数据集在多大程度上体现了科学图表所需的推理能力，又有哪些不足？2) 当无法直接关联到数据集时，大型多模态模型回答有关合成直方图问题的准确度如何？3) 大型多模态模型回答与直方图相关的统计问题时，其准确度与人类相比如何？

I. 科学图表视觉问答相关工作

视觉问答是一种研究过程，人类可以借助人工智能，通过自然语言查询 [4]来探究图示中以视觉形式呈现的信息。它将一系列任务规范化，这些任务交由人工智能模型处理，将有关视觉对象的选择题和开放式问题转化为自然语言答案。这些问题的复杂程度各不相同，同时针对图像的特定方面 [5]。

近年来，视觉问答被用于评估生成式人工智能的视觉理解能力。大型语言模型（LLM） [23]，、大型视觉模型（LVM） [24], [25] 和大型多模态模型 [26]的广泛涌现，促使研究人员探讨人工智能在视觉生成与理解中的作用。为此，视觉问答为评估更复杂的图像以及上下文中的图像 [27]提供了一个框架。

更具体地说，研究人员正在利用视觉问答来评估机器学习模型对绘图 [20], 图示 [28], 及

表 I：包含关联数据的图表视觉问答数据集汇总。

数据集名称	年份	Line	Bar	散点图	Pie	直方图	其他	边界框	关联数据	引用
ChartAssistant/ChartSFT	2024	✓	✓	✓	✓	-	✓	-	部分	[6]
图表转文本/Chart-to-text	2020/2022年	✓	✓	✓	✓	-	✓	部分	✓	[7]
巧克力	2024	✓	✓	-	-	-	-	-	✓	[8]
ChartQA/ChartQAPro	2022年/2025	✓	✓	-	✓	-	-	✓	✓	[9]
ChartCheck	2023	✓	✓	✓	✓	-	✓	-	✓	[10]
OpenCQA	2022	✓	✓	✓	✓	-	✓	✓	部分	[11]
MapQA	2022	-	-	-	-	-	✓	✓	✓	[12]
VisText	2023	✓	✓	-	-	-	✓	部分	✓	[13]
图表信息	2019-2024	✓	✓	-	-	-	✓	✓	部分	[14]
DocVQA	2021	✓	✓	-	✓	-	-	-	✓	[15]
图表事实核查	2023	-	✓	-	-	-	-	✓	✓	[16]
RealCQA	2023	✓	✓	✓	-	-	✓	✓	✓	[17]
StructChart	2023	✓	✓	-	✓	✓	✓	-	部分	[18]
LineEX	2023	✓	-	-	-	-	-	✓	✓	[19]
PlotQA	2020	✓	✓	✓	-	-	-	✓	✓	[20]
MatCha	2023	✓	✓	-	-	-	✓	-	✓	[21]
TQA	2017	-	-	-	-	-	✓	✓	✓	[22]

科学文章中的图表 [29] 进行问答的能力，运用自然语言推理，并使用合成生成的 [30] 和真实世界的数据集（例如 Statista 和 Pew， [11], [13]）。这些实验生成的问答数据集可用于评估新模型或文档集 [31]。

表 I 总结了图表视觉问答数据集，列出了图表类型，以及是否包含图表元素（数据标记、坐标轴标签等）的边界框。为简洁起见，表 I 仅列出包含某种形式关联数据的数据集。标记为“部分”的边界框表示仅标注了部分元素；标记为“部分”的关联数据则表示数据集中的部分图表提供了数据。需要指出的是，表 I 所指的边界框/关联数据是公开可用的 数据，因为若干视觉问答数据集使用的底层数据集并未公开提供（例如 ExcelChart400K、 [32]）。

我们的研究发现，直方图在当前视觉问答语料库中占比不足。尽管一些视觉问答数据集在分析中包含直方图 [18], [33],，但很少有数据集专门关注直方图；即使包含直方图，也缺少指向原始数据的链接 [33]，或者共享的数据已经过分箱处理 [18], ，即便图表对应真实世界数据也是如此 [11], [13]。由于缺乏科学文献语料库，我们提出了一个用于生成视觉问答分析合成数据的框架。

II. 图示与问题数据集生成

为测试人类和大型多模态模型的视觉问答能力，我们使用 Python 的 matplotlib 包生成合成科学图形 [20], [28], [29]。虽然使用合成数据训练模型可能会降低其在现实任务中的泛化能力 [3],，但对于使用此处呈现的基于分布的数据量化 LMM 在简单视觉问答任务上的准确性，合成数据能够支持更可控的研究。

发布的代码支持生成多种类型的绘图（直方图、折线图、散点图和等高线图），以及多种底层模型分布（线性分布、随机噪声、高斯混合模型）。代码还允许调整坐标轴标签、标题、颜色、字体、刻度（对数/线性）、误差棒、DPI、纵横比，以及每个图示中的绘图数量。

不过，为了降低本研究的复杂度，我们固定了除高斯混合模型参数以外的所有参数——高斯分布的数量设为 1 至 5，直方图的数据点范围设为 (-1,1)，并额外添加 5% 至 10% 的随机噪声。除了随机生成图示外，我们的流程还会针对包含一个或多个面板的图示中的特定面板生成问题；不过，本研究仅限于单面板图示。图 1 左侧面板展示了生成的直方图示例。

我们的图像采用 JPEG 格式，图示数据和问题则采用 JSON 格式。我们的大语言模型提示问题参考了 [34] 的研究。这些问题分为不同的“级别” [35], , 级别越高，回答问题所需的推理能力和复杂度也越高。问题以完整形式保存，同时也按“角色”（LMM 的角色）、“上下文”（回答问题时所处的上下文）、“问题”（最简形式的问题）和“格式”（答案的输出格式）拆分成句子 [36] , 以便未来的数据使用者能够在原始问题中替换自己的角色、上下文和格式。

III. 人工与大语言模型提示

我们利用生成的基于直方图的视觉问答数据，通过向一个LMM（ChatGPT-5）和两名人工标注者提出以下两个问题，来探讨我们的第二个和第三个研究问题：（1）“该图中面板中数据的中位数是多少？”；（2）“图中面板中的图所使用的数据是由多少个高斯分布生成的？”。在提示的“格式”部分增加了约束“请从1到5中选择一个整数。”，以模拟人工标注者已知高斯分布数量被限制在1至5范围内这一情况。在不施加这一限制的情况下进行的测试提高了LMM的错误率。

为了测试条形数量对准确率可能产生的影响，我们将高斯分布数量固定为 2，并将条形大小设为不同的值（10、20、45、60），每种条形数量绘制 10 张图，共计 40 个直方图示例。为了测试高斯分布数量对准确率可能产生的影响，我们将条形数量固定为 50，并将高斯分布数量设为不同的值

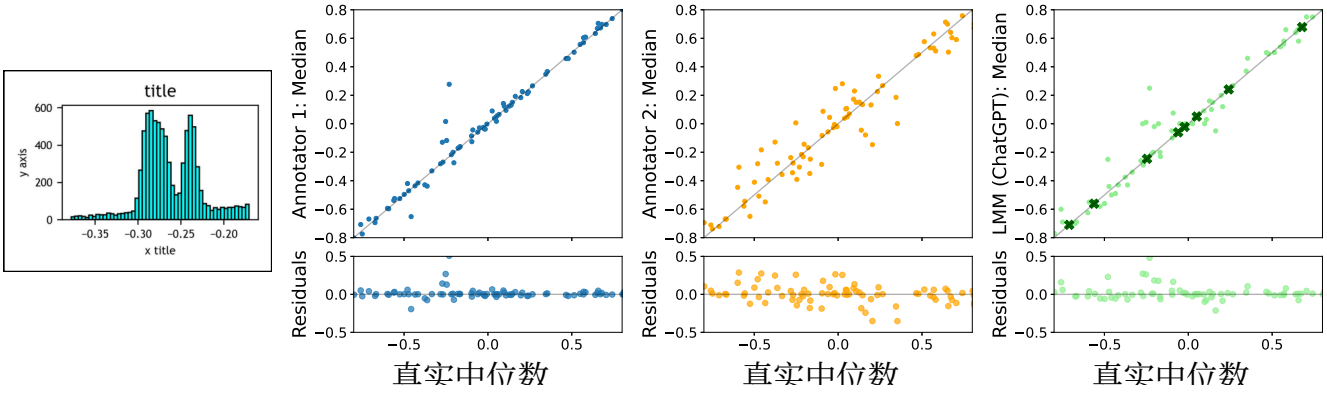


图 1：示例直方图（左），以及人工标注者（从左数第二个、从右数第二个）与 LMM（右）对 80 个直方图底层数据中位数的估计结果。底部面板显示了与真实答案的残差。LMM 图中的绿色“X”表示 LMM 在指定真实中位数处给出了格式错误的响应（LMM 拒绝回答，或其中位数与绘图范围相差约一个数量级）。

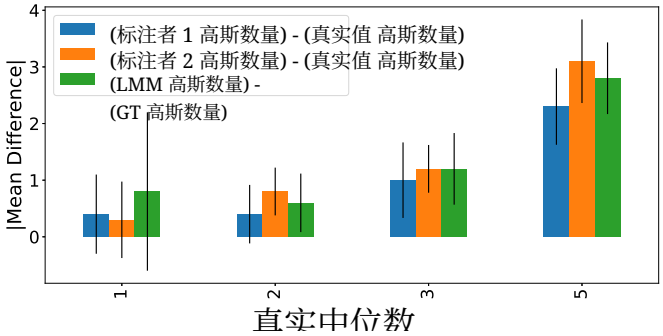


图 2：人工标注者和 LMM 对底层直方图分布中的高斯分布数量（#G）的估计与真实值（GT）之间的差异。

（1、2、3、5），每种高斯分布数量绘制 10 张图，额外生成共计 40 个直方图。

人工标注通过 Zooniverse 平台进行[37]，标注者会看到直方图图像，并被要求输入对中位数和高斯分布数量的数值估计。标注者还会在直方图上以垂直线“画出”中位数的位置。根据数据集中保存的边界框，将线的位置转换到 x 轴上。我们发现，通过绘图进行标注能更准确地估计中位数，因此下文讨论人工测量的中位数时，均指“绘制”的标注结果。

IV. 结果与讨论

图 1 展示了两位人工标注者（右侧、中间面板）和 LMM（GPT-5-nano，左侧面板）对中位数的估计结果。所有标注者的估计总体上都与真实值（实线）一致，残差（下方面板）通常聚集在真实值周围，且没有明显趋势。然而，与正态分布相比，所有标注者的残差分布峰度更高、尾部更长，并且所有残差均未通过夏皮罗–威尔克检验 [38] 的正态性检验。

因此，为了确定三个组的估计中位数是否彼此不同，我们使用非参数 Kruskal-Wallis 检验 [39]，结果显示，三个组的中位数残差之间没有显著差异。为了检验各组残差分布的方差是否存在显著差异，我们进行了 Levene 检验 [40]，发现组间方差存在显著差异。

在采用 Bonferroni-Holm 校正进行成对 Levene 检验后 [41]，我们发现标注者 2 的残差方差与标注者 1 和 LMM 均不相同。

这一结果与图 1 所示结果一致，其中标注者 2 的方差看起来大于标注者 1 和 LMM。这种差异可能源于标注者 1 与标注者 2 的背景不同——标注者 1 具有讲授统计学课程的背景，而标注者 2 没有。尽管标注者 1 与 LMM 的残差之间存在明显相似之处，但关于 LMM 与标注者 1 和标注者 2 相比是否具有同等“背景”，我们保留判断，留待更广泛的研究。

接下来，我们通过考察包含 40 个直方图的子集，检验直方图条柱数量对中位数估计的影响；该子集的高斯分布数量固定为 2，分箱大小可变，为 $n_{\text{bars}} = [10, 20, 45, 60]$。使用 Kruskal-Wallis 检验，在任意 n_{bars} 个分箱中，标注者之间残差的中位数或方差均无差异。

最后，我们使用分箱大小固定为 50 的 40 个直方图子集，检验高斯分布数量对标注者估计结果的影响。再次使用 Kruskal-Wallis 检验，对于四个“高斯数量”（高斯数量）分箱，我们发现，人类标注者或 *LMM* 标注者的中位数估计值之间均无显著差异。这与图 2 的结果一致：图中展示了标注者对四种高斯数量的估计——在给定的真实高斯数量分箱中，人类标注者（蓝色、橙色）与 LMM（绿色）对高斯分布数量的平均估计值之差，对每个分箱而言都远在误差棒（黑线）范围之内。不过，图 2 确实表明，高斯数量与所有标注者的

估计数量的能力之间存在关联：随着高斯数量增加，人类标注者和 LMM 的错误都会增加。为检验这一观察结果，我们使用“线性混合效应模型” [42]，发现即使将不同标注者的影响纳入考虑，随着真实高斯数量增加，对高斯数量的估计错误率仍显著上升。直观而言，这并不出乎意料——随着高斯分布数量增加，混合效应会使得在视觉上更难以区分它们对分布的影响。

V. 注意事项、结论与未来工作

在本研究中，我们创建并测试了一个基于分布的视觉问答数据集，并发现，对于其中一小部分问题，大型多模态模型（GPT-5）给出的答案比没有统计学背景的人类更接近具有统计学背景的人类所给出的答案。

本研究旨在呼吁创建并测试“基于分布”的视觉问答数据集；在这类数据集中，相关图表标记与生成它们所依据的数据并非一一对应。尽管本文介绍的数据集尚处于初步阶段，我们希望它能让人们关注到一种代表性不足的数据集类型，并展示大型多模态模型在估计统计问题答案方面能够达到、甚至有时超越人工标注者的能力。

为降低复杂度，我们聚焦于参数较少（例如数据范围）的直方图，并且只探讨了两个视觉问答问题。未来的研究方向包括使用微调模型、增加问题类型、采用不同数据范围和样式的直方图，以及涵盖文本和标记的所有边界框。

最后，我们的结果依赖于 GPT-5-nano，即 GPT-5 系列中最小的模型。我们对更大模型（如 GPT5-mini）的初步测试并未发现改进，反而增加了幻觉。我们将这些结果纳入数据和代码中，见 https: //github.com/ReadingTimeMachine/LLM VQA J CDL2025 – –

参考文献

[1] E. R. Tufte, 《定量信息的视觉呈现》。Graphics Press，1983年。[2] OpenAI, 《Introducing GPT-5》，2025。[3] K.-H. Huang 等, 《从像素到洞见：大型基础模型时代的自动图表理解综述》，2024。[4] A. Agrawal 等, 《VQA：视觉问答》，2016。[5] S. Antol 等, 《VQA：视觉问答》，载于*IEEE*国际计算机视觉会议论文集，2015年。[6] F. Meng等, 《ChartAssistant：通过图表转表格预训练和多任务指令微调实现的通用图表多模态语言模型》，载于计算语言学协会论文集：*ACL*，2024。[7] S. Kantharaj 等, 《Chart-to-text：用于图表摘要的大规模基准》，载于计算语言学协会年会论文集，2022年。[8] K.-H. Huang 等, 《大型视觉语言模型理解图表吗？分析并纠正图表说明生成中的事实错误》，2024, arXiv:2312.10160。[9] A. Masry、X. L. Do、J. Q. Tan、S. Joty 和 E. Hoque, 《ChartQA：一个结合视觉与逻辑推理的图表问答基准》，载于计算语言学协会论文集：*ACL*，2022年。[10] M. Akhtar、N. Subedi、V. Gupta、S. Tahmasebi、O. Cocarascu 和 E. Simperl, 《ChartCheck：针对真实世界图表图像的可解释事实核查》，2024，arXiv:2311.07453。[11] S. Kantharaj、X. L. Do、R. T. K. Leong、J. Q. Tan、E. Hoque 和 S. Joty, 《OpenCQA：基于图表的开放式问答》，2022年。[12] S. Chang、D. Palzer、J. Li、E. Fosler-Lussier 和 N. Xiao, 《MapQA：一个用于分级统计地图问答的数据集》，2022年。[13] B. Tang、A. Boggust 和 A. Satyanarayan, 《VisText：一个用于语义丰富的图表说明生成的基准》，载于计算语言学协会年会论文集，2023年。[14] K. Davila、F. Xu、S. Ahmed、D. A. Mendoza、S. Setlur 和 V. Govindaraju, 《ICPR 2022：从信息图表（图表信息图）中提取原始表格的挑战》，载于《国际模式识别会议论文集》，2022年。[15] M. Mathew、D. Karatzas 和 C. V. Jawahar, 《DocVQA：一个用于文档图像视觉问答的数据集》，2021年。[16] M. Akhtar、O. Cocarascu 和 E. Simperl, 《阅读并推理图表图像，以实现基于证据的自动事实核查》，载于《计算语言学协会论文集：EACL》，2023年。[17] S. Ahmed、B. Jawade、S. Pandey、S. Setlur 和 V. Govindaraju, 《RealCQA：作为一阶逻辑试验台的科学图表问答》，载于《文档分析与识别——ICDAR》，2023年。[18] R. Xia 等, 《StructChart：关于视觉图表理解中的模式、度量指标和增强》，2024，arXiv:2309.11268。[19] S. V. P、M. Y. Hassan 和 M. Singh, 《LineEX：从科学折线图中提取数据》，载于《IEEE/CVF冬季计算机视觉应用会议论文集》，2023年。[20] N. Methani、P. Ganguly、M. M. Khapra 和 P. Kumar, 《PlotQA：对科学绘图进行推理》，载于《IEEE/CVF冬季计算机视觉应用会议论文集》，2020年。[21] F. Liu 等, 《MatCha：通过数学推理和图表逆渲染增强视觉语言预训练》，2023，arXiv:2212.09662。[22] A. Kembhavi、M. Seo、D. Schwenk、J. Choi、A. Farhadi 和 H. Hajishirzi, 《你比六年级学生更聪明吗？面向多模态机器理解的教材问题回答》，载于《IEEE计算机视觉与模式识别会议论文集》，2017年。[23] M. Cai、Z. Huang、Y. Li、H. Wang 和 Y. J. Lee, 《利用大型语言模型实现可扩展矢量图形驱动的图像理解》，2023，arXiv:2306.06094。[24] L. Li 等, 《Multimodal ArXiv：用于提升大型视觉语言模型科学理解能力的数据集》，2024，arXiv:2403.00231。[25] J. Wang 等, 《大型视觉模型与视觉提示工程综述》，Meta-Radiol.，2023年。[26] M. Cai 等, 《ViP-LLaVA：让大型多模态模型理解任意视觉提示》，2024, arXiv:2312.00784。[27] K. Marino、M. Rastegari、A. Farhadi 和 R. Mottaghi, 《OK-VQA：一个需要外部知识的视觉问答基准》，2019, arXiv:1906.00067。[28] S. E. Kahou、V. Michalski、A. Atkinson、A. K´ad´ar、A. Trischler 和 Y. Bengio, 《FigureQA：一个用于视觉推理的标注图表数据集》，ICLR，2018年。[29] A. Wu 等, 《DCQA：面向复杂推理和常识理解的文档级图表问答》，2023年。[30] S.-H. Lim、S. Lee、S. S. Powers、M. Shankar 和 N. Imam, 《生成逼真的合成图的方法综述》，橡树岭国家实验室，技术报告，2016年。[31] A. Sundar、J. Xu、W. Gay、C. Richardson 和 L. Heck, 《cPAPERS：科学论文中情境化多模态交互对话的数据集》，2024年。[32] J. Luo、Z. Li、J. Wang 和 C.-Y. Lin, 《ChartOCR：通过深度混合框架从图表图像中提取数据》，2021年。[33] F. Liu 等, 《MMC：通过大规模指令微调推进多模态图表理解》，载于《北美计算语言学分会论文集》，2024年。[34] J. Seo、Y. Xia、B. Lee、S. McCurry 和 Y. J. Yam, 《MAIDR：通过多模态数据表示提升统计可视化的可访问性》，2024, arXiv:2403.00717。[35] A. Lundgard 和 A. Satyanarayan, 《通过自然语言描述实现可访问的可视化：四级语义内容模型》，《IEEE可视化与计算机图形学汇刊》，2022年。[36] 谷歌, 《Prompting Guide 101》，2024年。[37] Zooniverse, 《Zooniverse.org》，2009年。[38] S. S. Shapiro 和 M. B. Wilk, 《正态性方差分析检验（完整样本）》，Biometrika，1965年。[39] W. H. Kruskal 和 W. A. Wallis, 《在单因素方差分析中使用秩》，《美国统计协会杂志》，1952年。[40] H. Levene, 《方差齐性的稳健检验》，《概率与统计论文集》，1960年。[41] S. Holm, 《一种简单的序贯拒绝多重检验程序》，《斯堪的纳维亚统计学杂志》，1979年。[42] M. J. Lindstrom 和 D. M. Bates, 《用于重复测量数据的线性混合效应模型的Newton—Raphson和EM算法》，《美国统计协会杂志》，1988年。