arXiv:2601.12585v1 [cs.HC] 18 Jan 2026

多模态大语言模型看到的和我们一样吗？人工智能系统中的可视化素养障碍分析

Mengli (Dawn) Duan*、Yuhe (Sissi) Jiang*、Matthew Varona 和 Carolina Nobre

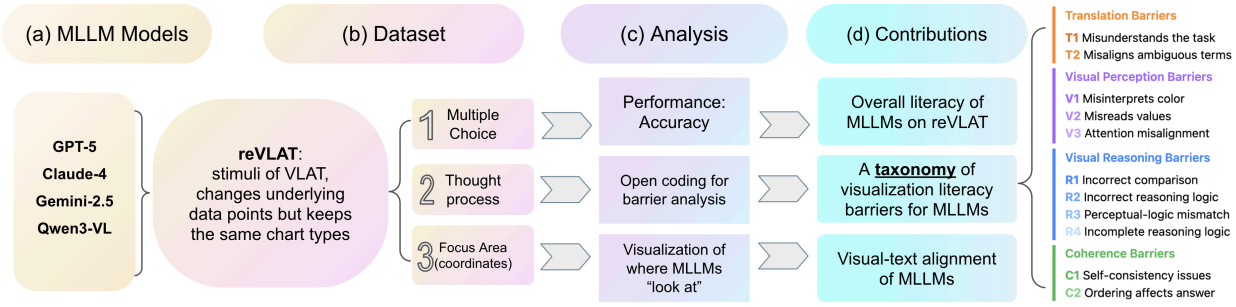


图 1：所提出的用于识别多模态大语言模型可视化素养障碍的框架概览，包括：(a) 参与评估的最先进的多模态大语言模型；(b) 数据集 reVLAT——采用合成底层数据的 VLAT 替代版本；(c) 收集用于分析的多模态大语言模型响应、推理依据和关注点；(d) 根据结果得出的关于多模态大语言模型推理依据的见解，以及对其障碍分类的概览。

摘要——多模态大语言模型（MLLMs）越来越多地用于解读可视化图表，但人们对它们为何会出错知之甚少。我们首次对多模态大语言模型中的可视化素养障碍进行了系统分析。我们采用基于合成数据的再生成版可视化素养评估测试（reVLAT）基准，并借鉴人类可视化素养研究中的以障碍为中心的策略，对四种最先进的模型产生的 309 条错误响应进行开放式编码。分析得出了多模态大语言模型失效分类体系，揭示了两种机器特有障碍，补充了以往的人类参与框架。结果表明，模型在简单图表上表现良好，但面对色彩密集型的基于分段的可视化时会遇到困难，且往往无法形成一致的比较推理。我们的研究发现可为未来可靠的人工智能驱动的可视化助手的评估与设计提供参考。

索引词—可视化素养、多模态大语言模型、评估研究

1 引言

数据可视化在现代数据分析和交流中发挥着基础性作用，既能将原始数据转化为易于理解的见解，也能促进研究人员与读者之间的交流。可视化素养是指阅读、解读视觉数据表示并进行推理的能力，这一能力已在各类人群中得到广泛研究 [1,7,20]。然而，随着人工智能系统的最新进展，一类新的读者应运而生：多模态大语言模型（MLLM），例如*GPT* 和*Claude*，它们能够同时读取文本输入和图像输入。

这些系统正越来越多地被用作专业、教育及一般分析场景中的助手，以提取数据、综合信息并解读复杂的可视化图表。近期研究对多模态大语言模型在基础视觉任务中的能力进行了初步探索，结果表明，它们理解数据可视化的表现具有竞争力，但仍不完美 [3,8,13,14]。由于这些模型高度复杂且不透明，了解其可视化素养能力及局限性，对于加强人机协作和可信度至关重要。

近期，[15] 建立了一套经验证的分类体系，用于分析人类的可视化素养障碍，为理解导致错误解读的认知差距奠定了基础，其中涉及人类参与者。然而，尽管近期研究 [13,14], 指出了一些失败模式，目前仍没有针对多模态大语言模型的同等系统框架。这自然引出一个问题：**多模态大语言模型是否会遇到与人类相似的障碍，还是会因其截然不同的感知与推理机制而呈现出独特的失败模式?** 在这项工作中，我们全面研究多模态大语言模型在可视化解读任务中为何以及如何失败，并将 [15]提出的分类体系扩展至多模态大语言模型常见的可视化素养障碍分类。

为弥补这一空白，我们对多模态大语言模型在可视化解读任务中的失败案例进行了全面研究。我们采用了受 [15]启发、以障碍为中心的开放式编码方法，并将其应用于多模态大语言模型，使用包含替代数据数值的可视化素养评估测试（VLAT）变体，以尽量减少数据泄漏，该变体称为 reVLAT [8]。我们收集并分析了四种最先进的多模态大语言模型的推理轨迹，据此构建了一个扩展分类体系，用于归纳多模态大语言模型的常见失败。我们的主要贡献包括：（a）对最先进的多模态大语言模型在 reVLAT 上的可视化素养进行全面评估，从而深入了解多模态大语言模型在视觉任务中的能力；（b）提出面向多模态大语言模型的扩展可视化素养障碍分类体系，对其在可视化[illegible]分类。[illegible]态大语言模型会出现与人类相似的错误，归为翻译障碍和视觉感知障碍；[illegible]败，[illegible]贯性障碍。

2 相关工作

本研究通过衔接针对人类参与者建立的成熟框架与近期研究，探讨多模态大语言模型在可视化方面面临的素养障碍

• *贡献相同。
• *Mengli (Dawn) Duan*、*Yuhe (Sissi) Jiang*、*Matthew Varona* 和 *Carolina Nobre* 均来自加拿大多伦多大学。电子邮件：*{mengli.duan, yuhe.jiang, }@mail.utoronto.ca*，*{varona, cnobre}@cs.toronto.edu*

稿件于 *201x* 年 *xx* 月 *xx* 日收到；于 *201x* 年 *xx* 月 *xx* 日接受。出版日期：*201x* 年 *xx* 月 *xx* 日；当前版本日期：*201x* 年 *xx* 月 *xx* 日。如需获取本文的抽印本，请发送电子邮件至：*reprints@ieee.org*。数字对象标识符：*xx.xxxx/TVCG.201x.xxxxxxx*

针对（多模态）大语言模型的评估流程[1]。本节回顾：（1）为人类参与者开发的可视化素养评估与障碍分析；（2）对（多模态）大语言模型完成可视化任务的表现进行的实证评估；以及（3）描述（多模态）大语言模型失败模式的新兴研究。

2.1 可视化素养评估

可视化素养是指阅读和解读数据可视化图表的能力 [10], , 相关研究已通过结构化评估进行了广泛探讨。早期一项具有重要影响力的研究是 [11], 提出的 VLAT，该测试涵盖 12 种图表类型和 53 道选择题，为对人类参与者的可视化素养进行标准化测试奠定了基础。在 VLAT 之后，还构建了其他一些基准，重点关注效率或批判性思维，例如 Mini-VLAT [18] 和 CALVI [5]。

除了定量评估，在视觉任务中探索失败案例也是学术研究的核心关注点。近期一项重要研究将 VLAT 题目扩展为开放式回答格式，以便对参与者的内部推理过程进行定性分析 [15]。研究人员通过对 120 名参与者提交的错误答案进行开放式编码，归纳出一套完整的人类可视化素养障碍分类体系，并将这些障碍分为概念与操作方面的差距。

受这项研究成功的启发，我们将这种以障碍为中心的方法调整并扩展至人工智能系统，旨在全面、系统地理解人工智能的失败模式。

2.2 人工智能系统在可视化任务中的能力

近年来，人工智能系统不断发展，尤其是大语言模型和多模态大语言模型（MLLMs）[2]，它们越来越多地被用作助手，以解读数据可视化并与之交互 [4]。因此，评估它们的可视化素养至关重要。

现有研究主要通过实证评估大语言模型和多模态大语言模型，并以准确率作为主要指标。 [3] 报告称，早期大语言模型在图表问答 [9]中的准确率达到 87%，前提是同时提供底层数据和可视化图表；但移除底层数据后，表现降至 31%。 [13] 对多模态大语言模型进行了比较研究，结果显示，这些模型在 VLAT 上的表现接近人类水平（模型为 58%，人类为 66%）。然而，这些结果引发了人们对数据泄漏的担忧，因为 VLAT 等开放数据集极有可能已包含在（多模态）大语言模型的训练语料库中。为降低这一风险， [8] 推出了 reVLAT 这一基准：使用一组新的合成数据重新生成 VLAT 可视化图表，同时保留相同的图表类型和问题。他们的研究结果显示，多模态大语言模型的准确率显著下降，表明它们在理解常见可视化图表方面可能存在缺陷。

尽管这些工作确立了多模态大语言模型在视觉任务中的基本能力，但它们通常缺乏对模型为何失效的完整分析。受 [15]启发并借助 reVLAT 流程，我们的工作直接弥补了这一空白，将重点从准确率测量转向对多模态大语言模型推理依据的定性分析，从而更深入地理解多模态大语言模型的可视化素养障碍。

2.3 理解多模态大语言模型失效模式

在上述工作中，研究者讨论了多模态大语言模型失效的若干潜在原因。例如，数值提取被认为是误差的主要来源之一 [13], , 因为多模态大语言模型难以从可视化图表中提取精确数值。涉及复杂空间或感知推理的特定图表类型也可能限制模型的表现 [8, 13]。随着 reVLAT 的发展， [14] 对回忆问题进行了更深入的研究，强调即使图像缺失，多模态大语言模型仍可能基于记忆或偏见持续输出答案。

尽管已有研究探讨了这些孤立问题，但目前尚无工作对多种最先进模型中的多模态大语言模型可视化素养障碍进行全面分类。我们的贡献在于填补这一空白：系统分析一系列最先进的多模态大语言模型在 reVLAT 上的表现，归纳出全面的失效分类体系，并探究多模态大语言模型是否面临与人类相同的障碍，还是具有独特的、机器特有的推理依据。

[1]大语言模型和多模态大语言模型[2]我们关注以**图像**和文本作为两种输入模态的多模态大语言模型。

3 方法

本节介绍实验设计、多模态大语言模型响应收集的配置，以及用于推导多模态大语言模型可视化素养障碍分类体系的分析计划。

3.1 研究设计

我们改编了 [8] 中的评估流程，以收集多模态大语言模型的响应，并开展受 [15]启发的定性分析。

模型。 如图 1所示，我们选取了四种最先进的多模态大语言模型，以全面了解当前的失败模式，包括 *GPT-5*（*GPT*） [16], *Claude-Sonnet-4*（*Claude*） [2], *Gemini 2*。*5* [6], 和 *Qwen3-VL*（*Qwen*） [19]。所有模型均通过 OpenRouter API [17]访问；并且可以直接输入图像。

数据集。 我们利用现有的 reVLAT GitHub 仓库 [8] 生成数据并收集多模态大语言模型的响应。reVLAT 包含 53 道题目，涵盖 12 种图表类型；每张图表都会使用给定的随机种子重新生成，从而创建合成的底层数据分布（与原始 VLAT 不同），以减轻数据泄漏问题。所有图像均为高分辨率（1233 × 600 像素）。随后，我们将问题和答案选项重新整理为 JSON 模式，并存储在本地，以提高可复现性。每道题目都会与相应的可视化刺激材料关联。我们统一了色彩调色板和字体大小，以复现原始 VLAT 中受控的感知条件。评估时，我们将数据集序列化为与多模态大语言模型提示兼容的 JSON 文件。整个流程使用 Python 实现；我们使用全局随机种子控制重新生成的数据。为提高一致性和可复现性，我们固定了随机种子。

回答增强。 受 [15], 启发，我们增强了标准的多项选择问答流程，以收集多模态大语言模型的三个不同回答组成部分，从而进一步对多模态大语言模型的推理依据进行定性分析：

- 除了原有的答案选项外，我们还添加了一个“我不确定”选项，让模型可以明确承认自己无法回答问题，而不是被迫猜测。
- 除了所选答案选项外，我们还要求模型用最多 5 句话说明支持其选择的推理依据。
- 最后，我们要求模型输出一组 (x,y) 坐标（最多 5 个），以标示其在问答阶段的关注点（即它“注视”的位置）[3]。

遵循 [8], 我们仅通过最少量的提示，便要求多模态大语言模型以结构化 JSON 格式输出这三类回答（详细的提示示例见附录7 ）。

3.2 试点实验

我们使用 Claude 4 对 reVLAT 中 53 道题目随机抽取的 20 道进行了一系列试点实验，并收集了相关见解，将其纳入完整评估中。

多模态大语言模型的温度。 温度参数（T）决定了多模态大语言模型的响应方差：较高的T 会增加生成文本的多样性和随机性，而较低的T 则会使输出更加确定 [12]。在试点实验中，我们测试了多模态大语言模型在$T = 0.0$，0。5，1。0时的表现，发现整体性能准确率（约79%）没有显著变化。因此，为了提高可复现性，我们在完整实验中将$T = 0.0$ 设为固定值。

[3]我们将直接绘制在 [20] 所收集图像上的草图 替换为关注点，因为多模态大语言模型无法稳定地输出高质量图像。

选项排序。 先前研究 [8]表明，选项顺序会对多模态大语言模型的响应产生不可忽视的影响。我们在试点实验中也观察到了类似趋势：不同选项排序下的评估结果显示，整体性能准确率介于67%至79%之间。因此，我们在完整实验中对多种选项顺序进行采样，以尽可能减少选项排序可能引入的副作用。

3.3 完整实验

根据试点实验的发现，我们开展了完整实验：为每个问题生成所有可能的选项排序，然后随机抽取 5 种排序（不重复），用于评估多模态大语言模型。每组问题与排序组合都在独立会话中运行，以防模型在不同试次间延续上下文或记忆。总共进行了 1060 次试次（53 个问题 $\times$ 5 种排序 $\times$ 4 个模型），每个模型进行了 265 次试次。计算成本共计 386 万个词元，费用为 11.79 美元。

3.4 多层次分析

我们对收集到的回答进行了多层次分析：

1. 性能分析：计算不同图表类型和模型家族的整体准确率指标；
2. 开放式编码：遵循 [15], 的方法，我们系统地对提取出的推理轨迹进行开放式编码，以建立一个扩展的多模态大语言模型可视化素养障碍分类体系（四个模型共计 309 个错误回答）；
3. 定量分析：以数值方式呈现模型层面和图表层面的障碍分布。

4 多模态大语言模型的可视化素养障碍

本节介绍我们通过对实验回答进行开放式编码，归纳出的多模态大语言模型可视化素养障碍分类体系。与以往关于人类可视化素养的研究一致， [15], 我们的分析揭示了可视化解释多个阶段中存在的障碍，从将文本查询转化为任务，到提取视觉信息并据此进行推理。与人类读者不同，多模态大语言模型会出现独特的推理和连贯性失误，这些失误源于其词元化生成过程 [21, 22], , 凸显了刻画这些障碍的重要性。

4.1

翻译障碍出现在对问题进行初步解释时，此时模型需要推断任务要求，并识别可视化中的相关元素。因此，这类障碍本质上属于概念层面，可能导致数值提取错误或推理错位。在我们的分类体系中，我们识别出两种主要的翻译失败形式：***T1***“误解任务”和***T2***“对含义模糊的术语理解错位”。

T1：误解任务 模型未能正确识别预期的视觉分析任务。例如，当被要求确定一个范围时，多模态大语言模型可能转而计算两个单独数值之间的差值，或将问题理解为要求判断趋势。与人类参与者类似 [15], , 误解任务会导致后续错误，例如提取错误的视觉元素或得出不正确的结论。

T2：误解含歧义的术语 这一障碍反映了与语言对齐或语义歧义相关的挑战，即模型对“large”和“small”等术语的理解可能与问题中的预期不同。模型可能难以采用一致的定义，或根据上下文解释描述性语言。此外，不清晰或缺乏具体说明的指令可能引发推理错误，导致模型关注不相关的图表元素或类别。

4.2

当模型难以从图表中正确提取颜色、数值或特定数据标记等低层次视觉信息时，就会产生视觉感知障碍。这与编码相关的问题类似在人类读者中观察到的 [15], 这些障碍反映的是解释感知线索的困难，而不是对任务本身的误解。这一阶段的错误往往会传递到后续推理步骤中，导致错误比较或结论，即使整体任务解释是正确的。因此，我们确定了三类常见的感知失败： ***V1*** “颜色误读”、***V2*** “数值误读”和***V3*** “注意力错位”。

V1：误解颜色 当模型误读可视化图表中的颜色编码时，就会发生颜色误解。这种错误尤其容易出现在使用渐变或量化色阶的可视化图表中，例如分级设色地图。模型可能正确定位目标区域，却误解对应的图例区间（例如，将其推断为 4% 至 6%，而不是 6% 至 8%）。因此，它会从色彩图中读取错误的数值。

V2：数值误读 当模型未能从可视化中正确提取数值时，就会发生数值误读。我们观察到这一障碍有两种表现形式，分别对应直接和间接的错误读取。直接错误读取是指模型明确给出与视觉证据不符的数值，例如纵轴上绘制的数值对应 50，模型却报告为 40。相反，间接错误读取是指模型没有明确说出误读的数值，但后续推理却反映出错误的感知。例如，错误比较或结论可能表明模型在内部误读了某个数值，即使它没有直接表述出这一错误读取。因此，这类错误是指模型读取了正确的数据元素，却误解了其数值刻度或与坐标轴的对应关系。

V3：注意力错位 即使理解了任务，当模型的“注意力焦点”落在错误的视觉区域或元素上时，就会发生注意力错位。由于“注意力焦点”偏移或视觉定位不够精准，模型可能会识别出错误的条形、区段或数据点。有时，提取到错误的“注意力焦点”区域会表现为不确定性（例如回答“无法确定”）。因此，这表明模型的视觉感知能力不足。

4.3

当模型能够准确感知相关视觉元素（如颜色、数值和空间关系），却无法据此构建正确的逻辑推断时，就会出现视觉推理障碍。与因误解视觉编码而导致的感知错误不同，推理障碍反映了模型在将感知输入整合为连贯结论方面的弱点。这类错误常见于需要进行比较、演绎或多步推理的任务中，即使初步视觉读取正确，也可能出现。正如多种错误模式所显示的，视觉推理障碍可能源于先前的感知不准确；但当模型执着于某种解释或采用有缺陷的逻辑结构时，这类障碍也可能独立出现。视觉推理障碍主要有四种形式：***R1***「错误比较」、***R2***「错误推理逻辑」、***R3***「感知—逻辑不匹配」和***R4***「不完整推理逻辑」。

R1：错误比较 当模型能够识别相关数值或视觉特征，却得出错误的比较结论（例如，认为条形较高的类别数值更小）时，就会出现错误比较。这种情况下，模型的感知能力仍然有效，但对元素之间关系的判断有误。这类失误在比较任务中很常见，因为模型必须推断相对大小，而不能只是读取数值。

R2：错误推理逻辑 错误推理逻辑反映了对视觉信息采用的因果或推断结构出现了问题。例如，模型可能会推断出“A 意味着非 B”，尽管它正确识别出可视化呈现的是相反的关系。这些错误源于逻辑推导不足，且往往是模型引入无依据的假设或对视觉内容采用不恰当的推理方式所致。

R3：感知—逻辑不匹配 当感知输入与逻辑推断出现偏差时，就会发生感知—逻辑不匹配错误。模型

可能正确读取数值，例如识别出“ A = 79 and B =60”，却得出“ A 小于 B”的结论。这表明，模型的推理过程可能受到某种内部启发式规则的影响，从而凌驾于感知证据之上。这类不匹配凸显了视觉解码与推理过程之间的脱节，与人类认知中观察到的冲突相似。

R4：不完整推理逻辑当模型得出结论时，若未纳入所有相关视觉信息，或未阐明中间推理步骤，就会出现不完整推理逻辑。这类错误源于模型缺乏关键比较。模型可能忽略全局结构，或未能整合多个数据组成部分。因此，模型会“草率下结论”，给出看似自信、却缺乏充分分析依据的答案。

4.4

当模型在整个推理过程中无法保持内部一致性时，就会出现连贯性障碍。与关注基于感知到的视觉信息进行推理是否正确的视觉推理障碍不同，连贯性障碍反映的是模型推理轨迹受到干扰。这类错误常见于模型生成的中间推理步骤与最终答案相矛盾的情况。此外，当选项排序等外部因素影响答案时，也会出现此类错误。由于多模态大语言模型的生成过程 [21, 22], 中后续推理步骤可能会覆盖或偏离先前步骤，因此连贯性障碍在多模态大语言模型中广为人知。我们归纳出两种主要的连贯性障碍： ***C1*“自洽性问题”**和***C2*“排序影响答案”**。

C1：自洽性问题 当模型生成的推理与最终答案相矛盾时，就会出现自洽性问题。例如，模型可能在中间推理中正确推断出“ A 大于 B”，却给出表明 B 更大的答案。这种不一致表明，推理链与决策输出并不一致，说明模型未能维持连贯的内部逻辑。当模型修正或重新评估其中间解释，得出与最初陈述不同的结论时，便会出现此类问题。在某些情况下，模型倾向于得出两种相互矛盾的内部判断，即“A > B”，同时又认为“B > A”。这表明其推理过程不稳定。

C2：排序影响答案 当信息或答案选项的呈现顺序影响模型的回答时，就会产生排序效应。在这种情况下，面对同一任务和相同的视觉内容，模型在一种选项排序下可能回答正确，在另一种排序下却回答错误。这意味着模型的决策过程受到编码结构变体的影响，而非以视觉推理为依据。因此，排序效应表明，多模态大语言模型可能依赖位置启发式，或对靠前答案选项存在潜在偏差，而不是基于可视化进行推理。

开放式编码示例 – 条形图比较错误

案例 ID:bar_chart_Q4_*Claude*问题：亚洲有多少个国家的平均网速比韩国慢？ 类型：进行比较（绝对值）正确答案：8 个国家模型答案：11 个国家

模型推理轨迹：
"我需要确定韩国的互联网速度，并统计有多少个国家的网速更慢。查看图表后，韩国的互联网速度似乎约为 14.5 兆比特每秒。逐一查看每个条形，我发现以下国家的网速低于韩国：中国（16.8）、香港（12）、印度（9.8）、印度尼西亚（13.2）、日本（11.8）、新西兰（10.2）、新加坡（12.8）、台湾（13.2）、泰国（10）和越南（16.8）。等等，让我重新确认一下——韩国的网速实际上似乎约为 14.5 兆比特每秒，而统计低于这一水平的国家，我数出有 11 个国家的网速更慢。"

模型坐标：
(693, 400)：“韩国的条形——估计网速值”，(400, 300)：“扫描所有条形以进行比较”，(200, 450)：“条形较低的国家”，(800, 350)：“右中部的条形组”，(600, 400)：“韩国附近的相对高度”。

为展示该分类体系的实际应用，我们 प्रस्तुत了一个来自*Claude* 的代表性示例，见示例4.4。问题是：“有多少个条形低于高亮显示的条形？ ” 正确答案是 8 个国家，而模型报告为 11 个。该答案同时体现了**视觉推理障碍**和**连贯性障碍**。

R1： 变映射为较慢网速：“查看图表可知，韩国的网速 秒。”模型还表现出**R4：不完整推理逻辑**，例如将中国（16.8）和越南（16.8）称为“较慢”，这表明其无法得出正确的比较结 出现，例如模型改口说“等等，让我重新检查一下”，这表明其决策不稳定。因此，该案例说明，尽管多模态大语言模型能够在视觉上定位图 象进行准确的比较推理时仍存在困难，从而导致错误的目标计数。

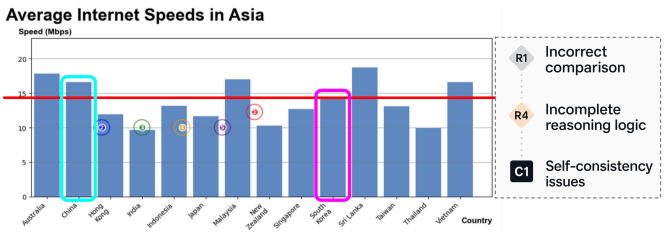


图 2：多模态大语言模型在条形图任务（Q4）中给出错误回答的示例。高亮区域表示模型报告的关注点及其误判的最大值。红色水平线表示正确识别速度超过 15 兆比特每秒的国家所需的阈值。

5 定量分析

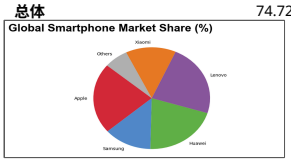
本节首先从准确率和图表类型两个方面总结模型的整体表现。随后，我们运用第 4 节介绍的分类体系进行障碍分析，考察错误模式，并通过 *Claude* 的示例回答展示关键失误。

5.1 整体表现

性能准确率。 在 53 道 reVLAT 题目及每道题的 5 种随机排序中，4 个模型的准确率各不相同。*Qwen* 的整体表现最佳，达到 86.42%；其次是*Claude* ，为 74.72%；而*GPT* 和*Gemini* 的准确率分别为 49.06% 和 73.21%，且在不同排序下的波动较大。

按图表类型划分的表现。 图3 展示了 reVLAT 中 12 种图表类型的准确率，揭示了不同模型之间若干一致的表现模式。 *Claude* 在条形图和直方图等常见图表类型上表现出色，但在饼图（46.70%）和堆叠条形图（72.00%）上的准确率有所下降，其总错误中近一半发生在这两类图表上。 *Qwen*作为总体表现最佳的模型，也呈现出类似趋势：虽然它在面积图、分级设色地图、直方图等图表上的准确率达到 100%，但在饼图（66.70%）和堆叠条形图（52.00%）上的表现明显下降。

图表类型	Claude 数量	GPT	Gemini	Qwen	平均准确率
100%堆积条形图	15	53.30%	86.70%	80.00%	66.70%	71.68%
面积图	20	85.00%	40.00%	100.00%	100.00%	81.25%
条形图	20	100.00%	75.00%	95.00%	75.00%	86.25%
气泡图	35	68.60%	31.40%	88.60%	100.00%	72.15%
分级设色地图	15	66.70%	53.30%	60.00%	100.00%	70.00%
直方图	15	100.00%	80.00%	100.00%	100.00%	95.00%
折线图	25	88.00%	80.00%	84.00%	80.00%	83.00%
饼图	15	46.70%	66.70%	40.00%	66.70%	55.03%
散点图	35	74.30%	25.70%	57.10%	88.60%	61.43%
堆叠面积图	30	70.00%	33.30%	63.30%	100.00%	66.65%
堆叠条形图	25	72.00%	12.00%	56.00%	52.00%	48.00%
树图	15	66.70%	73.30%	53.30%	100.00%	73.33%
总体		74.72%	49.06%	73.21%	86.42%	



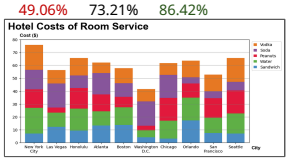


图 3：4 个多模态大语言模型按图表类型划分的准确率。单元格颜色从红色（准确率最低）到绿色（准确率最高）渐变。此可视化突显了各类图表的难度以及模型特有的弱点。

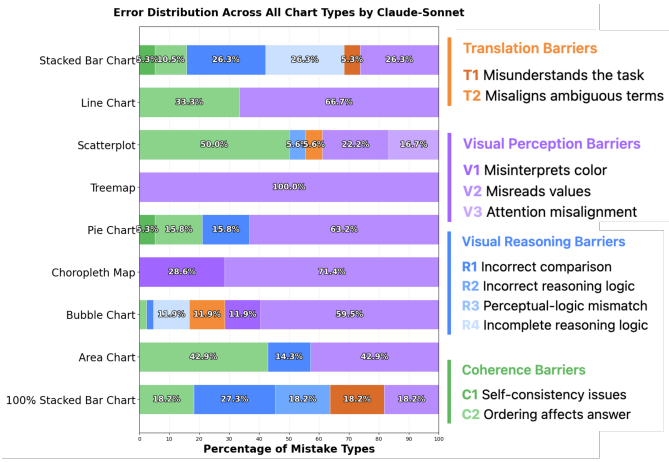


图4：所有图表类型中*Claude*的错误分布。条形表示每种图表中各类错误所占的百分比，并按照第4节所述的分类体系进行分组。

图3 呈现出一个规律：需要解读多个彩色分段的图表，例如饼图和堆叠条形图，在全部四个模型上的准确率都持续较低。我们的观察与既有研究结果一致，即当类别颜色编码与定量幅度估计相互作用时，多模态大语言模型往往会遇到困难。在这些图表中，模型必须同时解析用于类别识别的编码，以及用于数值估计的角度、高度或面积。我们推测，这种对颜色和幅度的双重依赖会在视觉嵌入空间中引入额外复杂性，使定量判断更加不稳定，也更容易出错。此外，这些图表类型在当前预训练语料库中的出现频率似乎也较低，这可能进一步限制模型对可靠回答此类问题所需视觉模式的熟悉程度。

洞见。图表类型的结果表明：(1) 与以往研究相比，现代多模态大语言模型在可视化素养方面已取得进展；(2) 错误不再均匀分布，而是集中在颜色丰富、基于分段的图表类型中；(3) 即使是 *Qwen* 等表现优异的模型，在需要联合解读颜色与数量时仍容易产生误解。这些发现进一步表明，评估多模态大语言模型时，不应只关注总体准确率，还应考察其面对不同视觉编码和设计选择时的稳健性。

5.2 各模型的障碍分析

我们总共对各模型的 309 条错误回答进行了开放式编码，其中*Claude*贡献了 67 条，*GPT*贡献了 135 条，*Gemini*贡献了 71 条，*Qwen*贡献了 36 条。***Claude***。图 4 展示了 *Claude*在所有图表类型中 67 条错误回答的障碍类型分布。颜色编码更丰富的图表，尤其是堆叠条形图和饼[illegible]式：V2 数值误读占错误的相当大比例。对于这些图表类型，这一障碍在错误分布中占主导地位。这种模式表明，当空间注意力定位准确时，多个彩色分段和图例的存在仍会妨碍模型解释或估算数值。我们推测，在视觉嵌入空间中引入多种色调会增加表征复杂性。因此，模型需要区分作为类别线索的颜色与高度、角度或面积等定量编码。相比之下，对于折线图和面积图等视觉上较简单的图表，错误则更均匀地分布在[illegible]之间。这表明，当必须从可视化中同时解读颜色和数量时，数值误读是一种敏感的失效模式。

6 局限性与未来方向

需要承认本研究存在若干局限。首先，多模态大语言模型对提示设计较为敏感，而我们采用的极简提示策略可能无法充分反映其在真实应用场景中的表现。未来研究可以采用更复杂的提示技术来复现我们的分析，并有可能揭示不同的失败模式。其次，最先进的多模态大语言模型在推理过程中难以直接绘制或标注可视化图表。尽管我们通过要求模型输出坐标，以文本形式捕捉空间注意力来应对这一局限，但我们的框架可以扩展，以适配未来具备多模态输出能力的多模态大语言模型。

未来，我们将重点拓展跨模型和任务条件的分析广度与深度。首先，扩大顺序操控的规模将使我们能够收集更丰富的行为数据，并更稳健地刻画呈现结构影响多模态大语言模型的响应的程度。其次，我们计划使用以大语言模型作为评判者的方法，自动化封闭式编码过程中的部分环节，从而在保持解释严谨性的同时，实现更高效、可复现且可规模化的标注 [23]。最后，我们计划设计有针对性的实验干预，直接检验每类障碍背后的假设机制，从而为多模态大语言模型为何会在可视化解读与推理的特定阶段失败提供更有力的因果证据。

7 结论

在本研究中，我们探究了多模态大语言模型是否以与人类相近的方式理解可视化图表，以及它们为何在解读图表类型时失败。通过将以障碍为中心的方法论 [15]应用于人工智能系统，我们首次为多模态大语言模型系统地建立了可视化素养障碍分类体系；该体系源自对4个最先进模型的309条错误响应进行的开放式编码。我们的分类体系拓展了以往针对人类的框架，识别出两类机器特有的障碍，即**视觉推理障碍**和**连贯性障碍**。

定量分析表明，错误并非均匀分布：模型在视觉上较为简单的图表上表现良好，但面对色彩密集、基于分段的可视化图表时则会遇到困难，因为这类图表需要同时处理类别编码和幅度估计。障碍分布进一步表明，即使多模态大语言模型能够准确定位相关视觉元素，也常常无法构建一致的比较推理，从而得出错误结论。

这些发现为理解多模态大语言模型在可视化解读中的认知局限提供了新见解，并对将其部署为分析助手具有启示意义。了解多模态大语言模型在何处以及为何会出错，为改进模型设计、评估方案以及视觉分析情境下的人机协作奠定了基础。展望未来，扩大行为采样规模、实现编码自动化，并设计有针对性的实验消融，将有助于为这些障碍背后的机制提供更有力的因果证据。我们的研究旨在助力制定相关准则，推动更可靠、更值得信赖地运用多模态大语言模型进行视觉推理。

致谢

作者衷心感谢 Matthew Varona 和 Carolina Nobre 教授提供的教学支持与指导。

参考文献

[1] B. Alper、N. H. Riche、F. Chevalier、J. Boy 和 M. Sezgin。小学阶段的可视化素养。载于《2017 年 CHI 人机交互系统中的人为因素会议论文集》，第 5485–5497, 2017 1 – 页。[2] Anthropic。Claude 发布介绍，5 月。4, 2025 2 [3] A. Bendeck 和 J. Stasko。对 GPT-4 多模态语言模型在可视化素养任务中的表现进行实证评估。IEEE 可视化与计算机图形学汇刊，2024 1, 2 。[4] K. Choe、C. Lee、S. Lee、J. Song、A. Cho、N. W. Kim 和 J. Seo。按需提升数据素养：大语言模型作为图表解读新手的指南。IEEE 可视化与计算机图形学汇刊，31 9 4712–4727, 2025 10 1109 2024 3413195 2 ( ): – 。doi: . /[5] TVCG. 。L. W. Ge、Y. Cui 和 M. Kay。Calvi：可视化素养的批判性思维评估。载于《2023 年 CHI 人机交互系统中的人为因素会议论文集》，第 1–18, 2023 2 – 页。[6] G. Gemini。Gemini 1.5 Flash，2 5 2025 2 。[7] L. Grammel、M. Tory 和 M.-A. Storey。信息可视化新手如何构建可视化图表。IEEE 可视化与计算机图形学汇刊，16(6):943–952，2010。doi: 10.1109/TVCG.2010.164 1[8] J. Hong、C. Seto、A. Fan 和 R. Maciejewski。大语言模型具备可视化素养吗？通过修改后的可视化图表评估其数据解读泛化能力。IEEE 可视化与计算机图形学汇刊，2025 1, 2, 3 。[9] D. H. Kim、E. Hoque 和 M. Agrawala。回答有关图表的问题并生成可视化解释。载于《2020 年 CHI 人机交互系统会议论文集》，CHI '20，共 13 页，第 1–13 页。计算机协会，美国纽约州纽约市，2020 年。doi: 10.1145/3313831.3376467 2[10] S. Lee、S.-H. Kim、Y.-H. Hung、H. Lam、Y. ah Kang 和 J. S. Yi。人们如何理解陌生的可视化图表？新手信息可视化意义建构的扎根模型。IEEE 可视化与计算机图形学汇刊，22(1):499–508，共 10 页，2016 年 1 月。doi: 10.1109/TVCG.2015.2467195 2[11] S. Lee、S.-H. Kim 和 B. C. Kwon。VLAT：可视化素养评估测试的开发。IEEE 可视化与计算机图形学汇刊，23 1 551–560, 2017 10 1109 2016 2598920 2 ( ): – 。[12] doi: . /TVCG. 。L. Li、L. Sleem、G. Nichil、R. State 等。探索温度对大语言模型的影响：冷热如何？Procedia Computer Science，264:242–251，2025。2[13] Z. Li、H. Miao、V. Pascucci 和 S. Liu。多模态大语言模型的可视化素养：一项比较研究。arXiv 预印本 arXiv:2407.10996，2024。1，2[14] Z. Li、H. Miao、X. Yan、V. Pascucci、M. Berger 和 S. Liu。看见还是回忆：检验视觉在多模态大语言模型完成可视化问答任务中的作用。arXiv 预印本 arXiv: *2504 09809*, 2025 1, 2 。。C.[15] Nobre、K. Zhu、E. Mörth、H. Pfister 和 J. Beyer。像素之间的玄机：探究可视化素养的障碍。载于《2024 年 CHI 人机交互系统会议论文集》，第 1–17 页，2024 年。1，2，3，5[16] 5, 2025 2 OpenAI。GPT-4o 发布介绍，8 月。[17] OpenRouter。OpenRouter API。https://openrouter.ai，2024 年。通过 API 进行模型推理。2[18] S. Pandey 和 A. Ottley。Mini-VLAT：一种简短且有效的可视化素养测量工具。载于《计算机图形论坛》，第 42 卷，第 1–11 页。Wiley Online Library，2023 2 。Qwen。Qwen：思考，更广泛的行动，9 月[19] 2025 2 。D. Statton Thompson、S. Beene、K. Greer、M. Wegmann、M. Fullmer、M. Murphy、S.[20] Schumacher 和 T. Saulter。图像激增：视觉素养的趋势、障碍与机遇。《视觉素养期刊》，41 2 113–131, 2022 1, 2 ( ): – 。X. Wang、J. Wei、D. Schuurmans、Q. Le、E. Chi、S. Narang、A. Chowdhery 和 D. Zhou。自[21] 一致性提升语言模型中的思维链推理，2023 3, 4 。J. Wei、X. Wang、D. Schuurmans、M. Bosma、B. Ichter、F. Xia、E. Chi、Q. Le 和 D. Zhou。思维链提示能够激发大语言模型的推理能力，2023 年。3, 4[22] L. Zheng、W.-L. Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang、Z. Lin、Z. Li、D. Li、E. P. Xing、H. Zhang、J. E. Gonzalez 和 I. Stoica。使用 MT-Bench 和 Chatbot Arena 评判 LLM 评判者，2023 年。 5

系统提示演示示例

系统提示

你是一位数据可视化分析专家。你将看到一张可视化图表，并需要回答与之相关的问题。

请按照以下 JSON 格式提供回答：

{ "choice": "<你选择的答案选项字母，例如 'a'、'b'、'c' 等>", "answer_value": "<你所选答案的实际文本本>", "rationale": "<用不超过 5 句话说明你选择此答案的推理>", "attention_regions": [{" x": <x 坐标>, "y": <y 坐标>, "description": "<你查看的内容>"}]}

重要说明：
1. 可视化图像的尺寸为 `1233 × 600`像素。2. 对于注意区域，请提供最多 5 个坐标点 (x, y)，标明你作答时关注的位置。3. x 范围：0（左侧） → `1233`（右侧）；y 范围：0（顶部）→ `600` （底部）。4. 每个区域都应简要描述你查看的内容。5. 请简明扼要地说明理由，不超过 5 句话。6. 如果不确定，请选择“无法确定”选项，并在理由中说明原因。7. 始终仅返回有效的 JSON（不要使用 Markdown 或添加额外说明）。