MathOPEval：用于数学推理中 MLLMs 视觉操作的细粒度评估基准

Xiaoyuan Li[1]*, Moxin Li[3], Wenjie Wang[1], Rui Men[2], Yichang Zhang[2], Fuli Feng[1], Dayiheng Liu[2], Junyang Lin[2]
University of Science and Technology of China[1] Alibaba Group[2] National University of Singapore[3]

Abstract

最近在多模态大型语言模型（MLLMs）方面的进展使得通过基于文本指令执行视觉操作来实现逐步的多模态数学推理成为可能。一种有前景的方法是使用代码作为中间表示，精确表达和操作推理步骤中的图像。然而，现有的评估主要集中在仅限文本的推理输出上，对于 MLLM 通过代码执行准确视觉操作的能力大多未被探索。本研究迈出了填补这一空白的第一步，通过评估 MLLM 在多模态数学推理中基于代码的能力。具体来说，我们的框架关注两个关键评估方面：（1）多模态代码生成（MCG）评估模型从头开始准确理解和构建可视化的能力。（2）多模态代码编辑（MCE）评估模型进行细粒度操作的能力，包括删除、修改和注释三种类型。为了评估上述任务，我们引入了一个涵盖五种最受欢迎的数学图形的数据集，包括几何图、函数图和三种类型的统计图，以提供对现有 MLLM 的全面和有效的评测。我们的实验评估涉及九个主流 MLLM，结果显示现有模型在执行细粒度视觉操作方面仍然明显落后于人类表现。

1 引言

近年来，多模态大语言模型（MLLMs）在视觉数学推理方面取得了显著进展。给定一张图片和一个文本问题描述，MLLMs 能够生成逐步推理以得到解决方案。这些推理步骤通常包括视觉操作，如标记角度、绘制辅助线以及识别关键元素（图 ?? ）。与仅文本的数学推理相比，这项任务需要更强的多模态能力，特别是在对齐文本和视觉信息、规划和执行视觉操作以达到正确答案方面。

最近，利用代码作为多模态数学推理的中间表示成为了一种有前景的方法，带来了显著的性能提升 (Hu et al., 2024a; Fu et al., 2025) 。在这一范式中，图像被翻译成相应的 Python 或 LaTeX 代码，这些代码可以精确地重建它们。视觉操作随后被表示为代码编辑。代码作为图像的精确和结构化的文本表示，与基于语言的推理很好地对齐，并减少了直接视觉操作中通常存在的歧义。然而，关于多模态数学推理的最新评估主要集中在仅有文本的输出 (Hu et al., 2024a; Cheng et al., 2024; Lee et al., 2024; Fu et al., 2025) 。目前尚不清楚现有的 MLLMs 能多有效地产生反映准确视觉操作的中间代码，这是实现精确、可解释的多模态数学推理的关键一步（图 ?? ）。

因此，在这项工作中，我们迈出了评估多模态数学推理的第一步，重点关注 MLLMs 在执行视觉操作时与代码相关的能力。我们关注两个主要方面：（1）多模态代码生成和（2）多模态代码编辑。多模态代码生成评估模型将图像输入转换为相应构建代码的能力。多模态代码编辑包含三种常见的视觉操作类型：1）删除测试模型识别和删除干扰元素的能力，简化图形以突出关键信息；2）修改：评估模型通过添加辅助线或更改视觉元素来更新图像的熟练程度；3）注释评估模型在适当位置添加数值、符号和其他标记的能力，通过清晰的注释增强视觉信息表达。

更进一步，我们手动整理了一个高质量的数据集。考虑到数学视觉推理任务的多样性，我们仔细选择了五种具有代表性的图形类型：(1) 几何图形，(2) 函数图，(3) 柱状图，(4) 折线图，和 (5) 饼图。对每个样本，我们进行详细的手动标注以构建如图 1 所示的初始数据集，包括：(1) 原始问题图像；(2) 四种视觉操作后的图像；(3) 实现这些视觉操作的代码和说明。为了确保全面和可靠的评估，我们将初始数据集转换为两种问题类型：10,293 道选择题测

*Work done when Xiaoyuan Li was intern at Alibaba Group.

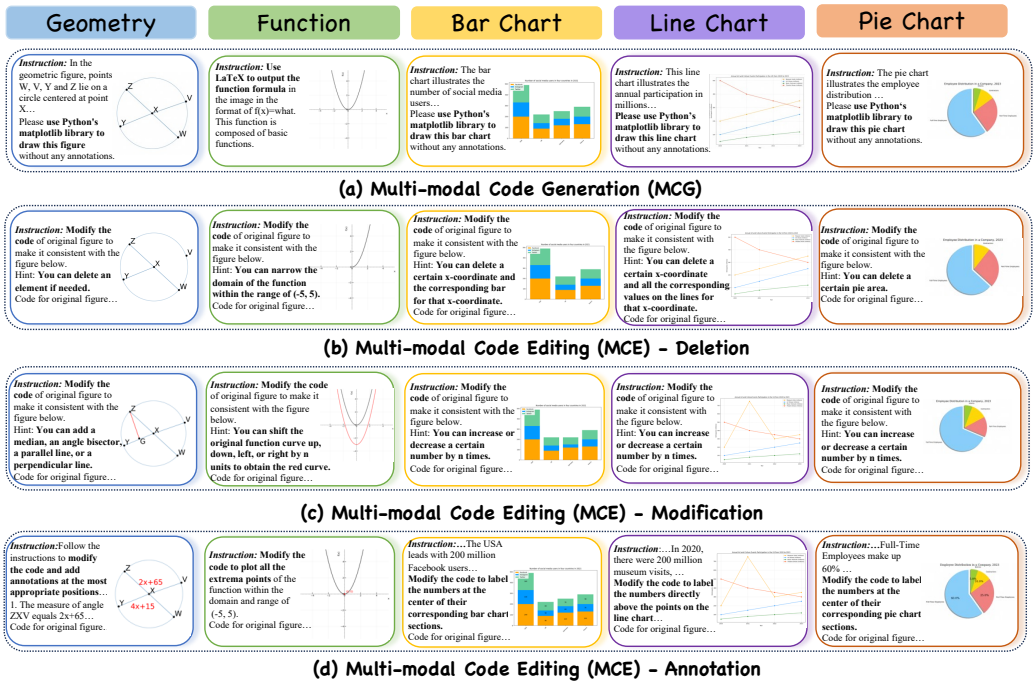

Figure 1: 针对五种可视化类型的四种视觉操作的初始数据集示例。数据集包括为不同任务和可视化类型构建的评估指令、代码和图像。详情请参阅 ?? 节。

试模型在视觉操作选择方面的能力，和 7,552 道开放式问题以评估模型生成视觉操作的能力。对于自由生成问题，我们还基于思维链（CoT）建立评价策略，以对模型在多个维度上的视觉操作进行细粒度评估，包括内容准确性、位置一致性、代码正确性和完整性等方面。

我们在九个主流的闭源和开源多模态大型语言模型上使用多样的提示策略进行了全面的实验。我们的实验结果揭示了以下关键发现：(1) 现有的多模态大型语言模型在我们视觉操作评估套件上的表现与人类水平相比，仍然存在显著的差距，这表明我们的基准对当前的多模态大型语言模型提出了重大挑战；(2) 在不同类型的视觉操作中，视觉修改被证实是最具挑战性的，而在不同类型的图表中，函数图是处理难度最高的；(3) 增加提示示例的数量可以适度提高模型在视觉操作中的理解和执行能力。

我们的主要贡献可以总结如下：

- 据我们所知，我们是第一个提出用于视觉操作的细粒度评估基准的人，这为深入评估 MLLMs 的视觉理解能力提供了新的视角和工具。
- 我们构建了一个精心标注的大规模视觉操作评估数据集，涵盖了各种数学图形类型和视觉操作类型。
- 通过广泛的实验验证，我们揭示了当前多模态语言模型在视觉操作中的能力边界和改进潜力，为未来的研究提供了重要参考。

2 任务表述

这项工作以多模态编码的形式评估了多模态大语言模型（MLLMs）的多模态视觉操作能力，聚焦于两个关键方面：多模态代码生成和多模态代码编辑。这些方面反映了多模态数学推理中的常见模式，其中模型必须将图像（例如，函数图或图表）翻译为代码，并根据图像的修改对代码进行精确、协调的编辑。

- 多模态代码生成 (MCG)：给定一幅图像 V_{in} ，MLLM 生成对应的代码 C_{in} ，可以精确构建这样的图像。代码可以是 Python Matplotlib 库的形式，或 LaTeX。正式来讲，将生成指令表示为 I_g 。

$$C_{in} = MLLM(V_{in}, I_g) \tag{1}$$

- 多模态代码编辑（MCE）：MLLM 被给定一个初始图像 V_{in} ，其构造代码为 C_{in} ，以及一个编辑后的图像 V_{out} 。期望 MLLM 生成相应的代码 V_{out} ，记为 C_{out} ，通

过编辑初始代码 C_{in} 。形式上，将编辑指令记为 I_e 。

$$C_{out} = MLLM(V_{in}, V_{out}, C_{in}, I_e) \tag{2}$$

对于 MCE，我们仔细研究了多模态数学推理示例，并彻底定义了三种类型的编辑操作。总的来说，我们定义了四个评估任务。

- MCE（删除）：V_{out} 是通过从 V_{in} 中删除某些元素获得的，例如删除饼图中的某个饼块，或删除折线图中的一条线。
- MCE（修改）：V_{out} 是通过修改特定的视觉元素从 V_{in} 得出的——例如，增加或减少柱状图中柱子的高度，或拉伸/压缩函数图中的曲线。
- MCE（注释）：通过定位并注释 V_{in} 中某些元素的值来获得 V_{out} 。与其他任务不同，我们故意从输入中省略了 V_{out} ，以评估模型在适当位置放置值注释的能力。此方法使我们可以更深入地评估模型对正确标注位置的理解程度。

实现这些评估任务的一个重大挑战是缺乏用于执行此类评估的数据集。为了解决这一限制，我们手动构建了一个满足我们特定评估需求的数据集 MathOPEval （视觉数学操作评估）。构建过程包括两个阶段：（1）创建一个包含代码、图像和四种任务类型说明的初始数据集 D_{init} ；（2）将此数据集转换为两种格式——自由生成 D_{gen} 和多项选择题 D_{mc} 。

2.1 初始数据集构建

我们的初始数据集 D_{init} 由图像 V 、相应的构建代码 C 以及文本指令 I 组成。对于每幅图像，我们构建了三个图像变体及其相应的代码和用于四个评估任务的指令。具体而言，

- 原始状态（V_{orig} ， C_{orig} ， I_{orig} ）：原始图像、其代码以及图像的文本描述。
- 已删除状态（V_{del} ， C_{del} ， I_{del} ）：从 V_{orig} 中删除元素后的图像，其代码以及可能的删除操作提示。
- 修改后的状态（V_{mod} , C_{mod} , I_{mod} ）：图像中元素从 V_{orig} 修改而来，其代码和关于可能修改的提示。
- 注释状态 (V_{ann} , C_{ann} , I_{ann})：注释的图像、其代码以及包含注释元素和数值的对应说明指令，如角度测量、坐标或数值。

为了确保图像类型的多样性，我们考虑了三个来源，五种类型的数学可视化：几何图形、函数图、包括柱状图、折线图和饼图的统计图表。针对每种类型的数学可视化，我们考虑其上下文领域的多样性。我们描述了数据来源和构建细节如下。

- 几何图形：从 Geometry3K 数据集的几何问题 (Lu et al., 2021) 开始，我们首先获得 V_{orig} ，并手动标注图像描述（I_{orig} ）和代码 C_{orig} 。对于标注状态，我们根据问题描述增强图形的点标签和线段测量。删除状态涉及移除指定的线段以简化可视化，而修改状态则包括添加辅助几何元素，如中线、垂线、角平分线和平行线，以促进问题解决。
- 函数图：我们开发了一种标准化方法，利用从互联网上收集的常见高中水平的函数[1] [2] 。我们将函数的定义域设为 $[-5, 5]$ ，并生成初始图和说明（V_{orig} 、$C_{\text{orig}}, I_{\text{orig}}$ ）。标注阶段集中于标注关键点，包括 x 轴和 y 轴截点以及极值点。在删除阶段，我们删除部分定义域。修改阶段则实现各种函数变换。
- 统计图表：我们考虑三种类型的图表：柱状图、折线图和饼状图。我们从 ChartX 数据集 (Xia et al., 2024) 中收集原始状态（V_{orig} , C_{orig} , I_{orig} ）。在标注状态中，我们添加数字标签以表示关键值：柱状图中的柱高、折线图中点的 y 坐标，以及饼状图中的扇区百分比。在删除状态中，我们移除可视化中的选定元素：从柱状图中移除柱子，从折线图中移除特定 x 坐标上的点，以及从饼状图中移除区域。在修改状态中，我们调整特定的视觉属性，如更改柱状图中的柱高、更改折线图中选定点的 y 坐标，以及修改饼状图中的扇区比例。

[1] https://homework.study.com
[2] https://mathspace.co/us

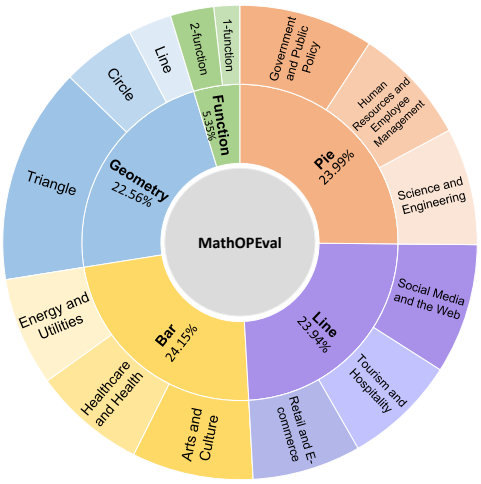

Figure 2: 五种可视化类型及其主要应用领域的分布。

Statistic	Number
Total samples	1,888
- Geometric figures	426
- Function plots	101
- Bar charts	456
- Line graphs	452
- Pie charts	453
Total questions	17,845
Total images	7,552
Multiple-choice questions	10,293
- Average question length	89.52
- Average choice length	132.12
Free-form questions	7,552
- Average question length	868.64
- Average answer length	783.77

Table 1: MathOPEval 的基本统计数据。

2.2 数据集格式转换

根据等式 equation 1 和等式 equation 2 中的定义，我们将初始数据集 D_{init} 转换为评估格式。具体来说，对于自由格式生成数据集（D_{gen}），

- MCG：$C_{\text{orig}} = MLLM(V_{\text{orig}}, I_{\text{orig}})$
- MCE（删除）：$C_{\text{del}} = MLLM(V_{\text{orig}}, V_{\text{del}}, C_{\text{orig}}, I_{\text{del}})$
- MCE（修改）：$C_{\text{mod}} = MLLM(V_{\text{orig}}, V_{\text{mod}}, C_{\text{orig}}, I_{\text{mod}})$
- 最大似然估计（注释）：$C_{\text{ann}} = MLLM(V_{\text{orig}}, C_{\text{orig}}, I_{\text{ann}})$

对于多项选择数据集（D_{mc}），我们为每个任务进一步增加了三个精心设计的错误选项，同时保持类似的格式和长度。

2.3 数据集统计

MathOPEval 的统计数据显示在表格 1 中。在初始数据集中 D_{init}，每个样本包含四组特定任务的代码、图像和指令。经过格式转换后，我们在生成数据集中 D_{gen} 得到 7,552 个问题，在多项选择数据集中 D_{mc} 得到 10,293 个问题，总共生成了 17,845 个问题。图 2 显示了不同类型可视化输入的比例统计，并列出了其主要的上下文领域。我们可以观察到，该数据集涵盖了多种领域，包括” 艺术与文化”、” 社交媒体和网络”，这表明其在不同上下文中的广泛适用性。

3 连锁思维评估策略

在构建评估数据集后，我们旨在开发一种有效且高效的策略来评估生成代码输出的正确性。虽然在多项选择格式中评估是直接的——通过提取答案选项并计算准确率——但在自由生成设置中评估变得更具挑战性。鉴于人工评估的高成本，我们利用具有强大语言和编码能力的最先进的 LLMs。具体来说，我们提出了一种自动化的链式思维（CoT）评估策略，利用 DeepSeek-V3 (DeepSeek-AI, 2024) 模型实现可靠且无偏的代码评估。我们的评估过程包括两个关键步骤：关键元素提取和分析评分。

关键元素提取为确保精确和高效的评估，我们指示评估模型首先从生成代码和参考代码中提取并比较关键元素，具体元素根据任务类型而有所不同。这种针对性的比较能够更可靠地评估代码正确性和任务特定的 MLLM 编码能力。

- 在 MCG 中，提取的元素因可视化类型而异：函数图的数学表达式，几何图形的点、线、形状等几何组件，以及统计图表的数据组件。对于柱状图，我们关注柱高、分

类标签、柱的数量和排列以及结构组织。折线图需要分析数据点、连接样式和点密度。饼图则根据切片比例、分类标签和切片顺序进行评估。我们从生成的和参考的 C_{out} 中提取这些元素以检查它们的身份。

- 在 MCE（删除）中，我们通过比较生成的和参考的 C_{out} 与 C_{in} 来分析被移除的元素。
- 在 MCE（修改）中，我们通过将生成的和参考 C_{out} 与 C_{in} 进行比较来检查修改后的元素。
- 在 MCE（注释）中，我们比较了 C_{in} 和 C_{out} 之间的额外注释代码段。

分析评分我们设计了一个综合的五级评分规则（1-2, 3-4, 5-6, 7-8, 9-10），用于从多个维度评估输出：内容准确性、位置一致性、代码正确性和完整性。对于每种可视化类型和任务，我们定义了详细的评估标准，对应于各个分数范围，如附录中所述。

两步合并为了说明完整的评估流程，图 5 展示了我们的几何图形 MCE（修改）任务过程的一个示例。根据完整的说明，评估模型首先识别生成代码和参考代码中相对于原始代码的关键元素。然后它在四个核心维度中进行全面分析。根据定义的评分标准，模型为代码质量分配一个整体评分。

人工评估为了评估我们的自动化评估的可靠性，我们随机抽取了 100 个模型生成的分数并进行了人工评估。结果表明，% 的分数被人工注释者认为是合理的，表明自动评估与人工判断之间有很强的一致性。关于人工评估标准的详细信息在附录中提供。

4 实验

4.1 实验设置

我们对多种模型进行全面评估，包括专有模型 GPT-4o (Hurst et al., 2024) 、Qwen-VL-Max (Yang et al., 2024) ，以及开源模型 Qwen2.5-VL (Yang et al., 2024) 、Gemma3 (Reid et al., 2024) 、LLaVA-NeXT (Liu et al., 2024) 系列和专门的推理模型 QVQ-72B (Yang et al., 2024) 。为了尽量减少模型输出的随机性，我们将温度参数设置为 0。对于选择题任务，我们研究了各种提示策略：（1）直接指示模型直接输出答案；（2）CoT (Wei et al., 2022) 引出逐步推理路径；（3）描述性思维链（DCoT）(Wu et al., 2023) 要求模型在回答之前生成描述；（4）思维可视化（VoT）(Wu et al., 2024) 提示模型想象推理路径，并指示“在每个推理步骤后可视化状态”。由于资源限制，我们评估了自由形式生成任务的直接提示。

4.2 主要结果

表 2 和 3 展示了多项选择题和自由生成任务的主要结果。对这些结果的分析揭示了几个关键发现：

模型表现：模型和人类基准在所有任务上存在显著的性能差距。人类直观理解的视觉操作对模型提出了重大挑战，平均性能差距为 52.39 %，而几何删除任务的差距最高达 67.13 %。在评估的模型中，GPT-4o 在自由生成任务中表现最佳，平均得分为 6.08，而 Qwen2.5-VL-72B-Instruct 在多选题中获得了最高的 68.02 % 准确率。在表现最低的另一端，Gemma3-4B-IT 在多选题和自由生成中分别得分为 21.00 % 和 3.30。结果表明，通常情况下，随着模型规模的增加，视觉操作理解能力有所提高，这与语言模型中的既定扩展规律 (Kaplan et al., 2020) 一致。

任务难度：不同任务类型的复杂性差异显著。在两种环境中，多模态代码修改任务最具挑战性，在选择题中平均正确率仅为 42.15 %，自由生成中的平均得分仅为 2.54。相比之下，多模态代码注释任务在两种环境中均获得最高分，这表明模型在理解和处理数字注释方面表现出色，但在需要对图像修改进行详细理解和操作的任务中则面临相当大的困难。

可视化挑战：在各种可视化类型中，函数图表现出最显著的挑战，在多项选择环境中仅达到 38.41 % 的准确率，而在自由形式生成中平均得分为 2.83。几何问题的难度等级其次。虽然饼图和折线图通常表现良好，平均准确率约为 50 %，折线图在多项选择准确率上意外地低于其他统计图，仅为 44.20 %，这揭示了理解和操作折线图修改的特定困难。

提示有效性：与先前的研究结果相反，直接提示以 43.33%的准确率优于其他策略，而 DCoT 的表现最低，为 37.59%。这一意外结果可能源于数学可视化中固有的更复杂的空间关系，

	MCG					MCE (Deletion)					MCE (Modification)					MCE (Annotation)					
	Geo	Func	Bar	Line	Pie	Geo	Func	Bar	Line	Pie	Geo	Func	Bar	Line	Pie	Geo	Func	Bar	Line	Pie	AVG
Human	93.33	91.67	98.33	96.67	96.67	100.00	100.00	100.00	100.00	100.00	96.67	98.33	98.33	100.00	98.33	98.33	100.00	96.67	98.33	96.67	97.92
GPT-4o																					
Direct	71.36	51.49	50.22	73.89	40.18	33.57	44.55	67.76	48.01	80.13	55.63	43.56	56.58	62.61	49.89	44.63	73.79	45.39	37.61	64.24	54.76
CoT	68.08	51.49	64.25	80.31	49.89	49.30	44.55	97.81	68.81	96.47	77.93	45.54	71.49	76.99	46.14	39.91	71.05	55.04	53.10	78.81	64.35
DCoT	78.17	50.50	58.77	83.85	50.99	48.59	22.77	97.81	67.26	96.91	77.00	48.51	69.30	75.88	47.68	54.59	75.15	69.74	66.59	80.35	66.02
VCoT	61.74	45.54	54.82	80.53	47.68	51.88	39.60	97.37	70.35	98.45	74.88	51.49	69.74	79.65	50.33	41.59	69.83	55.48	52.88	80.79	63.73
Qwen-VL-Max																					
Direct	84.51	43.56	41.45	58.85	43.93	52.82	59.41	91.89	69.91	98.45	68.54	44.55	67.11	60.62	54.30	50.06	70.43	51.75	50.22	71.52	61.69
CoT	79.11	46.53	84.43	90.93	49.67	45.77	68.32	97.15	72.35	98.90	81.92	47.52	76.54	78.98	48.79	48.71	67.62	61.62	47.57	62.25	67.73
DCoT	73.00	43.56	69.52	71.02	44.37	51.64	63.37	92.11	52.65	98.23	77.70	45.54	75.44	77.21	45.92	51.55	70.17	73.25	66.37	71.96	65.73
VCoT	77.46	45.54	85.96	86.06	48.34	50.94	63.37	96.71	72.79	98.90	81.46	46.53	77.63	78.32	51.43	48.06	66.06	64.47	50.66	66.67	67.87
Qwen2.5-VL-3B-Instruct																					
Direct	44.55	40.59	37.50	57.08	52.54	30.28	26.73	75.00	24.34	96.47	46.95	42.57	37.50	58.85	19.65	52.39	68.33	69.74	63.05	82.34	51.32
CoT	47.52	32.67	30.04	45.35	41.72	38.21	22.77	37.06	14.38	61.59	42.02	31.68	34.65	47.35	26.49	18.76	67.39	63.60	54.42	66.89	41.23
DCoT	41.58	39.60	32.68	38.27	45.03	31.46	31.68	69.30	21.68	92.05	49.06	41.58	30.70	50.22	19.87	41.91	61.22	74.78	69.69	73.95	47.82
VCoT	42.57	30.69	33.77	46.46	47.46	31.92	23.76	70.39	21.02	93.16	47.42	37.62	31.36	52.21	21.85	20.31	73.00	72.37	64.38	78.59	47.02
Qwen2.5-VL-72B-Instruct																					
Direct	91.55	44.55	83.77	95.13	71.96	41.55	56.44	99.56	75.22	99.78	77.93	52.48	63.82	72.79	35.98	72.70	79.62	92.11	91.15	97.79	74.79
CoT	77.70	47.52	87.94	93.58	24.06	32.63	50.50	96.93	41.37	98.90	82.16	48.51	62.94	79.20	26.05	50.06	68.68	77.63	54.65	85.87	64.34
DCoT	85.92	41.58	83.77	94.47	25.61	32.63	59.41	92.98	44.47	99.56	81.22	46.53	58.11	79.87	35.54	54.40	75.00	78.95	63.50	90.95	66.22
VCoT	78.40	42.57	88.82	93.81	61.59	33.33	53.47	96.71	53.76	98.90	81.69	46.53	59.87	78.98	27.59	48.19	70.52	78.07	54.42	86.75	66.70
Gemma3-4B-IT																					
Direct	32.86	26.73	30.48	24.78	28.48	34.04	14.85	25.66	20.35	37.53	31.22	24.75	33.77	35.40	33.55	29.43	9.89	33.11	29.42	30.24	28.33
CoT	10.09	21.78	26.75	24.56	23.84	25.59	17.82	37.94	19.69	23.84	28.40	31.68	41.89	44.03	41.28	9.77	23.07	25.88	14.60	29.14	26.08
DCoT	24.65	5.94	17.76	6.42	20.53	12.21	0.99	20.39	0.88	9.71	10.80	13.86	7.24	6.86	0.22	0.19	2.83	0.88	1.11	0.44	8.20
VCoT	10.09	18.81	18.86	9.29	22.52	14.79	9.90	8.55	21.90	22.30	23.71	30.69	38.38	34.96	32.45	10.03	19.70	28.29	24.34	28.70	21.41
Gemma3-27B-IT																					
Direct	44.84	17.82	21.05	26.55	34.00	35.45	24.75	22.15	33.19	25.39	29.58	22.77	28.29	24.56	43.71	24.26	4.94	28.51	25.66	31.79	27.46
CoT	15.26	23.76	31.36	28.10	36.87	16.90	26.73	18.42	21.24	27.59	27.70	13.86	33.99	37.61	44.15	11.32	10.96	29.39	22.57	32.01	25.49
DCoT	26.76	15.84	27.41	32.96	35.76	7.28	11.88	6.58	12.39	5.74	19.72	21.78	8.99	10.84	11.04	8.73	8.84	17.11	18.36	22.96	16.55
VCoT	5.87	24.75	27.41	25.88	35.54	25.59	24.75	11.40	15.27	25.17	31.46	22.77	41.01	40.93	35.98	18.76	15.26	24.12	23.67	32.23	25.39
LLaVA-NeXT-8B																					
Direct	42.96	26.73	23.03	30.09	40.18	25.35	41.58	83.55	51.55	68.21	42.49	41.58	31.80	31.86	33.11	43.40	51.19	51.32	35.84	50.55	42.32
CoT	30.05	12.87	24.34	11.50	20.53	20.66	39.60	28.07	12.83	39.07	43.43	12.87	15.13	15.27	22.96	10.35	43.29	44.74	21.90	29.58	24.95
DCoT	26.06	12.87	21.71	12.39	10.60	17.00	24.75	41.23	14.16	23.62	29.58	0.99	16.01	12.83	20.53	13.26	28.47	44.08	25.44	25.39	21.05
VCoT	35.92	17.82	25.44	10.40	26.27	25.82	34.65	44.96	30.53	36.20	31.22	5.94	17.98	18.36	28.26	13.13	41.79	33.77	19.69	27.37	26.28
LLaVA-NeXT-72B																					
Direct	73.24	34.65	35.75	36.95	33.33	36.62	41.58	84.65	33.85	83.89	50.47	30.69	30.04	38.94	35.76	58.21	63.49	63.38	49.56	69.98	49.25
CoT	68.54	26.73	30.92	41.37	19.43	34.98	43.56	86.40	44.91	83.00	46.24	24.75	36.18	45.58	29.80	52.20	56.25	67.98	53.32	71.96	48.21
DCoT	56.81	15.84	33.99	38.94	32.45	27.23	54.46	87.06	41.81	84.55	44.37	24.75	29.17	32.96	26.71	32.34	65.75	72.15	56.19	77.92	46.77
VCoT	64.55	27.72	33.55	38.94	32.01	35.68	46.53	85.96	44.03	82.34	44.84	23.76	26.75	35.84	34.00	49.68	56.46	69.30	55.97	72.85	48.04
AVG	52.24	32.15	44.30	49.65	37.42	32.87	37.16	64.67	38.65	68.28	51.21	33.38	43.11	49.27	33.78	35.09	50.94	53.69	44.31	58.53	45.53
AVG			43.15					48.33					42.15					48.51			45.53

Table 2: 五种可视化类型中的四种视觉操作的主要实验结果采用选择题形式。“Geo” 和 “Func” 分别是几何和函数的缩写。“AVG” 表示不同模型、可视化类型和视觉操作的平均准确率。粗体数字表示所有模型和提示方法中的最佳结果，而下划线数字表示各个模型在所有提示策略中的最佳准确率。

相比之下，在自然图像中，不精确的描述实际上会妨碍性能。值得注意的是，尽管 CoT 提示并不总是优于直接提示，但较大的模型更频繁地展示出增强的 CoT 能力，这表明随着模型规模的增加，推理能力可能会出现。

4.3 深入分析

本部分包含对视觉操作评估的进一步分析。

	MCG					MCE (Deletion)					MCE (Modification)					MCE (Annotation)					AVG
	Geo	Func	Bar	Line	Pie	Geo	Func	Bar	Line	Pie	Geo	Func	Bar	Line	Pie	Geo	Func	Bar	Line	Pie	
GPT-4o	3.18	2.59	8.29	8.34	8.12	4.78	4.27	9.57	7.28	9.32	5.42	4.44	2.59	3.41	2.44	4.35	5.87	9.12	8.37	7.99	6.08
Qwen-VL-Max	2.86	2.40	8.23	8.56	8.50	3.59	4.64	9.78	7.71	9.10	4.54	3.59	2.61	4.18	2.49	5.00	5.27	9.06	7.28	8.34	5.97
Qwen2.5-VL-3B-Instruct	2.19	1.88	6.71	8.05	8.80	1.18	3.63	3.30	2.38	6.42	2.45	2.70	1.80	1.60	1.81	3.09	4.17	5.27	5.18	3.65	3.83
Qwen2.5-VL-72B-Instruct	2.76	1.94	8.45	8.83	8.60	3.73	4.05	9.59	8.33	8.75	4.70	3.29	2.48	4.21	2.57	4.15	5.55	9.17	7.73	7.75	5.94
Gemma3-4B-IT	1.90	1.08	6.37	6.67	8.49	1.21	1.01	1.39	1.07	1.97	2.00	1.00	2.01	1.96	2.32	2.51	1.01	7.00	7.39	4.58	3.30
Gemma3-27B-IT	2.62	1.27	7.06	7.52	8.79	1.92	1.00	3.31	2.91	6.07	2.09	1.00	1.99	1.97	2.45	3.31	1.00	8.61	6.98	5.40	4.05
LLaVA-NeXT-8B	1.71	1.06	5.92	6.85	8.57	1.67	3.43	1.10	1.12	1.81	1.81	3.34	1.74	1.69	1.74	2.63	3.89	5.06	5.93	8.72	3.51
LLaVA-NeXT-72B	2.34	1.21	7.36	7.17	8.79	1.54	3.38	4.34	1.88	6.65	2.30	1.67	1.83	1.66	1.91	3.62	4.00	7.86	7.71	4.06	4.09
AVG	2.44	1.68	7.30	7.75	8.58	2.45	3.18	5.30	4.09	6.26	3.16	2.63	2.13	2.58	2.22	3.58	3.85	7.64	7.07	6.31	4.51
AVG			5.55					4.26					2.54					5.69			4.59

Table 3: 在生成格式中，五种可视化类型下四种视觉操作的主要实验结果。“Geo” 和 “Func” 分别是几何和函数的缩写。“AVG” 代表不同模型、可视化类型以及视觉操作的平均得分。加粗数字表示所有模型中的最佳结果。

推理增强模型与通用模型：为了研究推理增强模型在视觉操作中是否优于相似规模的通用模型，我们对 QVQ-72B-Preview 和 Qwen2.5-VL-72B-Instruct 进行比较分析。图 3 透露了一个有趣的模式：推理增强模型在数学函数的多模态代码生成和折线图的多模态代码删除上表现出色，但在其他任务上表现相当或者略逊。这种性能差异表明，尽管推理增强有效提升了模型的推理能力，但并不必然转化为在一般视觉操作中的更好表现。

Model	Angle	Line	AVG
GPT-4o	5.41	3.28	4.35
Qwen-VL-Max	5.19	4.81	5.00
Qwen2.5-VL-3B-Instruct	3.31	2.87	3.09
Qwen2.5-VL-72B-Instruct	5.42	2.87	4.15
Gemma3-4B-IT	2.62	2.39	2.51
Gemma3-27B-IT	3.97	2.64	3.31
LLaVA-NeXT-8B	2.69	2.56	2.63
LLaVA-NeXT-72B	4.48	2.76	3.62
AVG	4.14	3.02	3.58

Table 4: 自由形式生成的几何图形中角度和线条标注的比较。

Model	Backward	Forward	AVG
GPT-4o	62.82	45.18	54.00
Qwen-VL-Max	72.48	49.60	61.04
Qwen2.5-VL-3B-Instruct	57.81	33.34	45.58
Qwen2.5-VL-72B-Instruct	79.84	56.34	68.09
Gemma3-4B-IT	17.60	12.35	14.98
Gemma3-27B-IT	19.21	15.77	17.49
LLaVA-NeXT-8B	27.93	20.04	23.98
LLaVA-NeXT-72B	49.94	48.11	49.03
AVG	48.45	35.09	41.77

Table 5: 几何视觉注释中多选格式的正向问题和反向问题对比。

Model	X	Y	Extreme	AVG
GPT-4o	6.03	6.88	4.71	5.87
Qwen-VL-Max	5.26	6.50	4.06	5.27
Qwen2.5-VL-3B-Instruct	3.88	4.44	4.20	4.17
Qwen2.5-VL-72B-Instruct	5.16	6.52	4.98	5.55
Gemma3-4B-IT	1.01	1.01	1.00	1.01
Gemma3-27B-IT	1.00	1.00	1.00	1.00
LLaVA-NeXT-8B	4.04	3.33	4.32	3.89
LLaVA-NeXT-72B	4.22	4.40	3.40	4.00
AVG	3.82	4.26	3.46	3.85

Table 6: 函数类型视觉标注中自由形式生成的 x 截距、y 截距和极值点的比较。

Model	X	Y	Extreme	AVG
GPT-4o	74.65	60.15	82.57	72.45
Qwen-VL-Max	71.28	57.67	76.76	68.57
Qwen2.5-VL-3B-Instruct	69.68	64.11	68.66	67.48
Qwen2.5-VL-72B-Instruct	74.47	53.47	92.43	73.45
Gemma3-4B-IT	10.28	14.60	16.73	13.87
Gemma3-27B-IT	2.84	12.38	14.79	10.00
LLaVA-NeXT-8B	54.79	33.91	34.86	41.19
LLaVA-NeXT-72B	65.43	60.40	55.63	60.49
AVG	52.93	44.59	55.30	50.94

Table 7: 在选择题格式中比较功能可视化注释中的 x 截距、y 截距和极值点。

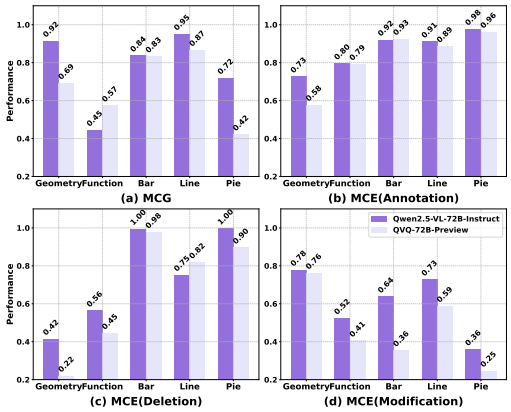

Figure 3: Comparison between reasoning-enhanced and general-purpose models on multiple-choice format using Direct prompt.

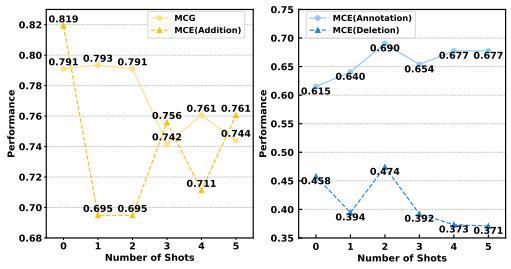

Figure 4: Analysis of shot count effects on Qwen-VL-Max's multiple-choice performance across four visual operations in geometric figures, using CoT prompt.

这样的观察引发了关于模型能力专业化的重要问题，以及在增强专注推理和一般视觉理解之间的权衡。

上下文学习的影响：我们使用 Qwen-VL-Max 对几何图形进行上下文学习（ICL）的有效性进行研究。如图 4 所示，与零样本学习相比，ICL 在注释和删除任务中显著提高了性能，这可以归因于这些操作的相对简单性。然而，当涉及到更复杂的任务时，比如包含多样的几何元素（例如三角形、矩形）和各类线型（包括平行线、垂线、角平分线和中线）的视觉创建和修改，ICL 可能引入意外的噪音。这种复杂性及可能操作的广泛变动性实际上可能导致性能下降。

射击计数分析：图 4 显示增加射击次数并不单调地提高表现。表现通常在 2 次射击时达到峰值，并在超过 3 次射击后下降。这一模式表明精心选择射击次数对优化表现至关重要。较高射击次数导致表现下降可能归因于两个因素。一方面，虽然初始射击有助于增强任务理解，但过多的示例会引入混乱，特别是在配置多样化的情况下。另一方面，模型有限的上下文窗口限制了其有效处理大量射击的能力，从而导致表现下降。

交点与极值点：我们对函数图多模态代码注释的分析显示了不同数学特征的不同性能模式。总体而言，模型在识别极值点方面表现最佳，其次是 x 截距，而 y 截距是最具挑战性的。有趣的是，我们观察到 Qwen2.5-VL-3B-Instruct 在 y 截距任务上表现优于 Qwen2.5-VL-72B-Instruct，而 Gemma3-4B-IT 在两种截距任务上都超过了 Gemma3-27B-IT。这一违反直觉

的发现表明，模型大小并不一定与对坐标交点的理解能力改进相关联。

前向与后向问题：在几何图形的多模态代码注释选择题中，我们比较了前向问题和后向问题的表现。前者基于几何元素识别注释，例如，角 ABC 的度数是多少，而后者基于注释识别几何元素，例如，哪个角的度数是 120 度。如表 5 所示，模型在后向问题上的表现始终更好。我们假设这种不对称可能源于这样一个事实：后向问题提供了明确的数值锚点，有助于收缩搜索空间，而前向问题则需要更全面的几何理解和空间推理。

点 vs. 线：关于几何问题中多模态代码注释的生成设置，我们进一步将注释分为两种类型：角度标记和线标签。如图 4 所示，角度注释的得分显著高于线注释。我们将这种性能差异归因于不同的放置要求的复杂性：角度标记只需放置在确定的顶点附近，而线标签需要放置在线段的中点。这种对线标签更严格的定位要求增加了准确放置的难度。

4.4 误差分析

我们对 100 个随机选择的包含推理过程的输出进行了人工分析，这些输出中有 50 个是正确的过程，50 个是错误的过程。我们的检查揭示了模型性能中令人担忧的模式：在正确答案中，只有 33 % 的输出同时展示了准确的结果和有效的推理过程，而 17 % 则是通过错误的推理得到了正确的答案——这表明正确的输出不一定反映出健全的问题解决能力。我们还识别出了四类不同的失败，其中视觉感知错误是主要问题，占所有错误的 86 % 。这一显著高的百分比突出了模型在准确处理和解释视觉元素方面的关键限制。指令理解错误和输出格式违反各占失败的 6 % ，而推理过程错误则构成了剩下的 2 % 。这一错误分布表明，视觉感知机制需要大幅改进以实现可靠的视觉问题解决。

5 相关工作

多模态数学评估：随着 MLLMs 的快速发展，已经出现了众多基准来评估视觉数学推理能力 (Wang et al., 2024; He et al., 2024; Yue et al., 2024; Zhang et al., 2024a; Wang et al., 2024; Sun et al., 2024; Qiao et al., 2024; Zhou et al., 2024; Xu et al., 2023; Masry et al., 2022) 。之前的工作 (Chen et al., 2022; Seo et al., 2015; Cao & Xiao, 2022) ，如 Inter-gps (Lu et al., 2021) 和 GeoQA (Chen et al., 2021) ，集中在几何问题上，而其他 (Lu et al., 2024; Zhang et al., 2024b) 则涵盖了更广泛的任务，包括函数图和统计图表。尽管最近的工作 (Hu et al., 2024a; Cheng et al., 2024; Lee et al., 2024) 引入了“多模态输入-多模态输出”范式，但现有评估仍只评估最终答案的准确性，忽视了推理过程中的中间视觉操作质量。

可视化编程：通过编程接口将复杂的视觉任务分解为可管理的步骤这一方法已展示出在提升视觉推理能力方面的显著潜力 (Yao et al., 2023; Yang et al., 2023; Zeng et al., 2023; Hu et al., 2024b) 。这种方法的核心思想是利用计算机程序的逻辑性和精确性，将传统上依赖人类直觉和经验的复杂视觉问题转化为一系列清晰的、逐步的操作。其中，Visprog (Gupta & Kembhavi, 2023) 和 ViperGPT (Surís et al., 2023) 展示了使用 MLLMs 生成用于顺序视觉操作的 Python 代码的有效性。SKETCHPAD (Hu et al., 2024a) 进一步基于中间结果实现动态视觉操作，特别是在通过代码绘制辅助线等视觉操作中，几何推理取得了成功。这一进展激励了我们的研究，促使我们超越仅评估最终数学解决方案，转向评估推理过程中中间视觉操作的质量和适用性。

在本文中，我们提出了首个针对数学推理任务中视觉操作的细粒度评估基准。我们的框架全面评估了五种数学可视化中的四种基本视觉操作，为评估大模型的能力提供了系统的方法。通过广泛的实验，我们发现了当前模型与人类之间的显著性能差距，特别是在函数图和多模态代码修改任务中。此外，我们观察到，虽然上下文学习可以提高性能，但其效果在不同任务中存在差异，并且需要仔细考虑样本数量。我们希望我们的框架和发现能够促进在视觉数学推理中开发更强的模型。

这项工作开始了对 MLLMs 在数学推理任务中视觉操作的评估，为未来的研究铺平了道路。有几个方向仍然开放：(1) 扩大评估范围，以涵盖更多的多模态推理技能，如在数学问题中的计划、反思和错误修正；以及 (2) 探索如何利用这一基准来对齐 MLLMs，增强其在数学场景中的多模态推理能力。

References

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022, pp. 1511–1520. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.130.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021, volume ACL/IJCNLP 2021 of Findings of ACL, pp. 513–523. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.FINDINGS-ACL.46. URL https://doi.org/10.18653/v1/2021.findings-acl.46.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pp. 3313–3323. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.218. URL https://doi.org/10.18653/v1/2022.emnlp-main.218.

Zihui Cheng, Qiguang Chen, Jin Zhang, Hao Fei, Xiaocheng Feng, Wanxiang Che, Min Li, and Libo Qin. Comt: A novel benchmark for chain of multi-modal thought on large vision-language models. CoRR, abs/2412.12932, 2024. doi: 10.48550/ARXIV.2412.12932. URL https://doi.org/10.48550/arXiv.2412.12932.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei A. F. Florêncio, and Cha Zhang. Refocus: Visual editing as a chain of thought for structured image understanding. CoRR, abs/2501.05452, 2025. doi: 10.48550/ARXIV.2501.05452. URL https://doi.org/10.48550/arXiv.2501.05452.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, pp. 14953–14962. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01436. URL https://doi.org/10.1109/CVPR52729.2023.01436.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.211. URL https://doi.org/10.18653/v1/2024.acl-long.211.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/fb82011040977c7712409fbdb5456647-Abstract-Conference.html.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 9590–9601. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.00916. URL https://doi.org/10.1109/CVPR52733.2024.00916.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. Gpt-4o system card. CoRR, abs/2410.21276, 2024. doi: 10.48550/ARXIV.2410.21276. URL https://doi.org/10.48550/arXiv.2410.21276.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. CoRR, abs/2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Jeongwoo Lee, Kwangsuk Park, and Jihyeon Park. VISTA: visual integrated system for tailored automation in math problem generation using LLM. CoRR, abs/2411.05423, 2024. doi: 10.48550/ARXIV.2411.05423. URL https://doi.org/10.48550/arXiv.2411.05423.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 26286–26296. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pp. 6774–6786. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.ACL-LONG.528. URL https://doi.org/10.18653/v1/2021.acl-long.528.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL https://doi.org/10.18653/v1/2022.findings-acl.177.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma Gongque, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? CoRR, abs/2407.01284, 2024. doi: 10.48550/ARXIV.2407.01284. URL https://doi.org/10.48550/arXiv.2407.01284.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. CoRR, abs/2403.05530, 2024. doi: 10.48550/ARXIV.2403.05530. URL https://doi.org/10.48550/arXiv.2403.05530.

Min Joon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton (eds.), Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015, pp. 1466–1476. The Association for Computational Linguistics, 2015. doi: 10.18653/V1/D15-1171. URL https://doi.org/10.18653/v1/d15-1171.

Kai Sun, Yushi Bai, Ji Qi, Lei Hou, and Juan-Zi Li. MM-MATH: advancing multimodal math evaluation with process evaluation and fine-grained classification. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, pp. 1358–1375. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-emnlp.73.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, pp. 11854–11864. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01092. URL https://doi.org/10.1109/ICCV51070.2023.01092.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), Advances

in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/a45296e83b19f656392e0130d9e53cb1-Abstract-Conference.html.

Yifan Wu, Pengchuan Zhang, Wenhan Xiong, Barlas Oguz, James C. Gee, and Yixin Nie. The role of chain-of-thought in complex vision-language reasoning task. CoRR, abs/2311.09193, 2023. doi: 10.48550/ARXIV.2311.09193. URL https://doi.org/10.48550/arXiv.2311.09193.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. CoRR, abs/2402.12185, 2024. doi: 10.48550/ARXIV.2402.12185. URL https://doi.org/10.48550/arXiv.2402.12185.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. Chartbench: A benchmark for complex visual reasoning in charts. CoRR, abs/2312.15915, 2023. doi: 10.48550/ARXIV.2312.15915. URL https://doi.org/10.48550/arXiv.2312.15915.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. CoRR, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: prompting chatgpt for multimodal reasoning and action. CoRR, abs/2303.11381, 2023. doi: 10.48550/ARXIV.2303.11381. URL https://doi.org/10.48550/arXiv.2303.11381.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, pp. 9556–9567. IEEE, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://doi.org/10.1109/CVPR52733.2024.00913.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S. Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net, 2023. URL https://openreview.net/forum?id=G2Q2Mh3avow.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, Chenghua Lin, Wenhao Huang, Wenhu Chen, and Jie Fu. CMMMU: A chinese massive multi-discipline multimodal understanding benchmark. CoRR, abs/2401.11944, 2024a. doi: 10.48550/ARXIV.2401.11944. URL https://doi.org/10.48550/arXiv.2401.11944.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII, volume 15066 of Lecture Notes in Computer Science, pp. 169–186. Springer, 2024b. doi: 10.1007/978-3-031-73242-3_10. URL https://doi.org/10.1007/978-3-031-73242-3_10.

Minxuan Zhou, Hao Liang, Tianpeng Li, Zhiyu Wu, Mingan Lin, Linzhuang Sun, Yaqi Zhou, Yan Zhang, Xiaoqin Huang, Yicong Chen, Yujing Qiao, Weipeng Chen, Bin Cui, Wentao Zhang, and Zenan Zhou. Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. CoRR, abs/2408.07543, 2024. doi: 10.48550/ARXIV.2408.07543. URL https://doi.org/10.48550/arXiv.2408.07543.

在不同的评估场景中，我们设计了特定的提示词来引导模型的回应。我们考虑了两种主要的问题类型：选择题和自由生成任务。

对于多项选择题，提示结构如下：

- 直接：

  Question: {question}
  Choices:
  {formatted_choices}
  Please directly output the correct option.

- 逐步思考（CoT）：

  Question: {question}
  Choices:
  {formatted_choices}
  Let's think step by step. Please equip the correct option with \boxed{} at the end of your response.

- DCoT：

  Question: {question}
  Choices:
  {formatted_choices}
  Describe the image information relevant to the question.
  Please equip the correct option with \boxed{} at the end of your response.

- VCoT：

  Question: {question}
  Choices:
  {formatted_choices}
  Visualize the state after each reasoning step.
  Please equip the correct option with \boxed{} at the end of your response.

对于自由形式生成任务，提示遵循类似的模式：

- 直接：

  Question: {question}
  Please directly output your code using markdown format.

- 共网 (CoT):

  Question: {question}
  Let's think step by step. Please wrap your code using markdown format at the end of your response.

- 双协同训练：

  Question: {question}
  Describe the image information relevant to the question.
  Please wrap your code using markdown format at the end of your response.

- VCoT：

  Question: {question}
  Visualize the state after each reasoning step.
  Please wrap your code using markdown format at the end of your response.

每个提示都旨在引出特定类型的回应，同时在不同的推理方法中保持一致的结构。

6 人工标注

起始数据集的标注是由计算机科学博士候选人的作者进行的，他们在机器学习和计算机视觉方面拥有丰富的专业知识。我们在该领域的理论基础和实践应用方面具备全面的认识，确保高质量的标注。

6.1 人类评价的主要实验

我们精心构建了一个由 400 个样本组成的测试子集，这些样本系统地从四种视觉任务和五个不同类别中选择（从每种组合中随机选择 20 个样本）。我们邀请了三位独立的评估人员，他们都有很强的计算机科学背景，但没有参与初始注释过程，以评估这些样本。这些评估人员是专门根据他们的编程专业知识来选择的，因为评估涉及代码评估组件。

评估过程的结构如下：

- 每位评估员独立评估了所有 400 个样本
- 结果取所有三位评估者的平均值，以尽量减少个人偏见

6.2 评估 LLM 的评分可靠性

为了评估大型语言模型 (LLM) 在评分任务中的可靠性，我们通过人工注释进行了评估。我们的分析重点在于检查 LLM 所分配的分数是否符合人类判断和既定标准。评估过程遵循了具有明确定义的评估标准的结构化方法。

以下标准用于判断 LLM 评分的合理性：

- 一致性：分数是否在相似的回答中保持一致性
- 理由：LLM 是否为其评分决策提供了逻辑解释
- 一致性：与人类专家评分的一致程度
- 语境理解：LLM 考虑语境因素的能力

我们为每个评分实例实现了一个二分类系统：

- 合理：分数符合评估标准
- 不合理：分数显示出与预期标准的显著偏离

在对评分样本进行彻底的人工审查后，我们发现 92 % 的 LLM 评分决定被认为是合理的。这个高比例表明，LLM 在评分方面表现出强大的能力。

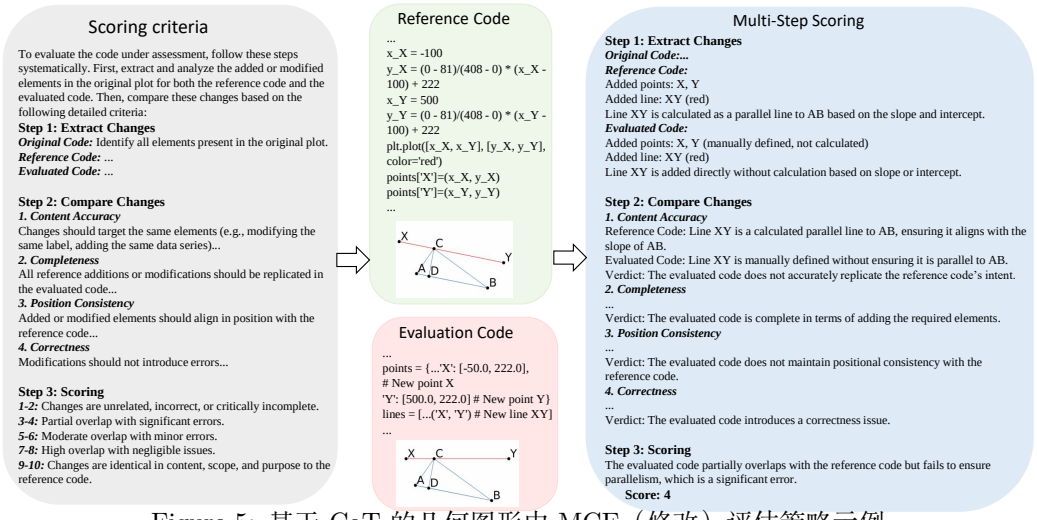

Figure 5: 基于 CoT 的几何图形中 MCE（修改）评估策略示例。

6.3 错误分析

我们的研究涉及分析 100 个随机选择的带有推理过程的输出，平均分为 50 个正确和 50 个错误案例。分析揭示了模型表现中的令人担忧的趋势：在正确答案中，只有 33 % 的答案和推理都是准确的，而 17 % 通过错误的推理得出了正确的结论——这表明正确的输出可能并不真正代表有效的问题解决能力。

我们识别出四个不同的错误类别，其中视觉感知错误最为显著，占所有错误的 86 %。这一显著高的比例表明模型在处理和解释视觉信息方面存在根本性的不足。指令理解错误和输出格式违规各占失败的 6 %，推理过程错误占剩下的 2 %。基于这种错误分布，很明显需要对视觉感知机制进行大幅改进，以实现可靠的视觉问题解决能力。

7 评分提示

图表示了不同任务类型和不同图像类型的评分提示。

Prompt for Scoring of Visual Creation on Bar Chart

Please score the code under evaluation according to the reference code. Focus on the following key data elements for bar charts:
1. Bar Heights: Specific numerical values for the heights (y-values).
2. Category Labels: Labels for the categories (x-axis).
3. Number and Order of Bars: Quantity and sequence of bars.
4. Stacked/Grouped Structure: Whether bars are stacked or grouped.
Score criteria:
1-2: Completely different heights/categories.
3-4: Partial match but significant differences.
5-6: Moderate similarity with noticeable differences.
7-8: High similarity with minor differences.
9-10: Nearly identical.
Reference Code:
Code Under Evaluation:
Please enclose the final score in \\boxed{{}}.

Figure 6: 柱状图多模态代码生成的评分提示。

Prompt for Scoring of Visual Creation on Line Chart

Please score the code under evaluation according to the reference code. Focus on the following key data elements for line charts:
1. Data Points: Specific x/y coordinates.
2. Connection Style: Straight lines or smooth curves.
3. Point Density: Number of data points.

Score criteria:
1-2: Completely different data points/trends.
3-4: Partial match with different trends.
5-6: Similar trends with noticeable differences.
7-8: Highly similar trends with minor differences.
9-10: Nearly identical.
Reference Code:
Code Under Evaluation:
Please enclose the final score in \\boxed{{}}.

Figure 7: 用于对折线图的多模态代码生成进行评分的提示。

Prompt for Scoring of Visual Creation on Pie Chart

Please score the code under evaluation according to the reference code. Focus on the following key data elements for pie charts:
1. Proportions: Numerical values for slices.
2. Category Labels: Names for each slice.
3. Slice Order: Sequence of categories.

Score criteria:
1-2: Completely different proportions/categories.
3-4: Partial match with major differences.
5-6: Moderate similarity with clear differences.
7-8: High similarity with minor differences.
9-10: Nearly identical.
Reference Code:
Code Under Evaluation:
Please enclose the final score in \\boxed{{}}.

Figure 8: 用于对饼状图的多模态代码生成进行打分的提示。

Prompt for Scoring of Visual Creation on Geometric Figure

Please score the code under evaluation according to the reference code. Focus on the following key data elements for geometric plots:
1. Points: Coordinates and distribution.
2. Lines: Start/end points and connection style.
3. Shapes: Type (circle/rectangle), position (center/vertices), and size (radius/side length).

Score criteria:
1-2: Completely different elements.
3-4: Partial match of elements with major differences.
5-6: Moderate similarity with clear differences.
7-8: High similarity with minor differences.
9-10: Nearly identical.
Reference Code:
Code Under Evaluation:
Please enclose the final score in \\boxed{{}}.

Figure 9: 用于几何图形多模态代码生成评分的提示。

Prompt for Scoring of Visual Creation on Function Plot

To evaluate the formula under assessment, focus exclusively on the mathematical expression.

Score Criteria:
- 1-2: Different function type (e.g., `cos` vs. `sin`, `x²` vs. `x³`).
- 3-4: Same function type but critical structural differences (e.g., `sin(2x)` vs. `sin(x)`, `x²` vs. `2x²`).
- 5-6: Same function type with moderate structural differences (e.g., `sin(x) + 1` vs. `sin(x)`, `x²` vs. `x² + 3`).
- 7-8: Same function type with minor structural differences (e.g., `sin(x + 0.1)` vs. `sin(x)`).
- 9-10: Exact match of the mathematical expression (e.g., `y = sin(2x)` vs. `y = sin(2x)`).

Input Data
Reference Formula:

Formula Under Evaluation:

Final Output
Enclose the final score in \\boxed{{}}.

Figure 10: 用于对函数图多模态代码生成进行评分的提示。

Prompt for Scoring of Visual Annotation

To evaluate the code under assessment, follow these steps systematically. First, extract and analyze the annotations added to the original plot in both the reference code and the evaluated code. Then, compare the annotations based on the following detailed criteria:
Step 1: Extract Annotations
- Reference Code: Identify all annotations (e.g., text labels, numerical values, or symbols) added to the plot by the reference code. Note their content, position, and purpose.
- Evaluated Code: Similarly, identify all annotations added to the plot by the evaluated code. Note their content, position, and purpose.
Step 2: Compare Annotations
Compare the annotations from both codes using the following criteria:
1. Annotation Content
- Both codes must label the same elements (e.g., angles, coordinates, proportions). For example, if the reference code labels "∠A = 60°", the evaluated code should label the same angle with equivalent content such as "Angle A: 60°".
- Differences in phrasing are acceptable as long as the meaning and intent remain consistent.
2. Annotation Position
- Annotations should be placed at approximately the same relative positions in the plot. For example, if the reference code places a label near `(x=1, y=10)`, the evaluated code can place it nearby, such as `(x=1, y=10.1)`.
- Exact pixel-perfect alignment is not required, but the annotations must be close enough to refer to the same visual element.
3. Annotation Completeness
- Both codes must label all critical elements in the plot. For example, if the reference code labels all peaks in a line chart, the evaluated code must also label all peaks.
- Missing any critical annotation in the evaluated code will result in a lower score.
4. Annotation Correctness
- Labels must show the correct numerical or textual values. For example, if the reference code shows `"y=15"`, the evaluated code must also show `"y=15"` or a value that is mathematically equivalent.
- Small numerical differences (e.g., rounding errors) are acceptable, but significant deviations are not.
Step 3: Scoring
Assign a score based on the comparison:
- 1-2: Annotations are unrelated or critically incorrect (e.g., labels wrong elements, major omissions).
- 3-4: Partial match with major differences in content, position, or completeness.
- 5-6: Moderate similarity with noticeable differences in completeness or correctness.
- 7-8: High similarity with only minor differences in style, position, or phrasing.
- 9-10: Annotations are nearly identical in content, position, and completeness.
Input Data
Original Code:
Reference Annotations Code:
Evaluated Annotations Code:
Final Output
Enclose the final score in \\boxed{{}}.

Figure 11: 用于多模态代码注释评分的提示。

Prompt for Scoring of Visual Deletion

To evaluate the code under assessment, follow these steps systematically. First, extract and analyze the deleted elements from the original plot in both the reference code and the evaluated code. Then, compare the deletions based on the following detailed criteria:

Step 1: Extract Deletions

- Original Code: Identify all elements present in the original code.
- Reference Code: Identify elements explicitly deleted compared to the original code.
- Evaluated Code: Similarly, identify elements deleted compared to the original code.

Step 2: Compare Deletions

Compare the deletions from both codes using the following criteria:

1. Element Matching

- Exact Match: Deleted elements in the evaluated code must align with the reference code (same content, type, and position).
- Partial Match: Deletions target the same general component (e.g., a label) but differ in specifics (e.g., wrong text or position).
- No Match: Deleted elements are unrelated to the reference code's deletions.

2. Completeness

- All reference deletions are replicated in the evaluated code.
- Missing deletions or extra deletions (not in the reference code) reduce the score.

3. Correctness

- Deletions must not introduce errors (e.g., removing critical structural code that alters the plot type).

Step 3: Scoring

Assign a score based on the comparison:

- 1-2: No meaningful overlap; deletions are incorrect or irrelevant.
- 3-4: Partial overlap with major errors (e.g., wrong elements deleted, significant omissions).
- 5-6: Moderate overlap with minor errors (e.g., correct elements deleted but with positional discrepancies).
- 7-8: High overlap with negligible issues (e.g., stylistic differences in deletion method).
- 9-10: Deletions are identical in content, scope, and purpose to the reference code.

Input Data

Original Code:

Reference Code:

Evaluated Code:

Final Output

Enclose the final score in \\boxed{{}}.

Figure 12: 用于多模态代码删除评分的提示。

Prompt for Scoring of Visual Modification

To evaluate the code under assessment, follow these steps systematically. First, extract and analyze the added/modified elements in the original plot for both the reference code and the evaluated code. Then, compare these changes based on the following detailed criteria:

Step 1: Extract Changes

- Original Code: Identify all elements present in the original plot.
- Reference Code: Identify elements added/modified compared to the original code. Note their content, position, and purpose.
- Evaluated Code: Similarly, identify elements added/modified compared to the original code. Note their content, position, and purpose.

Step 2: Compare Changes

Compare the modifications/additions from both codes using the following criteria:

1. Content Accuracy

- Changes must target the same elements (e.g., modifying the same label, adding the same data series).
- Content (text, values, colors) must match the reference code's intent.

2. Completeness

- All reference additions/modifications are replicated in the evaluated code.
- Missing changes or extra changes (not in the reference code) reduce the score.

3. Position Consistency

- Added/modified elements must align in position with the reference code.

4. Correctness

- Modifications must not introduce errors (e.g., altering the plot type or breaking the code structure).

Step 3: Scoring

Assign a score based on the comparison:

- 1-2: Changes are unrelated, incorrect, or critically incomplete (e.g., wrong elements modified, major omissions).
- 3-4: Partial overlap with significant errors (e.g., correct elements targeted but incorrect values/positions).
- 5-6: Moderate overlap with minor errors (e.g., correct changes but slight positional/style discrepancies).
- 7-8: High overlap with negligible issues (e.g., stylistic differences in label formatting).
- 9-10: Changes are identical in content, scope, and purpose to the reference code.

Input Data

Original Code:

Reference Code:

Evaluated Code:

Final Output

Enclose the final score in \\boxed{{}}.

Figure 13: 多模态代码修改评分的提示。