VL-ICL Bench：多模态情境学习基准测试细节中的魔鬼

Yongshuo Zong*, **Ondrej Bohdal***, **Timothy Hospedales**
School of Informatics
University of Edinburgh
{ yongshuo.zong, ondrej.bohdal, t.hospedales } @ed.ac.uk

Abstract

大型语言模型（LLM）以新兴上下文学习（ICL）而闻名，即使用作为提示提供的少量示例快速适应新任务的能力，而无需更新模型的权重。基于 LLM 构建的视觉大型语言模型（VLLM）在识别、推理和基础等领域取得了重大进展。然而，对多模态 ICL 的研究主要集中在小样本视觉问答（VQA）和图像标题上，我们将展示它们既没有利用 ICL 的优势，也没有测试其局限性。多式联运 ICL 的更广泛功能和局限性仍未得到充分探索。在这项研究中，我们引入了一个全面的多模态情境学习基准 VL-ICL Bench ，涵盖了涉及图像和文本作为输入和输出的广泛任务，以及来自 对推理和长上下文长度的感知 的不同类型的挑战。我们根据该基准测试套件评估了最先进的 VLLM 的能力，揭示了它们的不同优势和劣势，并表明即使是最先进的模型（例如 GPT-4）也发现这些任务具有挑战性。通过强调一系列新的 ICL 任务，以及现有模型的相关优势和局限性，我们希望我们的数据集能够激发未来在增强 VLLM 上下文学习能力方面的工作，并激发利用 VLLM ICL 的新应用。代码和数据集可在 https://github.com/ys-zong/VL-ICL 获得。

1 介绍

随着模型规模的扩大，大型语言模型（LLM）展示了著名的上下文学习（ICL） (**??**) 的新兴能力。这是指在单个前馈传递中从类比中学习的能力 - 因此，使模型能够使用一些输入输出示例来学习全新的任务，而无需对模型参数进行任何更新。ICL 的这种免培训特性使其在广泛的场景和应用中得到了快速和广泛的应用，如 (**????**) 等基准所示。

视觉大型语言模型（VLLM）通常建立在基础 LLM 上，通过通过某种拼接机制连接的视觉编码器和/或解码器来增强它 (**???????**) 。这些模型与 LLM 一起迅速发展，并因其在零样本识别、推理、接地和视觉问答（VQA）等功能方面的卓越多模态能力而备受关注。这些功能已经过一系列最近的基准测试套件的全面测试 (**????**) 。同时，VLLM 也被广泛认为从其基础 LLM 继承了情境学习（ICL）能力。然而，他们在这方面的能力没有得到充分的评估和理解。目前的 VLLM 研究主要报告了其通过上述基准测量的零样本能力，而 ICL 通常仅进行定性评估，或通过小样本视觉问答（VQA）或图像字幕 (**????**) 作为次要考虑因素，在更广泛的 ICL 任务中，定量评估存在明显缺陷。这大概是由于 VQA 和字幕基准测试基础设施的现成可用性。然而，我们将证明字幕和 VQA 任务对于 ICL 评估来说并不理想：它们既没有真正利用 ICL 的能力来提高示例的性能; 他们也没有测试 ICL 可以做的极限，以激励未来的 VLLM 研究更好地利用和揭示 LLM 的 ICL 能力的基础。

为了增强对多模态 ICL 的理解并评估最先进的 VLLM 的 ICL 能力，我们引入了一个新颖的基准套件 VL-ICL Bench （Figure **??** ），专为评估 VLLM 情境学习而量身定制。我们的基准测试套件包含文本输出和图像输出任务，旨在测试 VLLM 的各个方面，包括 细粒度感知、推理、规则归纳和上下文长度 . 我们对最先进的 VLLM 进行全面评估，这些 VLLM 能够处理交错的图像文本作为我们基准测试的输入。结果显示，尽管某些模型在特定任务上表现出

*Co-first authors

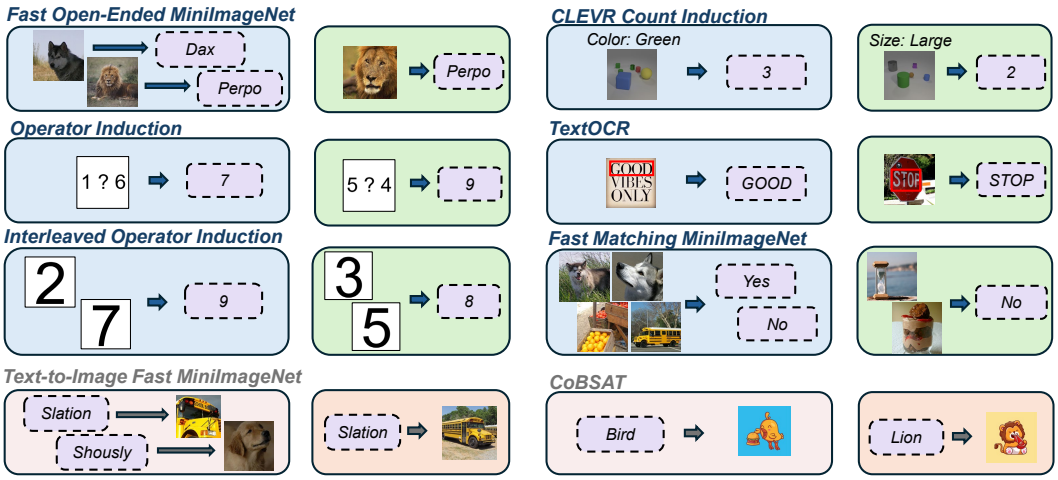


Figure 1: VL-ICL Bench 中不同任务的图示。图像转文本任务位于前三行，而文本转图像任务位于底行。第三行中的图像到文本任务对交错的图像文本输入进行推理。

合理的性能，但没有一个模型在整个任务范围内表现出一致的卓越性，并且一些模型在某些任务上表现得接近偶然水平。我们希望这种对多模态 ICL 不同机遇和挑战的系统研究，能够支持从业者了解目前在新的多模态任务的免培训学习方面什么是可能的，什么是不可能的，并激励 VLLM 模型开发人员研究如何尽可能多地将 LLM 的 ICL 能力暴露给多模态世界。

总结一下我们的贡献：（1）我们证明了通过 VQA 和字幕定量评估 VLLM ICL 的常见做法所固有的局限性。（2）我们推出了第一个全面和集成的 ICL 任务基准套件，涵盖各种挑战，包括 感知、推理、规则归纳、长上下文长度和文本到图像/图像到文本。（3）我们在基准套件中严格评估一系列最先进的 VLLM，并强调它们的不同优势和劣势，以及针对不同 ICL 挑战的解决方案的不同成熟度。

2 背景和动机

2.1 ICL 问题设置

给定一个预训练的 VLLM θ 、一个可选的文本指令 I 、一个查询示例 x 和标签 y 的上下文集 [2] $S = \{(x_i, y_i)\}$ ，以及一个测试示例 x^* ，ICL 模型估计

$$p_\theta(y^*|x^*, I, S) \tag{1}$$

带有前馈传递。对于 LLM，x 和 y 通常是文本。对于 VLLM，x 可以是文本和/或图像，y 可以是文本（图像到文本 ICL）或图像（文本到图像 ICL）。

此 ICL 设置与更简单的零样本场景形成鲜明对比，在该场景中，预训练模型仅根据 θ 中预先学习的知识来估计 $p_\theta(y^*|x^*, I)$ ，而 S 中没有提供额外的训练数据。零样本情景已经过各种基准测试的严格评估 (**????**) ，在下一节中，我们将讨论现有 ICL 评估的局限性，这些局限性激发了我们的基准。

2.2 ICL 评估的常见做法

在以前对多模态 ICL 进行定量评估的尝试中，最流行的基准是 VQA 和字幕。在本节中，我们将讨论的重点放在图像到文本模型上 (**?????**) ，因为上下文中的文本到图像模型 (**??**) 相对不太常见且不太成熟，因此目前还没有通用的评估实践。在标题的情况下，上下文集 S 包含图像 x 和标题 y 的示例; 而对于 VQA，上下文 S 包含图像问题对 x 和答案 y 。

图 **??** （a）在三个流行的基准上绘制了 六 最先进的 VLLM 的 ICL 性能 – MathVista VQA (**?**) 、VizWiz VQA (**?**) 和 COCO 字幕 (**?**) 不同数量的训练示例（镜头）。虽然不同型号的性能各不相同，但关键的观察结果是，显示的大多数线条只是微弱地增加。这意味着对于所有型号，与零镜头情况（镜头 $= 0$ ）相比，ICL（镜头 > 0 ）的改进有限。这是因为，虽然上

[2] 在本文中，我们可以互换使用上下文集和支持集。

Table 1: VL-ICL Bench 概述。它评估了 ICL 与 VLLM 的各种功能和挑战。同时，它结构紧凑，易于研究人员使用，没有过高的资源要求。

Dataset	Capabilities Tested	Train Set	Test Set	Size (GB)
Fast Open MiniImageNet	I2T, Fast Binding	5,000	200	0.18
CLEVR Count Induction	I2T, Fine Grained Perception, Induction	800	200	0.18
Operator Induction	I2T, Induction, Reasoning	80	60	0.01
Interleaved Operator Induction	I2T, Induction, Reasoning, Interleaving, Long-Context	80	60	0.01
TextOCR	I2T, Fine Grained Perception, Induction	800	200	0.98
Matching MiniImageNet	I2T, Induction, Interleaving, Long-Context	1,600	400	0.11
Text-to-image MiniImageNet	T2I, Fast Binding	5,000	200	0.18
CoBSAT	T2I, Induction	800	200	0.07
Total	T2I, I2T, Binding, Perception, Long-Context, Interleaving, Induction, Reasoning	14,160	1,520	1.72

下文集 S 说明了提问和回答问题或为图像添加字幕的概念，但基线 VLLM θ 在 VQA 和字幕方面已经相当出色。VLLM 字幕和 VQA 的限制因素是详细感知、常识知识等方面——所有这些都是对 VLLM 的根本挑战，而不是可以通过几个镜头支持集合理教授的方面。

鉴于上面的讨论，目前尚不清楚为什么性能应该通过投篮来提高？我们推测，这主要是由于 VLLM 学习了每个数据集的首选答案风格，而不是学习更好地解决多模态推理任务本身。例如，在字幕 零样本 VLLM 往往会产生比 COCO 中的基本事实更详细的标题，并且它们通过 ICL 学会了更简洁。同时，对于 VQA，有一种标准做法，即根据真值答案和模型提供的答案之间的字符串匹配进行评估。例如，VizWiz 有无法回答的问题，一些 VLLM 用 '' 我不知道" 回答这些问题，这些问题不会是与真实 '' 无法回答" 匹配的字符串。因此，一些模型从上下文集中学习答案格式（例如，首选术语; 避免使用任何可能不满足字符串匹配的前言或后置）。这确实是一种 ICL，但也许不是人们期望在 VQA 中学习的东西。为了验证这个猜想，我们重复了前面的评估，但使用软匹配来消除答案格式学习的影响。对于 VQA，我们使用预训练的 LLM 来确定预测在语义上是否等同于真实情况，而对于字幕，我们使用 LLM 以 1-10 分的等级对生成的标题的质量进行评分（附录中的详细信息）。图 **??** （b）显示曲线几乎完全变平，零样本性能有所改善。图 **??** （c）通过显示精确匹配和 LLM 匹配的平均改进率来量化这种差异。对 LLM 验证的更改几乎完全消除了 ICL 相对于零样本的任何好处。

与上述相反，语言领域中流行的 LLM ICL 基准通常表现出不平凡的 ICL 学习 (**??**) 。图 **??** 显示了三个最先进的 VLLM 及其相应的基础 LLM，在三个流行的 NLP 任务（AGNews (**?**) 、MIT Movies (**?**) 和 TREC (**?**) ）上进行了评估。我们可以看到，与 VQA/字幕基准相比，模型的零样本性能通常通过少镜头 ICL 得到显着改善。这一结果证实了 VLLM 中的 LLM 组件确实继承了其基础 LLM 的 ICL 能力。然而，它提出了一个问题，即我们如何在多模态环境中有意义地利用和测量 VLLM 的 ICL 能力。在下一节中，我们将介绍我们的基准测试 VL-ICL Bench ，它正是这样做的。

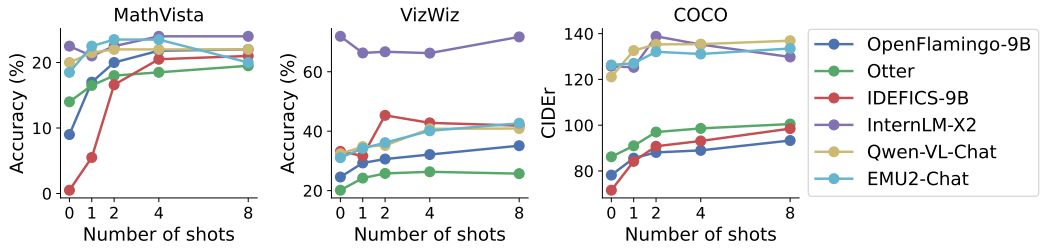
3 VL-ICL Bench

3.1 主要多式联运基准

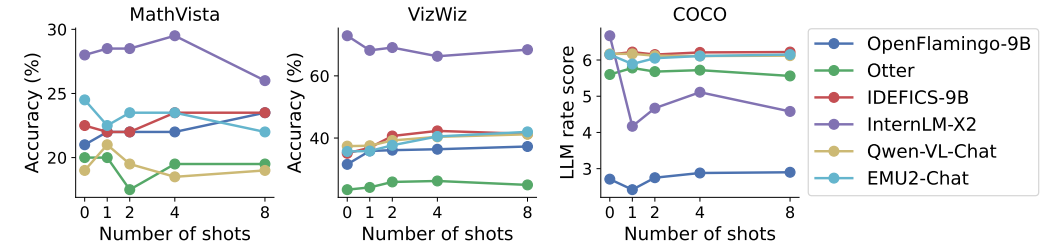
我们的 VL-ICL Bench 涵盖了许多任务，其中包括各种 ICL 功能，例如概念绑定、推理或细粒度感知。它涵盖了图像到文本和文本到图像的生成。我们的基准测试包括以下八个任务：快速打开 MiniImageNet、CLEVR 计数归纳、算子归纳、交错算子归纳、TextOCR、匹配 MiniImageNet、文本到图像 MiniImageNet 和 CoBSAT。我们在图 **??** 中提供了任务的图示，并在表 **??** 中总结了每个 VL-ICL Bench 任务测试的各种能力。此表还汇总了数据集统计信息，表明 VL-ICL Bench 紧凑且易于访问。我们遵循 ICL 社区 (**???**) [3] 的典型协议，并将每个数据集拆分为训练和测试拆分。然后，通过从训练拆分中抽取支持/上下文集以及从测试拆分中抽样测试/查询示例来执行/评估少量上下文学习。最终表现是许多此类 ICL 发作的平均值。

Fast Open MiniImageNet 我们使用 MiniImageNet 的变体，(**?**) (**?**) 中重新用于 ICL 的 ICL。在开放式分类中，VLLM 必须学会根据几个示例以开放式方式命名对象，而不是简单地将其分类为一组预定义的选项。因此，机会率实际上为零，而不是取决于类别的数量。快速绑定任务测试模型在不依赖先验知识的情况下将新名称或符号与概念相关联的能力。因此，(**?**) 为上下文/支持集中的对象类别（例如 dax 或 perpo）提供合成名称，并且模型必须学会将这些名称与图示的视觉概念相关联，以便正确命名测试图像。我们使用双向版本。

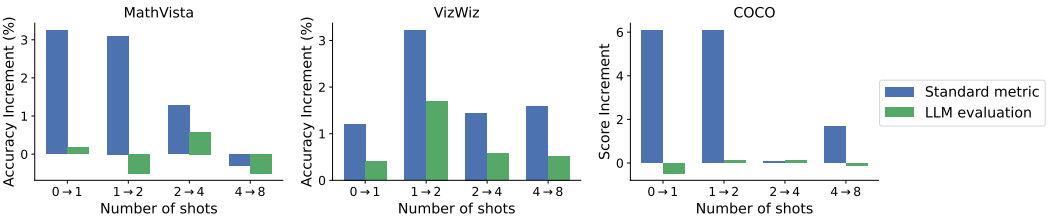
[3]这与 few-shot 元学习社区 (**??**) 不同，后者从同一池中采样支持/查询集。



(a) 用于评估的标准指标。



(b) 法学硕士作为评估法官。



(c) 添加更多支持示例时，准确性会增加。

Figure 2: VQA 和字幕是图像转文本 ICL 的较差基准。(a) 使用标准评估协议，在流行的图像转文本 ICL 基准测试（MathVista、VizWiz 和 COCO）的代表性示例上评估最先进的 VLLM。零击球表现高，ICL 表现仅微弱地依赖于击球，说明 ICL 学习不多。(b) 用基于 LLM 的评估对 VLLM 进行重新评估，进一步减少了对射击的依赖。(c) 当从传统的评估转向基于 LLM 的评估时，ICL 对绩效的影响从很小到可以忽略不计。这些基准的 ICL 主要学习答案风格/格式。

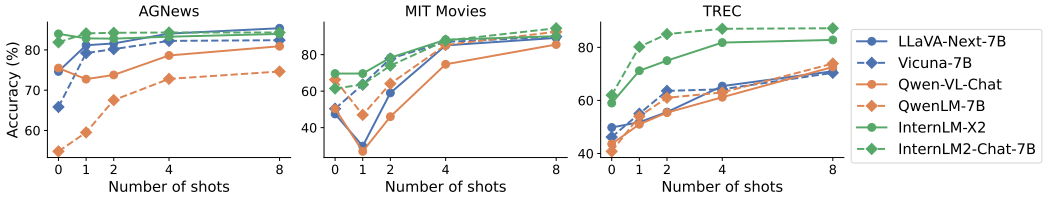


Figure 3: 在流行的文本转文本 ICL 基准测试中评估最先进的 VLLM/LLM 对。与图 **??** 中流行的图像转文本 VLLM 基准测试不同，小样本 ICL 通常可以显著提高零样本的性能，这表明有意义的上下文学习正在发生。

CLEVR Count Induction 在此数据集中，模型必须学会解决 “场景中有多少红色物体?” 类型的任务。但是，他们必须从示例中学习这一点，而不是被明确提示这样做。具体来说，我们从 CLEVR (**?**) 输入 x 场景图像，以及一个属性：值对，用于标识场景中特定类型的对象。有四种类型的属性：大小、形状、颜色、材料。所需的输出 y 是图像中满足提供的属性：值条件的对象计数。为了解决这项任务，该模型必须提供细粒度的感知，以支持计数和区

分对象类型，学习在图像中接地请求的属性，并且重要的是诱导所需的操作员将 x 映射到 y 正在 [4] 计算指定的对象类型。

Operator Induction 在此图像转文本任务中，模型必须求解给定训练示例（如 `1 ? 3 = 4`）的 `2 ? 7 = 9` 类型的任务。这意味着，除了解析图像 x 提取数字和算子外，模型还需要诱导未知算子的作用是加法，然后进行简单的算术推理，将诱导算子应用于解析的测试样本。可用的数学运算是加号、减号和时间，我们只考虑个位数。我们为此任务生成自己的图像。

Interleaved Operator Induction 此任务测试模型对 x 内的多个图像进行推理以生成单个答案 y 的能力。在这种运算符归纳的变体中，我们给模型两个查询图像作为包含表达式中每个数字的输入，而不是如上所述包含整个表达式的单个图像。否则，它与基本的操作员入职任务相同。从某种意义上说，分离图像使任务更容易，因为它大大简化了从单个图像解析表达式的感知任务。然而，当 VLLM 训练（例如字幕）通常训练模型以生成一次依赖于单个图像的输出时，它也更难，因为它需要 VLLM 在两个不同的图像之间执行归纳和推理。通过引入多个图像，它还引入了比基本运算符归纳更多的标记，从而强调了 VLLM 使用更大上下文长度的能力。

TextOCR 我们重新利用 TextOCR 数据集 (**?**) 创建一个任务，在该任务中，模型应学习输出红色矩形中显示的文本，灵感来自 (**?**) 。图像 x 包含突出显示文本窗口的图像，输出 y 是 OCR 文本。这项任务可以通过适当详细的零样本提示来实现，但与 (**?**) 不同的是，我们专注于评估该任务是否可以通过 ICL 进行示例诱导。因此，这项任务测试了细粒度的感知和感应能力。

Matching MiniImageNet 此任务是监督学习两个图像之间关系的最简单示例。对于关系学习，输入包含一个图像对 $x = \{x_1, x_2\}$ ，输出 y 指示它们之间是否 r 特定关系。VLLM 需要从它成立（$r(x_1, x_2) = true$ ）和不成立（$r(x_1, x_2) = false$ ）的示例中学习关系。我们重用 MiniImageNet (**??**) 数据集，学习的关系是两个图像是否来自同一个类 (**?**) ，在训练和测试之间考虑了不相交的类集。此任务测试归纳、处理多个交错图像的能力以及处理更大上下文长度的能力。

Text-to-Image MiniImageNet 这项新任务测试了文本到图像上下文中的快速概念绑定。我们引入了 MiniImageNet (**??**) 的进一步变体，它输入合成类别名称 x （用于快速绑定），并输出图像 y 。模型应从上下文集中学习，以将合成名称与图像分布相关联，从而学习在提示使用人工类别名称时生成相应类别的新图像（图 **??** ）。此任务测试映像生成和快速绑定。LLaVA-Next-7B 用于评估生成正确性。

CoBSAT 最后，我们还利用了最近的文本到图像 CoBSAT (**?**) 基准测试作为我们更大的 VL-ICL Bench suite 的一部分（图 **??** ）。这是一个文本到图像的任务，模型必须学习合成 y 指定文本概念 x 的图像（例如，对象类别），但此外，还有一个上下文集示例共有的潜在变量，必须在新的测试图像（例如，对象的共同颜色）中诱导并正确呈现。此任务测试图像生成和潜在变量归纳。

Capability Summary 上述 VL-ICL Bench 套件远远超出了任何单个现有的 ICL 基准测试，以测试多模态 ICL 的各种功能，包括（表 **??** ）：文本到图像和图像到文本的生成; 快速绑定 – 能够快速将新符号与视觉概念相结合，并在新数据的上下文中重用这些符号; 细粒度感知 - 根据计数或阅读文本的需要; 交错 – 在生成单个输出时对多个图像的内容进行推理的能力; 规则归纳法——从示例中引出数学运算符和潜在变量等非平凡概念; 简单的推理，如算术; 和长上下文 – VLLM 有效利用大量上下文令牌的能力。

3.2 文本变体

为了比较多模态的影响，我们还为我们的任务提供了文本版本的替代方案。对于像开放式 MiniImageNet 这样的数据集，我们提供图像标题而不是图像，并将其用于推理。例如，在 CLEVR 中，我们提供了场景中对象的枚举，包括它们的属性。请注意，文本版本并非适用于所有任务，特别是 TextOCR 很难翻译成合适的文本替代品。

[4] 此任务可以在适当详细的 VQA 提示下零点执行。但是，目标是测试模型是否可以从 ICL 的几个示例中学习任务。

4 结果

4.1 实验设置

Models 我们根据基准评估了具有各种大小（从 7B 到 80B 不等）和不同 LLM 主干的多样化最先进模型系列。具体来说，对于图像转文本 VLLM，我们选择 Open Flamingo （9B） (?)，IDEFICS （9B and 80B） (?)，Otter （9B） (?)，InterLM-XComposer2 （7B） (?)，LLaVA-Next （Vicuna-7B） (?)，Qwen-VL-Chat （9B） (?) 和 Emu2-Chat （34B） (?)。对于文本到图像的 VLLM，我们使用 GILL （7B） (?)，SEED-LLaMA （8B，14B） (?)，Emu1 （14B） (?)，Emu2-Gen （34B） (?)。我们还根据我们的基准测试评估 GPT4V (?)。我们使用官方发布的模型权重或 GPT4 API，并采用贪婪解码以提高可重复性。所有实验均使用三种不同的随机种子进行，我们报告了平均性能。A100-80GB GPU 用于实验。

Prompt 为了保持一致性，我们采用以下提示格式进行上下文学习。此外，我们还调查了各种提示格式的影响，并在补充材料中提供了详细的结果。

[Task Description]

Support Set : [Image] [Question] [Answer] (n-shot)

Query : [Image] [Question]

Prediction : [Answer]

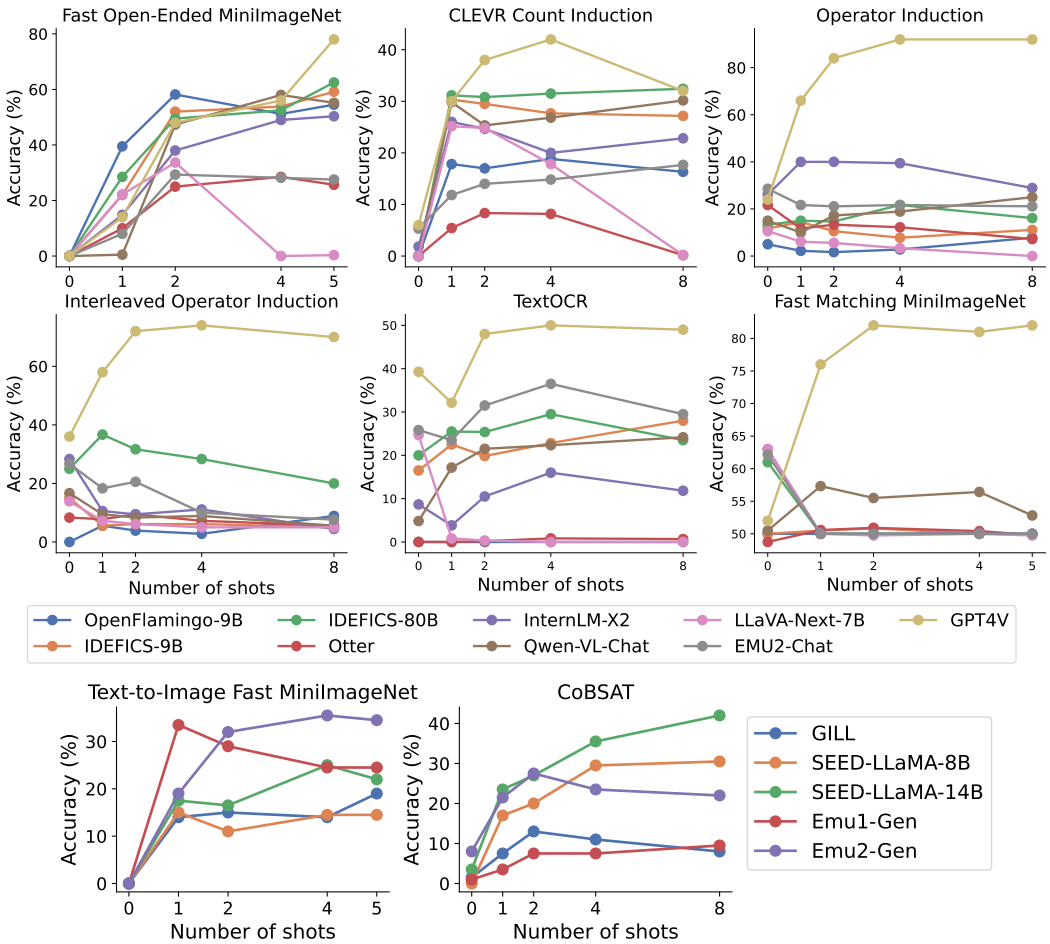


Figure 4: VL-ICL Bench 结果。前两行：Image-to-Text。底部：文本到图像任务。

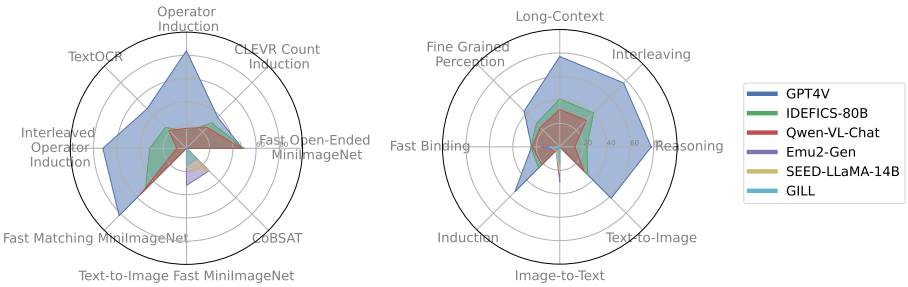


Figure 5: 说明最佳模型在我们的基准测试中的表现。左图：分别按数据集显示。右图：按能力评估，对数据集进行平均。

4.2 主要结果

VL-ICL Bench 的主要结果在图 **??** 中给出，包括对射击的细分，并在图 **??** 中总结为 2 发情况下的任务和能力的雷达图。我们得出以下观察结果：（1）VLLM 在 VL-ICL Bench 任务上展示了不平凡的上下文学习。与常见的 VQA 和字幕基准测试（图 **??** ）不同，我们的任务具有较低的零镜头性能，并且在每个任务中，至少有一个模型在镜头数量上显示出明显的性能改善。因此，ICL 能力现在确实正在得到展示和利用。（2）VLLM 通常难以使用更多的 ICL 示例。对于多个任务和模型，性能随着前几次拍摄而提高; 但这种增长并不是单调的。当我们转向更多的镜头时，性能通常会再次下降（例如，GPT4V CLEVR 计数感应;InternLM-XComposer2 操作员入职;IDEFICS-80B 交错式算子感应）。模型最终会被大量的图像和标记所迷惑，而不是利用它们来学习手头的任务。由于难以对上下文长度和输入图像数量进行外推，对于高镜头 ICL，这比用于 VLLM 训练的上下文长度和图像数量大 . 这显示了 ICL 当前最新技术的一个重要局限性：未来的模型必须支持更长的上下文和更多的图像，才能从更大的支持集中受益。（3）GPT4V 是最好的整体图像转文本模型。在所有型号中，GPT4V 是最强的（但被一些型号超越，例如 OpenFlamingo 在低镜头 MiniImageNet 中）。(4) 零样本性能并不能强烈表明 ICL 能力。LLaVA-Next-7B (**?**) 可能是 VL-ICL Bench 上最差的，这令人惊讶，因为它是主流零样本基准测试中最先进的开源模型。这是由于第（2）点：它的训练协议一次使用一个图像，并且每个图像使用大量的标记 - 因此 ICL 要求它在输入图像编号和令牌编号中进行大量推断，但它没有做到这一点。(5) 在文本到图像模型中没有明显的赢家。但是，文本到图像模型比图像到文本模型具有更一致的镜头缩放。这是由于使用更多样化的交错数据集进行训练，这些数据集为每个实例提供多个输入图像，并且每个图像使用更少的令牌以实现更好的缩放。

4.3 其他分析

接下来，我们使用 VL-ICL Bench 来分析影响 ICL 绩效的几个挑战和因素的作用。

Fast Concept Binding 在我们的开放 miniImageNet 任务中，我们遵循 (**?**) 要求快速绑定合成概念名称，以便纯粹测试模型的 ICL 能力，而不会混淆 VLLM 将视觉概念与名称关联的零样本能力。图 **??** 比较了快速和真实世界的 miniImageNet 识别，我们发现快速绑定的情况更具挑战性。

Direct Comparison of Multimodal and Text ICL 对于某些图像到文本的 VL-ICL Bench 任务，我们可以解开文本与图像输入的作用，在这些任务中，我们可以轻松地提供描述图像的语义等效文本输入，而不是图像标记。图 **??** 显示了计数归纳、算子归纳和交错算子归纳任务的图像输入与文本输入之间的比较。通过文本输入，性能会随着拍摄次数的增加而更加急剧和一致。这归因于（i）感知难度的降低，以及（ii）与图像输入相比，令牌总数的减少。

Scaling with Number of Shots 如 Sec. **??** 所述，不同的模型在拍摄次数方面表现出不同的缩放能力。我们通过对任务进行聚合并报告每个镜头增量的平均精度增量来总结它们的缩放能力 Fig. **??** 。显然，VLLM 在每次射击的精度增量以及它们从越来越多的射击中提取知识的能力方面各不相同。

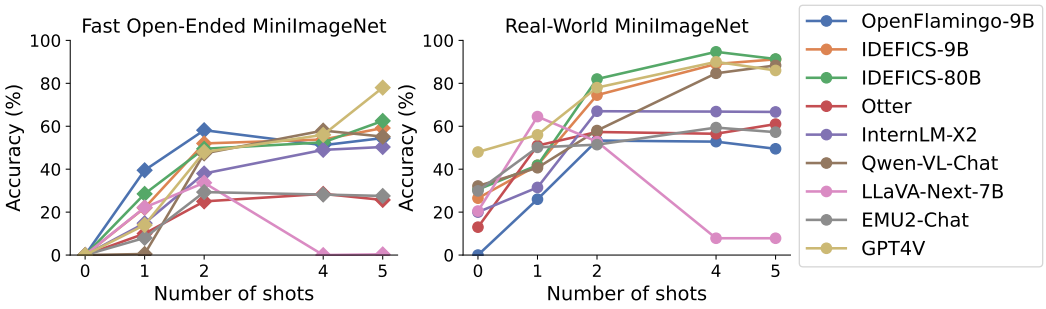


Figure 6: MiniImageNet 的快速绑定与实名版本的比较。与实名版本不同，快速绑定的零样本推理精度为零。因此，它的成功完全依赖于 ICL。

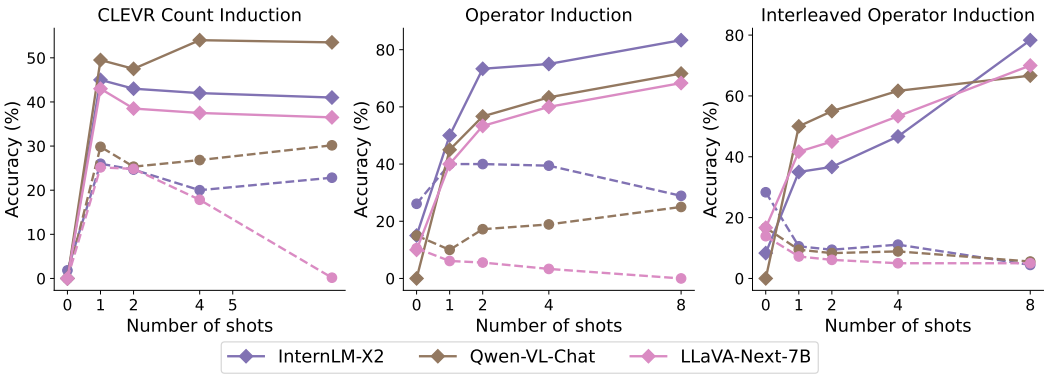


Figure 7: 多模态（虚线）和文本（实线）ICL 的比较。使用文本输入时，性能会更加显著且一致地提高。

4.4 定性分析

我们还包括定性分析，其中我们分析了使用更多支持示例对输出质量的影响。我们使用 Emu2-Gen 模型分析了 Fig. **??** 中的文本到图像任务。对于文本到图像的 MiniImageNet，模型应该从支持示例中学习，即人工名称分别对应于狮子和校车。Emu2-Gen 能够在一定程度上做到这一点，但可能会被其他支持示例所迷惑，因为更多的支持示例不一定有帮助。在 CoBSAT 中，支撑集诱导动物应该具有冰川和沙漠背景。在没有支持示例的情况下，模型仅显示动物，但通过更多支持示例，它了解到它应该在第一个示例中使用冰川背景。在第二个示例中，模型能够捕捉到它应该使用沙漠背景，但在显示所需的动物 - 斑马方面不太成功。生成的图像质量不一定随着支持示例的增加而更好。

对于图像到文本任务的定性分析，我们讨论了模型对每个任务的一些常见错误。

Open-Ended MiniImageNet 模型预测真实世界的类是相对常见的，即使它被要求使用支持集中的人工名称。有了更多的支持示例，随着模型学习使用人工名称，此类错误发生的可能性较小。

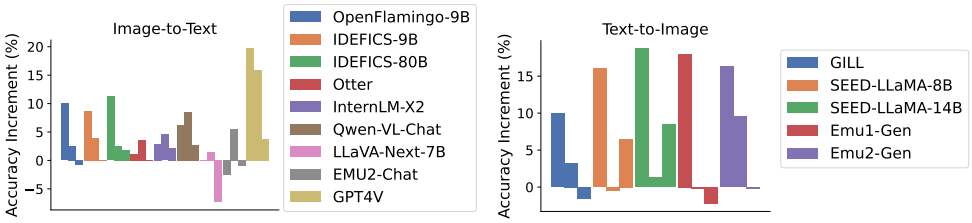


Figure 8: 跨数据集聚合分析，以查找添加更多镜头时的平均性能增量。一种颜色的单个条对应于 $0 \to 1$、$1 \to 2$ 和 $2 \to 4$ 的改进。

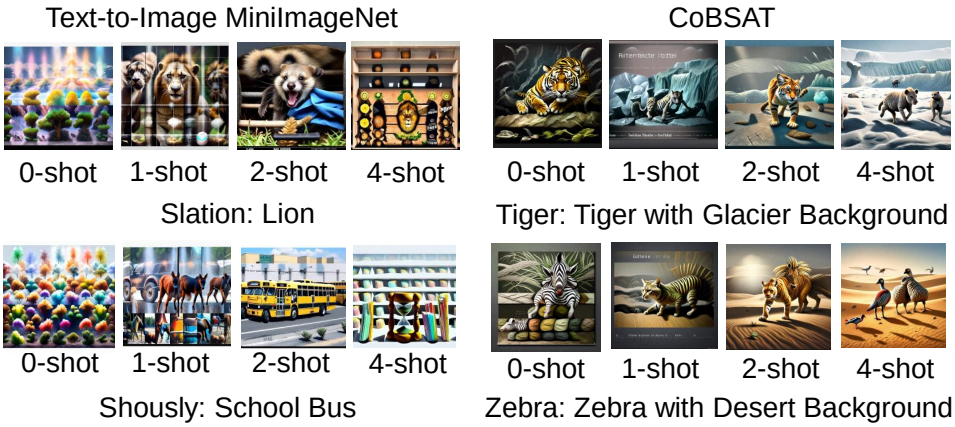


Figure 9: 定性分析：Emu2-Gen 生成的图像显示了学习支持示例诱导的概念的能力。

CLEVR Count Induction 在许多情况下，模型会重新表述问题，而在其他情况下，它会说例如，所描述的对象存在。这种行为在支持示例较少或没有的情况下更为常见。通过更多的支持示例，模型学习预测计数，但得到不正确的答案。这可能是因为某些物体更难识别，例如，如果一个物体部分覆盖了另一个物体。

Operator Induction 一个非常常见的错误是使用与支持示例中引入的运算符不同的运算符。例如，模型可能会猜测它应该将两个数字相加，而不是将它们相乘，反之亦然。

Interleaved Operator Induction 模型有时会预测第一个显示的数字或它们的直接组合，例如，如果两个数字是 1 和 2，则返回 12。在数字之间使用不正确的运算符也相对常见。

TextOCR 在许多情况下，模型返回的单词多于红色框中突出显示的单词，但将突出显示的单词作为其中之一。模型在文本中遗漏一个字母或返回一个与正确答案相似但不同的单词也很常见。但在某些情况下，答案可能与突出显示的内容大不相同，可能会在图像中返回不同的单词。

Matching MiniImageNet 模型通常描述其中一张图像上显示的内容。然而，在许多情况下，他们只是返回错误的答案，在应该是肯定的时候说不。

5 相关工作

VLLM Evaluation 随着 VLLM 的快速发展，研究人员正在创建评估基准，以从不同角度全面评估 VLLM 的能力。这些评估范围从零样本聚合基准，如 MME (**?**) 、MMbench (**?**) 和 MM-VET (**?**) 到专为在特定方面进行微调而设计的数据集，例如视觉推理 (**?**) 和基于知识的 QA (**?**) 。它们主要关注单图像场景，而对情境学习评估的探索不足。

In-Context Learning Evaluation 术语 “in-context” 以多种方式使用，包括描述具有交错输入的场景，例如多个视频帧或多回合对话 (**????**) 。尽管交错输入的研究提出了一个有趣的主题，但它与我们考虑遵循 (**???**) 的上下文学习的核心定义不一致，这涉及从 $x \to y$ 中学习函数的新兴能力，从少量支持输入输出对中学习函数。从这个意义上说，对 ICL (**????**) 的先前评估是有限的，并且存在严重的缺点，如第 Sec. **??** 节所述。在我们的工作同时，CobSAT (**?**) 引入了一个基准测试，旨在评估文本到图像模型中的上下文学习，特别关注潜在的变量归纳能力。我们的工作在此基础上进行了扩展，包括图像到文本和文本到图像生成的任务，评估了更广泛的能力（表 **??** ）。此外，我们还将 CobSAT 作为基准测试的一个子集。

Visual In-Context Learning 术语 “in-context” 也被用于纯视觉模型中，其目的是在没有特定任务预测头 (**???**) 的情况下执行各种图像到图像任务，例如语义分割、深度估计、目标检测等。但是，这些模型在配对的上下文输入输出数据上进行了显式训练，以便能够在推理期间执行可视化 ICL。在本文中，我们重点关注基于 LLMs 涌现能力的多模态视觉语言 ICL。

6 结论

我们推出了第一个全面的基准测试套件 VL-ICL Bench，用于使用 VLLM 进行多模态视觉和语言情境学习。该基准测试套件避免了现有主流但有限的图像转文本 ICL 评估方法的问题——ICL 与零样本推理相比提供有限的可证明优势，并且 VLLM 充其量只能学习答案格式，而不是任何真正的多模态功能。相比之下，VL-ICL Bench 测试了各种多模态功能，包括文本到图像和图像到文本的生成、细粒度感知、规则归纳、推理、图像交织、快速概念绑定、长上下文和镜头缩放。我们希望这个基准测试能够激励模型开发人员在 VLLM 开发中考虑所有这些功能，并告知从业者 VLLM ICL 随着该领域的发展可以做什么和不能做什么的演变。我们还计划在未来扩展我们的基准，以纳入更多的任务和模型。

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. NeurIPS , 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 , 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 , 2023.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. NeurIPS , 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. NeurIPS , 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 , 2021.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. arXiv preprint arXiv:2301.00234 , 2022.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394 , 2023.

Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. ICLR , 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In CVPR , pages 3608–3617, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. ICLR , 2021.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey. Meta-Learning in Neural Networks: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence , 2021. doi: 10.1109/TPAMI.2021.3079209.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR , 2019.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR , 2017.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. Generating images with multimodal language models. NeurIPS , 2023.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. NeurIPS , 36, 2023.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425 , 2023a.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 , 2023b.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. arXiv preprint arXiv:2307.16125 , 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV , 2014.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 , 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. NeurIPS , 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. TACL , 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? arXiv preprint arXiv:2307.06281 , 2023c.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In NeurIPS , 2022.

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 International Conference on Learning Representations (ICLR) , 2024.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. In NACL , 2022.

R OpenAI. Gpt-4 technical report. arXiv , pages 2303–08774, 2023.

Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In ICLR , 2022.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. TextOCR: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In CVPR , 2021.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. TMLR , 2023.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 , 2023a.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222 , 2023b.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In CVPR , 2018.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In NeurIPS , 2021.

Asahi Ushio and Jose Camacho-Collados. T-NER: An all-round python library for transformer-based named entity recognition. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations , 2021.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In NeurIPS , 2016.

Ellen M Voorhees and Dawn M Tice. Building a question answering test collection. In SIGIR , 2000.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In CVPR , 2023a.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In ICCV , 2023b.

Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. Generalizing from a few examples: A survey on few-shot learning. ACM Comput. Surv. , 53(3), June 2020. doi: 10.1145/3386252.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. NeurIPS , 2022.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490 , 2023.

Yuchen Zeng, Wonjun Kang, Yicong Chen, Hyung Il Koo, and Kangwook Lee. Can mllms perform text-to-image in-context learning? arXiv preprint arXiv:2402.01293 , 2024.

Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112 , 2023.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. NeurIPS , 28, 2015.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. arXiv preprint arXiv:2309.07915 , 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. arXiv preprint arXiv:2304.06364 , 2023.

1

附录

2

目录

3 实施和评估细节

VL-ICL Bench Evaluation Metrics 我们使用准确性作为基准测试中所有子集的指标。对于文本到图像的生成任务，我们使用最先进的 VLLM LLaVA-Next (**?**) 作为判断模型来判断生成的图像是否包含所需的对象或属性。

Models Configurations 我们还在 Table **??** 中总结了我们论文中基准测试的模型的配置，特别关注每张图像的标记数量和上下文长度。此信息有助于阐明为什么某些模型在镜头增加时表现出较差的可伸缩性，因为总长度超过了最大上下文窗口。

Table 2: 我们基准测试中使用的模型的详细配置。

Model	Connection Module	Image Tokens	Context Length (Train)	Context Length (Test)
OpenFlamingo-9B	Perceiver	64	2048	2048
IDEFICS-9B	Perceiver	64	2048	2048
Otter	Perceiver	64	2048	2048
InternLM-XComposer2	Perceiver	64	2048	4096
Qwen-VL-Chat	Cross-Attention	256	2048	8192
LLaVA-Next	MLP	576	2048	4096
Emu1	C-Former	512	2048	2048
Emu2	Linear layers	64	2048	2048
GILL	Linear layers	4	2048	2048
SEED-LLaMA	Q-Former	32	2048	4096

Prompts 我们在下面列出了用于特定实验的特定提示。

Prompt to judge image generation for Fast MiniImageNet and CobSAT dataset

User: Decide whether the image contains the following concept: { GT } . Answer with 'yes' or 'no'.

Prompt to judge the answer for Vizwiz VQA.

User: Based on the image and question, decide whether the predicted answer has the same meaning as the ground truth. Answer with 'yes' or 'no'. Question: { Question } Predicted answer: { Prediction } Ground Truth: { GT }

Prompt to rate the quality of COCO captioning (Main text, section 2.2)

User: Given the following image, you are to evaluate the provided generated caption based on its relevance, accuracy, completeness, and creativity in describing the image. Rate the caption on a scale from 1 to 10, where 10 represents an exceptional description that accurately and completely reflects the image's content, and 1 represents a poor description that does not accurately describe the image.

Generated Caption: { Prediction }

Ground Truth Caption: { GT }

Consider the following criteria for your rating:

1 (Very Poor): The caption does not correspond to the image's content, providing incorrect information or irrelevant descriptions. It misses essential elements and may introduce non-existent aspects.

3 (Poor): The caption only slightly relates to the image, missing significant details or containing inaccuracies. It acknowledges some elements of the image but overlooks key aspects.

5 (Fair): The caption provides a basic description of the image but lacks depth and detail. It captures main elements but misses subtleties and may lack creativity or precision.

7 (Good): The caption accurately describes the main elements of the image, with some attention to detail and creativity. Minor inaccuracies or omissions may be present, but the overall description is sound.

8 (Very Good): The caption provides a detailed and accurate description of the image, with good creativity and insight. It captures both essential and minor elements, offering a well-rounded depiction.

9 (Excellent): The caption delivers an accurate, detailed, and insightful description, demonstrating high creativity and a deep understanding of the image. It covers all relevant details, enhancing the viewer's perception.

10 (Exceptional): The caption offers a flawless description, providing comprehensive, accurate, and highly creative insights. It perfectly aligns with the image's content, capturing nuances and offering an enhanced perspective.

Please provide your rating. You should ONLY output the score number.

4 进一步分析

4.1 缩放到更多镜头

为了检查模型可以处理的最大镜头数以及模型是否仍然可以从更多镜头中受益，我们进一步将支撑集大小增加到 16、32 和 64 镜头。我们选择了三个模型进行实验：OpenFlamingo 9B (**?**) 、IDEFICS-9B-Instruct (**?**) 和 InternLM-XComposer2 (**?**) 。之所以选择这些模型，是因为它们处理的每个图像都转化为更少的标记（Table **??** ），确保它们在使用 64 次镜头进行评估时不会超过最大上下文长度。IDEFICS-9B-Instruct 在大多数数据集中与其他模型相比，表现出更好的缩放能力。此外，虽然 InternLM-XComposer2 在低射条件下具有很强的性能，但随着多次射程，性能会迅速下降。这可能是由于训练（4096）和测试（2048）上下文长度（Table **??** ）之间的不匹配，其中上下文长度的外推一直是众所周知的挑战性任务 (**??**) 。

4.2 思维链提示

为了研究是否有任何可以增强情境学习的策略，一种简单的方法是思路链（CoT）提示 (**?**) 。CoT 提示模型阐明其关于支持集中潜在变量的推理过程，从而可能提高其学习和推理能力。我们尝试了 Qwen-VL-Chat (**?**) 和 InternLM-XComposer2 (**?**) ，它们具有最先进的 LLM，具有很强的推理能力。以下是我们使用的具体提示。

[CoT Prompt]: Let's first think step by step and analyze the relationship between the given few-shot question-answer pairs. Give reasoning rationales.

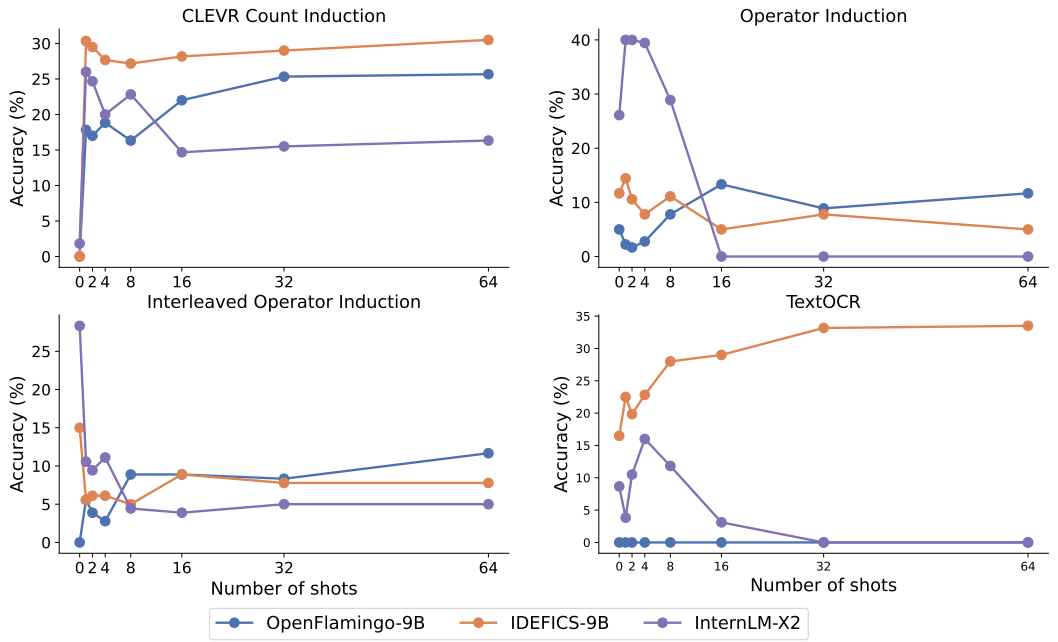


Figure 10: 缩放到多个镜头的结果（最多 64 张）。与其他模型相比，IDEFICS-9B-Instruct 在大多数数据集上表现出强大的缩放能力。此外，虽然 InternLM-XComposer2 在低镜头场景中表现出强大的性能，但随着镜头数量的增加，其性能会迅速下降。

User : [Task Description][Support Set][Query][CoT Prompt]
VLLMs : [Generated rationals]
User : [Task Description][Support Set][Query][Generated rationals]
VLLMs : Prediction

我们没有观察到思维链提示的持续改进：它有利于某些数据集的性能，而在其他数据集上则有损于性能。这些发现强调了情境学习任务的复杂性，表明模型开发的基本进步是必要的。这些任务不能用简单的提示技术（如 CoT）轻松解决。

4.3 重复支撑集

在本小节中，我们尝试了一个有趣的设置：我们多次复制相同的支持示例，以评估重复是否能提高性能。我们采用 Qwen-VL-Chat 模型进行这些实验，结果如图 **??** 所示。我们发现，在快速开放式 MiniImageNet 的 1 次拍摄场景中，复制镜头特别有益，尽管在其他数据集中并不一致地观察到这一点。可能的原因是，快速开放式 MiniImageNet 通过重复的例子加强了对概念的约束，而对于像运算符归纳这样的任务，需要不同的例子来促进学习过程。

4.4 指令微调的影响

我们研究了指令遵循微调如何影响上下文学习能力。我们比较了两个模型系列，每个模型系列都有一个预训练版本和一个遵循指令的微调版本：Qwen-VL 与 Qwen-VL-Chat（**?**）和 IDEFICS-9B 与 IDEFICS-9B-Instruct（**?**）。它们的性能差异如图 **??** 所示。尽管结果各不相同，但未使用指令进行微调的模型在拍摄次数方面表现出略高的可扩展性，如 TextOCR 数据集所示。需要进一步的研究来了解指令遵循微调是否会损害上下文能力。

4.5 不同级别的任务描述详细信息

我们在图 **??** 的提示描述中展示了不同细节层次的影响。结果表明，通常使用最详细的描述可以获得最佳结果，但并非所有设置都如此，在某些情况下，即使没有描述也可以更好。在不同细节级别上，性能通常相似，但在某些情况下，性能可能会明显更差，例如对于

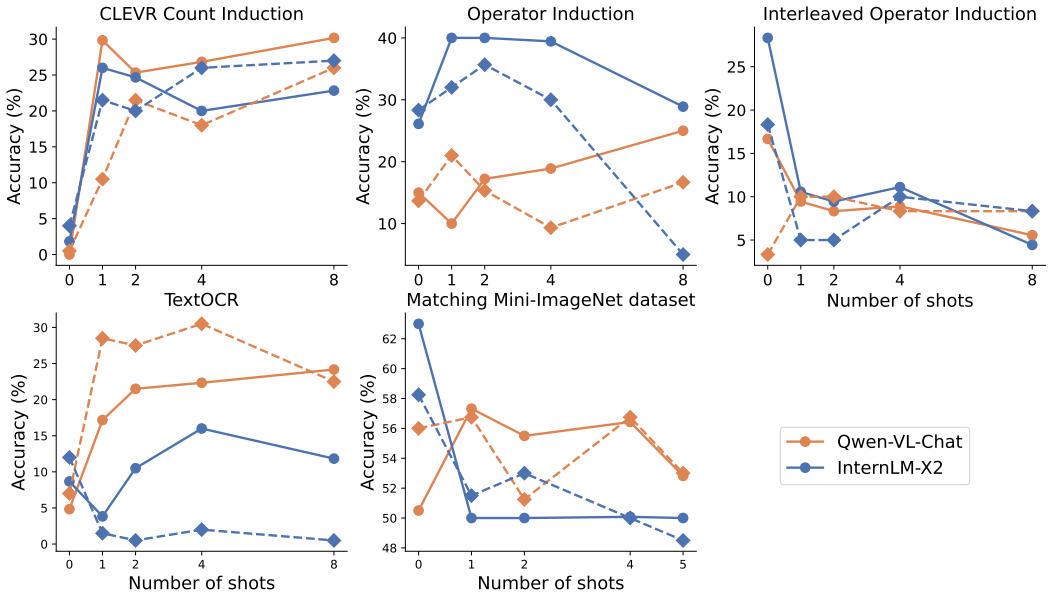


Figure 11: 在一系列数据集和模型中比较思路提示（虚线、菱形标记）与基线结果（实线、圆形标记）。思维链提示并不能始终如一地提高数据集的性能，这凸显了上下文学习任务的复杂性以及除了简单的提示技术之外对基本模型开发的需求。

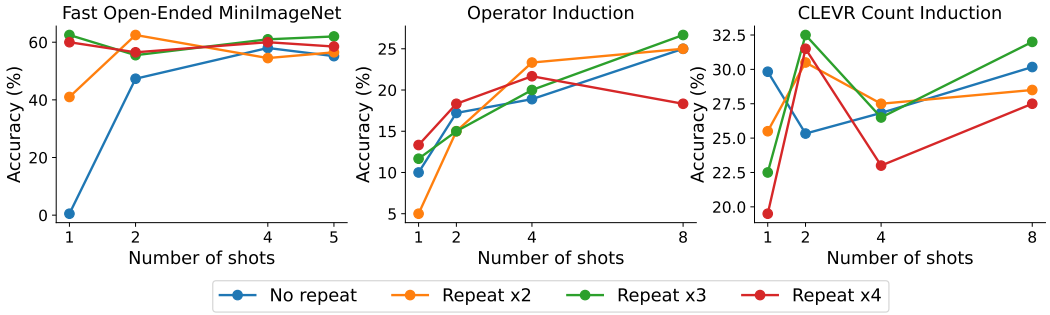


Figure 12: 使用 Qwen-VL-Chat 模型调查在选定的数据集中重复上下文示例的影响。X 轴表示唯一拍摄次数，而不是总拍摄次数。例如，1 次重复 x2 表示有一个唯一的镜头，并且重复两次。

TextOCR。我们还提供包含完整结果的表格。在我们的主要实验中，我们对所有数据集都采用了详细的任务描述。

我们用于不同数据集的任务描述如下：

Fast Open-Ended MiniImageNet

Detailed : Induce the concept from the in-context examples. Answer the question with a single word or phase.

Concise : Answer the question with a single word or phase.

CLEVR Count Induction

Detailed : The image contains objects of different shapes, colors, sizes and materials. The question describes the attribute and its value. You need to find all objects within the image that satisfy the condition. You should induce what operation to use according to the results of the in-context examples and then calculate the result.

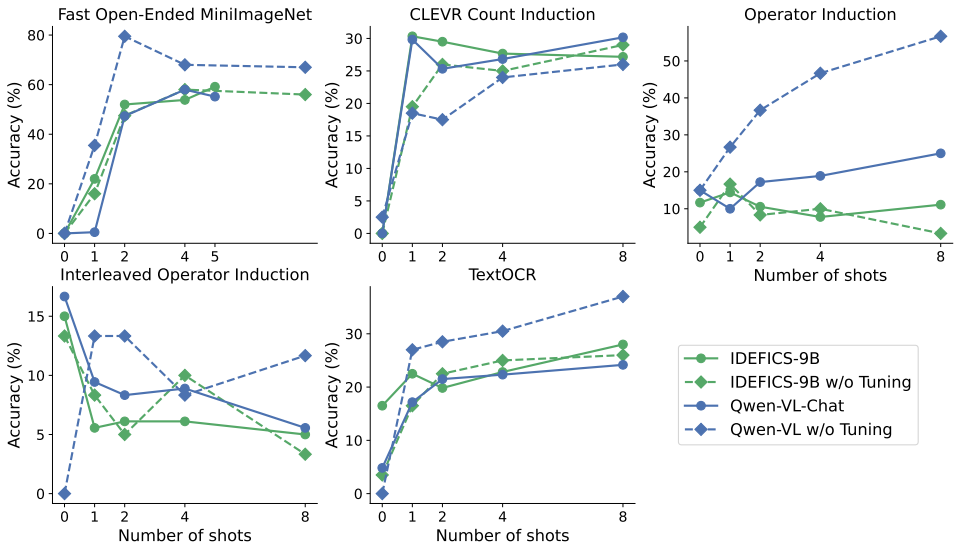


Figure 13: 使用（实线）和不使用指令调优（虚线）的比较。尽管结果各不相同，但未使用指令进行微调的模型在镜头数量方面表现出略高的可扩展性，正如 TextOCR 等数据集所证明的那样。

Concise : Find objects of the given type, induce what operation to use and calculate the result.

Operator Induction

Detailed : The image contains two digit numbers and a ? representing the mathematical operator. Induce the mathematical operator (addition, multiplication, minus) according to the results of the in-context examples and calculate the result.

Concise : Induce the mathematical operator and calculate the result.

TextOCR

Detailed : An image will be provided where a red box is drawn around the text of interest. Answer with the text inside the red box. Ensure that the transcription is precise, reflecting the exact characters, including letters, numbers, symbols.

Concise : Answer with the text inside the red box.

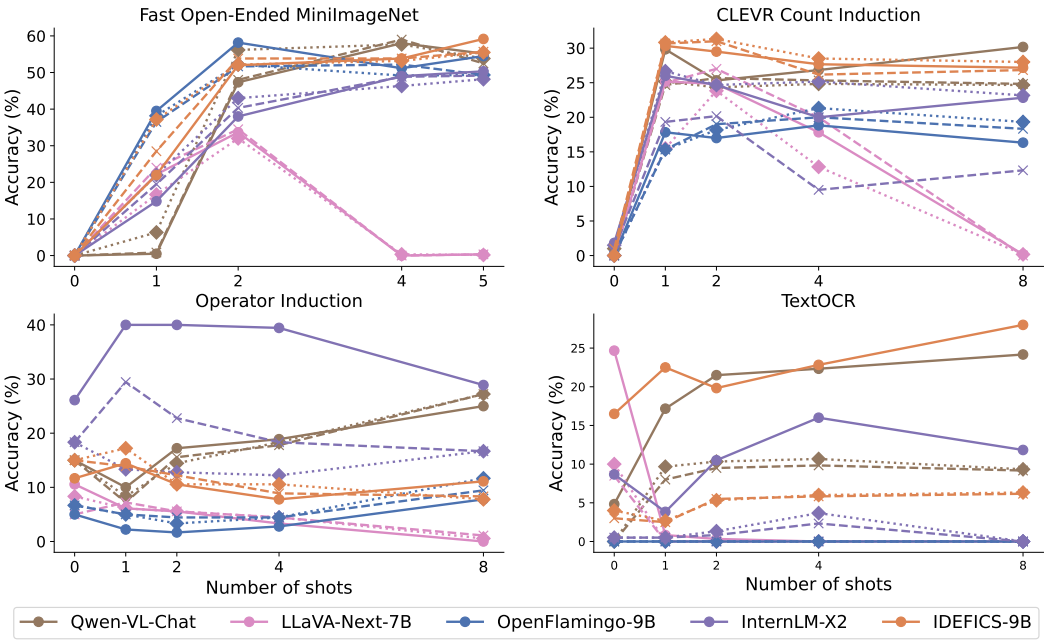


Figure 14: 在所选数据集和模型中比较详细的任务描述（实线、圆圈标记）、简洁的任务描述（虚线、x 标记）和无任务描述（虚线、菱形标记）。

5 完整结果

在本节中，我们在 Section **??** to **??** 中展示了正文中数字的原始结果，并在 Section **??** 中展示了补充材料的结果。

5.1 VQA 和图像标题的初步分析

Table 3: 使用字符串匹配的 MathVista 结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	9.00	17.00	20.00	21.80	22.00
Otter	14.00	16.50	18.00	18.50	19.50
IDEFICS-9B	0.50	5.50	16.60	20.50	21.00
InternLM-XComposer2	22.50	21.00	22.50	24.00	24.00
Qwen-VL-Chat	20.00	21.50	22.00	22.00	22.00
LLaVA-Next-Vicuna-7B	23.00	22.00	15.00	10.00	9.50
Emu2-Chat	18.50	22.50	23.50	23.50	20.00

Table 4: MathVista 使用 LLM 进行答案提取的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	21.00	22.00	22.00	22.00	23.50
Otter	20.00	20.00	17.50	19.50	19.50
IDEFICS-9B	22.50	22.00	22.00	23.50	23.50
InternLM-XComposer2	28.00	28.50	28.50	29.50	26.00
Qwen-VL-Chat	19.00	21.00	19.50	18.50	19.00
LLaVA-Next-Vicuna-7B	23.00	25.00	18.00	15.00	10.50
Emu2-Chat	24.50	22.50	23.50	23.50	22.00

Table 5: 使用精确匹配的 VizWiz 结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	24.57	29.31	30.62	32.14	35.11
Otter	20.13	24.21	25.78	26.33	25.71
IDEFICS-9B	33.20	31.67	45.33	42.80	41.87
InternLM-X2	71.93	66.33	66.73	66.27	71.73
Qwen-VL-Chat	32.40	34.80	35.20	40.80	40.90
LLaVA-Next-Vicuna-7B	54.12	28.13	10.20	6.60	0.40
Emu2-Chat	31.06	34.20	36.13	40.12	42.66

5.2 主要结果

我们呈现了 Table **??** to **??** 的主要结果。

5.3 其他分析

我们在表 **??** 到 **??** 中给出了附加分析的结果。

Table 6: VizWiz 使用 LLM 作为评判的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	31.54	35.85	36.10	36.4	37.28
Otter	23.40	24.12	25.88	26.19	24.93
IDEFICS-9B	35.10	36.90	40.66	42.30	41.35
InternLM-XComposer2	72.90	68.20	69.10	66.30	68.42
Qwen-VL-Chat	37.40	37.53	39.22	40.38	41.20
LLaVA-Next-Vicuna-7B	55.26	26.08	11.38	8.72	2.50
Emu2-Chat	35.68	35.83	37.67	40.51	41.99

Table 7: COCO 字幕（CIDEr）的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	78.2	85.6	88.1	89.0	93.3
Otter	86.2	91.0	97.0	98.6	100.5
IDEFICS-9B	71.6	84.20	90.8	93.1	98.5
InternLM-XComposer2	125.74	125.26	138.82	135.22	129.8
Qwen-VL-Chat	121.10	132.6	135.3	135.4	136.9
LLaVA-Next-Vicuna-7B	131.24	81.75	40.49	34.47	26.26
Emu2-Chat	126.3	127.0	132.06	131.10	133.5

Table 8: COCO 字幕的结果。分数由 LLaVA-Next 评分，从 1-10（越高越好）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	2.71	2.42	2.75	2.88	2.90
Otter	5.60	5.78	5.68	5.72	5.56
IDEFICS-9B	6.15	6.22	6.15	6.21	6.22
InternLM-XComposer2	6.67	4.17	4.67	5.11	4.58
Qwen-VL-Chat	6.17	6.17	6.12	6.11	6.12
LLaVA-Next-Vicuna-7B	6.34	3.54	3.75	3.43	3.57
Emu2-Chat	6.15	5.89	6.05	6.11	6.15

Table 9: AGNews 数据集上文本 ICL 的 VLLM 和 LLM 比较（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
LLaVA-Next-Vicuna-7B	74.66	81.16	81.61	84.05	85.38
Vicuna-7B	65.83	79.22	80.20	82.24	82.41
Qwen-VL-Chat	75.49	72.74	73.78	78.62	80.91
QwenLM-7B	54.80	59.51	67.53	72.80	74.64
InternLM-XComposer2	83.99	82.87	82.80	83.28	83.97
InternLM2-Chat-7B	81.89	84.11	84.25	84.32	84.33

Table 10: 麻省理工学院电影数据集上文本 ICL 的 VLLM 和 LLM 比较（Accuracy %）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
LLaVA-Next-Vicuna-7B	47.47	29.88	59.04	85.06	89.16
Vicuna-7B	50.36	63.61	77.83	86.99	89.88
Qwen-VL-Chat	50.36	26.99	46.02	74.70	85.54
QwenLM-7B	66.27	46.99	64.10	85.30	92.53
InternLM-XComposer2	69.64	69.64	78.31	88.19	90.12
InternLM2-Chat-7B	61.45	63.61	73.98	87.71	94.46

Table 11: TREC 数据集上文本 ICL 的 VLLM 和 LLMs 比较（精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
LLaVA-Next-Vicuna-7B	49.80	51.80	55.60	65.40	71.00
Vicuna-7B	46.20	55.00	63.60	64.20	70.40
Qwen-VL-Chat	43.60	51.00	55.40	61.20	72.60
QwenLM-7B	40.80	54.00	61.00	63.00	73.90
InternLM-XComposer2	59.00	71.20	75.00	81.80	82.80
InternLM2-Chat-7B	62.00	80.20	85.00	87.00	87.20

Table 12: 不同模型在 Fast Open-Ended Mini-ImageNet（精度%）上的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
OpenFlamingo-9B	0.00 ± 0.00	39.50 ± 1.22	58.17 ± 3.57	51.17 ± 0.85	54.50 ± 5.66
IDEFICS-9B	0.00 ± 0.00	22.00 ± 0.41	52.00 ± 2.94	53.83 ± 0.94	59.17 ± 6.20
IDEFICS-80B	0.00 ± 0.00	28.50 ± 0.27	49.50 ± 1.28	52.47 ± 3.25	62.50 ± 2.00
Otter	0.00 ± 0.00	10.00 ± 0.71	25.00 ± 1.22	28.50 ± 2.86	25.67 ± 2.25
InternLM-X2	0.00 ± 0.00	14.83 ± 1.03	38.00 ± 1.78	49.00 ± 1.78	50.33 ± 3.86
Qwen-VL-Chat	0.00 ± 0.00	0.50 ± 0.41	47.33 ± 2.49	58.00 ± 2.83	55.17 ± 2.25
LLaVA-Next-7B	0.00 ± 0.00	22.17 ± 4.03	33.67 ± 2.25	0.00 ± 0.00	0.33 ± 0.24
Emu2-Chat	0.00 ± 0.00	8.00 ± 1.87	29.33 ± 1.84	28.18 ± 4.26	27.54 ± 5.12
GPT4V	0.00	14.00	48.00	56.00	78.00

Table 13: 不同模型在实名 Mini-ImageNet（精度%）上的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
OpenFlamingo-9B	0.00 ± 0.00	26.00 ± 2.86	53.33 ± 3.27	52.83 ± 0.94	49.50 ± 1.22
IDEFICS-9B	26.50 ± 0.00	41.83 ± 2.25	74.50 ± 2.27	89.00 ± 0.41	91.17 ± 1.89
IDEFICS-80B	30.50 ± 0.00	41.83 ± 1.18	82.00 ± 2.68	94.67 ± 0.62	91.33 ± 1.43
Otter	13.00 ± 0.00	51.00 ± 2.16	57.33 ± 3.09	56.50 ± 1.08	61.00 ± 1.87
InternLM-X2	20.00 ± 0.00	31.50 ± 1.63	67.00 ± 1.47	66.83 ± 0.24	66.67 ± 1.89
Qwen-VL-Chat	32.17 ± 0.24	40.67 ± 1.03	58.00 ± 0.71	84.67 ± 1.03	88.33 ± 2.05
LLaVA-Next-7B	20.50 ± 0.00	64.50 ± 0.82	52.83 ± 1.25	7.83 ± 1.65	7.83 ± 1.55
Emu2-Chat	29.89 ± 0.00	50.17 ± 1.44	51.43 ± 1.52	59.38 ± 2.03	57.25 ± 3.06
GPT4V	48.00	56.00	78.00	90.00	86.00

Table 14: 不同模型在算子归纳数据集上的结果（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	5.00 ± 0.00	2.22 ± 3.14	1.67 ± 1.36	2.78 ± 0.79	7.78 ± 2.08
IDEFICS-9B	11.67 ± 0.00	14.44 ± 0.79	10.56 ± 2.08	7.78 ± 2.08	11.11 ± 1.57
IDEFICS-80B	13.33 ± 0.00	15.00 ± 2.72	14.67 ± 2.36	21.67 ± 1.36	16.11 ± 2.08
Otter	21.67 ± 0.00	11.67 ± 2.36	13.33 ± 1.36	12.22 ± 1.57	7.22 ± 1.57
InternLM-X2	26.11 ± 3.14	40.00 ± 10.80	40.00 ± 4.91	39.44 ± 7.49	28.89 ± 19.83
Qwen-VL-Chat	15.00 ± 0.00	10.00 ± 1.36	17.22 ± 3.14	18.89 ± 1.57	25.00 ± 2.72
LLaVA-Next-7B	10.56 ± 1.57	6.11 ± 1.57	5.56 ± 2.08	3.33 ± 2.72	0.00 ± 0.00
Emu2-Chat	28.56 ± 1.57	21.67 ± 5.93	21.11 ± 1.57	21.67 ± 0.00	21.11 ± 5.50
GPT4V	24.00	66.00	84.00	92.00	92.00

Table 15: 不同模型在 TextOCR 数据集上的结果（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
IDEFICS-9B	16.50 ± 0.00	22.50 ± 1.08	19.83 ± 0.62	22.83 ± 1.31	28.00 ± 1.63
IDEFICS-80B	20.00 ± 0.00	25.50 ± 2.18	25.38 ± 2.78	29.50 ± 2.89	23.50 ± 3.47
Otter	0.00 ± 0.00	0.00 ± 0.00	0.17 ± 0.24	0.83 ± 0.47	0.67 ± 0.24
InternLM-X2	8.67 ± 4.01	3.83 ± 0.62	10.50 ± 0.71	16.00 ± 2.48	11.83 ± 2.95
Qwen-VL-Chat	4.83 ± 6.84	17.17 ± 1.43	21.50 ± 1.08	22.33 ± 1.31	24.17 ± 0.24
LLaVA-Next-7B	24.67 ± 2.25	0.83 ± 0.24	0.33 ± 0.24	0.00 ± 0.00	0.00 ± 0.00
Emu2-Chat	25.83 ± 0.24	23.50 ± 1.47	31.50 ± 1.87	36.50 ± 1.87	29.50 ± 1.78
GPT4V	39.29	32.14	48.00	50.00	49.00

Table 16: 不同模型在 CLEVR 数据集上的结果（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	0.00 ± 0.00	17.83 ± 2.25	17.00 ± 2.27	18.83 ± 1.03	16.33 ± 1.43
IDEFICS-9B	0.00 ± 0.00	30.33 ± 2.25	29.50 ± 1.47	27.67 ± 2.05	27.17 ± 2.87
IDEFICS-80B	0.00 ± 0.00	31.16 ± 2.10	30.82 ± 1.59	31.50 ± 1.00	32.43 ± 3.62
Otter	0.00 ± 0.00	5.42 ± 1.06	8.33 ± 2.24	8.17 ± 1.44	0.17 ± 0.24
InternLM-X2	1.83 ± 0.24	26.00 ± 1.63	24.67 ± 5.25	20.00 ± 2.94	22.83 ± 0.85
Qwen-VL-Chat	0.00 ± 0.00	29.83 ± 4.55	25.33 ± 3.47	26.83 ± 3.06	30.17 ± 2.95
LLaVA-Next-7B	0.00 ± 0.00	25.17 ± 6.64	24.83 ± 4.90	17.83 ± 4.59	0.17 ± 0.24
Emu2-Chat	5.33 ± 0.24	11.83 ± 2.72	14.00 ± 3.49	14.83 ± 1.89	17.67 ± 1.03
GPT4V	6.00	30.00	38.00	42.00	32.00

Table 17: 不同模型在交错算子归纳法上的结果（精度% ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
OpenFlamingo-9B	0.00 ± 0.00	5.56 ± 1.57	3.89 ± 2.83	2.78 ± 0.79	8.89 ± 3.42
IDEFICS-9B	15.00 ± 0.00	5.56 ± 2.08	6.11 ± 0.79	6.11 ± 1.57	5.00 ± 2.36
IDEFICS-80B	25.00 ± 0.00	36.67 ± 1.21	31.67 ± 2.46	28.33 ± 3.13	20.00 ± 2.77
Otter	8.33 ± 0.00	7.78 ± 1.57	9.44 ± 3.14	7.22 ± 2.83	5.56 ± 2.83
InternLM-X2	28.33 ± 0.00	10.56 ± 2.83	9.44 ± 2.83	11.11 ± 3.93	4.44 ± 2.83
Qwen-VL-Chat	16.67 ± 0.00	9.44 ± 0.79	8.33 ± 1.36	8.89 ± 2.83	5.56 ± 0.79
LLaVA-Next-7B	13.89 ± 1.57	7.22 ± 2.83	6.11 ± 3.14	5.00 ± 0.00	5.00 ± 2.72
Emu2-Chat	26.67 ± 0.00	18.33 ± 2.72	20.56 ± 3.42	10.00 ± 0.00	7.62 ± 1.83
GPT4V	36.00	58.00	72.00	74.00	70.00

Table 18: 不同模型在匹配 MiniImageNet 数据集上的结果（精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
OpenFlamingo-9B	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
IDEFICS-9B	50.00 ± 0.00	50.50 ± 0.35	50.83 ± 0.85	50.00 ± 0.20	49.92 ± 0.12
IDEFICS-80B	61.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
Otter	48.75 ± 0.00	50.58 ± 0.31	50.92 ± 0.42	50.42 ± 0.12	49.83 ± 0.31
InternLM-XComposer2	63.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.08 ± 1.65	50.00 ± 0.00
Qwen-VL-Chat	50.50 ± 0.50	57.32 ± 1.82	55.50 ± 1.50	56.43 ± 1.17	52.82 ± 0.49
LLaVA-Next-7B	63.00 ± 0.00	50.00 ± 0.00	49.75 ± 0.00	50.00 ± 0.00	49.75 ± 0.00
Emu2-Chat	62.15 ± 3.28	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
GPT4V	52.00	76.00	82.00	81.00	82.00

Table 19: CobSAT 上不同模型的结果：总精度（精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
GILL	2.67 ± 0.24	12.33 ± 1.31	9.33 ± 0.24	11.50 ± 1.47	8.00 ± 1.63
SEED-LLaMA-8B	0.50 ± 0.41	15.83 ± 1.65	21.83 ± 1.65	27.83 ± 2.36	33.67 ± 2.32
SEED-LLaMA-14B	5.50 ± 0.71	26.83 ± 1.65	33.33 ± 3.32	40.83 ± 1.65	43.83 ± 2.87
Emu1-Gen	0.33 ± 0.47	4.83 ± 0.47	6.17 ± 2.72	8.67 ± 1.18	9.67 ± 0.24
Emu2-Gen	8.67 ± 0.62	23.00 ± 3.24	28.67 ± 2.01	27.33 ± 2.72	20.83 ± 0.85

Table 20: CobSAT 上不同模型的结果：潜在精度（精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
GILL	7.67 ± 0.24	47.67 ± 1.43	53.17 ± 1.65	67.33 ± 1.03	72.33 ± 0.85
SEED-LLaMA-8B	8.00 ± 0.41	38.50 ± 1.47	44.33 ± 0.62	49.50 ± 0.82	56.00 ± 1.41
SEED-LLaMA-14B	15.17 ± 0.62	41.50 ± 1.41	52.17 ± 2.09	53.67 ± 1.93	57.17 ± 1.84
Emu1-Gen	8.00 ± 0.41	55.50 ± 2.55	71.00 ± 0.71	77.00 ± 0.82	82.00 ± 0.00
Emu2-Gen	18.00 ± 1.63	43.83 ± 4.33	72.33 ± 1.25	81.50 ± 0.41	78.33 ± 1.25

Table 21: 不同模型在 CobSAT 上的结果：非潜在精度（精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
GILL	19.33 ± 0.24	21.33 ± 1.65	16.17 ± 1.65	19.83 ± 2.01	15.83 ± 2.78
SEED-LLaMA-8B	12.33 ± 1.03	47.33 ± 4.03	52.00 ± 1.87	58.83 ± 1.93	63.33 ± 2.01
SEED-LLaMA-14B	82.67 ± 0.24	76.33 ± 0.94	75.83 ± 1.70	78.33 ± 0.85	80.83 ± 0.85
Emu1-Gen	26.50 ± 0.41	11.17 ± 0.47	13.33 ± 2.25	16.00 ± 0.71	17.00 ± 0.71
Emu2-Gen	62.00 ± 0.41	49.17 ± 4.29	42.33 ± 2.62	35.67 ± 2.05	29.33 ± 1.43

Table 22: 不同模型在 Text-to-Image Fast Mini-ImageNet 上的结果（精度%）

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
GILL	0.00 ± 0.00	16.00 ± 2.27	15.17 ± 2.72	14.83 ± 0.24	14.33 ± 2.25
SEED-LLaMA-8B	0.00 ± 0.00	15.00 ± 3.27	12.67 ± 1.18	16.00 ± 2.12	16.50 ± 1.87
SEED-LLaMA-14B	0.75 ± 0.25	17.25 ± 2.75	16.75 ± 1.75	21.25 ± 1.75	21.00 ± 3.00
Emu1-Gen	0.50 ± 0.41	31.50 ± 1.87	22.83 ± 2.72	25.00 ± 0.71	23.17 ± 1.03
Emu2-Gen	0.00 ± 0.00	24.33 ± 3.30	30.67 ± 1.31	37.00 ± 1.22	34.50 ± 0.00

Table 23: 不同模型在文本版的算子归纳（精度%）上的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
InternLM-XComposer2	15.00	50.00	73.33	75.00	83.33
Qwen-VL-Chat	0.00	45.00	56.67	63.33	71.67
LLaVA-Next-Vicuna-7B	10.00	40.00	53.33	60.00	68.33

Table 24: 不同模型在交错算子归纳法文本版上的结果（精度% ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
InternLM-XComposer2	8.33	35.00	36.67	46.67	78.33
Qwen-VL-Chat	0.00	50.00	55.00	61.67	66.67
LLaVA-Next-Vicuna-7B	16.67	41.67	45.00	53.33	70.00

Table 25: 不同模型在 CLEVR 数据集文本版本上的结果（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
InternLM-XComposer2	0.00	45.00	43.00	42.00	41.00
Qwen-VL-Chat	0.00	49.50	47.50	54.00	53.50
LLaVA-Next-Vicuna-7B	0.00	43.00	38.50	37.50	36.50

Table 26: 不同模型在文本版的文本版本上的结果快速 Mini-ImageNet （精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
GILL	0.00	18.50	20.00	20.50	18.50
SEED-LLaMA-8B	0.00	16.30	15.20	16.50	14.20
SEED-LLaMA-14B	1.50	23.00	20.00	22.50	15.50
Emu1-Gen	0.50	28.60	29.10	24.20	20.00
Emu2-Gen	0.20	32.40	38.80	40.50	42.10

Table 27: 不同模型在 CobSAT 文本版本上的结果：总精度（%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
GILL	6.00	13.00	20.50	22.50	23.50
SEED-LLaMA-8B	0.50	14.50	15.50	30.50	32.00
SEED-LLaMA-14B	6.00	13.50	28.00	34.00	40.50
Emu1-Gen	2.50	11.00	19.50	23.50	20.00
Emu2-Gen	7.50	19.50	32.50	46.50	45.00

Table 28: 不同模型在 CobSAT 文本版本上的结果：潜在精度（%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
GILL	6.50	33.00	39.00	37.50	38.00
SEED-LLaMA-8B	4.00	17.50	18.50	35.00	46.00
SEED-LLaMA-14B	6.50	60.00	55.50	60.00	66.00
Emu1-Gen	6.00	24.00	31.50	43.50	42.00
Emu2-Gen	12.00	74.00	86.00	92.50	88.50

Table 29: 不同模型在 CobSAT 文本版本上的结果：非潜在精度（%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
GILL	86.00	44.50	62.50	67.00	71.50
SEED-LLaMA-8B	21.00	80.00	83.50	80.50	74.50
SEED-LLaMA-14B	90.00	21.50	56.50	63.00	67.50
Emu1-Gen	30.00	33.50	52.00	48.50	45.50
Emu2-Gen	68.50	22.50	37.50	50.50	49.00

Table 30: 在快速开放式 MiniImageNet 数据集上进行和不进行指令跟随微调的上下文学习能力的比较。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
IDEFICS-9B	0.00	16.00	47.50	58.00	56.00
IDEFICS-9B-Instruct	0.00	22.00	52.00	53.83	59.17
Qwen-VL	0.00	35.50	79.50	68.00	67.00
Qwen-VL-Chat	0.00	0.50	47.33	58.00	55.17

Table 31: 在 TextOCR 数据集上进行和不进行指令跟随微调的上下文学习能力的比较（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
IDEFICS-9B	3.50	16.50	22.50	25.00	26.00
IDEFICS-9B-Instruct	16.50	22.50	19.83	22.83	28.00
Qwen-VL	0.00	27.00	28.50	30.50	37.00
Qwen-VL-Chat	4.83	17.17	21.50	22.33	24.17

Table 32: 在 CLEVR 数据集上比较有和没有指令跟随微调的上下文学习能力（Accuracy % ）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
IDEFICS-9B	0.00	19.50	26.00	25.00	29.00
IDEFICS-9B-Instruct	0.00	30.33	29.50	27.67	27.17
Qwen-VL	2.50	18.50	17.50	24.00	26.00
Qwen-VL-Chat	0.00	29.83	25.33	26.83	30.17

Table 33: 在操作员归纳数据集（Accuracy %）上比较有和没有指令跟随微调的上下文学习能力。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
IDEFICS-9B	5.00	16.67	8.33	10.00	3.33
IDEFICS-9B-Instruct	11.67	14.44	10.56	7.78	11.11
Qwen-VL	15.00	26.67	36.67	46.67	56.67
Qwen-VL-Chat	15.00	10.00	17.22	18.89	25.00

Table 34: 在交错运算符归纳数据集（Accuracy %）上比较有和没有指令跟随微调的上下文学习能力。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
IDEFICS-9B	13.33	8.33	5.00	10.00	3.33
IDEFICS-9B-Instruct	15.00	5.56	6.11	6.11	5.00
Qwen-VL	0.00	13.33	13.33	8.33	11.67
Qwen-VL-Chat	16.67	9.44	8.33	8.89	5.56

5.4 补充结果

下面我们介绍附录中数字的原始结果。

5.4.1 缩放到多个镜头

表 **??** 到 **??**

5.4.2 思维链提示

表 **??** 到 **??**

5.4.3 重复支撑集

表 **??** 到 **??** 。

5.4.4 不同级别的任务描述

表 **??** 到 **??** 。

Table 35: CLEVR 数据集上许多镜头的结果。

Model	16-Shot	32-Shot	64-Shot
OpenFlamingo-9B	22.00 ± 1.47	25.33 ± 1.65	25.67 ± 2.39
IDEFICS-9B-Instruct	28.17 ± 2.66	29.00 ± 1.08	30.50 ± 1.78
InternLM-X2	14.67 ± 1.70	15.50 ± 1.08	16.33 ± 1.03

Table 36: Operator Induction 数据集上许多镜头的结果。

Model	16-Shot	32-Shot	64-Shot
OpenFlamingo-9B	13.33 ± 3.60	8.89 ± 1.57	11.67 ± 1.36
IDEFICS-9B-Instruct	5.00 ± 3.60	7.78 ± 1.57	5.00 ± 1.36
InternLM-X2	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 37: 交错算子归纳数据集上多次拍摄的结果。

Model	16-Shot	32-Shot	64-Shot
OpenFlamingo-9B	8.89 ± 3.42	8.33 ± 3.60	11.67 ± 3.60
IDEFICS-9B-Instruct	8.89 ± 2.08	7.78 ± 2.08	7.78 ± 2.83
InternLM-X2	3.89 ± 0.79	5.00 ± 1.36	5.00 ± 1.36

Table 38: TextOCR 数据集上许多镜头的结果。

Model	16-Shot	32-Shot	64-Shot
OpenFlamingo-9B	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
IDEFICS-9B-Instruct	29.00 ± 1.22	33.17 ± 0.85	33.50 ± 1.47
InternLM-X2	3.11 ± 0.58	0.00 ± 0.00	0.00 ± 0.00

Table 39: 在 Operator Induction 数据集上使用思路链提示的结果。

Model	0-shot	1-shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat	13.67	21.00	15.33	9.33	16.67
InternLM-X2	28.33	32.00	35.67	30.00	5.00

Table 40: 交错算子归纳数据集上具有思路链提示的结果。

Model	0-shot	1-shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat	3.33	10.00	10.00	8.33	8.33
InternLM-X2	18.33	5.00	5.00	10.00	8.33

Table 41: 在 TextOCR 数据集上使用思路链提示的结果。

Model	0-shot	1-shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat	7.00	28.50	27.50	30.50	22.50
InternLM-X2	12.00	1.50	0.50	2.00	0.50

Table 42: CLEVR 数据集上具有思维链提示的结果。

Model	0-shot	1-shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat	0.50	10.50	21.50	18.00	26.00
InternLM-X2	4.00	21.50	20.00	26.00	27.00

Table 43: 在匹配的 Mini-ImageNet 数据集上具有思维链提示的结果。

Model	0-shot	1-shot	2-Shot	4-Shot	5-Shot
Qwen-VL-Chat	56.00	56.75	51.25	56.75	53.00
InternLM-X2	58.25	51.50	53.00	50.00	48.50

Table 44: Qwen-VL-Chat 在 Fast Open-Ended MiniImageNet 数据集上的结果，并带有重复的上下文示例。

Model	1-shot	2-Shot	4-Shot	5-Shot
No Repeat	0.50	47.33	58.00	55.17
Repeat x2	41.00	62.50	54.50	56.50
Repeat x3	62.50	55.50	61.00	62.00
Repeat x4	60.00	56.50	60.00	58.50

Table 45: Qwen-VL-Chat on Operator Induction 数据集的结果，并带有重复的上下文示例。

Model	1-shot	2-Shot	4-Shot	8-Shot
No repeat	10.00	17.22	18.89	25.00
Repeat x2	5.00	15.00	23.33	25.00
Repeat x3	11.67	15.00	20.00	26.67
Repeat x4	13.33	18.33	21.67	18.33

Table 46: Qwen-VL-Chat 在 CLEVR 数据集上的结果，带有重复的上下文示例。

Model	1-shot	2-Shot	4-Shot	8-Shot
No Repeat	29.83	25.33	26.83	30.17
Repeat x2	25.50	30.50	27.50	28.50
Repeat x3	22.50	32.50	26.50	32.00
Repeat x4	19.50	31.50	23.00	27.50

Table 47: 不同模型在 Fast Open-Ended MiniImageNet 数据集上的结果（精度%）。

Model	0-Shot	1-Shot	2-Shot	4-Shot	5-Shot
Qwen-VL-Chat – Detailed	0.00 ± 0.00	0.50 ± 0.41	47.33 ± 2.49	58.00 ± 2.83	55.17 ± 2.25
Qwen-VL-Chat – Concise	0.00 ± 0.00	0.83 ± 0.62	48.00 ± 2.45	59.00 ± 0.41	52.50 ± 2.68
Qwen-VL-Chat – None	0.00 ± 0.00	6.33 ± 0.47	56.17 ± 1.65	57.67 ± 0.85	53.83 ± 2.78
LLaVA-Next-7B – Detailed	0.00 ± 0.00	22.17 ± 4.03	33.67 ± 2.25	0.00 ± 0.00	0.33 ± 0.24
LLaVA-Next-7B – Concise	0.00 ± 0.00	24.00 ± 0.71	34.50 ± 2.68	0.00 ± 0.00	0.33 ± 0.24
LLaVA-Next-7B – None	0.00 ± 0.00	16.67 ± 2.01	32.00 ± 2.55	0.33 ± 0.24	0.17 ± 0.24
OpenFlamingo-9B – Detailed	0.00 ± 0.00	39.50 ± 1.22	58.17 ± 3.57	51.17 ± 0.85	54.50 ± 5.66
OpenFlamingo-9B – Concise	0.00 ± 0.00	36.50 ± 0.41	51.67 ± 2.78	52.17 ± 0.62	49.33 ± 1.25
OpenFlamingo-9B – None	0.00 ± 0.00	38.17 ± 1.03	52.17 ± 2.46	49.17 ± 0.85	49.33 ± 1.25
InternLM-X2 – Detailed	0.00 ± 0.00	14.83 ± 1.03	38.00 ± 1.78	49.00 ± 1.78	50.33 ± 3.86
InternLM-X2 – Concise	0.00 ± 0.00	19.50 ± 1.47	40.33 ± 1.89	48.83 ± 0.85	49.17 ± 1.93
InternLM-X2 – None	0.00 ± 0.00	22.00 ± 2.04	43.00 ± 2.16	46.33 ± 3.06	48.17 ± 0.62
IDEFICS-9B – Detailed	0.00 ± 0.00	22.00 ± 0.41	52.00 ± 2.94	53.83 ± 0.94	59.17 ± 6.20
IDEFICS-9B – Concise	0.00 ± 0.00	28.50 ± 1.78	53.83 ± 4.09	53.83 ± 0.94	55.67 ± 2.09
IDEFICS-9B – None	0.00 ± 0.00	37.17 ± 4.29	52.17 ± 4.48	53.17 ± 1.25	55.50 ± 1.47

Table 48: 使用不同级别的任务描述（精度%）的 CLEVR 计数归纳数据集上不同模型的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat – Detailed	0.00 ± 0.00	29.83 ± 4.55	25.33 ± 3.47	26.83 ± 3.06	30.17 ± 2.95
Qwen-VL-Chat – Concise	0.00 ± 0.00	24.67 ± 2.32	25.67 ± 0.85	25.33 ± 1.65	24.83 ± 2.32
Qwen-VL-Chat – None	1.00 ± 0.00	25.17 ± 2.72	24.33 ± 1.31	24.83 ± 1.31	24.67 ± 2.36
LLaVA-Next-7B – Detailed	0.00 ± 0.00	25.17 ± 6.64	24.83 ± 4.90	17.83 ± 4.59	0.17 ± 0.24
LLaVA-Next-7B – Concise	0.00 ± 0.00	25.00 ± 3.49	27.00 ± 3.89	20.00 ± 2.48	0.00 ± 0.00
LLaVA-Next-7B – None	0.00 ± 0.00	15.50 ± 2.12	23.83 ± 2.87	12.83 ± 1.70	0.17 ± 0.24
OpenFlamingo-9B – Detailed	0.00 ± 0.00	17.83 ± 2.25	17.00 ± 2.27	18.83 ± 1.03	16.33 ± 1.43
OpenFlamingo-9B – Concise	0.00 ± 0.00	15.33 ± 2.39	19.00 ± 2.27	20.00 ± 0.71	18.33 ± 3.09
OpenFlamingo-9B – None	0.00 ± 0.00	15.33 ± 0.94	18.17 ± 1.03	21.33 ± 1.89	19.33 ± 2.78
InternLM-X2 – Detailed	1.83 ± 0.24	26.00 ± 1.63	24.67 ± 5.25	20.00 ± 2.94	22.83 ± 0.85
InternLM-X2 – Concise	1.00 ± 0.00	19.33 ± 2.25	20.17 ± 1.31	9.50 ± 1.41	12.33 ± 2.32
InternLM-X2 – None	1.50 ± 0.00	26.67 ± 2.09	24.67 ± 2.01	25.17 ± 1.18	23.17 ± 2.25
IDEFICS-9B – Detailed	0.00 ± 0.00	30.33 ± 2.25	29.50 ± 1.47	27.67 ± 2.05	27.17 ± 2.87
IDEFICS-9B – Concise	1.00 ± 0.00	30.67 ± 1.84	31.00 ± 3.94	26.17 ± 1.55	26.83 ± 0.62
IDEFICS-9B – None	0.00 ± 0.00	30.83 ± 1.43	31.33 ± 2.95	28.50 ± 1.78	28.00 ± 0.41

Table 49: 使用不同级别的任务描述（精度%）的操作员归纳数据集上不同模型的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat – Detailed	15.00 ± 0.00	10.00 ± 1.36	17.22 ± 3.14	18.89 ± 1.57	25.00 ± 2.72
Qwen-VL-Chat – Concise	15.00 ± 0.00	7.22 ± 2.08	15.56 ± 3.42	17.78 ± 2.08	27.22 ± 0.79
Qwen-VL-Chat – None	15.00 ± 0.00	8.33 ± 2.36	14.44 ± 2.83	18.33 ± 2.72	27.22 ± 0.79
LLaVA-Next-7B – Detailed	10.56 ± 1.57	6.11 ± 1.57	5.56 ± 2.08	3.33 ± 2.72	0.00 ± 0.00
LLaVA-Next-7B – Concise	5.00 ± 0.00	7.22 ± 0.79	5.56 ± 2.08	4.44 ± 2.08	1.11 ± 0.79
LLaVA-Next-7B – None	8.33 ± 0.00	6.11 ± 0.79	5.56 ± 1.57	4.44 ± 1.57	0.56 ± 0.79
OpenFlamingo-9B – Detailed	5.00 ± 0.00	2.22 ± 3.14	1.67 ± 1.36	2.78 ± 0.79	7.78 ± 2.08
OpenFlamingo-9B – Concise	6.67 ± 0.00	5.00 ± 3.60	4.44 ± 3.14	4.44 ± 1.57	9.44 ± 1.57
OpenFlamingo-9B – None	6.67 ± 0.00	5.00 ± 3.60	3.33 ± 2.36	4.44 ± 2.08	11.67 ± 3.60
InternLM-X2 – Detailed	26.11 ± 3.14	40.00 ± 10.80	40.00 ± 4.91	39.44 ± 7.49	28.89 ± 19.83
InternLM-X2 – Concise	18.33 ± 0.00	29.44 ± 3.42	22.78 ± 2.83	18.33 ± 1.36	16.67 ± 2.36
InternLM-X2 – None	18.33 ± 0.00	13.33 ± 2.36	12.78 ± 2.83	12.22 ± 2.08	16.67 ± 2.72
IDEFICS-9B – Detailed	11.67 ± 0.00	14.44 ± 0.79	10.56 ± 2.08	7.78 ± 2.08	11.11 ± 1.57
IDEFICS-9B – Concise	15.00 ± 0.00	13.89 ± 2.83	12.22 ± 0.79	8.89 ± 0.79	8.33 ± 3.60
IDEFICS-9B – None	15.00 ± 0.00	17.22 ± 2.83	10.56 ± 0.79	10.56 ± 2.08	7.78 ± 3.93

Table 50: 使用不同级别的任务描述（精度%）的 TextOCR 数据集上不同模型的结果。

Model	0-Shot	1-Shot	2-Shot	4-Shot	8-Shot
Qwen-VL-Chat – Detailed	4.83 ± 6.84	17.17 ± 1.43	21.50 ± 1.08	22.33 ± 1.31	24.17 ± 0.24
Qwen-VL-Chat – Concise	0.00 ± 0.00	8.00 ± 0.82	9.50 ± 0.41	9.83 ± 0.62	9.17 ± 0.24
Qwen-VL-Chat – None	0.00 ± 0.00	9.67 ± 0.62	10.33 ± 0.47	10.67 ± 0.47	9.33 ± 0.47
LLaVA-Next-7B – Detailed	24.67 ± 2.25	0.83 ± 0.24	0.33 ± 0.24	0.00 ± 0.00	0.00 ± 0.00
LLaVA-Next-7B – Concise	8.50 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LLaVA-Next-7B – None	10.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
OpenFlamingo-9B – Detailed	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
OpenFlamingo-9B – Concise	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
OpenFlamingo-9B – None	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
InternLM-X2 – Detailed	8.67 ± 4.01	3.83 ± 0.62	10.50 ± 0.71	16.00 ± 2.48	11.83 ± 2.95
InternLM-X2 – Concise	0.50 ± 0.00	0.50 ± 0.41	0.83 ± 0.47	2.33 ± 1.03	0.00 ± 0.00
InternLM-X2 – None	0.50 ± 0.00	0.50 ± 0.41	1.33 ± 0.47	3.67 ± 2.09	0.00 ± 0.00
IDEFICS-9B – Detailed	16.50 ± 0.00	22.50 ± 1.08	19.83 ± 0.62	22.83 ± 1.31	28.00 ± 1.63
IDEFICS-9B – Concise	3.00 ± 0.00	2.50 ± 0.41	5.50 ± 0.41	5.83 ± 0.24	6.17 ± 0.47
IDEFICS-9B – None	4.00 ± 0.00	2.67 ± 0.62	5.33 ± 0.47	6.00 ± 0.41	6.33 ± 0.62