

用视觉原语进行推理

Thinking with Visual Primitives

Ruijie Lu, Yiyang Ma, Xiaokang Chen, Lingxiao Luo, Zhiyu Wu, Zizheng Pan, Xingchao Liu, Yutong Lin, Hao Li, Wen Liu, Zhewen Hao, Xi Gao, Shaoheng Nie, Yixuan Wei, Zhenda Xie, Ting Chen, Gang Zeng

摘要

近期多模态大语言模型 (MLLMs) 在缩小'感知鸿沟'方面取得了进展,但在复杂的结构推理任务上仍然表现不佳。本研究发现核心瓶颈在于'指代鸿沟 (Reference Gap)'——自然语言过于模糊,无法精确指向密集的空间布局,导致推理过程中出现逻辑崩溃和幻觉。本项目提出了一种范式转变:模型不再只是'看得更清楚',而是学会'边指边思考'。通过将空间标记(点和边界框)直接嵌入推理轨迹作为最小思维单位,模型能够将抽象的语言概念锚定到具体的物理坐标上。该方法基于 DeepSeek-V4-Flash 架构,实现了极致的视觉 token 效率,在计数和空间推理基准上匹敌 GPT-5.4、Claude-Sonnet-4.6、Gemini-3-Flash 等前沿模型。

1. Introduction

已探索将边界框融入思维链过程，但它们主要将视觉定位视为一种事后验证机制，以增强感知密集型任务的性能。此外，它们对劳动密集型监督的依赖进一步限制了可扩展性。在此类任务中，视觉标记必须作为思维的内在媒介，而不仅仅是可验证的证据。在本研究中，我们提出一种范式转变：基于视觉基元的思维（Thinking with Visual Primitives）。我们不再将视觉定位视为次要任务或最终输出。使其直接交错嵌入模型的推理轨迹中。该机制的灵感来源于人类的认知过程。当在复杂迷宫中导航或清点密集物体时，人类会自然地使用指示性指针（如手指手势）来降低认知负荷并保持逻辑一致性。的协同机制，从而将抽象的语言思维有效锚定到具体的空间坐标上。长上下文的多模态交互而设计。将每²个视觉词元的键值（KV）缓存压缩为单一条目。模型架构与训练流程。开发，我们的模型在预训练阶段获得了基础的视觉基元生成能力。随后进入后训练阶段，采用专家级专业化与整合范式。建议在屏幕上查看以获得最佳效果。

通过广泛的基准测试，我们证明了“基于视觉原语的思维（Thinking with Visual Primitives）”在推理准确率上实现了显著跃升。我们的模型展现出极具竞争力的性能，在一系列极具挑战性的空间推理与视觉问答任务中，表现媲美甚至超越了 GPT、Claude 和 Gemini 的最新版本（见图 1）。我们的研究表明，多模态智能的未来不仅在于“看见”更多的像素，更在于开发更精确、歧义更少的指代机制，以弥合语言与视觉世界之间的鸿沟。

通过广泛的基准测试，我们证明了“基于视觉原语的推理”（Thinking with Visual Primitives）在推理准确率上实现了显著跃升。我们的模型展现出极具竞争力的性能，在一系列极具挑战性的空间推理与视觉问答任务中，表现媲美甚至超越了 GPT、Claude 和 Gemini 的最新版本（见图 1）。研究表明，多模态智能的未来不仅在于“看见”更多的像素，更在于开发更精确、歧义更少的指代机制，以弥合语言与视觉世界之间的鸿沟。

通过广泛的基准测试，我们证明了“基于视觉原语的推理”（Thinking with Visual Primitives）在推理准确率上实现了显著跃升。我们的模型展现出极具竞争力的性能，在一系列极具挑战性的空间推理与视觉问答任务中，表现媲美甚至超越了 GPT、Claude 和 Gemini 的最新版本（见图 1）。研究表明，多模态智能的未来不仅在于“看见”更多的像素，更在于构建更精确、歧义更少的指代机制，从而弥合语言与视觉世界之间的鸿沟。

通过广泛的基准测试，我们证明了“基于视觉原语的推理”（Thinking with Visual Primitives）在推理准确率上实现了显著跃升。我们的模型展现出极具竞争力的性能，在一系列极具挑战性的空间推理与视觉问答任务中，表现媲美甚至超越了 GPT、Claude 和 Gemini 的最新版本（见图 1）。研究表明，多模态智能的未来不仅在于“看见”更多的像素，更在于开发更精确、歧义更少的指代机制，以弥合语言与视觉世界之间的鸿沟。

通过广泛的基准测试，我们证明了“基于视觉原语的推理”（Thinking with Visual Primitives）在推理准确率上实现了显著跃升。我们的模型展现出极具竞争力的性能，在一系列极具挑战性的空间推理与视觉问答任务中，表现媲美甚至超越了 GPT、Claude 和 Gemini 的最新版本（见图 1）。研究表明，多模态智能的未来不仅在于“看见”更多的像素，更在于构建更精确、歧义更少的指代机制，从而弥合语言与视觉世界

2. Method

本节首先介绍模型架构。接下来，我们将详细阐述训练流程（如图2所示），并描述在预训练和后训练阶段所使用的相应数据。符合学术规范，流畅。类似的标准架构。语言标记的交错序列。该序列随后被输入到大语言模型（LLM）中以生成响应。这是一个混合专家（MoE）模型，包含2840亿总参数，在推理期间激活130亿参数。类似的标准架构。语言标记（tokens）的交错序列。该序列随后被输入至大语言模型（LLM）以生成响应。这是一个混合专家（MoE）模型，总参数量为2840亿（284B），推理时激活参数量为130亿（13B）。这是一种从头训练的内部ViT，支持任意分辨率的输入。的patch大小对输入图像进行划分，以生成patch标记。的空间标记压缩（沿通道维度将每9个相邻的patch标记压缩为单个标记）。此外，利用基础LLM中集成的压缩稀疏注意力（CSA）机制，存储在键值（KV）缓存中的视觉标记进一步被压缩了4倍。即压缩4倍。该模型从头训练，支持任意分辨率的输入。的图像块（patch）大小对输入图像进行划分，生成图像块标记。的空间标记压缩（沿通道维度将每9个相邻的图像块标记压缩为1个标记）。此外，借助基础LLM中集成的压缩稀疏注意力（CSA）机制，存储在键值（KV）缓存中的视觉标记进一步被压缩了4倍。个像素的输入图像。个图像块标记供ViT使用。个视觉标记被输入到LLM中。最终，CSA机制将其进一步减少至KV缓存中仅81个视觉KV条目。倍的整体压缩率。数字格式统一为中文字符或保留英文逗号。

抽象视觉指代，例如追踪运动轨迹或解决拓扑推理问题。提供了相对准确的边界框或点标注，但它们存在规模不足和多样性显著缺乏的问题。范式的泛化能力，必须构建具有丰富语义和高多样性的超大规模网络数据。筛选。更一致。标注的确定性：边界框紧密包围目标物体，使其标注具有相对较高的确定性。相反，点标注具有高度歧义性；物体边界内的任意坐标均可作为有效参考点，导致缺乏严格的标准答案（ground truth）。在涉及遮挡的极端场景中，原本指向背景物体的点可能落在前景遮挡物上，从而产生显著的歧义。任务泛化性：训练输出边

界框的模型可以轻松地泛化至基于点的格式。由于边界框可由两个点（左上角和右下角坐标）定义，因此它天然包含了点的表示形式。信息丰富度：与点相比，边界框能够支持更广泛的下游任务。点仅提供空间定位信息，而边界框则封装了详细的几何信息（如宽度和高度）。框架内执行更复杂的推理。为平衡性能与计算成本，我们将 ViT 输出的视觉标记（visual tokens）数量限制在 81 至 384 之间。分辨率超出此范围的图像将在保持原始宽高比的前提下进行缩放。原始数据获取。我们通过跨多个网站进行大规模网络抓取，收集了大量与边界框定位（box grounding）相关的互联网数据。标签的任务数据。提供了相对准确的边界框或点标注，但仍面临规模不足与多样性匮乏的显著局限。整理大规模网络数据。更贴合后文标题。

模型评估过程中的数据污染（即数据泄露）。文件，自动将多样化的数据集结构转换为我们预定义的统一存储格式。相关的数据源。人脸）到特定领域实体（如 CT 扫描中的病灶区域或特定动漫角色）的广泛范围。然而，这些原始框标注仍存在语义歧义和几何不精确等各种问题，因此需要进一步的严格过滤。为此，我们设计了一个两步过滤流程，如下所示。步骤一：基于语义的审查。由多模态大语言模型（MLLM）驱动的语义审查机制。传统的数据过滤主要关注边界框的几何精度，而本阶段旨在确保语义标签的有效性。由于这些标签缺乏人类可读的自然语言语义，强制模型学习此类映射会严重损害其语言生成能力。因此，这些数据被直接剔除。的视觉特征无法泛化为通用概念），此类数据会被严格过滤。相反，广为人知的名人或公众人物则予以保留。往往缺乏具体的视觉描述性。之间毫无视觉关联。模型评估过程中的数据污染（即数据泄露）。

Draft: 对于每个数据集，我们采样三张图像，并提示模型根据上述标准计算质量评分（范围为0到10）。*Refinement:* 针对每个数据集，我们随机抽取三张图像，并提示模型依据前述标准计算质量评分（范围为0至10）。(Added "随机" for sampling context, but "采样" is fine. "依据前述标准" is good.) *Draft:* 模型随后输出明确的“保留”或“丢弃”决定，并附带清晰的理由。*Refinement:* 随后，模型输出明确的“保留（KEEP）”或“丢弃（DISCARD）”决策，并附带清晰的判定依据。*Draft:* 该审查阶段从最初的97,984个数据源中保留了43,141个，随后进入下一个过滤阶段。*Refinement:* 经过此轮审查，初始的97,984个数据源中保留了43,141个，并进入下一阶段的过滤流程。*Draft:* 第二步：视觉几何质量审查。我们进一步评估边界框的几何质量和标注完整性，以确保模型学习精确的区域-文本对齐。*Refinement:* **步骤二：视觉与几何质量审查。** 我们进一步评估边界框（bounding boxes）的几何质量与标注完整性，以确保模型能够学习精确的区域-文本对齐关系。*Draft:* 严重漏标（低召回率）：指图像

中存在多个与给定标签对应的实例，但仅标注了少数几个的情况。如果在采样过程中检测到大规模漏标问题（例如漏标率>50%），则该数据集将被立即丢弃。*Refinement:* **严重漏标（低召回率）**：指图像中实际存在多个与给定标签对应的实例，但仅对其中少数进行了标注的情形。若在采样过程中发现大规模漏标问题（如漏标率 > 50%），则该数据集将被直接剔除。*Draft:* 严重截断和偏移：当边界框未能合理包围目标对象时发生。

在实践中，我们采用差异容忍策略：略微宽松的框（包含少量背景噪声）被认为是可接受的；然而，严重截断物体关键视觉特征的框（例如切断头部或车轮）是严格不可接受的。*Refinement:* **严重截断与偏移**：指边界框未能合理包裹目标对象的情形。实际操作中，我们采用分级容忍策略：略微宽松的框（包含少量背景噪声）可予接受；但若截断严重至切割物体关键视觉特征（如切掉头部或车轮），则严格不予接受。*Draft:* 超大框问题：如果边界框无意义地覆盖了超过90%的图像区域，通常表明图像分类数据被强制转换为检测数据。如果这在采样批次中仅偶尔发生，则被视为可接受的噪声。然而，如果此类全局框在所有三张采样图像中持续出现，则认为该数据集缺乏有意义的定位信息并被丢弃。*Refinement:* **超大框（Mega Boxes）问题**：若边界框无意义地覆盖了超过90%的图像区域，通常表明该数据系由图像分类数据强行转换而来。若此类情况在采样批次中仅偶发出现，则视为可接受的噪声；但若在所有三张采样图像中均频繁/持续出现此类全局框，则认为该数据集缺乏有效的定位信息，予以剔除。*Draft:* 该审查阶段从剩余的43,141个数据源中进一步保留了31,701个。*Refinement:* 经过此轮审查，剩余的43,141个数据源中进一步保留了31,701个。*Draft:* 为了实现数据集平衡，我们设计了一种基于类别的采样策略。*Refinement:* 为实现数据集类别平衡，我们设计了一种基于类别的采样策略。*Draft:* 对于每个数据集中的每个类别，我们随机采样与该类别相关的张图像（如果某个类别的可用图像总数少于，则全部保留）。*Refinement:* 针对每个数据集中的各个类别，我们随机抽取与该类别相关的 N 张图像（若某类别的可用图像总数不足 N 张，则全部保留）。

Draft: 由于单张图像可能同时属于多个类别，我们在每个类别选择后对聚合集进行全局去重。*Refinement:* 鉴于单张图像可能同时属于多个类别，我们在完成各类别抽样后，对汇总集合执行全局去重操作。

对于通用多模态数据，我们主要采用大规模网络爬取数据，而非通过模型蒸馏生成的合成数据（例如合成图像描述）。原始数据经过精心筛选与整理，我们避免使用大语言模型（LLM）对数据内容进行重写。我们为边界框定位（box grounding）和点数据制定了统一的格式标准。在图像中定位TARGET并报告其边界框坐标

。其中TARGET作为查询对象的占位符。为词表中的特殊标记。分别表示边界框的左上角和右下角坐标。这些坐标被归一化为0到999之间的离散整数。在存在多个实例的场景中，边界框按从左到右的顺序排列。帮我找到TARGET。请给出每个实例的中心点坐标。的提示模板。表示点坐标。值得注意的是，与边界框定位格式不同，点任务的响应范式无需输出对象名称。这一设计旨在将基于点的表示扩展到更抽象的概念，例如利用点序列来表示轨迹。最终，整个预训练阶段共消耗了数亿个多模态词元 (tokens)。后训练的冷启动数据。及随后的统一RFT) 需要一个规模较小但高精度的冷启动数据集，以在我们的视觉基元输出接口下引导指令遵循与奖励学习。

(边界框/点) 或程序化生成，(ii) 尽可能使用自动验证器 (例如基于规则的检查器) 以减少标签噪声。我们选择了能够从基于视觉基元的推理 (通过边界框或点) 中受益的代表性任务，并从四个关键维度设计了我们的冷启动数据：计数、空间推理与通用视觉问答、迷宫导航以及路径追踪。多模态大语言模型在准确计数方面始终面临挑战，尤其是在密集场景中。与人类通常采用的系统性扫描与累加策略不同，基于语言的模型在对象数量较多时，往往难以建立精确的对象对应关系。我们通过采用边界框作为视觉基元来提供明确的参考锚点，从而解决这一根本性瓶颈。任务分解。我们将计数任务分为两类：粗粒度计数和细粒度计数。前者侧重于对通用类别 (例如“狗”) 进行计数，而后者则需要根据特定属性或空间约束 (例如“白色的狗”或“左边的狗”) 来区分对象。粗粒度计数。我们聚合了来自多个密集检测数据集的数据，包括[2, 9, 14, 22, 28, 29, 35]。为确保数据质量，我们实施了一个基于三个主要标准的过滤流程：避免对象密度过高、确保边界框足够大以便于清晰识别，以及保持对真实边界框标注的高召回率。对于过滤后的样本，我们提示多模态大语言模型根据图像和边界框标注生成推理过程和简洁的最终回答。推理过程的生成遵循结构化的三步流程：(1) 意图分析，模型在此步骤中识别目标类别；(2) 批量定位，模型利用视觉基元同时定位所有候选对象 (我们发现批量定位在粗粒度任务中效率更高，因为它利用了模型固有的定位优势，同时避免了重复枚举)；以及(3) 统计求和，基于视觉基元进行。为消除冷启动训练期间的噪声，我们实施了严格的验证机制，以确保推理内容中的所有边界框视觉基元与元数据坐标严格对齐，遵循预定义语法，并与最终的数值计数相匹配。The opening fragment: I will translate it exactly as provided, assuming it continues from a previous bullet point. “(边界框/点) 或程序化生成，(ii) 尽可能采用自动验证器 (例如基于规则的检查器) 以降低标签噪声。”

细粒度计数。由于公开可用的细粒度计数专用数据集十分稀缺，我们设计了一套专门的数据构建流程。(1) 问题生成：利用 GQA [10] 中的图像与场景图元数

据，我们提示多模态大语言模型 (MLLM) 筛选信息量丰富的细粒度计数问题。无法生成有意义问题的样本将被丢弃。对于每个有效样本，我们记录真实目标 ID、被排除的负候选 ID，以及 QA 对构建的底层依据。(2) 思维内容合成：利用图像...

细粒度计数
基于视觉原语的推理
[Trigger_Placeholder]
地面上有多少只熊？
原始图像
含视觉原语的图像

1. **我在寻找什么**

问题要求我统计地面上的熊。我需要扫描整个场景，定位所有熊，并检查它们的位置，确保只统计站在地面或岩石上的熊，排除可能处于高处或攀爬状态的个体。

2. **评估每只熊的位置**

从图像上方开始扫描，我发现一只熊[[452,23,804,411]]
垂直攀附...

是否存在一个与灰色金属物体大小相同的紫色橡胶物体？用户要求验证场景中是否存在一个与“灰色金属物体”尺寸相同的“紫色橡胶物体”。这需要执行多步属性验证与比较：首先，我必须定位灰色金属物体，以确定其在场景中的相对尺寸；其次，我需要扫描整个场景，寻找与该特定尺寸相匹配的紫色橡胶物体。扫描场景中的物体后，我在背景偏右上方的位置发现了一个灰色、高反光的球形物体。该物体为`小灰色金属球[[502,209,555,288]]`。将其与场景中的大立方体、大球体和大圆柱体进行对比，可明确将其归类为小型物体。

和黄色物体，但绝对没有紫色物体。因此，场景中不存在与灰色金属物体尺寸相同的紫色橡胶物体。回答 错误，不存在与灰色金属物体尺寸相匹配的紫色橡胶物体。图 4 | 空间推理的冷启动数据示例。模型执行意图分解，并利用视觉原语锚定所有相关实体，从而促进复杂的多跳逻辑推理。建议在屏幕上放大查看。我们提示 MLLM 设计以空间关系与物体交互为中心的问题，以及相应的 t...

(如有)。若存在路径，则在末尾显示 \boxed{True}，否则显示 \boxed{False}。

原始图像
含视觉原语的图像
基于视觉原语的推理

我将采用试错策略探索该迷宫。首先定位起点：[[309,512]]，以及终点：[[652,644]]。

****开始探索**：**

****步骤1**：**向右上移动到到达 [[357,369]] 后，我面临 2 个可选方向（右下、左）。右下方向可能也正确，但我先探索左方向。

****步骤2**：**向左移动：...

成功到达目的地：[[652,644]]！
该迷宫有解。验证后的路径为：图5 | 迷宫导航任务的冷启动数据示例。模型首先识别起点和终点，随后以深度优先搜索 (DFS) 的方式探索可能的路径。建议在屏幕上放大查看效果最佳。领域。尽管多模态大语言模型 (MLLMs) 在解决高级科学问题方面已展现出卓越能力，但一种稳健的拓扑推理范式仍难以捉摸。纯语言形式的思维链 (CoT) 难以准确描述不规则形状的轨迹。为弥补这一不足，“基于视觉基元的思考” (Thinking with Visual Primitives) 方法将点作为认知单元，特别适合应对此类挑战。我们首先引入一项迷宫导航任务，要求模型判断迷宫是否有解——这一过程需要对空间连通性与可达性有基本的理解。我们通过合成数据生成来构建冷启动数据，具体细节如下。设计方法。我们采用深度优先搜索 (DFS)、普里姆 (Prim) 和克鲁斯卡尔 (Kruskal) 算法来生成有解且非平凡的迷宫。这三种算法均能生成具有挑战性的迷宫，其中任意两个单元格之间仅存在极少路径，从而确保解法无法被轻易猜测。我们设计了三种迷宫拓扑结构：矩形网格、由同心环与角扇区组成的圆形迷宫，以及六边形（蜂窝状）网格。为提升模型鲁棒性，我们还额外设计了一系列无解迷宫。我们首先生成一个有解迷宫并获取其解路径，随后在该路径中段附近刻意添加若干墙壁——避开距离起点或终点过近的区域。这种设计以较不明显的方式破坏了连通性，使迷宫乍看之下似乎有解，但实际上需进行完整搜索才能确认不存在有效路径。我们采用了多样化的视觉风格，包括渐变与超粗墙壁、变化的背景图案、多种标记类型以及随机小角度旋转，以防止模型过拟合于特定视觉模式。图像分辨率随机化，宽高比连续采样，网格尺寸按比例调整……（注：原文在此处截断）

难度控制。迷宫导航的难度很大程度上取决于模型需要串联多少个视觉推理步骤。我们通过改变网格大小来控制这一点。随着网格变大，模型必须解析更多单元格、在更长距离上追踪连通性，并处理更多需要回溯的死胡同。这些因素都会增加整体推理复杂度。具体而言，简单迷宫仅要求模型串联少量局部连通性检查，而噩梦级迷宫则要求持续、长距离地组合数百个此类原语操作...

原语来决定哪条分支延续目标曲线。为确保该原语得到真正检验，我们严格防止任何端点与无关线条重叠或被其穿过，并丢弃和重新生成违反这些约束的配置。我们还进一步包含一种统一样式模式，其中每条线共享

相同的颜色与线宽。

11

路径追踪

[Trigger_Placeholder]

皇冠图标连接到哪里？将目标图标名称放入 \boxed{}

基于视觉原语的推理

我找到了您提到的起点，它位于此处：

[[176,758]]...

全部随机化，以防止浅层模式匹配。思维内容合成。我们将路径追踪过程显式地表示为沿目标曲线采样的一系列坐标，这反映了模型在图像中关注并跟随路径的方式。该过程首先定位查询的起点，随后通过一系列中间路径点沿曲线行进，最终确定到达的终点。重要的是，这些路径点的密度会根据曲线的局部几何形状进行自适应调整。直线段用较少的点表示，而高曲率区域或密集交叉点则用更细粒度的坐标进行描述，这模拟了人类在视觉复杂区域会放慢速度并更加专注的机制。个冷启动样本。示例如图 6 所示。的策略，具体细节如下。数据组成。节构建的两类冷启动数据分别进行 SFT：框（结合定位进行思考）和点（结合指代进行思考）。这种分离设计在专用数据量相对较小时，能够有效避免模式冲突。经过该训练阶段后，我们获得两个专用模型，分别记为 FTwG 和 FTwP。随后，我们分别对 FTwG 和 FTwP 独立应用强化学习 (RL)。我们采用组相对策略优化 (GRPO) 算法，并沿用其超参数设置。鉴于我们冷启动数据中的思维内容所包含的视觉基元（如框和点）已经过严格验证，在 RL 阶段，我们不对模型思考过程中生成的视觉基元进行显式监督。该设计提升了 RL 训练数据的可扩展性。问题和最终答案，这显著拓宽了可获取数据的范围。思考内容生成。中间点。更准确。形状。更好。边界框与点视觉基元。更学术。

针对每个任务，我们从三个角度进行约束：格式约束、质量约束和准确性约束。前两种约束在不同任务间共享，而最后的准确性约束则需要针对任务类型进行专门设计。格式 RM。该 RM 基于规则评估输出，生成分值为 0 到 1 的奖励。具体而言，它验证模型生成的视觉原语表示格式是否正确。对于带定位的推理，该 RM 还会额外检查模型输出中的冗余，例如生成重复...

分别为缩放系数和衰减率。在实践中，我们设定 $\alpha = 0.7$ 且 $\beta = 3$ ，这些参数经经验选取，旨在提供稳定且平滑的学习信号。

空间推理与通用 VQA 的准确性 RM。针对这些任务，我们设计了一个基于 LLM 的 GRM。我们将模型的思维内容、最终回复、用户查询和真实答案输入 GRM，独立评估并打分思维过程与回复。最终奖励为两项分数的平均值。

迷宫导航的准确性 RM。为鼓励模型探索迷宫，我们设计了一个基于规则的...

这种分解确保奖励信号稠密且信息丰富：模型因每个正确应用的视觉原语而获得得分，而不仅因最终二值答案。路径追踪的准确性 RM。为强制模型沿线条追踪，我们提出一个基于规则的 RM 来评判生成的...

冷启动模型 (FTwG 或 FTwP) 在数据池上执行 rollout，为每个样本生成 条 rollout。随后，基于 RM 分数，我们统计每个样本在 条 rollout 中正确回复的数量，并将数据池划分为三个难度级别：简单级：全部 条 rollout 均正确。正常级：正确 rollout 数量 满足 $1 \leq$

库勒-莱布勒 (KL) 散度损失，其中表示学生模型。在我们的 OPD 实现中，我们采用全词表 logit 蒸馏。实践中，我们使用了两个教师模型，包括 ETwG 和 ETwP。

3. Experiments

进行训练与评估，该框架是一个构建于 PyTorch 之上的轻量级高效分布式训练框架。在预训练阶段，我们采用 64K 的序列长度和 FP8 精度；在后训练阶段，序列长度扩展至 256K。为最大化领域专家模型的性能，我们在专项监督微调 (Specialized SFT) 和专项强化学习 (Specialized RL) 阶段采用 FP8 精度，随后在统一强化微调 (Unified RFT) 和 OPD 阶段应用 FP4 (MXFP4) 量化。我们的评估框架整合了广泛使用的公开基准与精心构建的内部测试集。能力。为弥补这一不足，我们的内部测试集引入了更多样化且更具挑战性的评估维度，作为对公开数据集的重要补充。公开基准。使用其官方测试集划分。等基准。内部基准。多跳空间推理以及拓扑推理。与前沿模型的对比。为确保公平对比，我们使用相同的提示词 (prompts) 通过各模型对应的 API 对所有模型进行了评估。最佳结果以粗体显示，次佳结果以下划线标注。

细粒度计数 (Fine-grained Counting)：现有的细粒度计数基准 (如 TallyQA [1]) 通常存在标注错误和歧义问题，难以严格评估模型的细粒度计数能力。为此，我们引入了 DS_Finegrained_Counting 评估集。具体而言，我们提示多模态大语言模型 (MLLM) 生成受特定属性或空间位置约束的计数查询，并刻意确保包含困难负样本 (即与查询目标类别相同但属性不同的物体)。经过严格的人工验证以确保数据质量后，我们最终保留了 600 个高质量测试用例。多跳空间推理 (Multi-hop Spatial Reasoning)：我们从 CLEVR [13] 的验证集中抽取了 1,000 道判断题和 1,000 道开放式问题。为实现标准化和自动化评估，我们利用 MLLM 为开放式查询生成合理的干扰选项，从而将其转

换为多选题格式。该重组后的评估套件记为 DS_Spatial_Reasoning。拓扑推理 (Topological Reasoning)：遵循第 2.4.3 节和第 2.4.4 节的方法，我们构建了两个不同的评估集：DS_Maze_Navigation 和 DS_Path_Tracing，每个集合均包含 2,000 个样本。为确保公平比较，我们在所有模型上采用统一的评估协议。鉴于部分早期公开基准包含低分辨率图像，我们应用了预处理步骤以保证数据质量。具体而言，对于总像素数低于 640,000 的图像，我们在严格保持原始宽高比的前提下进行上采样，使其达到该像素阈值。对于支持可配置推理或思考预算的前沿模型 (例如 GPT 和 Gemini-3Flash)，我们在所有评估中统一将思考预算设置为“低” (low)，以确保比较的公平性和一致性。对于其他所有基准，我们遵循官方的评估协议和指标。结果如表 1 所示。得益于利用视觉基元进行思考的能力，我们的模型在这些任务上取得了具有竞争力的性能，并展现出卓越的 Token 效率。值得注意的是，所有前沿模型在拓扑推理任务上的表现均不尽如人意，这表明该领域仍有巨大的改进空间…… (原文在此处截断)

Translation: 多模态大语言模型。
Translation: 如图7至图9所示，我们的模型通过“结合定位的思考” (thinking with grounding) 在粗粒度和细粒度计数任务上表现出强劲的性能，同时也展现出涌现的能力协同效应。
Translation: 例如，该模型能够整合世界知识以进行视觉问答，执行反事实推理，并根据用户的日常需求提供带有空间坐标的可操作建议。
Translation: 尽管我们关于视觉基元的后训练数据中未包含任何中文语料，但得益于从基座模型继承的多语言能力，该模型仍能够以中文进行思考和响应。
Translation: 如图10所示，我们的模型通过“结合指点的思考” (thinking with pointing) 展现出拓扑推理能力，能够为迷宫生成逐步探索轨迹，并为路径追踪生成顺序跟踪轨迹。
Translation: 在领域内实例上，模型具备识别并跟随路径的能力，这得益于在专门的强化学习过程中通过缓解冷启动数据问题并设置奖励机制所实现的强化。Better: “在领域内实例上，模型能够识别并沿路径行进，这一能力是通过在专门的强化学习过程中缓解冷启动数据问题并引入奖励机制得以强化的。” Or: “在领域内样本上，模型具备识别并跟随路径的能力，该能力通过在专门的强化学习过程中缓解冷启动数据问题并施加奖励机制而得到强化。” Let's adjust to be more precise academically: “在领域内实例上，模型具备识别并跟随路径的能力，这一能力通过在专门的强化学习过程中缓解冷启动数据问题并引入奖励机制而得到强化。” 多模态大语言模型。

4. Limitations

尽管取得了上述令人鼓舞的成果，本研究仍存在一定局限性。首先，受限于输入分辨率，模型在细粒度场景下的表现仍有待提升，导致视觉基元的输出偶尔不够精确。未来可通过将本框架与针对“感知鸿沟” (Percepti

on Gap) ”的现有方法相融合，以发挥互补优势，从而缓解该问题。其次，当前“基于视觉基元进行推理”的能力仍需依赖显式触发词进行激活。未来，我们致力于使模型能够根据具体上下文，自主判断是否调用该机制。第三，以点作为视觉基元来解决复杂的拓扑推理问题仍是一项严峻挑战，且当前模型在跨场景泛化方面表现有限。探索提升该技术适用性与鲁棒性的途径，将是未来研究的重要方向。 尽管取得了上述令人鼓舞的成果，本研究仍存在一定局限性。首先，受限于输入分辨率，模型在细粒度场景下的表现尚不理想，导致视觉基元的输出偶尔不够精确。该问题有望通过将本框架与针对“感知鸿沟 (Perception Gap) ”的现有方法相结合，以发挥互补优势来加以解决。其次，当前“基于视觉基元的思考”能力仍需依赖显式触发词进行激活。未来，我们致力于使模型能够根据具体上下文，自主判断是否调用该机制。第三，以点作为视觉基元来解决复杂的拓扑推理问题仍是一项严峻挑战，且当前模型的跨场景泛化能力有限。探索提升该技术适用性与鲁棒性的途径，将是未来研究的重要方向。

5. Conclusion

为应对多模态大语言模型 (MLLM) 在复杂推理过程中固有的“指代鸿沟” (Reference Gap)，我们提出了“基于视觉原语的推理” (Thinking with Visual Primitives)，一种新型推理框架。超越单纯依赖提高感知分辨率的常规做法，我们将空间标记——如点与边界框——提升为“思维的最小单元”，并直接交织进模型的思维过程。该机制赋予模型“边推理边指向”的能力，将抽象语言概念精确锚定到物理图像坐标。此外，l...

出现在第一行 (第二列和第四列)、第二行 (第一列和第三列) 以及第三行 (第二列和第四列)。它们大多呈浅黄褐色或棕白色，有着大大的黑眼睛和大耳朵。
3. **确认最终计数** 观察已识别的犬只，我总共数出 6 只吉娃娃。 回答 图片中有 6 只吉娃娃。含视觉原语的图像 图 7 | 带定位推理的展示案例。示例包括细粒度计数与反常识视觉问答。建议在屏幕上放大查看。 19 [Trigger_Pla...

结合定位 (grounding) 的思考展示。示例包括基于世界知识的问答以及寻求可操作的建议。建议在屏幕上放大查看效果最佳。这为什么好笑？为何令人发笑？这是一张左右对比图。左侧展示了一块水果 (可能是梨或苹果)，上面有类似人脸的凹陷和深色斑痕。右侧则是一只猫脸的特写。看起来像是有人将水果雕刻成了与猫脸一模一样的样子。我正在寻找视觉上的相似之处。在其浅色的果肉中，有一些天然的深色斑点，看起来很像眼睛。完美地模拟了瞳孔。

站在中间微微前倾以便被看到的人，以及坐在最前排椅子上的人。统计我刚刚在整个群体中定位到的不同个体，共计29人。图像中共有29人。结合定位的推理

示例。密室逃脱游戏指引以及计数。建议在屏幕上放大查看。是否可以从绿色菱形出发，穿过迷宫到达红色标记处？如果可以，请说明路径。好的，让我逐步分析这个迷宫的结构。向外方向)。逆时针方向也可通行，但从迷宫布局来看，向外方向的路径看起来更宽。逆时针)。虽然逆时针方向也可通行，但在观察路径结构后，我判断顺时针方向更有可能通向出口。向外)。让我先尝试向外方向。

Translation: 通道特征更符合主路径的特征。/ 该通道的特征更类似于主路径。*Refinement:* 通道特征更类似于主路径。*Translation:* **步骤8** : 到达[[532,720]]时，我面临2个分岔口 (顺时针方向、向外方向)。向外方向也可通行，但根据迷宫布局，顺时针方向的路径看起来更宽。*Refinement:* **步骤8** : 到达[[532,720]]时，前方出现2个分岔口 (顺时针方向、向外方向)。向外方向同样可通行，但结合迷宫布局判断，顺时针方向的路径显得更宽阔。*Translation:* **步骤9** : 在[[471,722]]处，我注意到有2个方向可供选择 (顺时针方向、向内方向)。虽然向内方向也可通行，但在观察路径结构后，我判断顺时针方向更可能通向出口。*Refinement:* **步骤9** : 位于[[471,722]]时，发现前方有2个方向可选 (顺时针方向、向内方向)。尽管向内方向同样可走，但经观察路径结构后，我判断顺时针方向更有可能通向出口。*Translation:* **步骤10** : 到达[[415,683]]时，我面临2个分岔口 (顺时针方向、向外方向)。虽然向外方向也可通行，但在观察路径结构后，我判断顺时针方向更可能通向出口。*Refinement:* **步骤10** : 到达[[415,683]]时，前方出现2个分岔口 (顺时针方向、向外方向)。尽管向外方向同样可通行，但经观察路径结构后，我判断顺时针方向更有可能通向出口。*Translation:* **步骤11** : 到达[[371,637]]时，我面临2个分岔口 (向内方向、向外方向)。与向外方向相比，向内方向的通道特征更类似于主路径。*Refinement:* **步骤11** : 到达[[371,637]]时，前方出现2个分岔口 (向内方向、向外方向)。

与向外方向相比，向内方向的通道特征更符合主路径的特征。*Translation:* **步骤12** : 在[[378,567]]处，我注意到有3个方向可供选择 (顺时针方向、向内方向、向外方向)。向内和向外方向也可通行，但根据迷宫布局，顺时针方向的路径看起来更宽。*Refinement:* **步骤12** : 位于[[378,567]]时，发现前方有3个方向可选 (顺时针方向、向内方向、向外方向)。向内与向外方向同样可通行，但结合迷宫布局判断，顺时针方向的路径显得更宽阔。*Translation:* **步骤13** : 在[[396,467]]处，我注意到有2个方向可供选择 (向外方向、向外方向)。排除死路后，向外方向是最佳选择。*Note:* "(outward, outward)" seems like a typo in the original, but I'll translate it literally as "(向外方向、向外方向)" or adjust slightly for clarity if

needed. I'll stick to literal:
(向外方向、向外方向)。*Refinement:* **步骤13*
* : 位于[[396,467]]时,发现前方有2个方向可选(向
外方向、向外方向)。排除死路后,向外方向为最佳选
择。*Translation:* **步骤14** : 到达[[368,397]]时
,我面临2个分岔口(顺时针方向、向外方向)。向外
方向也可通行,但根据迷宫布局,顺时针方向的路径看
起来更宽。*Refinement:* **步骤14** : 到达[[368,3
97]]时,前方出现2个分岔口(顺时针方向、向外方向
)。向外方向同样可通行,但结合迷宫布局判断,顺时
针方向的路径显得更宽阔。

Translation: **步骤15** : 在[[330,355]]处,
我注意到有2个方向可供选择(向外方向、向外方向)
。排除死路后,顺时针方向是最佳选择。*Note:*
Again, "(outward, outward)" but conclusion says
"clockwise". This is likely a typo in the source, but I
will translate faithfully:
(向外方向、向外方向)。排除死路

C连接到哪个数字?请展示路径并输出

\boxed{终点数字}。

根据视觉追踪结果,终点为5。[\boxed{5}] 图10 | 基于
指向/指代的思考示例。示例包括迷宫导航和路径追踪
。建议在屏幕上放大查看效果最佳。 [1] M. Acharya,
K. Kafle, and C. Kanan. Tallyqa : 回答复杂计数问题。
收录于《人工智能协会 (AAAI) 会议论文集》,第33
卷,第8076–8084页,2019年。 [4] M. Deitke, et al.
Molmo与Pixmo : 面向最先进视觉语言模型的开放权重
与开放数据。收录于《计算机视觉与模式识别会议 (CV
PR) 论文集》,第91–104页,2025年。 [5] M. Du, et
al. EmbSpatial-Bench : 使用大型视觉语言模型对具身
任务的空间理解进行基准测试。收录于《计算语言学协
会 (ACL) 第62届年会论文集 (第2卷: 短文) 》,第
346–355页,2024年。

[9] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu. 基于空
间正则化区域提议网络的无人机目标计数。见《IEEE
国际计算机视觉会议论文集》,第 4145–4153
页,2017年。

[10] D. A. Hudson and C. D. Manning. GQA : 面向
真实世界视觉推理与组合式问答的新数据集。见《IEEE
/CVF 计算机视觉与模式识别会议论文集》,第
6700–6709页,2019年。

[11] M. Jia, Z. Qi, S. Zhang, W. Zhang, X. Yu, J.
He, H. Wang, and L. Yi. OmniSpatial : 面向全面的视觉空间推理基准...