



Seed1.5-VL 技术报告

ByteDance Seed

See Contributions and Acknowledgments section for a full author list.

Abstract

我们提出 Seed1.5-VL，这是一个视觉语言基础模型，旨在推动通用多模态理解和推理。Seed1.5-VL 由一个 532M 参数的视觉编码器和一个 20B 活动参数的专家混合（MoE）大语言模型组成。尽管其架构相对紧凑，但它在广泛的公共 VLM 基准和内部评估套件中表现出色，在 60 个公共基准中取得了 38 个的最先进性能。此外，在以智能体为中心的任务中，如 GUI 控制和游戏玩法，Seed1.5-VL 超过了领先的多模态系统，包括 OpenAI CUA 和 Claude 3.7。除了视觉和视频理解之外，它还展示了强大的推理能力，使其在视觉谜题等多模态推理挑战中特别有效。我们认为这些能力将推动更广泛的应用于各种任务。在本报告中，我们主要提供了我们在模型设计、数据构建和各个阶段训练方面的经验的全面回顾，希望本报告能激发进一步的研究。Seed1.5-VL 现已在 Volcano Engine^a 上可用。

Date: 2025 年 5 月 13 日

Correspondence: shiguang.sg@bytedance.com

^a模型 ID: doubao-1-5-thinking-vision-pro-250428

Contents

1 引言	3
2 架构	4
2.1 视觉编码器	4
2.1.1 架构	5
2.1.2 ViT 预训练阶段	5
2.2 视频编码	6
3 预训练	7
3.1 预训练数据	7
3.1.1 通用图像-文本对和知识数据	7
3.1.2 光学字符识别 (OCR)	8
3.1.3 视觉定位与计数	9
3.1.4 三维空间理解	10
3.1.5 视频	10
3.1.6 科学、技术、工程和数学 (STEM)	11
3.1.7 图形用户界面 (GUI)	11
3.2 训练食谱	11
3.3 缩放律	12
4 后训练	14
4.1 监督微调	14
4.1.1 SFT 数据构建	15
4.1.2 训练食谱	15
4.2 来自人类反馈的强化学习	15
4.2.1 偏好数据	15
4.2.2 视觉语言模型作为奖励模型	16
4.2.3 强化学习的数据监管	16
4.3 具有可验证奖励的强化学习	17
4.3.1 视觉 STEM	17
4.3.2 视觉感知与推理	17
4.4 混合强化学习	18
4.5 通过拒绝采样微调的迭代更新	19
5 训练基础设施	20
5.1 大规模预训练	20
5.1.1 混合并行性	20
5.1.2 工作量平衡	20
5.1.3 并行感知数据加载	20

5.1.4 容错能力	21
5.2 后训练框架	21
6 评估	21
6.1 公共基准	21
6.1.1 视觉编码器作为 zero-shot 分类器	21
6.1.2 视觉任务评估	21
6.1.3 视频任务评估	24
6.2 多模态智能体	26
6.3 内部基准	27
6.3.1 动机与设计原则	28
6.3.2 与最先进的技术对比	29
6.3.3 分布外泛化	29
6.4 局限性	31
7 结论与下一步计划	31
8 贡献与致谢	47
A 定性示例	50
A.1 推理案例：视觉推理	51
A.2 推理案例：地理定位预测	52
A.3 视觉推理：解决谜语谜题	53
A.4 视觉推理：表情符号测验	54
A.5 视觉推理：文字游戏 I	55
A.6 视觉推理：词语游戏 II	56
A.7 视觉推理：视觉模式识别	57
A.8 视觉谜题：找不同	58
A.9 几何学	59
A.10 在复杂场景中计数	60
A.11 空间理解：深度排序	61
A.12 视频时间定位	61
A.13 OCR 解析与文档理解	62
A.14 多语言 OCR 解析	63
A.15 生成新格式图的代码	64
A.16 图像引导的创意写作	65
A.17 失败案例：三维空间想象	66
A.18 失败案例：幻觉（知识先验）	67
A.19 失败案例：组合搜索 I	68
A.20 失败案例：组合搜索 II	69

B 评估细节	70
B.1 内部基准结构	70
B.2 内部基准的综合比较	72
B.3 能力与基准任务	73
B.4 评估提示	74

1 引言

视觉-语言模型（VLMs）已成为一种基础范式，使通用人工智能能够在开放式的虚拟和物理环境中感知、推理和行动。通过在统一模型中对齐视觉和文本模态，VLMs 在多模态推理 [96, 129, 141]、图像编辑 [35, 97]、GUI 智能体 [5, 98, 105]、自动驾驶 [103, 131, 157] 和机器人技术 [31, 55, 63] 等领域迅速推进了研究前沿，同时也在教育、医疗、聊天机器人和可穿戴设备等现实世界应用中得到了应用。

然而，尽管取得了显著进展，当前的视觉语言模型（VLM）仍然无法达到人类水平的通用性，特别是在需要 3D 空间理解、物体计数、想象性视觉推理和交互式游戏的任务中。这些限制凸显了 VLM 开发中的固有挑战。与大型语言模型（LLM）不同，后者受益于丰富且高质量的文本语料库，这些语料库涵盖了广泛的人类知识，而 VLM 缺乏同样丰富多样的视觉-语言标注，尤其是涉及低层次感知现象的概念。此外，多模态数据的异质性在训练和推理中引入了额外的复杂性，使数据流水线设计、并行训练策略和评估协议变得更加复杂。

在本报告中，我们分享了在开发 Seed1.5-VL 过程中所做的努力，Seed1.5-VL 是我们最新的一款用于视觉-语言理解的多模态基础模型。为了解决高质量标注数据稀缺的问题，我们开发了一套针对关键能力的多样化数据合成流水线，包括预训练期间的光学字符识别（OCR）、视觉定位、计数、视频理解以及长尾知识，以及后训练期间的视觉谜题和游戏。Seed1.5-VL 在跨越多种模态（即图像、视频、文本和人机交互数据）的数万亿多模态词元上进行预训练，以获得广泛的视觉知识并掌握核心视觉能力。我们还分享了预训练阶段的缩放行为。在后训练阶段，我们结合人类反馈和可验证的奖励信号，进一步增强其通用推理能力。

我们还解决了以不对称架构训练大规模多模态模型的挑战，特别是视觉编码器和语言模型之间的不平衡问题。我们的贡献包括（1）一种为这种不对称性优化的新型混合并行方案，以及（2）一种视觉词元重新分配策略以平衡 GPU 工作负载。此外，我们实现了一个定制的数据加载器，在 3D 并行下最小化 I/O 瓶颈。这些创新结合标准的系统级优化（e.g.，内核融合，选择性活性值检查点，卸载），共同提高了整体训练吞吐量。

为了全面了解当前视觉语言模型（VLM）能力的现状，从而为未来模型改进的研究方向提供参考，我们在一系列广泛的公共和内部基准测试中对豆包进行评估，这些基准测试涵盖了视觉推理、基础、计数、视频理解和计算机使用等广泛的任务。具体来说，我们报告了在 60 个公共基准测试中的结果，其中豆包在 38 个基准测试中取得了领先的性能，包括 34 个视觉语言基准测试中的 21 个，19 个视频基准测试中的 14 个，以及 7 个图形用户界面（GUI）智能体任务中的 3 个。除了基准测试性能外，我们还将豆包部署在内部聊天机器人系统中，以监测其在动态、交互式环境中的真实世界和分布外（OOD）性能。

尽管其强大的能力，Seed1.5-VL 保持了紧凑高效的架构，具有 532 亿参数的视觉编码器和 200 亿活跃参数的语言模型。这种简化的设计降低了推理成本和计算需求，使该模型非常适合交互式应用。Seed1.5-VL 的效率通过 API 服务提高了更广泛用户群体的可访问性，并有助于在 Doubao 聊天机器人中提供更流畅的用户体验。Seed1.5-VL 的访问权限即将在火山引擎 API 平台上线¹。

本报告的其余部分组织如下。我们首先介绍模型架构并详细说明图像和视频编码方法（section 2）。Section 3 描述了数据监管策略和预训练过程，包括多模态模型缩放定律和指标预测的初步发现——这是一

¹<https://www.volcengine.com>

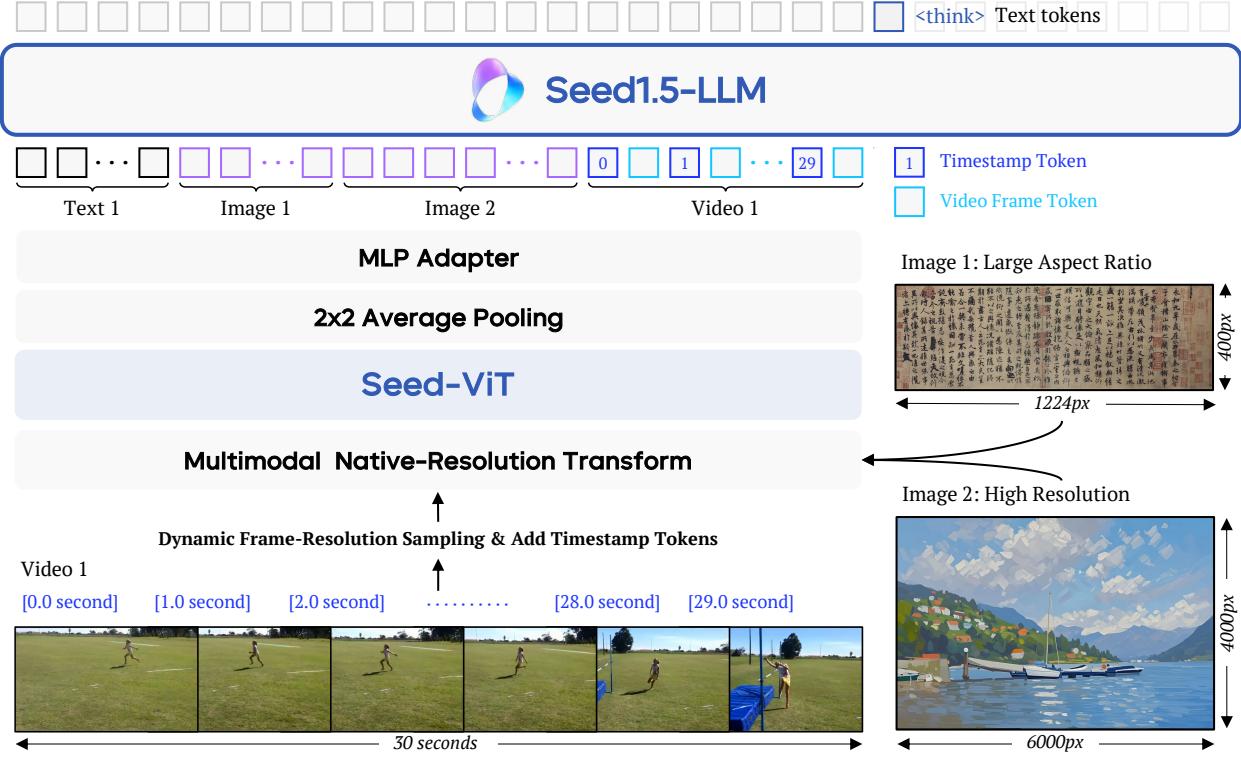


Figure 1 Seed1.5-VL 的架构. 所提出的 Seed1.5-VL 包含三个主要组件：(1) SeedViT 用于编码图像和视频，(2) 一个 MLP 适配器将视觉特征投影到多模态词元，以及 (3) 一个大型语言模型来处理多模态输入。Seed1.5-VL 接受不同分辨率的图像，并使用原生分辨率变换来保留最大图像细节。对于视频输入，我们提出了动态帧分辨率采样策略，该策略动态调整采样帧率和分辨率。此外，在每一帧前添加一个时间截词元以增强模型的时间感知能力。

个相对未被探索的领域。Section 4 详细介绍了后训练阶段使用的数据和技巧，以增强与人类偏好的对齐并提高推理能力。Section 5 详述了为实现可扩展的预训练和后训练而开发的必要基础设施创新。最后，section 6 在公开基准上展示了全面的评估结果，通过定性示例展示模型能力，讨论当前多模态模型的局限性，并提出未来研究的方向。

2 架构

Seed1.5-VL 的架构包含三个组件：视觉编码器、MLP 适配器和大型语言模型（LLM）。视觉编码器原生支持动态图像分辨率，并采用 2D RoPE [126] 进行位置编码，从而能够灵活适应任意维度的图像。为了提高计算效率，该架构在相邻的 2×2 特征块上应用平均池化；随后，两层 MLP 处理这些池化后的特征，然后再输入到 LLM 中。不考虑无编码器的架构 [1, 23, 127]，因为视觉编码器提供了高效的图像压缩，能够在较少的词元下实现高分辨率图像表示。整体架构如 figure 1 所示。

2.1 视觉编码器

许多当代视觉-语言模型（VLMs）[2, 5, 7, 16, 37, 54, 71, 78, 104, 128, 141] 通常整合为固定输入分辨率设计的预训练视觉编码器，通常是方阵图像。虽然这种方法简化了模型架构，但在处理高分辨率图像、视频或需要精细细节的任务（如 OCR）时，可能会无意中丢弃细粒度的视觉信息。最近的努力，例如 Qwen2-VL [141] 和 InternVL-2.5 [16]，已探索对预训练视觉编码器进行微调以适应动态分辨率输入，从

而部分缓解了这一限制。然而，这些方法仍然主要依赖于适应现有固定分辨率架构，并需要调整位置嵌入（e.g.，从 1D 展平位置嵌入过渡到 2D RoPE [16, 141] 或将 1D 位置嵌入插值到各种形状 [99, 135]），这可能无法在适应后完全保留视觉细节和准确率。此外，我们在预训练阶段引入视频数据，使模型不仅能够学习图像中的空间特征，还能学习时空动态，从而增强其处理动态场景和复杂视觉内容的能力。

针对固定分辨率处理带来的挑战，我们开发了 Seed-ViT，这是一种专为原生分辨率特征抽取设计的视觉编码器。基于广为人知的视觉 Transformer (ViT) 架构 [26]，Seed-ViT 包含 5.32 亿个参数。它在跨多个领域的通用视觉感知方面表现出强大的能力。值得注意的是，在 zero-shot 分类基准上，Seed-ViT 的表现可与参数量显著更多的模型（如 InternVL-C (60 亿参数)）相媲美，突显了其效率。进一步的架构细节和我们的 Seed-ViT 预训练方法分别在 sections 2.1.1 和 2.1.2 中提供。

2.1.1 架构

Seed-ViT 的架构参数可以在 table 1 中找到。

Patch size	Pos embed	Head dim	Num heads	Embed dim	MLP ratio	Depth
14	2D RoPE	64	20	1280	4.0	27

Table 1 Seed-ViT 的架构超参数。

我们的视觉编码器设计用于适应不同维度的输入图像。首先，输入图像经过一个涉及双线性插值的预处理步骤，以将其分辨率调整为 28×28 像素的最近倍数。随后，每张图像被分割成一系列非重叠的块，每个块为 14×14 像素。按照 NaViT [20] 中概述的方法，我们将多个输入图像的块序列连结成一个统一的序列。这些原始的块序列然后通过一个线性块嵌入层投影到嵌入空间中的词元，然后输入到 Transformer 块中。为了确保属于一张图像的词元不会在批处理序列中关注其他图像的词元，我们在 Transformer 块内的自注意力计算中使用适当的注意力掩码。最后，在将输出块嵌入传递到后续的 MLP 适配器和 LLM 之前，应用一个 2×2 平均池化操作，如上所述。

2.1.2 ViT 预训练阶段

Categories	Unlabeled image	Image-text pairs	Video-audio-text tuples
Training samples	2.2B	4.8B	65M
Token percentages	4.0%	91.2%	4.8%
Batch sizes	55,296	32,768	1,024
LR warm up steps	1,692	2,000	12,800
Maximum LR	7.06×10^{-3}	1.0×10^{-4}	5.0×10^{-5}
Minimum LR	1.05×10^{-5}	1.2×10^{-6}	2.02×10^{-7}

Table 2 三个 ViT 预训练阶段使用的训练设置和超参数。

我们的视觉 Transformer，Seed-ViT，在与 LLM 集成之前会经过专门的预训练流水线。基于实证证据，我们建立了三条预训练方法的关键指导原则：

- **通过 ViT 预训练提高训练效率。** 大多数成功的 VLMs [7, 16, 141] 遵循拥有视觉编码器 (e.g., CLIP 或 SigLIP [171]) 的设置, 而少数工作 [1, 24] 尝试完全去除视觉编码器, 并直接在仅解码器的 LLMs 中传递图像块, 但结果参差不齐。Beyer et al. [10] 同样得出结论, 无编码器的 VLM 可能是一个有前景的未来方向, 但在训练效率方面仍存在问题。
- **原生分辨率建模的早期整合。** 我们在预训练流水线中优先引入原生分辨率建模。Seed-ViT 的架构在 ViT 预训练和 VLM 阶段始终保持一致。这确保了防止由于架构修改而导致的性能下降, 并消除了需要大量微调来补偿这些差异的必要性。
- **全面的数据利用。** 预训练阶段利用了用于 VLM 训练的全部数据, 包括未标记的图像、图像-文本对以及带有视觉和音频字幕的视频。

根据上述指南, ViT 预训练流水线分为三个阶段: (i) 掩码图像建模 (MIM) [145] 与 2D RoPE, (ii) 原始分辨率对比学习, 以及 (iii) 全模态预训练。以下, 我们提供每个阶段的更多细节。

带有 2D RoPE 的 MIM. 在第一阶段, 我们的目标是通过 MIM 增强视觉感知能力以及对视觉几何和结构意识的理解。我们采用 EVA02-CLIP-E [29] 作为教师模型, 学生模型则按照 table 1 中定义的架构进行随机初始化。在训练过程中, 我们随机遮蔽 75% 的图像块和相应的 RoPE 嵌入, 并使用教师模型生成的 CLIP [107] 特征作为重构目标。此过程通过学生和教师输出中被遮蔽块之间的简单余弦相似度损失进行优化。我们发现学生和教师模型之间视觉位置嵌入的差异不会影响性能, 因为教师采用可学习的位置嵌入, 而学生使用 2D RoPE。相反, 2D RoPE 使学生具备了强大的原生动态分辨率识别能力。随着这种 MIM 过程的扩展, VLM 在图表/文档理解和 OCR 方面的能力得到了显著提升。

原生分辨率对比学习. 在对比学习阶段, 视觉编码器使用我们 MIM 训练的学生模型进行初始化, 而文本编码器则使用 EVA-02-CLIP-E 中的文本编码器进行初始化。对于每个给定的图像-文本对, 我们通过注意力池化将视觉编码器提取的块特征聚合为一个 1280 维的图像嵌入。然后通过联合优化 SigLIP 损失 [171] 和 SuperClass 损失 [52] 来实现图像和文本嵌入之间的对齐。

全模态预训练. 此阶段采用 MiCo 框架 [174], 从视频数据中构建包含视频帧、音频、视觉字幕和音频字幕的对齐元组。ViT 对视频帧和音频进行编码, 而分离的文本编码器处理字幕。通过这些嵌入的对齐, ViT 学习统一的全模态表示。尽管仅消耗了分配给整个 ViT 预训练过程的 4.8% 的词元预算, 但此阶段显著提升了 ViT 在图像和视频理解任务上的性能。

Table 2 总结了每个阶段使用的训练设置和超参数。

2.2 视频编码

有效编码视频, 超越静态图像表示, 仍然是一个核心挑战。模型解读时间序列、适应不同帧率以及感知绝对时间的能力对于理解动态视觉内容至关重要。Seed1.5-VL 通过引入 **动态帧分辨率采样** 解决这些挑战, 这是一种新的策略, 通过在时间 (帧) 和空间 (分辨率) 维度上联合优化采样, 以平衡语义丰富性和计算效率。

在这一动态帧分辨率采样策略下, 视频被处理为图像帧的序列。时间维度通过动态帧采样进行管理。而不是统一的速率, Seed1.5-VL 根据内容复杂度和任务需求调整帧采样频率。默认采样率为每秒 1 帧 (FPS), 适用于捕捉视频内容的一般理解。对于任务 [73, 139] 需要详细的时间信息, 帧采样率提高到 2

FPS。对于视频计数 [27] 或运动跟踪 [48] 等任务，速率提高到 5 FPS。为了明确地将每一帧置于视频的时间轴中，我们在每个帧前添加时间戳词元（即 [1.5 second]）。这种明确的时间标注显著增强了模型的时间意识，并使其能够有效处理现实场景中常见的可变帧率。

考虑到处理长视频序列的计算约束，采样的空间维度通过动态调整每个选定帧分配的分辨率来控制，在每段视频中最大预算为 81,920 个词元。模型动态调整空间分辨率，通过分层分配系统为每帧分配词元，提供六个预定义级别：{640, 512, 384, 256, 160, 128}。这允许灵活的权衡，即对较少的帧使用更高分辨率或降低分辨率以容纳更长视频中的更多帧。当视频特别长，即使使用最低的词元分配（每帧 128 个词元）仍超过最大编码长度时，将触发回退机制。此时，模型通过在整个视频上进行均匀采样减少总帧数。虽然这会降低时间密度，但可以确保整个视频被表示，平衡了处理效率与重要时间信息的保留。

这种灵活的策略使 Seed1.5-VL 能够高效准确地处理不同长度和帧率的视频，同时保持对于各种视频理解任务至关重要的时间细节。

3 预训练

本节描述了 section 3.1 和 section 3.2 的数据监管过程以及 section 3.3 中我们模型的缩放行为。

3.1 预训练数据

The Seed1.5-VL 预训练语料库包含 3 万亿种多样化、高质量的源词元。这些数据根据目标能力进行分类，每个类别的整理过程在以下小节中详细说明。

3.1.1 通用图像-文本对和知识数据

网络来源的图像-文本对数据，包括替代文本、图像标题和周围文本，在规模上前所未有（数十亿个实例），在视觉和文本概念方面表现出高度的多样性。然而，这些数据本质上是噪声的（例如，不相关或不准确的文本），并且通常表现出类别不平衡。

为了缓解这些挑战，我们首先采用一系列过滤技术，包括图像-文本相似度评分（例如，CLIP 得分阈值处理），基于图像的标准（例如，去除尺寸过小的图像或具有极端宽高比的图像），基于文本的标准（例如，过滤过短或过长的文本），去重策略（例如，去除确切和近似重复的图像），以及基于 URL/域名的过滤。

此外，原始图像-文本对中视觉概念的分布遵循长尾模式。为了实证测试这一观察结果，我们使用 Biotrove [159] 进行沙盒实验，这是一个用于物种分类的大规模数据集，包含跨越 366,600 个物种的 161.9 亿张图像。我们使用三种不同的数据分布训练了我们的 VLM 的一个 11 亿活跃参数变体：

- **随机-46M**。从训练集随机选择的 4600 万样本。
- **Max1k-46M**。选择 4600 万份样本，每种最多 1000 份样本，确保稀有物种的纳入。
- **Max100-15M**。1500 万个样本，每种最多 100 个样本，从而增加对稀有物种的相对暴露程度。

我们在两个特别过滤的测试集上评估了模型，这些测试集源自原始数据集：Balanced10k（从 BioTrove-Balanced 采样，代表常见物种）和 Rare2k（从 BioTrove-Unseen 采样，代表稀有物种）。我们的实验显示在 table 3 中，Random-46M 配置在稀有物种识别上表现不佳。相反，限制每个常见物种的最大样本

数 (Max1k-46M) 显著提高了在稀有物种上的性能。进一步限制常见物种的表示 (Max100-15M) 增强了对稀有物种的记忆，但对常见物种识别产生了不利影响。因此，有效捕捉视觉知识需要保持常见视觉概念的多样化示例，同时确保稀有视觉概念有足够的训练迭代。

	Training tokens	Balanced10k	Rare2k	Average
Random-46M (1 epoch)	12B	78.92	10.46	44.69
Max1k-46M (1 epoch)	12B	79.17	44.85	62.01
Max100-15M (3 epochs)	12B	60.31	89.41	74.86

Table 3 在三种训练数据分布 (Random-46M、Max1k-46M 和 Max100-15M) 下，Balanced10k 和 Rare2k 的性能比较。评估是通过一个开放式的问答系统 (QA) 任务进行的，回答由大语言模型评委自动评分。所有模型都使用固定预算的 120 亿个词元进行训练。

为了解决从图像-替代文本对中常见和罕见视觉知识获取之间的不平衡问题，我们提出了一种目标预处理框架。最初，该框架利用我们 VLM 的前期版本自动用相关语义领域（例如，地标、食物、商品、生物学）和相关命名实体（例如，产品品牌、物种名称）标注数据。表现出低语料库频率的命名实体被识别为罕见视觉知识的实例。为了缓解数据稀疏性，我们识别出其表示不足平均领域频率 50% 的领域。随后，这些欠代表领域的替代文本会被复制。通过将这个增强子集（包含来自较少频繁领域的样本）合并回原始语料库，我们实现了视觉概念更平衡的分布。这种再平衡旨在增强视觉知识学习组件，这对我们的预训练方法至关重要。

3.1.2 光学字符识别 (OCR)

为了增强视觉语言模型 (VLM) 的光学字符识别 (OCR) 能力，特别是对于多语言文本、特殊符号以及结构复杂文档的分析，如 [figure 2](#) 所示，我们采用大量标注和合成数据来训练 Seed1.5-VL。

我们构建了一个包含超过 10 亿个样本的内部 OCR 训练数据集，涵盖文档、场景文字、表格、图表和流程图。对于文档数据，我们从各种来源收集了大量页面，并应用我们的内部工具来提取内容和版式信息。此外，我们整理了一组多样化的字体，包括艺术字体、手写体和非拉丁字母脚本，并随后利用 SynthDog [62] 和 LaTeX 等工具合成超过 2 亿张文本密集型图像（参见 [figure 2\(a\)](#) 中的示例）。为了提高模型在理解图像中文本内容方面的鲁棒性，我们对合成数据应用了多种数据增强技术，包括模糊处理、添加 moiré 模式以及图像扭曲。[Figure 2\(c\)](#) 展示了应用基于扭曲的数据增强后的文档图像示例。

我们的图表数据集结合了现有的开源数据集（例如，FigureQA [58]）和新生成的合成数据。合成图表是使用传统工具 (ECharts [70], Matplotlib [53]) 和一种新的基于大语言模型的流水线生成的。在我们的流水线中，大语言模型生成文本图表组件（标题、图例等），然后通过大语言模型将其转换为用于渲染的 LaTeX 或 Python 代码 ([figure 2\(b\)](#))。图表图像通过执行此代码获得。这种多方面的策略产生了超过 1 亿张图表的大规模数据集。

对于表格数据，我们从各种来源提取 HTML、LaTeX 和 Markdown 格式的文本，包括网页 HTML、GitHub README 文件以及 arXiv 的 LaTeX 文件。利用这些文本，我们渲染了超过 5000 万张表格图像，创建了一个全面的表格解析数据集。该数据集使我们的模型能够高效地将表格转换为 HTML、LaTeX 和 Markdown 等格式。

为了进一步增强模型对图像中文本内容的理解，我们构建了一个视觉问答 (VQA) 数据集，以补充结构

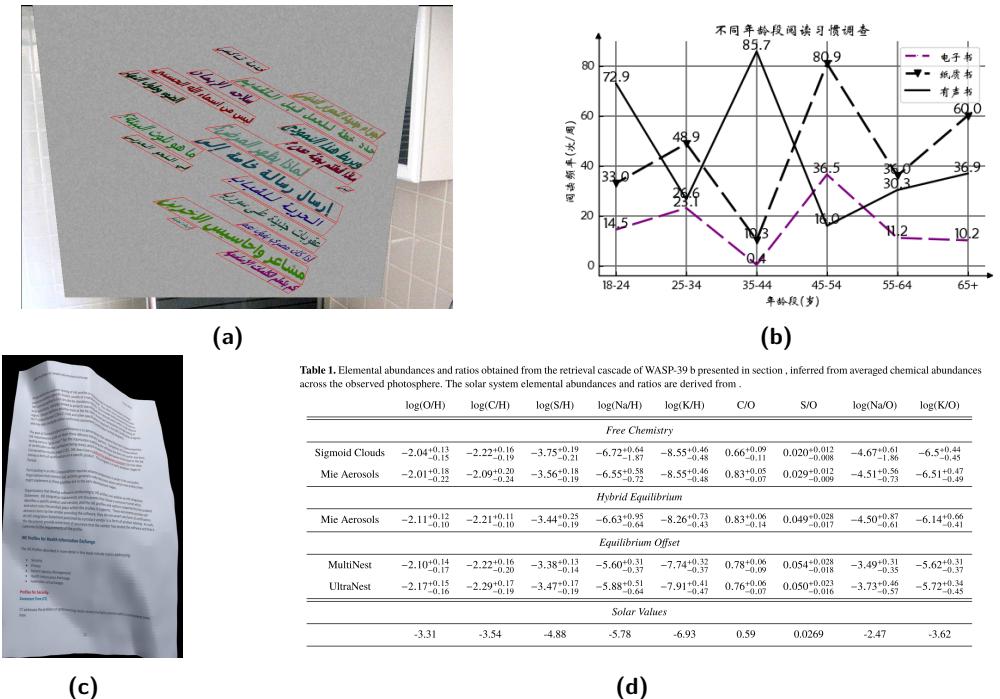


Figure 2 (a) 由 SynthDog 生成的图像和相应的文本标注以以下格式组织: `<text>...</text><polygon>...</polygon>`; (b) 合成的图表数据包括两种类型的标注: 图表到文本解析和问答对; (c) 原始文档图像经过变换以模拟现实世界的失真, 例如透视偏移、弯曲和褶皱。这些增强操作提高了模型的鲁棒性, 并增强了其在多样且具有挑战性条件下识别文本的能力; (d) 为上述合成表格图像生成的一个问答示例: 问题: *Sigmoid Clouds* 的 $\log(C/H)$ 值是多少? 给出分析步骤。答案: 我们查找标记为 “Sigmoid Clouds” 的行和标记为 “ $\log(C/H)$ ” 的列。该单元中的值是 $-2.22^{+0.16}_{-0.19}$

化的图像-文本表示。具体来说, 我们使用了我们 VLM 的先前版本, 通过条件化于 OCR 输出、图表内容、表格文本和图像本身, 采用 few-shot 提示方法生成问题-答案对。**Figure 2(d)** 给出了一个输入表格图像和相应的生成 QA 对的例子。随后, 我们应用了一个内部 LLM 来筛选生成的问题-答案对, 去除问题和答案之间语义相关性较低的实例。我们的实验表明, 包含这个 VQA 数据集显著提高了模型理解图像中文本信息的能力。

3.1.3 视觉定位与计数

目标定位是多模态模型的基本能力, 涉及将用户指令进行解释, 以识别和定位图像中的特定目标区域。在本工作中, 我们为 Seed1.5-VL 采用两种主要的定位表示: 边界框和中心点。在此定位基础之上, 我们将 Seed1.5-VL 的能力扩展至目标计数。相应地, 我们的训练策略主要利用三种数据类型: 边界框标注、点标注和计数数据。

边界框数据. 首先, 我们采用广泛使用的开源数据集进行通用物体定位, 包括 Objects365 [118], Open-Images [66], 以及 RefCOCO/+g [60, 92, 164]。而不是直接将这些数据集用于训练, 我们过滤开源数据集中的低质量样本并构建多样化的定位任务。具体来说, 我们将每个类别的所有物体边界框渲染到图像上, 并采用我们 VLM 的前一版本进行数据检查, 这使我们能够过滤掉具有错误标注、缺失标签或冗余标注的样本。此外, 我们使用这些开源数据集构建多样化的多任务训练数据, 包括: (1) 通用的 2D 定位, (2) 关于空间关系的问答系统, 以及 (3) 带有视觉提示的问答系统, 这产生了约 4800 万样本和 410 亿词元。考虑到开源定位数据集在数据领域和类别方面的多样性有限, 我们开发了一个高效的通用多物

体定位自动标注流水线，使用大规模图像-文本对。具体来说，我们遵循之前的工作 [17]，从描述中提取名词短语和实体，然后采用 Grounding DINO [14, 80] 来标注网络图像中的多样化开放词表物体。我们使用 CLIP [106] 和启发式度量标准过滤掉低质量的标注，例如非最大值抑制。自动标注流水线带来了约 2 亿样本和 2000 亿词元。

点数据. 最初，我们使用了 PixMo-Points 提供的公开数据 [21]。认识到可用 PixMo 数据在多样性和数量上的局限性，我们开发了一个专用的流水线来生成额外的指向数据。该流水线使用 Molmo [21] 和 CountGD [3] 对大量网络图像中的物体中心点进行标注。值得注意的是，CountGD 在密集图像场景中对物体的标注表现尤为有效。标注完成后，过滤掉了低质量的数据样本，最终得到一个包含约 1.7 亿条指令和 1100 亿个词元的数据集。

计数数据. 我们进一步从上述边界框和点数据中采样以构建计数数据集，包含大约 800 万个样本和 130 亿个词元。具体来说，我们开发了两种变体：基于边界框的计数和基于点的计数，遵循两阶段流水线：1) 检测或定位，然后 2) 根据边界框或点的数量生成计数结果。

在训练过程中，我们采用相对坐标并规范化所有坐标值，使得输出的边界框和点落在范围 [0, 999] 内，这使得 Seed1.5-VL 能够准确预测相应的边界框和点，而不管输入图像的分辨率如何。我们将这种规范化策略应用于所有相关数据，包括光学字符识别（OCR）和图形用户界面（GUI）。

3.1.4 三维空间理解

为了使模型具备从单张图像中理解三维空间的能力，我们构建了针对以下三个任务的数据：相对深度排序、绝对深度估计和三维定位。为了生成**相对深度排序**数据，我们采用了 DepthAnything V2 [160] 来推断从 200 万张互联网图像中采样的物体之间的深度关系。此过程产生了一个包含 32 亿个词元的数据集组件，与该任务相关。特别是，我们选择了相对深度差异超过 20% 的物体的平均深度。

用于**绝对深度估计**的数据来源于公开的数据集。对于每个由语义掩码标识的实体，我们使用相应的标注深度图确定其绝对深度。此过程产生了 1800 万条指令对（e.g., 查询/深度值），并为我们的预训练语料库贡献了 280 亿个词元。

对于**3D 定位**数据，我们使用了互联网上公开的数据集。这些数据集随后被处理并重新整理为问答（QA）对。具体来说，我们的重新整理涉及提示特定类别物体的 3D 位置。此过程生成了一个包含 77 万条指令遵循对的数据集，共计 13 亿个词元。

3.1.5 视频

这一部分数据用于提升模型对视频中多帧时间序列图像的理解。它包括三个主要类别。首先，通用视频理解数据，这一部分涵盖多种任务，包括视频字幕生成、视频问答系统、动作识别、动作定位和多图像理解。数据来源于公开数据集和内部收集的视频-字幕对。其次，我们包含多个公开可用的数据集用于视频时间定位和时刻检索，以增强模型的时间意识。具体来说，Seed1.5-VL 直接根据用户提示预测开始和结束时间戳，默认格式为秒。时间定位能力有助于视频中的复杂推理任务。最后，视频流数据对于理解动态和连续的视频内容至关重要。数据来源于各种来源，并结构化为三个主要组成部分：

- **交叉的标题/问答数据.** 首先，我们通过直接对分割的视频片段添加标题或按时间顺序构建多轮问答对来构造交叉的视频文本序列。这些标题和问答对被插入到视频中的相应时间戳处，以增强实时视频

理解。

- **主动推理数据**。第二，我们将基于视频的问答和稠密描述数据重构为逐帧响应格式。此数据要求模型持续监控视频流并主动确定生成响应的合适时间戳。
- **实时评论数据**。第三，我们利用自然的时间同步视频评论数据，以提供视频帧和文本的细粒度交叉和对齐。这种结构使模型能够根据视频流处理中断并动态更新响应。

这些数据集共同构成了有效视频训练的全面基础。

3.1.6 科学、技术、工程和数学（STEM）

为了在预训练过程中增强模型的推理能力，我们整合了跨多个 STEM 领域的问题解决数据，这些数据通过爬取和人工标注获得。这项工作最终形成了围绕两个主要组成部分的全面 STEM 数据集：**图像理解数据** 和 **问题解决数据**。

图像理解数据 包含多个子集。我们在数学、物理、化学和生物的 300 个类别中收集了 320 万份高质量的教育基础样本。此外，我们还合成 1000 万张结构化表格，格式多样，生成 450 万张化学结构图，并制作 150 万张合成坐标系图，包括函数图和位置图。一个特定的子集，K12 Caption 数据，包括 10 万个人工标注的教育图片标题，100 万个视觉问答（VQA）对，100 万个使用自动化流水线生成的机器生成标题，以及数十万张与几何相关的标题。

对于 **问题解决数据** 组件，我们通过严格的数据清洗和重新表述过程处理了超过 1 亿条 K12 级别的练习题。这还补充了数千万条精心挑选的中文成人教育问题和数百万条与图像相关的英语问题。

这些数据集的构建采用了混合采集策略，结合了人工标注、自动合成和严格的质量控制措施。这种方法确保了跨数学、物理和化学等核心 STEM 领域文本、视觉和图表表示的多模态覆盖。

3.1.7 图形用户界面（GUI）

对于 GUI 数据，我们主要包含从 UI-TARS [105, 116] 整理的数据。具体来说，为了支持鲁棒的 GUI 感知、定位和推理，我们在网页、应用和桌面环境中整理了一个大规模的数据集。每张截图都配有结构化元数据—元素类型、边界框、文本和深度—这些数据通过自动解析和人工辅助探索收集而来。对于**感知**，我们构建了包括元素描述、稠密标注和状态转移标注的任务。这些任务教会模型识别小型 UI 组件，理解整体布局，并检测跨帧的细微视觉变化。还叠加了视觉标记（Set-of-Mark）以加强空间对应关系。对于**定位**，我们训练模型根据文本描述预测元素坐标。边界框在不同分辨率下是规范化的。对于**推理**，我们收集了多步骤任务轨迹，每个轨迹都带有观察结果、中间想法和动作的注释。这种结合内部数据和标准化开源迹线的数据，使模型能够学习逐步规划、纠正和反思。

3.2 训练食谱

大型多模态模型通常从一开始就通过联合多模态学习进行训练 [54, 128]，或者在语言模型预训练后通过事后适应进行训练 [16, 141]。Seed1.5-VL 目前采用后者以便灵活的消融实验和快速迭代开发。

如 section 2 所述，我们提出的模型包含三个主要模块：视觉编码器、MLP 适配器和语言模型。在视觉语言模型预训练阶段之前，视觉编码器会进行独立的训练过程，如 section 2.1 所述。语言模型从一个内部

Stages	Stage 0	Stage 1	Stage 2
Training budget (tokens)	16B	3T	240B
Sequence length	32,768	32,768	131,072
Trainable components	MLP adaptor	all	all
Batch sizes (tokens)	8.4M	71M	71M
LR warmup steps	100	500	0
Maximum LR	2.52×10^{-4}	5.22×10^{-5}	5.22×10^{-6}
Minimum LR	4.50×10^{-5}	5.22×10^{-6}	5.22×10^{-6}

Table 4 预训练三个阶段的训练设置和超参数

预训练模型初始化，该模型具有大约 200 亿个活跃参数。该语言模型采用仅解码器的专家混合 (MoE) 架构 [119]，并已在包含数万亿高质量纯文本词元的大规模语料库上进行了训练。我们的视觉语言模型预训练方法分为三个不同的阶段，如 table 4 所述：

1. 在阶段 0 中，我们仅通过训练 MLP 适配器将视觉编码器与语言模型对齐，同时保持视觉编码器和语言模型冻结。省略此阶段会导致损失略有增加且性能变差。
2. 在第一阶段，所有模型参数都是可训练的。此阶段专注于知识积累，通过在包含 3 万亿词元的多模态语料库上进行训练，使模型掌握视觉定位和 OCR 能力，该语料库主要由描述文本、交错的图文对、视觉定位和 OCR 数据组成。实证发现，添加少量纯文本词元（例如，5%）可以保持模型的纯语言能力。此外，添加少量指令遵循数据可以得到更可靠的评估结果，这使得我们可以将预训练开发与后训练解耦。
3. 在第二阶段，我们在不同任务之间创建了一个更平衡的数据混合，并添加了来自新领域（如视频理解、编程和 3D 空间理解）的数据。此外，我们将序列长度从 32,768 增加到 131,072，这能更好地适应视频中的长距离依赖关系和复杂推理问题。与第一阶段相同，所有模型参数都是可训练的。

我们还尝试了另一种训练策略，类似于 [16, 141] 所采用的方法，在阶段-0 时同时训练 MLP 适配器和视觉编码器，而语言模型保持冻结。然而，实验评估表明，我们的训练方案能产生更好的性能。我们假设这种差异可能源于视觉编码器尝试弥补冻结的 LLM 可能存在的不足，这可能会因此损害其感知能力。

我们在所有三个阶段的训练中使用 AdamW 优化器 [64]，其中 $\beta_1 = 0.9$ 、 $\beta_2 = 0.95$ ，以及权重衰减为 0.1。偏置和规范化参数不参与权重衰减，其他训练超参数可在 table 4 中找到。阶段-0 和阶段-1 的训练遵循完整的余弦衰减学习率调度，而阶段 2 的初始学习率等于阶段 1 的结束学习率，并在整个训练过程中保持不变。在阶段 2 中，我们从阶段 1 加载优化器状态，因此不使用学习率预热。

3.3 缩放律

视觉-语言模型 (VLMs) 如 Seed1.5-VL 的预训练与大型语言模型 (LLMs) 的标准实践有本质区别，后者通常涉及所有模型参数的随机初始化。相比之下，Seed1.5-VL 建立在预训练组件之上，包括一个视觉编码器、一个 MLP 适配器和一个语言模型。本节重点在于理解 Seed1.5-VL 在预训练第一阶段的缩放行为。基于对 LLM 缩放定律的先前工作 [45, 46, 59]，平均负对数似然损失 L 被建模为模型参数 N 和训

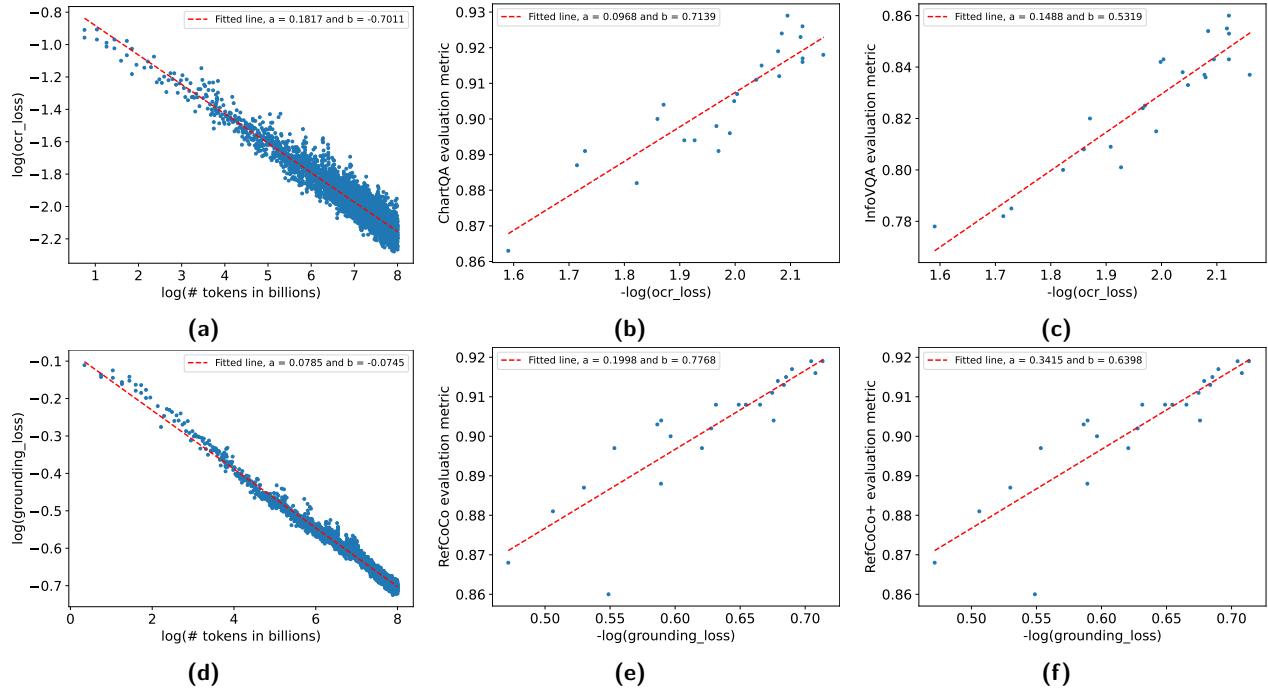


Figure 3 大多数子类别的训练损失与训练词元之间的关系遵循幂律 [46]。此外，在局部邻域内，子类别的训练损失与相应的下游评估指标之间的关系似乎呈对数线性关系 (e.g., $\text{metric} \sim \log(\text{loss})$)。(a) OCR 相关数据集的训练损失作为训练词元的函数；(b) ChartQA [88] 上的 Top-1 准确率作为训练损失的函数；(c) InfographicVQA [90] 上的 Top-1 准确率作为训练损失的函数；(d) 定位相关数据集的训练损失作为训练词元的函数；(e) RefCOCO [60, 164] 上的 Precision@IoU=0.5 作为训练损失的函数；(e) RefCOCO+ [60, 164] 上的 Precision@IoU=0.5 作为训练损失的函数。请注意，本图中显示的评估指标代表预训练后的性能，因此与通过强化学习 (RL) 获得的最终结果不直接可比，如第 6 节所述。

训练词元 D 的函数：

$$\hat{L} \sim \frac{A}{N^\alpha} + \frac{B}{D^\beta}. \quad (1)$$

鉴于我们的模型架构以及因此参数数量在此阶段是固定的，equation (1) 简化为主要依赖于训练数据的规模：

$$\hat{L} \sim \frac{B}{D^\beta}. \quad (2)$$

为了便于分析，我们通过取两边的对数，在对数-对数空间中考察这种关系：

$$\log(\hat{L}) \sim \log(B) - \beta \log(D) = -a \log(D) + b. \quad (3)$$

我们把预训练数据集组织成与特定能力相对应的不同类别（如 section 3.1 所详细描述）。我们观察到，这些数据子类别的训练损失表现出对 equation (3) 定义的缩放关系的明显遵循。如 figure 3 (a) 和 (d) 所示，OCR 和定位相关数据集的训练损失可以建模如下：

$$\begin{aligned} \log(\hat{L}_{\text{ocr}}) &\approx -0.1817 \log(D) - 0.7011 \\ \log(\hat{L}_{\text{grounding}}) &\approx -0.0785 \log(D) - 0.0745. \end{aligned}$$

除了训练损失的缩放规律之外，我们的分析表明，在特定数据子类别上实现的训练损失可以作为相关下

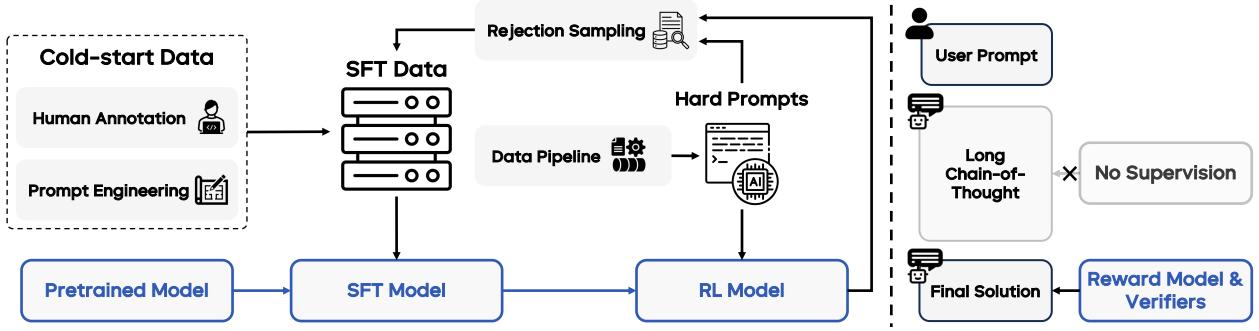


Figure 4 对 Seed1.5-VL 的后训练概述。对 Seed1.5-VL 的后训练包括结合拒绝采样和在线强化学习的迭代更新。我们构建了一个数据流水线，包括收集和整理困难提示，以增强后训练数据。我们强化学习实现的一个关键方面是，通过奖励模型和规则验证器中介的监督仅应用于最终生成的输出。我们有意不监督详细的思维链推理本身，这一区别在图示的右侧部分中突出显示。

游任务性能的预测器。我们发现，子类别训练损失与其对应的下游指标之间的关系近似为对数线性。然而，需要注意的是，这种对数线性关系可能仅在性能值的局部邻域内可持续，因为典型评估指标（如准确率、F1 得分）的值域本身是固有受限的，通常在 0 到 1 之间。如 figure 3 (b) 和 (c) 所示，ChartQA 和 InfographicVQA 数据集上的 top-1 准确率与 OCR 训练损失的对数显示出明显的相关性 ($\log(\text{loss}_{\text{OCR}})$)，这由以下近似线性模型所捕获：

$$\text{Acc}_{\text{ChartQA}} \approx -0.0968 \log(\text{loss}_{\text{ocr}}) + 0.7139$$

$$\text{Acc}_{\text{InfoVQA}} \approx -0.1488 \log(\text{loss}_{\text{ocr}}) + 0.5319$$

类似地，figure 3 (e) 和 (f) 详细描述了模型在训练期间的定位损失与其在 RefCOCO 评估基准上的性能之间的估计关系。性能预测仍然是一个活跃的研究领域，先前的工作已经使用 Sigmoid 函数来建模 LLM 性能与损失 [37, 151] 或计算 [101] 之间的关系。

4 后训练

后训练阶段通过监督微调 (SFT) 和强化学习 (RL) 的结合，赋予 Seed1.5-VL 强大的指令遵循和推理能力。如 figure 4 所示，这一过程始于在精心策划的冷启动数据上训练的 SFT 模型。一个关键组成部分是我们的数据流流水线，它持续收集困难且多样化的提示，这些提示被用于强化学习，并通过拒绝采样来改进 SFT 数据。后训练过程是迭代进行的：SFT 模型通过提炼强化学习模型在多样化提示上的学习成果而逐步得到增强。这种迭代优化会一直持续，直到提示池耗尽且性能指标收敛。最终，这一过程产生了 Seed1.5-VL，它能够生成快速、简洁的回答，以及包含长思维链 (LongCoT) 推理的深入回应 [56]。我们将在以下小节中讨论每个组件的细节。

4.1 监督微调

监督微调 (SFT) 阶段对于在强化学习之前赋予 Seed1.5-VL 基础的指令遵循和推理能力至关重要。我们的 SFT 数据集包含两个主要组成部分，分别针对不同的能力。第一部分是通用指令数据，用于训练 Seed1.5-VL 处理多样化的复杂指令，强调生成简洁准确的回答。第二部分是长思维链 (LongCoT) 数据，专注于生成详细、分步骤的推理过程。此数据通过提示工程和拒绝采样 (受 [134] 启发) 生成，主

要使用 Seed1.5-VL 的高质量输出；具体细节见 section 4.5。此外，每种数据类型都关联着一个不同的系统提示，允许用户在推理过程中动态切换 LongCoT 推理。SFT 数据集的构建方法以及 Seed1.5-VL 的 SFT 训练方案的具体内容分别在 sections 4.1.1 和 4.1.2 中进一步详述。

4.1.1 SFT 数据构建

在 SFT 数据构建的初始阶段，我们旨在赋予模型处理广泛应用场景的能力。为此，我们开发了一个模型能力分类体系，该体系基于传统视觉任务的分类以及视觉-语言模型的经验应用需求。在该分类体系的指导下，我们通过众包方式从互联网上收集图像，并生成大约 13000 条高质量的指令微调数据，每条数据包括一个提示和相应的回答。这些初始回答被设计为与人类偏好具有较强的对齐性。

为了进一步提升模型的性能，我们整合了另外 30,000 个高质量数据样本，这些样本来源于研究社区。这些样本是从我们精心收集的包含约 150 万条记录的开源仓库中筛选出来的。最初，我们使用了一个专有的图像-文本嵌入模型，将图像-文本对聚类到特定任务的类别中。这种聚类实现了有针对性的下采样，确保数据集在各种任务中保持高度的多样性。随后，我们利用与人类偏好对齐的微调模型，在此采样子集上进行了多次运行。生成的响应通过 LLM-as-a-judge [177] 进行过滤，这可以以原始真实值为参考来验证模型生成响应的正确性。在此基础上，我们进一步采用了奖励模型 (section 4.2.2) 从保留的结果中筛选出最符合人类偏好的响应，从而获得最终的拒绝采样微调数据 [134]。最后，我们将 SFT 数据中的开源数据量从 150 万压缩到约 30,000 个高质量数据。其他开源数据则提前用于预训练阶段。

在预训练过程中获得的增强能力基础上，包括复杂图表理解、STEM 相关推理、常识推理和 3D 感知，以及视频分析，我们迭代地增加了微调数据和指令的复杂性。这涉及到减少那些可以仅凭个体能力轻松解决的简单提示的比例，并引入了之前暴露了预训练模型局限性的更具挑战性的问题。利用自指导方法 [143]，我们通过根据各种逻辑结构组合多个更简单的提示，生成了新的复杂提示及其对应的模型响应。通过自指导和拒绝采样生成的响应经过人工二次验证过程，以识别和修正错误。与直接人工标注相比，这种精炼模型生成响应的方法显著提高了人工标注效率。此外，它还能够排除超出模型当前能力的数据，从而降低幻觉的风险。

4.1.2 训练食谱

在 SFT 阶段，我们构建了一个简洁且高质量的数据集，包含约 50,000 个样本。这个多模态 SFT 数据与内部的纯文本 SFT 数据集进行了整合。结合如 section 4.5 所述的长思维链 (LongCoT) SFT 数据，该综合语料库用于进行两轮训练。在 SFT 过程中，视觉编码器的参数被冻结，而所有其他模型参数保持可训练。训练时使用了 131,072 个词元的序列长度和相当于序列长度 16 倍的批量大小。我们使用了 AdamW 优化器 [64] 进行训练，超参数设置为 $\beta_1 = 0.9$ 、 $\beta_2 = 0.95$ ，并设置了 0.1 的权重衰减。训练过程包括一个持续总步骤数 10% 的预热阶段，之后学习率按照余弦衰减计划从峰值 2×10^{-5} 衰减到 2×10^{-6} 。

4.2 来自人类反馈的强化学习

为了进一步提升人工评估性能和多模态理解能力，我们进行从人类反馈中强化学习 (RLHF) [180]，这包括偏好数据采集、奖励模型训练以及使用强化算法进行最优化。

4.2.1 偏好数据

为了训练奖励模型，我们通过人工标注和启发式合成收集列表级多模态偏好数据集来进行奖励建模。

人工标注. 人工标注的偏好数据涉及使用 5 分制评分系统比较多个候选模型响应。生成偏好数据的提示涵盖了所有通用视觉理解能力，并在所有能力上保持平衡的评分。我们利用当前性能最佳的内部模型，通过核采样 [47] 随机采样响应。为确保响应的多样性，我们在选择用于人工标注的响应之前应用了过滤技术——如编辑距离、语义相似度和长度平衡策略。除了按质量对响应进行排序外，我们还指示人工标注者选择一个需要最少编辑来纠正或改进其质量的模型响应，这进一步弥补了有限响应采样中多样性不足的问题。标注者还被要求识别并突出显示响应中的问题——如幻觉、帮助性、信息量等——并对这些问题提供详细解释。为了进一步提高标注过程的效率，我们采用最新的奖励模型进行预标注排名，为人工标注者提供初步指导。这种方法不仅简化了标注工作流，还确保了更一致和客观的评估。

合成数据. 虽然一些近期方法 [172, 179] 采用了故意引入错误来合成偏好对，但多项研究 [4, 75, 162] 表明，这种合成数据通常难以有效泛化，因为奖励模型倾向于学习编辑后响应与原始响应之间的固有模式。相反，我们汇集了一组多样化的多模态提示，并具有明确的真值，同时实施格式约束，例如“最终答案:”。对于每个提示，我们生成模型响应 K 次，并使用现有的视觉-语言模型根据真值评估其正确性和格式遵循情况。因此，我们建立了具有明确排名的列表偏好：正确且格式清晰的响应排名最高，其次是格式清晰但错误的响应，最后是不遵循格式的错误响应。此外，我们遵循 FeedQuill [162] 生成图像描述偏好对，这有助于减少幻觉。所有合成偏好数据均按照偏好强度进行精炼 [137]。

4.2.2 视觉语言模型作为奖励模型

我们使用经过指令微调的视觉语言模型（VLM）初始化奖励模型。然后，按照 [86, 120]，我们促使模型 π_ϕ 充当生成式分类器，该分类器在给定提示 x 的情况下，直接输出关于两个响应 y_1 和 y_2 之间偏好的答案指示词元 \hat{I} 。这个过程可以表述为 $\hat{I} \sim \pi_\phi(I|x, y_1, y_2)$ 。

我们发现，由于直接处理词元概率和响应比较，这种方法相比传统的 Bradley-Terry 奖励模型 [100] 能够生成更稳健且优越的奖励模型。为了减轻视觉-语言模型 [176] 中固有的位置偏差，我们计算了两种可能的响应顺序的概率，即 (x, y_1, y_2) 和 (x, y_2, y_1) 。这确保了模型的偏好判断是公平的，不会受到响应呈现顺序的影响。此外，在训练过程中，我们应用了迭代学习策略，以在标注原则标准不断演变的情况下保持一致性。该策略包括持续更新训练数据和标注指南，以反映最新和最准确的标准。通过这样做，我们确保奖励模型保持可靠并能适应不断变化的要求。这种方法有助于提高模型的泛化能力，并长期维持高质量的性能。

4.2.3 强化学习的数据监管

我们的在线强化学习实现采用了一种近端策略优化（PPO）算法的变体 [155]。在这种方法中，奖励信号来自于奖励模型对生成的答案词元分配的概率。此外，在 PPO 训练期间，将真实值响应或 SFT 模型的最佳 N 个响应作为参考答案提供给奖励模型。

用于强化学习训练的提示语是从偏好数据集中获取的。观察到提示语分布的覆盖范围对强化学习性能有关键影响。因此，我们的数据采集策略旨在模仿偏好数据的分布。然而，收集到的提示语在质量上表现出显著的异质性，表现为任务难度和能力类别上的高度偏斜分布。为了解决这些问题，实施了一个多

阶段的数据精炼流水线。首先，训练一个标注模型，以将能力类别标签分配给提示语，然后进行分层采样，以确保不同能力类别之间的平衡表示。随后，对于每个提示语，使用最先进的内部模型生成 K 个响应，并使用我们奖励模型的最新迭代进行评估。根据奖励得分方差应用了过滤准则：如果 K 个响应中最大奖励与平均奖励之间的差异低于预定义阈值，则排除这些提示语。这一步确保保留奖励模型表现出显著区分能力的提示语。最后，在强化学习训练的初始阶段，那些同时快速增加奖励和 KL 散度的提示语（表明任务难度较低）会被下采样。

4.3 具有可验证奖励的强化学习

除了人类反馈外，带有可验证奖励的强化学习 (RLVR) [68] 成为一种高效的训练方法，适用于各种任务 [39, 69]，例如数学推理和编程，在这些任务中我们仅使用答案匹配或约束验证来训练模型，而不是利用基于模型的奖励估计。在本节中，我们设计了几种视觉任务，其最终解决方案可以通过规则或外部执行器精确验证，这些任务将在以后纳入 RLVR 训练中。

4.3.1 视觉 STEM

STEM (科学、技术、工程和数学) 问题通常有唯一且可验证的答案，这使其适合 RLVR。我们收集了超过一百万个包含图像的 STEM 领域问题，主要集中在数学上，来源于开源资源 [85] 和内部的 K-12 教育资料库。

为了准备训练数据，最初通过去除选项将多项选择题转换为开放式格式，从而迫使模型生成正确答案的内容并防止随机猜测。随后，基于 SFT 模型的性能通过拒绝采样选择难题。我们仔细删除仅通过文本或文本和字幕就能回答的问题，确保在强化学习中不会强化文本或表面视觉元素上的捷径。具体而言，每个问题生成 16 个回答，并且丢弃在 SFT 模型上准确率达到 0% 或大于 75% 的问题。这种筛选隔离了适合于基于强化学习的视觉推理探索的具有挑战性的提示 ($0\% < \text{accuracy} \leq 75\%$)，同时去除了潜在错误或琐碎的问题。最后，在提示前加上一个前言指令，指示模型使用指定的 LaTeX 标识符（例如，`\boxed{answer}`）来格式化最终答案，以便能够直接进行自动提取。

我们的 STEM 验证器将预测的答案转换为 sympy 表达式，并与真实值进行匹配。为了确保验证器的准确性，我们还移除了包含多个问题或其真实值为复杂短语的提示。

4.3.2 视觉感知与推理

验证者反馈也可以通过各种视觉任务收集，以增强 VLM 的感知和推理能力。在这里，我们展示了一些关于定位、视觉谜题和与感知相关的游戏的早期探索。

定位. 定位任务旨在评估模型将文本描述准确地（“定位”）与输入图像中的相应视觉元素关联的能力。为了便于答案提取，我们在提示中添加了一个指令，以鼓励模型输出用 `<bbox>` 和 `</bbox>` 词元包围的预测边界框。奖励是根据预测边界框与真实边界框之间的交并比 (IoU) 计算的。我们还以类似的方式优化了指向能力，并将物体的中心点位置放在 `<point>` 和 `</point>` 之间。

视觉指令遵循. 通过合成数据和基于规则的验证器可以提高指令遵循能力 [25, 161]。遵循这一想法，我们合成多种视觉指令，其结果可以通过相应的正则表达式进行验证，以进一步增强视觉指令遵循能力。

视觉谜题 & 游戏. 视觉谜题是需要模型从视觉场景中收集信息并应用抽象推理、归纳推理和演绎推理

等推理技术的任务。类似于 [18, 132]，我们为 RLVR 合成超过 20k 个视觉谜题及其对应的解答。我们仔细地用现有的视觉谜题基准，如 PuzzleVQA [18] 对我们的合成训练数据进行去污染处理。我们还涉及图推理 [146] 和模式识别的谜题。类似于 STEM 验证器，我们提示模型将谜题的最终答案包含在 \boxed{answer} 中，并通过字符串匹配算法验证预测结果。

除了生成自然语言响应外，我们正在探索允许直接与图像内容交互或操作的 VLM 输出格式，旨在通过更直观和吸引人的交互方式促进更广泛的 VLM 应用。例如，想象一下，增强型 AI 眼镜可以直接在用户的视野上叠加导航路线，而不是仅仅依赖文本或语音——这可能是一种更直观的方法。作为开发这些交互能力的初步步骤，我们专注于视觉游戏，这些游戏是合适的测试平台，因为它们需要强大的感知技能，并且有明确可验证的结果来表明成功。具体来说，我们针对“找不同”游戏，让模型识别两张图片之间的差异。至关重要的是，模型不仅要使用自然语言解释这些差异，还要输出精确定位差异区域的边界框，直接在图像上显示。我们使用两种方法利用合成生成的数据来训练这种能力：(1) 我们从开源数据集中获取图像，随机遮罩部分区域，使用扩散模型进行修复（参见 figure 5 示例），然后过滤掉修复内容与原始内容过于相似的图像对；(2) 为了确保模型能够感知到线条宽度或物体大小等细微差异，我们通过系统地修改来自开源数据集的 SVG 属性来生成额外的图像对。



Figure 5 用于“找不同”游戏训练的合成图像对示例，差异部分在左图中用红色方框标出。

4.4 混合强化学习

Seed1.5-VL 模型是利用一种从 PPO 算法变体中衍生出的混合 RL 框架进行训练的。该框架结合了如 [156] 所述的生成式 RM，并整合了近期 RL 研究中的若干改进和探索技术 [121, 165, 167, 170]。具体而言，我们的训练结合了 RLHF 和 RLVR。我们如下文所示提供更详细的实现方法：

格式化奖励. 我们预先定义了一种 `<think>{想法}</think>{解决方案}` 的响应格式，以确保模型在给出最终解决方案之前提供全面的思考。如果模型的响应不符合此格式，我们将奖励设置为零。如果响应未能遵循不同任务中不同验证器的格式要求，我们也会施加惩罚。

混合奖励. 我们的训练提示根据任务分为通用提示和可验证提示，分别由 RM 和验证器进行奖励。每个轮次中，提示都会随机打乱。因此，每个批量中都混合了通用提示和可验证提示。我们截断思考过程，只保留响应奖励模型的解决方案。因此，RM 将忽略 CoT 思考过程，只专注于为最终解决方案提供奖励。这种修改可以减轻对思考过程的约束，并鼓励模型探索更有效的 CoT 思考过程。

共享评论员. 采用单个评论员模型架构来估计对应于两个奖励源（即奖励模型和验证器）的价值函数。由于两个奖励信号都在相同的规范化值域 $[0, 1]$ 内运行，这种统一的方法是可行的。具体而言，奖励模型本质上在此区间内生成输出，而从所有验证器得出的结果会被明确缩放以符合相同的 $[0, 1]$ 值域。评论员模型的参数使用预训练奖励模型的权重进行初始化。随后，评论员经历由 100 个训练步骤组成的初始热身阶段，利用由 SFT 模型生成的轨迹数据（展开）。

KL 系数. 我们为通用提示和可验证提示使用不同的 KL 散度系数。具体来说，通用提示应用了系数 1×10^{-5} ，而可验证提示则使用了系数 0。对通用提示应用较小的 KL 系数旨在减轻潜在的奖励黑客问题。相反，在没有 KL 散度项的情况下训练可验证任务可以提高模型的探索能力。

训练方法. 混合 RL 训练的上下文长度和最大输出长度分别为 8,192 和 16,384。我们在每个回合中采样 4,096 次滚动。对于训练更新，我们使用 512 个样本的最小批量大小，每个回合执行 8 次梯度步骤。训练的 PPO 剪切值域为 0.2。演员和评论员的学习率分别为 6×10^{-7} 和 7.5×10^{-7} 。每个提示的滚动次数不同，因为更难的提示需要更全面的探索。我们仅对奖励模型奖励的每个提示进行一次采样，而对验证器奖励的对应提示进行 4 次或 8 次采样。值得注意的是，尽管我们在 RL 阶段仅使用 LongCoT 响应来训练 Seed1.5-VL，但我们仍然观察到常规响应在没有扩展推理的情况下有显著提升。

4.5 通过拒绝采样微调的迭代更新

在本工作中，我们采用一种迭代训练策略，在强化学习阶段提升 Seed1.5-VL。该过程从一个冷启动的 SFT 模型开始，用于 LongCoT，该模型最初在少量通过基础模型的上下文提示生成的低质量 LongCoT 样本上进行训练，这些样本是通过一小部分人工标注的例子生成的。观察到更强的冷启动 SFT 自然会导致 LongCoT RL 之后更强的最终模型，我们采用拒绝采样微调方法以获得改进的起点。具体来说，每次发布 LongCoT RL 模型的迭代后，我们通过我们的数据流水线收集更多的挑战性提示，并在这些提示上评估最新的 RL 模型。正确回答的响应随后被收集，类似于拒绝采样，并纳入后续 SFT 发布的数据中。在 RL 阶段使用的相同验证器用于确认这些响应的正确性。此外，我们实现了手动设计的基于正则表达式的过滤器，以去除不良模式，如无限重复、过度思考和其他语言缺陷。当前的 Seed1.5-VL 版本已经经历了四轮这样的迭代，表现出持续的改进，这种迭代优化预计将进一步提高其性能。

5 训练基础设施

5.1 大规模预训练

为了加速和稳定预训练，我们开发了多项训练最优化方法，包括混合并行、工作负载平衡、感知并行的数据加载和鲁棒训练。我们还应用高性能注意力内核用于上下文并行、选择性活性值检查点和卸载、内核融合以及细粒度通信重叠 [13, 173]。预训练阶段总共消耗了 130 万 GPU 小时²。

5.1.1 混合并行性

视觉语言模型的训练面临独特的挑战，这是由于数据的异质性，数据包括视觉数据和自然语言数据，而模型则由一个较小的视觉编码器和一个显著更大的语言模型组成。现有的训练框架主要针对顺序单峰值任务设计，在视觉语言模型训练中表现不足。它们要么将编码器视为语言模型数据的预处理步骤，要么完全将编码器与语言模型分离，导致工作负载不平衡、设备停滞时间延长和可扩展性差。为了解决这些挑战，我们开发了一种混合并行方法 [30]，该方法以不同方式对视觉编码器和语言模型进行并行处理。对于视觉编码器和 MLP 适配器，我们利用 ZeRO 数据并行 [109]，而对于语言模型，我们使用标准的 4-D 并行，该并行结合了专家并行 [65, 123]、交错流水线并行 [50, 93, 94]、ZeRO-1 数据并行 [109] 和上下文并行 [77] 用于上下文扩展。为了效率和简单性，我们分离了编码器/适配器和语言模型的并行策略——在不引入流水线级不平衡的情况下，将编码器和适配器集成到 4-D 并行中是具有挑战性的。我们的混合并行方法简单高效，只需对模型代码进行最小修改即可显著加速训练。

5.1.2 工作量平衡

视觉样本包含不同数量的图像，导致 GPU 之间的计算不平衡。我们采用一种经典的贪心算法重新分配视觉数据，以实现视觉编码器和适配器的负载均衡。首先，我们根据每个图像的计算强度（定义为处理每个图像所需的浮点运算次数 (FLOPS)）对图像进行降序排序。其次，我们按排序后的顺序扫描这些图像，并将每张图像分配给总计算强度最低的 GPU。此外，我们利用分组平衡来减少数据重新分配的开销。我们不是在所有 GPU 之间平衡视觉数据，而是将它们分成大小相等的组，并且仅在每组内部平衡视觉数据。经验上，我们将组的大小设置为 128-256 个 GPU。

5.1.3 并行感知数据加载

为减少多模态数据 IO 开销，我们还构建了一个并行感知的数据加载器。例如，非数据并行组内的 GPU 应消耗同一组训练样本。从分布式文件系统中重复读取相同数据会显著增加数据读取和预处理开销，从而延缓微批量的准备。我们通过一个并行感知的数据加载器来解决这个问题。例如，在 PP 组内仅有一个 GPU 加载数据，而其他 PP 秩则通过广播从该 GPU 获取必要的元数据。此外，由于我们对视觉编码器使用纯数据并行，每个 GPU 仅处理加载图像数据的一部分。在将训练批次移动到 GPU 之前，我们过滤掉不必要的图像，以减少 PCIe 流量。为了隐藏这些数据广播和传输成本，我们使用预取器确保 IO 和计算完全重叠。

²为保持一致性，本报告中提到的所有计算成本均以 H800 为基础归一化为 GPU 小时。

5.1.4 容错能力

为了处理训练过程中的各种硬件和软件故障，我们使用鲁棒训练框架 MegaScale [57] 来实现容差。一旦鲁棒训练框架检测到故障，它将触发恢复过程并从最后一个成功的检查点继续训练。我们利用 ByteCheckpoint [136] 实现高效的检查点保存和恢复。

5.2 后训练框架

我们使用基于 verl 的 [122] 框架对 Seed1.5-VL 进行结合人类反馈 (RLHF) 和验证器反馈 (RLVF) 的混合强化学习。它结合了一个用于管理跨 RL 角色数据流的单一控制器和多个用于管理内部 RL 角色数据和模型并行的控制器。验证器被部署在基于进程的服务中以隔离潜在的验证器故障。这种设计大大简化了各种实验的部署和开发。我们使用与预训练阶段相同的训练系统和优化技术来进行高效的演员和评论员更新，并使用 vLLM [67] 进行展开的自回归生成。具体来说，演员和评论员训练采用 3D 并行 [50, 93, 109, 123]；展开生成和奖励/参考模型推理使用副本，每个副本都配置了张量并行 [115]。Seed1.5-VL 的强化学习阶段耗时 60k GPU 小时。奖励模型使用与 Seed1.5-VL 预训练阶段相同的框架进行训练，需要 24k GPU 小时。后训练阶段也利用 ByteCheckpoint [136] 进行高效的检查点保存和恢复。

6 评估

本节结构如下。公开基准的定量结果在 section 6.1 中展示，随后在 section 6.2 中评估智能体任务的性能。我们内部基准的设计以及我们的模型与行业领先模型的比较随后在 section 6.3 中详细说明。模型的局限性在 section 6.4 中讨论。定性示例在 section A 中提供，全面的评估情景在 section B 中描述。

6.1 公共基准

6.1.1 视觉编码器作为 zero-shot 分类器

我们使用 zero-shot 图像分类基准对 Seed-ViT 进行评估，包括 ImageNet-1K [22]，ImageNet-V2 [112]，ImageNet-A [44]，ImageNet-R [43]，ImageNet-S [138]，以及 ObjectNet [8]。如 table 5 中详细所述，Seed-ViT 在这些数据集上的平均 zero-shot 准确率为 82.5，与 InternVL-C-6B [16] 相当，尽管 Seed-ViT 的参数数量仅为 InternVL-C-6B 的 9%。令人印象深刻的是，与具有 $30\times$ 更多参数的 EVA-CLIP-18B 相比，Seed-ViT 在大多数 ImageNet 变体上实现了相当的准确率。此外，与 DFN-5B-CLIP-H/14++ [28] 相比，Seed-ViT 在 ObjectNet（包含具有挑战性背景、旋转和视角的图像）和 ImageNet-A（包含自然对抗样本）上表现出更优越的性能，表明 Seed-ViT 对现实世界的变化具有更高的鲁棒性。

6.1.2 视觉任务评估

我们评估了 Seed1.5-VL 在一系列公共图像基准上的性能，将其与包括 Gemini 2.5 Pro (0325 版本)、OpenAI o1、Claude 3.7 Sonnet、OpenAI GPT-4o 和 Qwen 2.5-VL 72B 在内的多个最先进多模态模型进行了比较。我们选择将 Seed1.5-VL 与 Gemini 2.5 Pro (Preview 03-25) 进行比较，而不是 Gemini 2.5 Pro (Preview 05-06)，因为 Gemini 2.5 Pro (Preview 03-25) 在开放视觉语言基准中表现出更强的能力 (MMMU 中为 81.7_{Preview 03-25} 对比 79.6_{Preview 05-06})³。评估涵盖了从多模态推理和通用视觉问答到文

³<https://deepmind.google/technologies/gemini/pro/>

Models	Seed-ViT	OpenCLIP-G/14	DFN-5B-CLIP-H/14++	InternVL-C	EVA-CLIP-18B
#Param	532M	1.8B	632M	6B	17.5B
ImageNet-1K	83.6	80.4	84.3	83.2	83.8
ImageNet-V2	77.6	73.6	78.3	77.3	77.9
ImageNet-A	85.5	69.3	79.6	83.8	87.3
ImageNet-R	95.2	92.8	94.9	95.7	95.7
ImageNet-S	74.1	69.9	73.6	74.3	74.7
ObjectNet	79.2	73.0	78.0	80.6	82.2
Avg.	82.5	76.5	81.4	82.5	83.6

Table 5 预训练的 Seed-ViT (在与 LLM 集成之前) 与现有参数更多的竞争者在常见 zero-shot 基准上的比较。

档理解、定位和空间推理的各种能力。Table 6 展示了详细结果，每个基准中最高得分用粗体标出，次高得分用下划线标出，除了 FSC-147 和 NYU-Depth V2，其中较低的得分更好。我们报告了 Seed1.5-VL 在其标准的“非思考”模式和增强的“思考”模式下的结果，后者结合了长思维链以提高推理能力。

多模态推理。 在复杂的多模态推理任务中，Seed1.5-VL 在思考模式和非思考模式下均表现出强大的能力。值得注意的是，它在 MathVista (85.6 思考)、V* (89.5 非思考)、VLM are Blind (92.1 思考)、ZeroBench (sub) (30.8 思考) 和 VisuLogic (35.0 思考) 上达到了最先进的 (SOTA) 性能。在 MathVista 和 VLM are Blind 上，Seed1.5-VL 显著优于所有列出的对手。虽然 Gemini 2.5 Pro 在 MMMU (81.7 比 Seed1.5-VL 的思考模式 77.9)、MMMU-Pro (68.8 比 67.6)、MathVision (73.3 比 68.7) 和 OlympiadBench (69.8 比 65.0) 等基准测试中领先，但 Seed1.5-VL 仍具有竞争力，位居第二。对于 ZeroBench (main)，Seed1.5-VL 在思考模式下解决了 2 个案例，与 OpenAI o1 并列第二，仅次于 Gemini 2.5 Pro 和 Claude 3.7 Sonnet。Seed1.5-VL 在非思考模式下也显著优于其非思考模式的对手。

我们观察到，在我们的第一次 LongCoT RL 训练中，模型自然地表现出多种以视觉为中心的策略，例如“让我再看一遍图像”和“在识别位置之前分析细节”，如 figure 9 和 figure 10 所示，尽管当时我们还没有对相关 SFT 数据进行标注。

通用视觉问答。 在通用视觉问答基准测试中，Seed1.5-VL 表现出强大的性能。它在 RealWorldQA (78.4 思考) 和 SimpleVQA (63.4 思考) 上取得了最先进结果。在 MMStar 上，Seed1.5-VL (77.8 思考) 在比较的模型中也获得了最高得分。同样，在 MMBench-en (89.9 思考) 和 MMBench-cn (89.1 思考) 上，Seed1.5-VL 的得分接近 Gemini 2.5 Pro 和 Qwen 2.5-VL 72B 等顶级表现者。在 HallusionBench 上，Seed1.5-VL (60.3 思考) 获得第二高的得分，略低于 Gemini 2.5 Pro (63.7)。

文档和图表理解。 Seed1.5-VL 在文档和图表理解任务中表现出色。它在 TextVQA (84.2 非思考) 上设立了新的 SOTA 基准，在 InfographicVQA (91.2 思考) 和 DocVQA (96.9 非思考) 上也设立了新的 SOTA 基准，在这些领域超越了 Qwen 2.5-VL 72B 和 Gemini 2.5 Pro 等强大模型。在 ChartQA 上，Seed1.5-VL (89.1 思考) 取得了第二高的得分，仅落后于 Qwen 2.5-VL 72B (89.5)。它在 AI2D (88.5 非思考) 和 OCRBench (881 非思考) 上也表现出色，排名与 Qwen 2.5-VL 72B 和 Gemini 2.5 Pro 竞争激烈。对于 CharXiv (DQ)，Seed1.5-VL (92.6 思考和非思考) 排名第二，仅次于 Gemini 2.5 Pro (94.4)。然而，在 CharXiv (RQ) 上，其表现 (60.2 思考) 落后于领先者 Gemini 2.5 Pro (69.9) 和 Claude 3.7 Sonnet (68.9)。

Capability	Benchmark	Seed 1.5-VL	Seed 1.5-VL	Gemini 2.5 Pro	OpenAI o1	Claude 3.7 Sonnet	OpenAI GPT-4o	Qwen 2.5-VL 72B
		thinking	non-thinking	thinking	thinking	thinking	non-thinking	non-thinking
Multimodal reasoning	MMMU	77.9	73.6	81.7	77.6	75.2*	70.7*	70.2
	MMMU-Pro	<u>67.6</u>	59.9	68.8*	66.4*	50.1*	54.5*	51.1
	MathVision	<u>68.7</u>	65.5	73.3*	63.2*	58.6*	31.2*	38.1
	OlympiadBench	<u>65.0</u>	60.4	69.8*	48.5*	54.2*	25.9*	35.9
	MathVista	85.6	<u>83.0</u>	82.7*	71.8	74.5*	63.8*	74.8
	V*	<u>89.0</u>	89.5	79.1*	69.7*	86.4*	73.9*	86.4
	VLM are Blind	92.1	<u>90.8</u>	84.3*	57.0*	69.0*	50.4*	69
	ZeroBench (main)	<u>2</u>	0	3*	0*	3*	0*	0
	ZeroBench (sub)	30.8	<u>29.0</u>	26.0*	20.2*	20.4*	19.6*	13.0
	VisuLogic	35.0	33.0	31.0*	29.0*	24.8*	26.3*	28.0
General visual question answering	RealWorldQA	78.4	77.0	<u>78.0*</u>	77.1*	67.8*	76.2*	75.7
	SimpleVQA	63.4	<u>63.1</u>	62.0*	58.8*	50.1*	52.4*	52.4
	MMStar	77.8	76.2	<u>77.5*</u>	67.5*	68.8*	65.1*	70.8
	MMBench-en	89.9	88.0	90.1*	83.8*	82.0*	84.3*	88.6
	MMBench-cn	<u>89.1</u>	88.1	89.7*	81.3*	82.7*	82.0*	87.9
	MMVP	69.3	70.7	<u>70.7*</u>	—†	—†	70.7*	66.7
	HallusionBench	<u>60.3</u>	60.0	63.7*	55.6*	58.3*	56.2*	55.2
Document and chart understanding	TextVQA	81.8	84.2	76.8*	66.2*	62.4*	81.4*	83.5
	AI2D	87.3	<u>88.5</u>	88.4*	79.5*	82.1*	84.9*	88.7
	ChartQA	<u>89.1</u>	87.4	83.3*	83.1*	56.5*	86.7*	89.5
	InfographicVQA	91.2	<u>89.3</u>	84.3*	65.4*	66.5*	79.2*	87.3
	DocVQA	96.9	<u>96.7</u>	94.0*	81.6*	87.4*	66.2*	96.4
	OCRBench	861	<u>881</u>	866*	750*	793*	806*	885
	CharXiv (RQ)	60.2	59.8	69.9*	55.1*	<u>68.9*</u>	52.0*	49.7*
	CharXiv (DQ)	92.6	92.6	94.4*	88.9*	92.0*	86.5*	87.4*
Grounding & counting	BLINK	72.1	70.2	<u>70.6*</u>	66.1*	62.5*	65.9*	64.4
	LVIS-MG	<u>72.5</u>	73.8	63.8*	—†	—†	—†	—†
	VisualWebBench	<u>87.3</u>	88.0	<u>87.3*</u>	80.9*	85.9*	80.2*	82.3*
	RefCOCO-avg	<u>91.3</u>	91.6	74.6*	—†	—†	—†	90.3
	CountBench	93.7	93.5	91.0*	86.6*	86.1*	85.7*	93.6
	FSC-147 ↓	17.9	<u>18.6</u>	24.5*	34.3*	33.4*	46.8*	28.6*
3D Spatial understanding	DA-2K	<u>91.7</u>	91.9	73.0*	72.3*	40.1*	66.9*	69.6*
	NYU-Depth V2 ↓	<u>13.6</u>	11.6	27.5*	82.1*	92.4*	73.8*	35.5*
	All-Angles Bench	58.6	59.0	53.4*	54.0*	50.0	49.1*	55.7

* Results self-collected via API in April 2025.

† Invalid results due to failures in following format requirements.

Table 6 与先前模型相比，豆包在公共视觉语言基准测试 (section B.3) 上的性能。除了 Claude - 3.7 Sonnet 推荐使用默认采样模式外，所有基准测试均采用贪心解码进行评估。我们在这些基准测试中报告 Pass@1。对于 FSC - 147 和 NYU - Depth V2，分别使用平均绝对误差 (MAE) 和绝对相对误差 (AbsRel) 作为指标，因此数值越低越好。对于所有其他基准测试，数值越高越好。每个基准测试中的最高分用**加粗**标记，第二高分用下划线标记。

定位与计数. 这一类别突显了 Seed1.5-VL 的显著优势。它在所有列出的定位与计数基准中均实现了 SOTA 性能。具体而言，Seed1.5-VL 在 BLINK (72.1 思考)、LVIS-MG (73.8 非思考)、VisualWebBench (87.8 非思考)、RefCOCO-avg (91.6 非思考)、CountBench (93.7 思考) 和 FSC-147 (17.9 思考，越低越好) 上表现领先。值得注意的是，Seed1.5-VL 在 LVIS-MG 上的表现优于传统检测器，*i.e.*, Grounding DINO-L [14, 80]，其获得的 F1 得分是 54.4，这表明 Seed1.5-VL 在多对象定位方面具有强大的能力。在这些多样化任务中持续的顶级表现，凸显了 Seed1.5-VL 在目标定位、细粒度视觉理解和计数方面的卓越能力。

三维空间理解. 我们选择深度估计、三维目标检测和多视角推理作为三个任务来评估 Seed1.5-VL 在三维空间理解方面的能力。特别是对于深度估计，我们在两个公开基准 DA-2K [160] 和 NYU-Depth V2 [95] 上报告结果。在 DA-2K 中，我们遵循 [160] 并报告两个像素之间相对深度估计的准确率 (*e.g.*, 哪个像素更近)。在 NYU-Depth V2 中，我们报告标准的绝对相对误差，计算方式为 $|\text{dist}_{\text{pred}} - \text{dist}_{\text{gt}}| / \text{dist}_{\text{gt}}$ ，其中 $\text{dist}_{\text{pred}}$ 和 dist_{gt} 分别为预测距离和真实值。如 table 6 所示，Seed1.5-VL-thinking 在 DA-2K 上的得分为 91.7，在 NYU Depth V2 上的错误率为 0.136，远远超过了之前的 VLMs。在非思考模式下，Seed1.5-VL 在 DA-2K 和 NYU-Depth V2 上的得分分别为 91.9 和 0.116。对于三维目标检测，我们在 SUN-RGBD [125] 上报告结果。在非思考模式下，我们的模型在 SUN-RGBD 上的 AP@15 得分为 33.5，超过了 Gemini 2.0 Pro Experimental 的 32.5 AP@15 [129]。然而，我们观察到在此任务中使用思考模式会导致性能下降。也就是说，结果下降至 32.0 AP@15。对于多视角推理，我们在 All-Angles Bench [163] 上进行评估。Seed1.5-VL 在非思考模式下的得分为 59.0，在思考模式下的得分为 58.6，这显著超过了之前模型。

总之，Seed1.5-VL 在广泛的视觉语言基准测试中表现出最先进的或高度具有竞争力的性能。它在定位、计数、3D 空间理解、文档理解 (TextVQA、DocVQA、InfographicVQA) 以及某些推理任务 (MathVista、VLM are Blind 等) 方面表现尤为出色，确立了其作为强大且多功能的多模态模型的地位。

6.1.3 视频任务评估

我们对 Seed1.5-VL 在视频理解方面的熟练程度进行评估，从五个维度评估其能力：短视频、长视频、流媒体视频、视频推理和视频定位。Table 7 将 Seed1.5-VL 与最先进 (SOTA) 模型进行基准测试。由于 API 限制 (例如，网络超时、视频处理错误)，我们无法在所有基准上对某些专有模型 (如 Gemini 2.5 Pro) 进行评估。因此，表格中报告的是从公开报告或通过 API 自行收集的最高得分。

对于短视频理解，Seed1.5-VL 在 MotionBench、TVBench、Dream-1K 和 TempCompass 上实现了最先进性能，展示了其在处理简短视频片段的时间动态和运动模式方面的卓越能力。对于长视频理解，它也取得了强劲的结果，具有 128K 词元的上下文窗口 (最多 640 帧)。我们认识到扩展时间理解的重要性，并计划未来的工作重点是扩大此上下文窗口容量，以进一步提升长视频的理解能力。关于流式视频理解，我们在 OVBBench [51]、OVOBench [74]、StreamBench [153] 以及 StreamingBench 的主动子任务 [76] 上进行了评估。Seed1.5-VL 在所有这些基准上都实现了最先进性能，表明其在实时应用 (如交互式视频对话系统) 方面具有强大的潜力。在视频推理 (Video-MMMU [49]、MMVU [175]) 中，Seed1.5-VL 分别获得了 81.4 和 70.1 的得分，目前仍落后于 Gemini 2.5 Pro 等顶级模型。此外，Seed1.5-VL 在视频定位任务中表现出色，这些任务专门用于定位与文本描述相对应的视频时间片段。它在 Charades-STA [34] 和 TACoS [114] 上实现了最先进性能，展示了精确的定位能力。

Capability	Benchmark	Seed1.5-VL thinking	Seed1.5-VL non-thinking	Prior SOTA
Short video	MotionBench [48]	68.4	68.4	62.8 GLM-4V
	MVBench [73]	74.4	74.3	76.4 InternVL-2.5
	TOMATO [117]	44.7	44.2	46.9* Gemini 2.5 Pro
	TVBench [19]	63.6	61.5	62.6* Gemini 2.5 Pro
	Dream-1K [139]	43.9	42.6	42.0 Tarsier2
Long video	TempCompass [82]	83.7	83.1	75.8* Gemini 2.5 Pro
	LongVideoBench [147]	74.0	74.4	66.7 GPT-4o
	LVBench [142]	64.6	64.0	69.2* Gemini 2.5 Pro
	MLVU [178]	82.1	81.8	81.2* Gemini 2.5 Pro
	VideoMME(w/o sub) [32]	77.9	77.6	87.0* Gemini 2.5 Pro
Streaming video	TemporalBench [12]	79.8	78.9	73.3 GPT-4o
	OVBench [51]	60.0	59.6	54.9 PMB [51]
	OVOBench [74]	72.3	72.0	67.7 Gemini1.5-Pro
	StreamBench [153]	72.8	71.2	68.7 GPT-4o
	StreamingBench(proactive) [76]	68.0	82.8	64.7 Claude 3.5 Sonnet
Video reasoning	Video-MMMU [49]	81.4	72.1	76.7 Kimi-K1.6
	MMVU [175]	70.1	70.1	75.8* Gemini 2.5 Pro
Video grounding [†]	Charades-STA [34]	64.0	64.7	60.7 SG-DETR [36]
	TACoS [114]	49.6	47.8	42.4 SG-DETR [36]

* Results self-collected via API in April 2025.

[†] We adopt mIoU as the main metric for video grounding tasks.

Table 7 Seed1.5-VL 在公开视频基准上的性能与之前模型的对比。对于所有基准，数字越高越好。评估帧率分别为 MotionBench、MVBench、TOMATO 和 TVBench 的 2 FPS，Dream-1K 的 3 FPS，以及其他所有数据集的 1 FPS。

6.2 多模态智能体

多模态智能体是通过视觉输入感知世界、理解自然语言指令并采取动作以完成任务的系统。评估此类智能体的两个关键场景是 GUI 交互和游戏，它们测试现实世界的可用性和复杂推理能力。GUI 智能体通过在桌面、浏览器和移动设备上的屏幕界面进行感知和操作来模拟人机交互。这些任务需要精确的视觉定位和多步骤执行。游戏智能体在视觉丰富且交互性强的环境中运行，需要战略规划、实时决策和常识推理。我们在 GUI 操作和游戏这两个领域对 Seed1.5-VL 进行了基准测试，使用了多样化的评估方法。结果如 tables 8 和 9 所示，我们报告了 Seed1.5-VL 在思考模式下的表现。

Capability	Benchmark	Seed	OpenAI	Claude	UI-TARS	Kimi	Qwen 2.5
		1.5-VL	CUA [98]	3.7 Sonnet [6]	1.5 [116]	VL-A3B [130]	VL 72B [7]
GUI Grounding	ScreenSpot-V2 [149]	95.2	87.9	87.6	<u>94.2</u>	92.8	-
	ScreenSpot-Pro [72]	<u>60.9</u>	23.4	27.7	61.6	34.5	43.6
Computer Use	OSWorld [152]	36.7	<u>38.1</u>	28.0	42.5	8.2	8.8
	Windows Agent Arena [11]	<u>39.6</u>	-	38.9	42.1	10.4	-
Browser Use	WebVoyager [42]	87.2	<u>87.0</u>	84.1	84.8	-	-
	Online-Mind2Web [158]	76.4	71.0	62.9	<u>75.8</u>	-	-
Phone Use	Android World [111]	<u>62.1</u>	-	-	64.2	-	<u>35.0</u>

Table 8 Seed1.5-VL 在公共 GUI 在线基准上的性能与之前模型的比较。

GUI 识别. GUI 识别指的是模型理解并定位界面元素的能力——这是基于视觉的智能体的基本技能。我们在 ScreenSpot Pro [72] 上评估这种能力，该数据集专注于专业情景中的专家标注任务，以及 ScreenSpot v2 [149]，该数据集涵盖了桌面、移动和网页界面的识别。Seed1.5-VL 展现出强大的识别性能，在 ScreenSpot Pro 上达到 60.9，在 ScreenSpot v2 上达到 95.2，优于 OpenAI CUA 和 Claude 3.7 Sonnet。作为多模态交互的基础，GUI 识别使智能体能够感知可操作元素，并实现感知与控制之间的衔接。

GUI 智能体. 对于 GUI 智能体能力评估，我们在涵盖计算机使用、浏览器使用和手机使用的不同 GUI 情景下，将 Seed1.5-VL 与强大的基准模型如 OpenAI CUA [98] 和 Claude 3.7 Sonnet [6] 进行比较。如 table 8 所示，Seed1.5-VL 在多个关键基准上始终优于之前模型。例如，在 OSWorld [152] 和 Windows Agent Arena [11] 上，Seed1.5-VL 分别取得 36.7% 和 39.6%，超过 Claude 3.7 Sonnet 的 28.0% 和 38.9%。在浏览器使用中，Seed1.5-VL 在 WebVoyager [42] 上得分为 87.2%，在 Online-Mind2Web [158] 上得分为 76.4%，优于 OpenAI CUA 和 Claude 3.7 Sonnet，创造了新的最先进结果。在 AndroidWorld [111] 这个具有挑战性的移动界面任务中，Seed1.5-VL 也取得了 62.1% 的高分。总体而言，在所有基础 VLMs (即 Claude 3.7 Sonnet、Kimi VL-A3B 和 Qwen 2.5-VL) 中，Seed1.5-VL 在 GUI 智能体任务中表现出显著更好的性能。这些结果突显了 Seed1.5-VL 在执行 GUI 任务方面的卓越能力，以及其在多样化环境和设备中的强大泛化能力，牢固确立了其在 GUI 领域的领先地位。

智能体. 游戏玩法为多模态模型提供了一个严格的基准，结合了视觉丰富的环境和复杂的逻辑，挑战模型处理复杂推理、顺序决策和快速适应的能力。游戏中的成功取决于直观的常识推理、长期战略规划以及适应动态挑战的能力，这使其成为展示最先进多模态智能体高级认知能力的理想测试平台。

Game	Seed1.5-VL	UI-TARS-1.5	OpenAI CUA	Claude 3.7 Sonnet
2048 (score)	870.6	721.3	611.2	800.0
Cubinko (level)	2.0	0.0	0.0	0.0
Energy (level)	2.3	1.8	0.8	1.0
Free-The-Key (level)	1.0	0.0	0.0	0.0
Gem-11 (score)	35.1	10.8	8.7	0.0
Hex-Frvr (score)	1414.0	1583.7	651.6	523.1
Infinity-Loop (level)	1.4	0.7	0.4	0.1
Laser-Maze-Puzzle (level)	2.6	2.2	1.4	1.4
Maze:Path-of-Light (level)	1.3	0.3	0.3	0.8
Shapes (level)	2.2	1.5	0.9	0.2
Snake-Solver (level)	1.3	0.2	0.2	0.2
Tiles-Master (level)	2.3	1.7	1.5	1.6
Wood-Blocks-3d (score)	864.0	213.3	18.1	0.0
Yarn-Untangle (level)	6.0	5.7	5.1	1.6

Table 9 Seed1.5-VL 在 14 个 Poki 游戏中的表现，包括得分或完成的关卡。模型经过多次运行评估，允许最多 100 步。对于所有游戏，数字越高越好。

我们从 Poki.com 汇集了一个包含 14 种不同游戏的基准测试集⁴，评估 Seed1.5-VL 在定位、感知和推理方面的能力。如 table 9 所示，Seed1.5-VL 在多个游戏中都优于之前的模型。例如，Seed1.5-VL 在 2048 中获得 870.6 分，超过 OpenAI CUA (611.2) 和 Claude 3.7 Sonnet (800.0)，在 Hex-Frvr 中获得 1414.0 分，明显领先于 OpenAI CUA (651.6) 和 Claude 3.7 Sonnet (523.1)。这些结果突显了 Seed1.5-VL 在完成游戏关卡和获得高得分方面的卓越表现。此外，游戏玩法的长时域特性使其特别适合评估推理时的缩放行为。如 figure 6 所示，Seed1.5-VL 展现出强大的可扩展性，在交互轮次增加时保持更高的性能。这展示了其稳健的设计和先进的推理能力，确保随着任务复杂性的增长，性能仍能持续提升。

6.3 内部基准

除了公开基准外，我们还建立了内部基准以全面评估我们的模型。我们在 section 6.3.1 中展示了内部基准的动机和设计原则，在 section 6.3.2 中展示结果，并在 section 6.3.3 中演示模型的分布外 (OOD) 泛化能力。

⁴<https://poki.com>

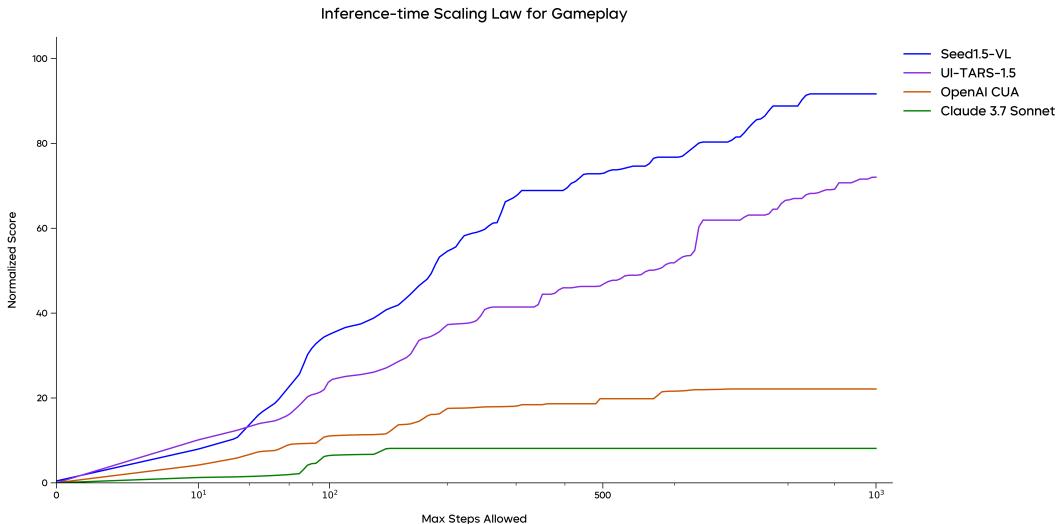


Figure 6 对于每场比赛，我们使用规范化的参考得分计算每个模型的缩放曲线，并将其平均以生成整体的推理时间缩放趋势。

6.3.1 动机与设计原则

除了利用公开基准进行详尽评估外，我们还开发了一套内部基准测试套件，以解决现有资源中的一些固有局限性。首先，公开基准中英语占主导地位，这需要创建全面的基准来专门评估模型在中文环境下的性能，以符合运营需求。其次，多模态研究的快速发展导致许多公开基准出现饱和，降低了其对模型性能微小提升的敏感度，并阻碍了顶尖模型之间的有效区分。最后，与公开数据集中普遍采用的基于规则的评估方法相关的局限性，包括答案解析中的挑战以及标签错误等潜在的数据质量问题，突显了需要定制化的内部基准，可能具有更稳健的评估协议和精心筛选的数据。

因此，我们制定了内部基准，这些基准遵循若干核心原则：

- **关注核心能力而非用户对齐**：基准优先评估基础模型能力（例如，感知、推理），而不是表面的对齐特性，例如对响应冗长性的偏好。这种方法可以最小化对齐调优对迭代模型改进评估的混淆影响。
- **全面的范围（原子能力和集成能力）**：评估套件涵盖了对特定的原子能力（例如，细粒度的视觉识别）以及跨越不同应用领域的复杂集成多模态任务的评估。
- **评估准确率和方法**：我们使用大型语言模型（LLMs）作为评判者，超越了传统的基于规则的指标。这些“评估者”模型所使用的提示和参考答案不断优化，以确保高评估保真度。当前评估者的准确率对于选择题或简单回答问题（例如，单个词/数字的回答）平均高于 95%，对于开放性问题则超过 90%（更多细节见 section B.1）。
- **基准过拟合的缓解**：为了防止由于模型对基准数据过拟合而导致的性能得分虚高，我们实施了一个严格的数据去重流水线。此外，基准中的任务类型和数据源会定期更新。
- **任务和输入多样性**：认识到多样性在 VLM 中的关键作用，我们的基准测试强调任务类型和输入图像的多样性。当可行时，图像来源优先选择非公开爬取的数据。我们将基准测试结构化为多个不同的

维度，结果在超过 100 个任务和超过 12,000 个样本来自不同的源和领域。这包括一个专门的 Out-of-Distribution (OOD) 分类，包含非传统任务，旨在探测模型泛化能力。一个详细的靶向能力分类学提供在 section B.1。

6.3.2 与最先进的技术对比

Level-1 Capabilities	Level-2 Capabilities	Weight	Seed 1.5-VL thinking	Gemini 2.5 Pro thinking	OpenAI o1 thinking	OpenAI o4-mini w/o tool use	Claude 3.7 Sonnet thinking
Overall		1.0	<u>59.3</u>	61.6	54.0	55.4	48.6
Vision Capabilities	Perception	0.1	<u>63.0</u>	64.4	51.6	56.8	48.4
	Recognition	0.1	<u>72.4</u>	74.8	<u>74.5</u>	64.8	55.7
	OCR	0.1	<u>67.2</u>	70.7	<u>55.7</u>	64.4	57.1
	Caption & Counterfactual	0.05	<u>47.7</u>	54.9	43.6	27.6	34.1
Integrated Capabilities	OOD	0.15	44.1	<u>43.1</u>	42.3	38.4	35.9
	STEM	0.04	<u>63.3</u>	64.0	56.1	55.0	45.2
	Knowledge	0.06	64.9	73.6	<u>68.5</u>	57.8	50.8
	Reasoning	0.1	47.6	<u>52.4</u>	44.9	57.4	39.6
	Document & Diagram Understanding	0.1	<u>73.1</u>	75.5	66.3	70.9	64.7
	Agent	0.1	63.1	63.1	53.2	52.9	53.2
	Atomic Instruction Following	0.03	69.6	<u>69.2</u>	63.8	68.7	50.5
	Code	0.05	44.0	<u>43.7</u>	39.9	60.6	<u>54.6</u>
	ToB	0.02	<u>47.1</u>	54.7	30.2	39.8	29.1

Table 10 评估结果比较 Seed1.5-VL 与最先进模型在内部基准上的表现。总体得分是根据定义的子类别中的性能计算的加权平均值。其他模型的数据通过 2025 年 4 月的 API 访问获取。平均值的权重设置是为了最小化评估的方差并突出每个类别的重要性。最高得分用 **粗体** 标记，第二高得分用 下划线 标记。

我们在思维模式下，于 table 10 中，将豆包与行业领先模型 (Gemini 2.5 Pro、OpenAI o1、OpenAI o4 - mini、Claude 3.7) 进行比较。领先得分 61.6 (Gemini 2.5 Pro) 表明在此基准上仍有很大的改进空间，这与 table 6 中许多接近饱和的公共基准得分超过 80 不同。在 section B.2 中可以找到包括非思维模型的更全面比较。

Seed1.5-VL 取得了总体第二高的得分。它在分布外 (OOD)、智能体、原子指令遵循等类别中达到了最先进的性能，并在 STEM 以及文档与图表理解方面展现出强大的能力。与表现最佳者相比，其主要弱点体现在知识、推理、代码以及字幕生成/反事实任务方面。我们认为这一差距部分归因于当前模型的规模，该模型使用了一个约有 200 亿个活跃参数的语言模型。在 figure 3 中展示了进一步缩放可能带来的收益的证据，其中训练损失在处理 3 万亿词元后仍无饱和迹象，且评估指标与损失密切相关。因此，我们预计随着模型大小和训练计算量的增加，性能差距将逐渐缩小。

按参数数量严格分组模型具有挑战性，因为许多模型的具体参数细节未公开。我们的模型大小与最近发布的 Llama 4 Maverick [91] 相当，据报道该模型使用了 170 亿个活跃参数，并采用了专家混合 (MoE) 架构。我们的评估表明，Seed1.5-VL 在此基准上的表现明显优于 Llama 4 Maverick (figure 29)。

6.3.3 分布外泛化

如图 table 10 所示，我们的模型在基准测试的 Out-of-Distribution (OOD) 类别中表现出与行业领先模型 Gemini Pro 2.5 和 OpenAI o1 相当的性能。除了标准化评估外，还开发了一个内部聊天机器人平台，以评估模型在复杂现实场景中整合多种原子能力的能力。突出展示了三个代表性例子。首先，该模型成

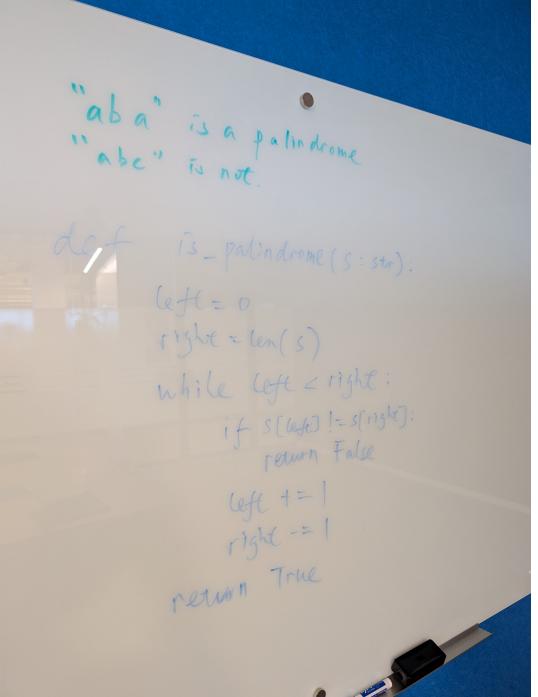
功解决了图 figure 7 中显示的 Rebus 谜题，利用了其 OCR、知识检索和推理能力。图 section A.3 中提供了更多示例，其中一些可能对人类来说具有挑战性。在第二个例子（figure 8）中，它处理从具有挑战性的视觉输入（一个闪亮的白板）中捕获的代码，分析并纠正错误，并使用用户提供的示例验证更正后的代码。在第三个例子（figure 23）中，该模型解析了一种未见过的格式的图表，执行“逆向图形”以生成潜在的 Mermaid 代码，同时遵守特定的空间约束（“南部区域”）。

Get it
Get it
Get it
Get it

Prompts: can you solve this Rebus puzzle?

Model Responses: <think>...</think>The Rebus puzzle features the phrase “Get it” repeated four times. The number of repetitions (four) combined with the phrase ”Get it” creates a homophonic play: “**Four get it**” sounds like “**Forget it**”. This wordplay uses the repetition count (four) to form the first part of the phrase, which phonetically matches “forget”. Thus, the solution is **Forget it**.

Figure 7 为了解决这个谜语图谜题，我们的模型必须正确识别图像中的字符，计算重复次数，检索关于谜语图谜题的知识，并进行推理以获得最终答案。为简洁起见，省略了思考词元。



The image shows a whiteboard with handwritten Python code for checking if a string is a palindrome. The code defines a function `is_palindrome` that takes a string `s` and initializes two pointers, `left` at 0 and `right` at `len(s)`. It then enters a loop where it compares characters at `s[left]` and `s[right]`. If they are not equal, it returns `False`. Otherwise, it moves the `left` pointer to the next character and the `right` pointer to the previous character, repeating until the pointers meet or cross. At the end of the loop, it returns `True`. Above the code, there are two examples: “aba” is a palindrome and “abc” is not.

Prompts: There is a bug in the code on the whiteboard. Can you spot it and return the correct code?

Model Responses: <think>...</think>The bug is in the initialization of the right pointer. It should be `len(s) - 1` (since Python uses zero-based indexing, the last character is at index `len(s)-1`, not `len(s)`). Here's the corrected code:

```
def is_palindrome(s: str) -> bool:
    left = 0
    right = len(s) - 1 # Fixed: use last index (len(s)-1)
    instead of len(s)
    while left < right:
        if s[left] != s[right]:
            return False
        left += 1
        right -= 1
    return True
```

Explanation:

- Right Pointer Initialization:** `right = len(s) - 1` ensures we start at the last character (correct index), avoiding an `IndexError`.
- Loop Logic:** Compares characters from the start (`left`) and end (`right`), moving pointers inward. If any pair mismatches, returns `False`. If the loop completes (all pairs match), returns `True`.

Testing:

- “aba”: `left=0, right=2` → match (`'a' == 'a'`), then `left=1, right=1` → loop ends, return `True` (correct).
- “abc”: `left=0, right=2` → `'a' != 'c'` → return `False` (correct).

This fixes the index error and correctly implements the palindrome check using two pointers.

Figure 8 一个使用其 OCR 和编码功能在白板上编写代码的示例。

一个独立的基准测试被用来评估新模型与用户偏好的契合度，作为内部评估的补充。此次评估通过为模型添加搜索工具来实现，可作为其对斗罗实用性的代理指标。⁵ 用户基础。Seed 1.5VL 的有用性率为 62.6%。同样，我们还测试了 Gemini 2.5 pro 0325，其有用性率达到了 57.4%。

⁵<https://www.doubaocom/chat/>

6.4 局限性

尽管在许多基准上表现出色，Seed1.5-VL 在细粒度视觉感知和复杂推理方面仍存在一定的局限性。

在复杂的视觉感知任务中，Seed1.5-VL 在物体不规则排列、颜色相似或部分遮挡时难以准确计数。识别图像之间的细微差异也存在挑战，有时会导致遗漏细节或描述不准确。此外，类似于一些同时期的模型（例如 OpenAI GPT-4o 和 Gemini 2.5 Pro），在精确解释复杂的空间关系方面可能会遇到困难，尤其是在不同视角下，以及准确响应视觉提示时，可能会错误地识别指定区域内的内容或数量，或出现定位错误。

超越感知，更高层次的推理也面临挑战，这从开放和内部基准中均可看出。Seed1.5-VL 在人类轻松完成的推理任务上表现不佳，例如解决 Klotski 拼图或导航简单迷宫，这表明未来需要探索诸如视觉思维链 (CoT) 等技术，其中模型可能通过辅助视觉工具（e.g., 线条）来增强其过程。需要组合搜索的推理对许多现有 VLM 构成重大挑战。[Figures 27 and 28](#) 提供了属于此类问题的两个示例。尽管对当前 VLM 架构具有挑战性，但组合搜索任务通常可以通过编程或基于代码的方法更容易解决。因此，将代码使用和其他外部工具纳入 VLM 框架代表了未来研究的重要方向，旨在增强此类推理能力。

大多数 VLM 在 3D 空间推理任务中存在局限性。这些挑战包括涉及 3D 物体操作或关于 3D 物体投影的推理的任务。[figure 25](#) 中提供了说明此类失败情况的具体实例。解决这一限制的一种潜在方向是将图像生成能力纳入基础模型，这可能进一步实现视觉思维链机制。这种方法仍然是未来研究的主题。

此外，视觉语言模型有时仍然会产生不正确的推理，尤其是在视觉谜题等任务中。这些推理错误可能源于潜在的感知错误（错误地解释形状或条件）或逻辑演绎过程本身的局限性。在需要规划或遵循复杂指令的任务中，Seed1.5-VL 可能会忽略特定条件或引入没有根据的假设，这可能导致不完整或无效的回答。

我们的内部评估还揭示了模型在时间推理能力方面的不足，因为模型在辨别连续动作的时间顺序或从物体的前后状态推断顺序方面遇到困难。多图像推理的能力有限，对于需要在多个图像中综合具有强逻辑关联线索的任务，性能会下降。

最后，幻觉仍然是所有 VLMs 的一个重大挑战。[Figure 26](#) 展示了这样一个特别显著的案例，其中视觉输入与从语言模型组件中获得的先验知识发生冲突。在这些情况下，模型往往更倾向于优先考虑这种获得的知识，从而有效地覆盖或使所感知的视觉信息符合学成的先验知识。

7 结论与下一步计划

在本文中，我们介绍了 Seed1.5-VL，这是我们最新推出的多模态基础模型，展示了在推理、OCR、图表理解、视觉定位、3D 空间理解以及视频理解方面的强大能力。尽管其规模相对适中，Seed1.5-VL 在 60 个评估基准中的 38 个上取得了最先进的结果，包括在 MMMU 基准上获得 77.9 分，该基准被广泛视为多模态推理能力的关键指标。

除了基准性能之外，Seed1.5-VL 表现出显著的集成能力和对与其训练数据不同的任务的泛化能力。例如，包括解决复杂的视觉推理任务，如谜语拼图，从白板图像中解释和纠正手写代码，以及作为计算机交互和游戏的智能体。有必要进一步探索这些涌现能力。

我们的缩放分析表明，模型性能尚未出现饱和迹象，这表明增加模型参数和训练计算量是一个有前景的

即时方向。通过我们的评估，我们还发现了当代视觉语言模型（VLMs）普遍存在的局限性，例如稳健的三维空间推理、幻觉缓解以及复杂的组合搜索。解决这些挑战是我们当前研究的核心部分，其中包括将现有模型能力与图像生成统一起来（可能实现视觉思维链）以及引入稳健的工具使用机制。

此处提出的进展建立在人工智能研究社区的大量先验工作之上，利用了如 Transformer 和视觉 Transformer 架构等基础性发展。为了促进未来的发展，我们在本报告中详细介绍了我们的模型架构、数据合成流水线、训练方法、训练框架创新以及内部评估设计。

References

- [1] Fuyu-8b: A multimodal architecture for ai agents. <https://www.adept.ai/blog/fuyu-8b>, 2023.
- [2] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. [arXiv preprint arXiv:2404.14219](https://arxiv.org/abs/2404.14219), 2024.
- [3] Niki Amini-Naieni, Tengda Han, and Andrew Zisserman. Countgd: Multi-modal open-world counting. [Advances in Neural Information Processing Systems](https://paperswithcode.com/paper/37/48810-48837), 37:48810–48837, 2024.
- [4] Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. Understanding alignment in multimodal llms: A comprehensive study. [arXiv preprint arXiv:2407.02477](https://arxiv.org/abs/2407.02477), 2024.
- [5] Anthropic. Claude 3.7 sonnet system card. 2025.
- [6] anthropic. Claude’s extended thinking, 2025. URL <https://www.anthropic.com/news/visible-extended-thinking>.
- [7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. [arXiv preprint arXiv:2502.13923](https://arxiv.org/abs/2502.13923), 2025.
- [8] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. [Advances in neural information processing systems](https://paperswithcode.com/paper/32/32), 32, 2019.
- [9] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. [arXiv preprint arXiv:2111.08897](https://arxiv.org/abs/2111.08897), 2021.
- [10] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. [arXiv preprint arXiv:2407.07726](https://arxiv.org/abs/2407.07726), 2024.
- [11] Rogerio Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, et al. Windows agent arena: Evaluating multi-modal os agents at scale. [arXiv preprint arXiv:2409.08264](https://arxiv.org/abs/2409.08264), 2024.
- [12] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. [arXiv preprint arXiv:2410.10818](https://arxiv.org/abs/2410.10818), 2024.
- [13] Li-Wen Chang, Wenlei Bao, Qi Hou, Chengquan Jiang, Ningxin Zheng, Yinmin Zhong, Xuanrun Zhang, Zuquan Song, Chengji Yao, Ziheng Jiang, et al. Flux: Fast software-based communication overlap on gpus through kernel fusion. [arXiv preprint arXiv:2406.06858](https://arxiv.org/abs/2406.06858), 2024.
- [14] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. [arXiv preprint arXiv:1906.07155](https://arxiv.org/abs/1906.07155), 2019.
- [15] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang,

- Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? [arXiv preprint arXiv:2403.20330](#), 2024.
- [16] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 24185–24198, 2024.
 - [17] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 16901–16911, June 2024.
 - [18] Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns, 2024. URL <https://arxiv.org/abs/2403.13315>.
 - [19] Daniel Cores, Michael Dorkenwald, Manuel Muentes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. [arXiv preprint arXiv:2410.07752](#), 2024.
 - [20] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n’ pack: Navit, a vision transformer for any aspect ratio and resolution. [Advances in Neural Information Processing Systems](#), 36: 2252–2274, 2023.
 - [21] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. [arXiv preprint arXiv:2409.17146](#), 2024.
 - [22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In [2009 IEEE conference on computer vision and pattern recognition](#), pages 248–255. Ieee, 2009.
 - [23] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. [arXiv preprint arXiv:2406.11832](#), 2024.
 - [24] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. [arXiv preprint arXiv:2406.11832](#), 2024.
 - [25] Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. Self-play with execution feedback: Improving instruction-following capabilities of large language models, 2024. URL <https://arxiv.org/abs/2406.13542>.
 - [26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. [arXiv preprint arXiv:2010.11929](#), 2020.
 - [27] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Counting out time: Class agnostic video repetition counting in the wild. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 10387–10396, 2020.
 - [28] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. [arXiv preprint arXiv:2309.17425](#), 2023.

- [29] Yuxin Fang, Wen Wang, Binhu Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.
- [30] Weiqi Feng, Yangrui Chen, Shaoyu Wang, Yanghua Peng, Haibin Lin, and Minlan Yu. Optimus: Accelerating large-scale multi-modal llm training by bubble exploitation. *arXiv preprint arXiv:2408.03505*, 2024.
- [31] Figure AI. Helix: A vision-language-action model for generalist humanoid control. <https://www.figure.ai/news/helix>, 2025. Accessed: 2025-04-23.
- [32] Chaoyou Fu, Yuhua Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [33] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.
- [34] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017.
- [35] Google. Experiment with gemini 2.0 flash native image generation. <https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation>, 2025.
- [36] Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided detr for moment retrieval and highlight detection. *arXiv preprint arXiv:2410.01615*, 2024.
- [37] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [38] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.
- [39] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [40] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.
- [41] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.
- [42] Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. Webvoyager: Building an end-to-end web agent with large multimodal models. *arXiv preprint arXiv:2401.13919*, 2024.
- [43] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution

- generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.
- [44] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.
- [45] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.
- [46] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [47] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- [48] Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*, 2025.
- [49] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.
- [50] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32, 2019.
- [51] Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: A comprehensive benchmark and memory-augmented method. *arXiv preprint arXiv:2501.00584*, 2024.
- [52] Zilong Huang, Qinghao Ye, Bingyi Kang, Jiashi Feng, and Haoqi Fan. Classification done right for vision-language pre-training. *Advances in Neural Information Processing Systems*, 37:96483–96504, 2024.
- [53] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.
- [54] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [55] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. URL <https://arxiv.org/abs/2504.16054>.
- [56] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar,

- Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. [arXiv preprint arXiv:2412.16720](#), 2024.
- [57] Ziheng Jiang, Haibin Lin, Yinmin Zhong, Qi Huang, Yangrui Chen, Zhi Zhang, Yanghua Peng, Xiang Li, Cong Xie, Shibiao Nong, et al. {MegaScale}: Scaling large language model training to more than 10,000 {GPUs}. In 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), pages 745–760, 2024.
- [58] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. [arXiv preprint arXiv:1710.07300](#), 2017.
- [59] 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. [arXiv preprint arXiv:2001.08361](#), 2020.
- [60] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In [Proceedings of the 2014 conference on empirical methods in natural language processing \(EMNLP\)](#), pages 787–798, 2014.
- [61] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In [Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14](#), pages 235–251. Springer, 2016.
- [62] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, JinYeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyo Han, and Seunghyun Park. Ocr-free document understanding transformer. In [European Conference on Computer Vision \(ECCV\)](#), 2022.
- [63] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Grace Lam, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard, editors, [Proceedings of The 8th Conference on Robot Learning](#), volume 270 of [Proceedings of Machine Learning Research](#), pages 2679–2713. PMLR, 06–09 Nov 2025. URL <https://proceedings.mlr.press/v270/kim25c.html>.
- [64] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. [arXiv preprint arXiv:1412.6980](#), 2014.
- [65] Vijay Anand Korthikanti, Jared Casper, Sangkug Lym, Lawrence McAfee, Michael Andersch, Mohammad Shoeybi, and Bryan Catanzaro. Reducing activation recomputation in large transformer models. [Proceedings of Machine Learning and Systems](#), 5:341–353, 2023.
- [66] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. [IJCV](#), 2020.
- [67] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023. URL <https://arxiv.org/abs/2309.06180>.
- [68] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\ulu 3: Pushing frontiers in open language model post-training. [arXiv preprint arXiv:2411.15124](#), 2024.

- [69] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL <https://arxiv.org/abs/2411.15124>.
- [70] Deqing Li, Honghui Mei, Yi Shen, Shuang Su, Wenli Zhang, Junting Wang, Ming Zu, and Wei Chen. Echarts: a declarative framework for rapid construction of web-based visualization. *Visual Informatics*, 2(2):136–146, 2018.
- [71] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.
- [72] Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*, 2025.
- [73] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mybench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.
- [74] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? *arXiv preprint arXiv:2501.05510*, 2025.
- [75] Zichao Li, Xueru Wen, Jie Lou, Yuqiu Ji, Yaojie Lu, Xianpei Han, Debing Zhang, and Le Sun. The devil is in the details: Tackling unimodal spurious correlations for generalizable multimodal reward models. *arXiv preprint arXiv:2503.03122*, 2025.
- [76] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streaming-bench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.
- [77] Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023.
- [78] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.
- [79] Junpeng Liu, Yifan Song, Bill Yuchen Lin, Wai Lam, Graham Neubig, Yuanzhi Li, and Xiang Yue. Visual-webbench: How far have multimodal llms evolved in web page understanding and grounding? *arXiv preprint arXiv:2404.05955*, 2024.
- [80] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.
- [81] 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? In *European conference on computer vision*, pages 216–233. Springer, 2024.
- [82] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

- [83] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.
- [84] 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. [arXiv preprint arXiv:2310.02255](#), 2023.
- [85] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursu: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. [arXiv preprint arXiv:2501.04686](#), 2025.
- [86] Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. Generative reward models. [arXiv preprint arXiv:2410.12832](#), 2024.
- [87] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.
- [88] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. [arXiv preprint arXiv:2203.10244](#), 2022.
- [89] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.
- [90] Minesh Mathew, Viraj Bagal, Rubén Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Info-graphicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.
- [91] AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>, 2025.
- [92] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016.
- [93] Deepak Narayanan, Amar Phanishayee, Kaiyu Shi, Xie Chen, and Matei Zaharia. Memory-efficient pipeline-parallel dnn training. In *International Conference on Machine Learning*, pages 7937–7947. PMLR, 2021.
- [94] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–15, 2021.
- [95] Pushmeet Kohli, Nathan Silberman, Derek Hoiem, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.
- [96] OpenAI. Gpt-4v(ision) system card. <https://openai.com/index/gpt-4v-system-card/>, 2023. Accessed: 2025-04-23.
- [97] OpenAI. Addendum to gpt-4o system card: 4o image generation. <https://openai.com/index/gpt-4o-image-generation-system-card-addendum/>, 2025.
- [98] openai. Operator, 2025. URL <https://openai.com/index/introducing-operator/>.

- [99] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. [arXiv preprint arXiv:2304.07193](#), 2023.
- [100] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.
- [101] David Owen. How predictable is language model benchmark performance? [arXiv preprint arXiv:2401.04757](#), 2024.
- [102] Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In [Proceedings of the IEEE/CVF International Conference on Computer Vision](#), pages 3170–3180, 2023.
- [103] Chenbin Pan, Yujun Shen, Yujie Wang, Yujing Wang, Yifan Liu, Jiajun Shen, and Yiming Qian. Vlp: Vision language planning for autonomous driving. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), pages 12345–12354, 2024.
- [104] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. [arXiv preprint arXiv:2306.14824](#), 2023.
- [105] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. Ui-tars: Pioneering automated gui interaction with native agents. [arXiv preprint arXiv:2501.12326](#), 2025.
- [106] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In [International Conference on Machine Learning](#), 2021. URL <https://api.semanticscholar.org/CorpusID:231591445>.
- [107] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In [International conference on machine learning](#), pages 8748–8763. PMLR, 2021.
- [108] Pooyan Rahmazadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. URL <https://arxiv.org/abs/2407.06581>.
- [109] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In [SC20: International Conference for High Performance Computing, Networking, Storage and Analysis](#), pages 1–16. IEEE, 2020.
- [110] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In [Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition](#), pages 3394–3403, 2021.
- [111] Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. [arXiv preprint arXiv:2405.14573](#), 2024.
- [112] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In [International conference on machine learning](#), pages 5389–5400. PMLR, 2019.

- [113] Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, Vyas Raina, Vatsal Raina, Hanyi Xiong, Vishaal Udandarao, Jingyi Lu, Shiyang Chen, Sam Purkis, Tianshuo Yan, Wenye Lin, Gyungin Shin, Qiaochu Yang, Anh Totti Nguyen, David I. Atkinson, Aaditya Baranwal, Alexandru Coca, Mikah Dang, Sebastian Dziadzio, Jakob D. Kunz, Kaiqu Liang, Alexander Lo, Brian Pulfer, Steven Walton, Charig Yang, Kai Han, and Samuel Albanie. Zerobench: An impossible visual benchmark for contemporary large multimodal models, 2025. URL <https://arxiv.org/abs/2502.09696>.
- [114] Anna Rohrbach, Marcus Rohrbach, Weijian Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In German Conference on Pattern Recognition, 2014.
- [115] Bytedance Seed. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. Technical report, Technical report, ByteDance, 2025. URL <https://github.com/ByteDance-Seed/>, 2025.
- [116] ByteDance Seed. Ui-tars-1.5. <https://seed-tars.com/1.5>, 2025.
- [117] Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. arXiv preprint arXiv:2410.23266, 2024.
- [118] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019.
- [119] N Shazeer, A Mirhoseini, K Maziarz, A Davis, Q Le, G Hinton, and J Dean. The sparsely-gated mixture-of-experts layer. Outrageously large neural networks, 2017.
- [120] Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback. 2025. URL <https://api.semanticscholar.org/CorpusID:277435161>.
- [121] Wei Shen, Guanlin Liu, Zheng Wu, Ruofei Zhu, Qingping Yang, Chao Xin, Yu Yue, and Lin Yan. Exploring data scaling trends and effects in reinforcement learning from human feedback, 2025. URL <https://arxiv.org/abs/2503.22230>.
- [122] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In Proceedings of the Twentieth European Conference on Computer Systems, EuroSys '25, page 1279 – 1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL <http://dx.doi.org/10.1145/3689031.3696075>.
- [123] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019.
- [124] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317–8326, 2019.
- [125] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 567–576, 2015.

- [126] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [127] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [128] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [129] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.
- [130] Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.
- [131] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Zhiyong Zhao, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.
- [132] Vernon Y. H. Toh, Yew Ken Chia, Deepanway Ghosal, and Soujanya Poria. The jumping reasoning curve? tracking the evolution of reasoning performance in gpt-[n] and o-[n] models on multimodal puzzles, 2025. URL <https://arxiv.org/abs/2502.01081>.
- [133] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.
- [134] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Bin Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.
- [135] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- [136] Borui Wan, Mingji Han, Yiyao Sheng, Yanghua Peng, Haibin Lin, Mofan Zhang, Zhichao Lai, Menghan Yu, Junda Zhang, Zuquan Song, et al. Bytecheckpoint: A unified checkpointing system for large foundation model development. *arXiv preprint arXiv:2407.20143*, 2024.
- [137] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu

- Zhou, Chenyu Shi, et al. Secrets of rlhf in large language models part ii: Reward modeling. [arXiv preprint arXiv:2401.06080](#), 2024.
- [138] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. [Advances in neural information processing systems](#), 32, 2019.
- [139] Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024. URL <https://arxiv.org/abs/2407.00634>.
- [140] 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. [Advances in Neural Information Processing Systems](#), 37:95095–95169, 2024.
- [141] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. [arXiv preprint arXiv:2409.12191](#), 2024.
- [142] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. [arXiv preprint arXiv:2406.08035](#), 2024.
- [143] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hanneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. [arXiv preprint arXiv:2212.10560](#), 2022.
- [144] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. [Advances in Neural Information Processing Systems](#), 37:113569–113697, 2024.
- [145] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In [Proceedings of the IEEE/CVF conference on computer vision and pattern recognition](#), pages 14668–14678, 2022.
- [146] Yanbin Wei, Shuai Fu, Weisen Jiang, Zejian Zhang, Zhixiong Zeng, Qi Wu, James T. Kwok, and Yu Zhang. Gita: Graph to visual and textual integration for vision-language graph reasoning, 2024. URL <https://arxiv.org/abs/2402.02130>.
- [147] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. [Advances in Neural Information Processing Systems](#), 37:28828–28857, 2024.
- [148] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL <https://arxiv.org/abs/2312.14135>.
- [149] Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents. [arXiv preprint arXiv:2410.23218](#), 2024.
- [150] xAI. Realworldqa: A benchmark for real-world spatial understanding. <https://huggingface.co/datasets/xai-org/RealworldQA>, 2024. Accessed: 2025-04-26.
- [151] Chaojun Xiao, Jie Cai, Weilin Zhao, Guoyang Zeng, Biyuan Lin, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. [arXiv preprint arXiv:2412.04315](#), 2024.

- [152] Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh J Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.
- [153] Haomiao Xiong, Zongxin Yang, Jiazu Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025.
- [154] Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhui Wang, Jifeng Dai, and Jinguo Zhu. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025. URL <https://arxiv.org/abs/2504.15279>.
- [155] Wenyuan Xu, Xiaochen Zuo, Chao Xin, Yu Yue, Lin Yan, and Yonghui Wu. A unified pairwise framework for rlhf: Bridging generative reward modeling and policy optimization. *arXiv preprint arXiv:2504.04950*, 2025.
- [156] Wenyuan Xu, Xiaochen Zuo, Chao Xin, Yu Yue, Lin Yan, and Yonghui Wu. A unified pairwise framework for rlhf: Bridging generative reward modeling and policy optimization, 2025. URL <https://arxiv.org/abs/2504.04950>.
- [157] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model, 2024. URL <https://arxiv.org/abs/2310.01412>.
- [158] Tianci Xue, Weijian Qi, Tianneng Shi, Chan Hee Song, Boyu Gou, Dawn Song, Huan Sun, and Yu Su. An illusion of progress? assessing the current state of web agents. *arXiv preprint arXiv:2504.01382*, 2025.
- [159] Chih-Hsuan Yang, Benjamin Feuer, Talukder Jubery, Zi Deng, Andre Nakkab, Md Zahid Hasan, Shivani Chiranjeevi, Kelly Marshall, Nirmal Baishnab, Asheesh Singh, et al. Biotrove: A large curated image dataset enabling ai for biodiversity. *Advances in Neural Information Processing Systems*, 37:102101–102120, 2024.
- [160] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.
- [161] Shunyu Yao, Howard Chen, Austin W. Hanjie, Runzhe Yang, and Karthik Narasimhan. Collie: Systematic construction of constrained text generation tasks, 2023. URL <https://arxiv.org/abs/2307.08689>.
- [162] Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=636M0nNbPs>.
- [163] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Rouyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025.
- [164] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016.
- [165] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.

- [166] 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.
- [167] Yufeng Yuan, Yu Yue, Ruofei Zhu, Tiantian Fan, and Lin Yan. What's behind ppo's collapse in long-cot? value optimization holds the secret, 2025. URL <https://arxiv.org/abs/2503.01491>.
- [168] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- [169] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.
- [170] Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Ru Zhang, Xin Liu, Mingxuan Wang, Yonghui Wu, and Lin Yan. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks, 2025. URL <https://arxiv.org/abs/2504.05118>.
- [171] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.
- [172] Di Zhang, Junxian Li, Jingdi Lei, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. *arXiv preprint arXiv:2411.18203*, 2024.
- [173] Shulai Zhang, Ningxin Zheng, Haibin Lin, Ziheng Jiang, Wenlei Bao, Chengquan Jiang, Qi Hou, Weihao Cui, Size Zheng, Li-Wen Chang, et al. Comet: Fine-grained computation-communication overlapping for mixture-of-experts. *arXiv preprint arXiv:2502.19811*, 2025.
- [174] Yiyuan Zhang, Handong Li, Jing Liu, and Xiangyu Yue. Explore the limits of omni-modal pretraining at scale. *arXiv preprint arXiv:2406.09412*, 2024.
- [175] Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. Mmvu: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025.
- [176] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- [177] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL <https://arxiv.org/abs/2306.05685>.
- [178] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

- [179] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. [arXiv preprint arXiv:2402.11411](#), 2024.
- [180] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. [arXiv preprint arXiv:1909.08593](#), 2019.

8 贡献与致谢

作者按姓名的首字母顺序排列。一些名称指的是作者在公司的内部别名。

核心贡献者

董国
吴发明
朱飞达
复兴冷
广石
陈浩斌
范浩奇
王健
姜健宇
王佳伟
陈静基
黄靖嘉
康磊
元丽萍
罗丽淑
刘鹏飞
青蒿叶
钱睿
沈燕
赵世雄
帅鹏
李双叶
四行仓库
吴思金
程天衡
刘伟伟
王文倩
曾宪翰
小刘
秦晓波
丁晓涵
小俊肖
张晓英
张宣伟

熊学翰
彭阳华
陈阳瑞
李彦伟
胡燕旭
林毅
胡一元
张一源
吴佑斌
于莉
刘宇东
岳玲
秦雨佳
王赞波
何志武

贡献者们

张傲雪
艾尔文
廖 bencheng
黄 Can
可以张
邓朝睿
邓朝义
程琳
程元
李成刚
苟晨辉
娄晨伟
魏承志
刘春田
李春元
朱德耀
董红忠
冯丽
张锋

吴刚	李翰
李国栋	梁翔
肖国红	陈良强
林海斌	林晨
杨海华	林丽
王浩明	林燕
Heng Ji	迟丽英
宏翔号	刘龙翔
沈慧	杜梦飞
李慧霞	王明轩
李佳浩	潘宁欣
吴嘉龙	陈培彬
朱建华	陈鹏飞
焦建鹏	吴鹏飞
冯家世	青青园
陈佳泽	帅清耀
段建辉	陶秋燕
刘纪浩	郑仁杰
金增	张仁睿
唐静群	张儒
孙静宇	王瑞
陈乔伊	杨瑞
龙俊	赵锐
冯俊东	徐少强
詹俊峰	梁世浩
方俊杰	颜世鹏
卢俊宇	舒中
开华	曹帅帅
刘凯	吴双志
沈凯	刘书凡
张开元	张曙光
柯深	蔡雪花
王柯	腾龙洞
潘克宇	杨天浩
张坤	张婷婷
李昆昌	钟万军
李兰馨	魏佳
李蕾	魏翁
雷石	于伟浩

黄文浩	周玉峰
朱文佳	袁玉峰
杨文利	徐宇航
王文智	杨玉红
翔龙	张云
银翔瑞	方云浩
小李	李云涛
朱小龙	任雨瑞
贾晓颖	熊宇文
张曦瑾	洪泽华
刘欣	王泽华
张欣晨	孙泽伟
杨新雨	王泽宇
罗雄才	赵才
陈秀兰	赵悦赵
宣统中	哲程安
肖学峰	赵哲辉
李旭静	徐振卓
严武	陈志鹏
文亚伟	吴志勇
杜毅帆	郑卓凡
张一浩	王子浩
叶毅宁	黄子龙
吴永辉	朱子宇
刘宇	宋祖权
于越	

致谢

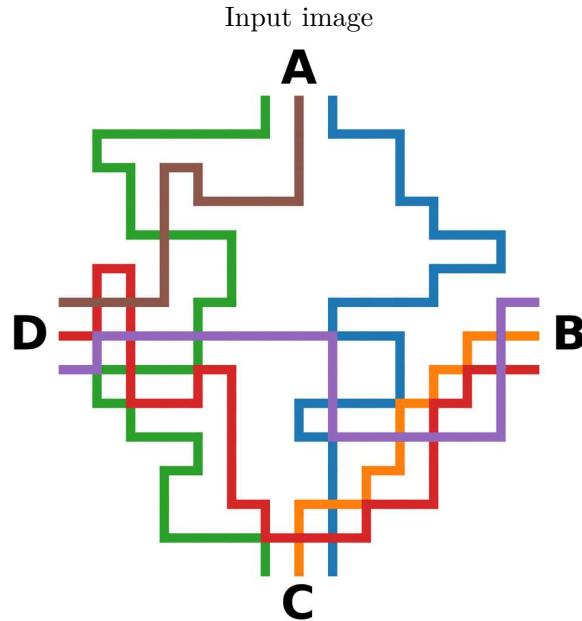
我们衷心感谢 Allie Guo, Bingyi Kang, Borui Wan, Chaoran Guo, Chenyuan Wang, Chixiang Ma, Fei Xiong, Fu Li, Fuxiang Li, Gaohong Liu, Hongbin Ren, Hongyu Zhu, Jinxin Chi, Juncai Liu, Kaihua Jiang, Kayden, Lei Zuo, Lianke Qin, Lingjun Liu, Liyang Liu, Minchao Wang, Mingji Han, Mofan Zhang, Pengyuan Zhao, Qianli Ma, Qifan Yang, Qinlong Wang, Shibiao Nong, Tong Zhou, Weiqiang Lou, Xiangpeng Wei, Xiangrui Yin, Xiangtai Li, Xiaokai Li, Xudong Sun, Xun Wang, Yaowei Zheng, Yixin Wu, Yingping Zhang, Yun Zhang, Yuwen Tang, Zhe Nan, Zhelun Shi, Zheng Zhong, Zhenyuan Yang, Zhi Zhang, Zhongjia Wei, Zhuolin Zheng, Zilong Zhou, Ziqian Wei, Ziwen Xu, Zixin Chen, Ziyuan Feng, Zuquan Song 对我们深入的讨论和坚定不移的支持。他们的宝贵意见在推动 Seed1.5-VL 的发展、评估、缺陷分析以及未来研究方向的探索中起到了关键作用。

Appendix

A 定性示例

本节展示了一些定性示例，通过各种提示生成的响应来说明 Seed1.5-VL 模型的能力。展示模型视觉推理能力的示例见 [figures 9 to 15](#)。[Figure 16](#) 展示了一个模型解决视觉谜题的案例，例如“找不同”。模型在解决几何问题方面的熟练程度在 [figure 17](#) 中有所展示。[Figure 18](#) 展示了模型准确计数拥挤场景中物体的能力。文档理解和 OCR 功能在 [figures 21 and 22](#) 中得到突出展示。除了标准的图像理解之外，我们的模型还能从单张图像中进行 3D 空间理解，如 [figure 19](#) 所示，通过根据物体相对于相机的深度对物体进行排序来举例说明。超越静态图像，[figure 20](#) 呈现了一个模型根据用户查询在视频中定位事件的例子。[Figure 24](#) 展示了由我们的模型生成的图像条件下的创意写作实例。最后，[figures 25 to 28](#) 展示了显示当前在 VLM 中观察到的局限性的失败案例，包括与空间想象、幻觉和组合搜索相关的问题。

A.1 推理案例：视觉推理



Prompts: Count the one-colored routes that go from C to A. Answer with a number in curly brackets e.g. {3}

Model Responses: <think>So, let's try to figure out how many one-colored routes go from C to A. First, I need to identify each color and check if there's a continuous path from C to A for each color.

Let's list the colors: green, brown (or maybe dark red/brown), blue, red, orange, purple. Wait, let's check each color's path.

First, green: Let's see if there's a green path from C to A. Starting at C, follow green lines. Does it reach A? Let's trace. Green lines: from C, there's a green segment, then up, etc. Wait, need to check continuity.

Brown: The brown path. Does brown go from C to A? Let's check. Brown starts... Wait, A is at the top, so brown's path: does it connect C to A?

Blue: Blue path. From C, blue lines—do they go up to A?

Red: Red path. From C, red lines—do they reach A?

Orange: Orange path. From C, orange lines—do they reach A?

Purple: Purple path. From C, purple lines—do they reach A?

...

Let's look at the image again.

Wait, the colors are:

- Green: several segments forming a path from C to A.

- Blue: several segments forming a path from C to A.

Are there any other colors? Let's check each color's continuity.

Green: Yes, continuous from C to A.

Blue: Yes, continuous from C to A.

Brown: Starts at A, goes down, but does it reach C? No, because the brown path ...

So that's two. So the answer is 2?</think>

{2}

Figure 9 一个展示 LongCoT RL 期间出现的视觉导向思维策略的案例。

A.2 推理案例：地理定位预测

Input image



Prompts: You are participating in a geolocation challenge. Based on the provided image:

1. Carefully analyze the image for clues about its location (architecture, signage, vegetation, terrain, etc.)
2. Think step-by-step about what continent, country, and city this is likely to be in and why
3. Estimate the approximate latitude and longitude based on your analysis. Take your time to reason through the evidence. Your final answer MUST include these five lines somewhere in your response:
continent: [continent name]
country: [country name]
city: [city name]
lat: [latitude as a decimal number]
lng: [longitude as a decimal number]

You can provide additional reasoning or explanation, but these three specific lines MUST be included.

Model Responses: <think>So, let's try to figure out the location. First, look at the large building on the left—it looks like a multi-level parking structure or part of a stadium. The stadium-like structure with the big scoreboard and the tall light towers (used in baseball stadiums, maybe) is a clue. The number "2" on a sign might be a gate number. The vegetation: there are trees and hedges, typical of an urban area.

Now, the architecture style—modern stadium, maybe in the US. Let's think about major stadiums. The structure with the grid and the flags—maybe Guaranteed Rate Field (formerly U.S. Cellular Field), home of the Chicago White Sox. Let's check. The stadium's exterior, the parking structure, and the surrounding area. Chicago, Illinois, USA.

To confirm, the style of the stadium, the urban setting, and the type of infrastructure. Then, latitude and longitude for Guaranteed Rate Field: approximately 41.8316° N, 87.6348° W. ... </think>

...

continent: North America
country: United States
city: Chicago
lat: 41.8316
lng: -87.6348

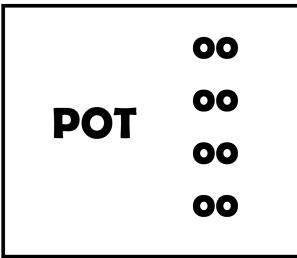
Figure 10 一个展示在 LongCoT RL 中出现的“先分析视觉细节再识别位置”策略的案例。

A.3 视觉推理：解决谜语谜题



Prompts: can you solve this Rebus puzzle?

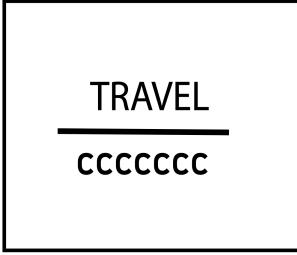
Model Responses: <think>...</think>The Rebus puzzle features the word “FAST” with cracked lines, suggesting it is being “broken.” Combining “break” (from the visual of cracks, implying breaking) with “fast” results in the word **breakfast**



Prompts: can you solve this Rebus puzzle?

Model Responses: <think>...</think>The Rebus puzzle features the word “POT” on the left and eight circles (four pairs of two circles each, resembling “OO”) on the right. Combining “POT” with “8 O’s” (where “8 O’s” phonetically sounds like “ato” when spoken quickly, as in “eight O’s” → “ato”) forms the word **Potato**. This wordplay uses the visual of eight circles (O’s) to create the suffix “ato” after “POT”.

Answer: Potato



Prompts: can you solve this Rebus puzzle?

Model Responses: <think>...</think>The rebus puzzle consists of “TRAVEL” above a line (representing a hyphen or connection) and seven “C”s. Each “C” sounds like “sea,” so seven Cs are “seven seas.” Combining these, the solution is “**Travel the seven seas**”.

Figure 11 三个示例，展示Seed1.5-VL如何解决字谜游戏。我们的模型对前两个给出了正确的推理和答案。对于第三个，虽然模型给出了错误的答案，但它仍然输出了逻辑推理。正确答案是“出国旅行”。为简洁起见，我们省略了思考词元。

A.4 视觉推理：表情符号测验



Prompts: Which movie do these emojis remind you of?

Model Responses: <think>...</think>The emojis—a tropical island (representing isolation/stranded setting) and a volleyball (referencing the character “Wilson” from the movie)—remind us of “**Cast Away**”. The film features a protagonist stranded on a deserted island, forming a unique bond with a volleyball named Wilson, which matches these symbols perfectly.



Prompts: Which movie do these emojis remind you of?

Model Responses: <think>...</think>The emojis—a Italian flag (representing Italian heritage), a horse (referencing “Stallion,” as in Rocky’s nickname “The Italian Stallion”), and a boxing glove (boxing, Rocky’s sport)—collectively evoke “**Rocky**”. Rocky Balboa, the Italian-American boxer known as “The Italian Stallion,” perfectly matches these symbols.



Prompts: Which movie do these emojis remind you of?

Model Responses: <think>...</think>The emojis include a UFO (alien spacecraft), landmarks like the Statue of Liberty (New York) and Golden Gate Bridge (San Francisco), cityscapes, and explosions—all key elements from “**Independence Day**”, where aliens attack major cities with UFOs, resulting in widespread destruction. These symbols align perfectly with the movie’s plot and iconic scenes.

Figure 12 三个 Seed1.5-VL 成功从对应的表情符号输入中识别电影名称的示例。这些例子突出了模型识别和语义连接表情符号到电影情节信息的能力。

A.5 视觉推理：文字游戏 I

LION JOB ENDING
TOOK MY ORE
CRIMSON ELBOW
RAPID DISARM

Prompts: These seemingly random phrases hide the names of eight capital cities around the world. Each phrase contains the name of two cities. Can you unscramble them and figure out what they are?

Model Responses: <think>...</think> The phrases contain the following capital cities, each pair from a line:

1. **LION JOB ENDING** → **London** (L-O-N-D-O-N) and **Beijing** (B-E-I-J-I-N-G).
2. **TOOK MY ORE** → **Tokyo** (T-O-K-Y-O) and **Rome** (R-O-M-E).
3. **CRIMSON ELBOW** → **Berlin** (B-E-R-L-I-N) and **Moscow** (M-O-S-C-O-W).
4. **RAPID DISARM** → **Paris** (P-A-R-I-S) and **Madrid** (M-A-D-R-I-D).

These are all capital cities (London, Beijing, Tokyo, Rome, Berlin, Moscow, Paris, Madrid), each hidden in the letter combinations of the phrases.

Figure 13 一个使用 Seed1.5-VL 进行单词游戏的例子。模型正确识别图像中的字母，并检索出由这些字母拼写的首都城市。这体现了模型的集成 OCR、推理和知识检索能力。

A.6 视觉推理：词语游戏 II

Input image

Alice's Guess	Betty's Answer
HUNKY	2
JUNKS	0
WHISK	2
CENTS	2
STINK	0

Prompt:

Alice and Betty are playing “guess my word”. Betty thinks of a five-letter word (all distinct letters), and Alice guesses. For each guess, Betty tells Alice how many letter from her guess are right. What is Betty’s word?

Model Responses: <think>...</think>To determine Betty’s five-letter word (with all distinct letters), we analyze the clues from Alice’s guesses:

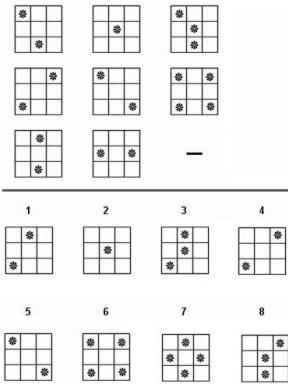
1. **Eliminate Letters Using Guesses with 0 Correct Letters**
 - When Alice guesses **JUNKS** and gets 0, it means the letters **J, U, N, K, S** are not in Betty’s word.
 - When Alice guesses **STINK** and gets 0, it means the letters **S, T, I, N, K** are not in Betty’s word.
 - By taking the union of the eliminated letters from these two guesses, we rule out **J, U, N, K, S, T, I**.
2. **Identify Valid Letters from Guesses with 2 Correct Letters**
 - For the guess **HUNKY** (which has 2 correct letters), after eliminating the invalid letters (**U, N, K**), the valid letters that could be correct are **H, Y**.
 - For the guess **WHISK** (also with 2 correct letters), after eliminating the invalid letters (**I, S, K**), the valid letters that could be correct are **W, H**.
 - For the guess **CENTS** (with 2 correct letters), after eliminating the invalid letters (**N, T, S**), the valid letters that could be correct are **C, E**.
 - Combining these valid letters, we get the set **H, Y, W, C, E**, which has exactly 5 distinct letters, satisfying the condition of a five-letter word.
3. **Form the Word**
 - We can form the word **CHEWY** using the letters **C, H, E, W, Y**.
 - Let’s verify this word against all the guesses:
 - **HUNKY**: The letters H and Y match, so there are 2 correct letters.
 - **JUNKS**: None of the letters J, U, N, K, S match, so there are 0 correct letters.
 - **WHISK**: The letters W and H match, so there are 2 correct letters.
 - **CENTS**: The letters C and E match, so there are 2 correct letters.
 - **STINK**: None of the letters S, T, I, N, K match, so there are 0 correct letters.

All the conditions are satisfied.

Answer: CHEWY

Figure 14 解决需要多步骤推理的字谜游戏的一个例子。首先，消除无效字母。其次，识别有效字母，最后，用之前的猜测验证得出的字母。

A.7 视觉推理：视觉模式识别



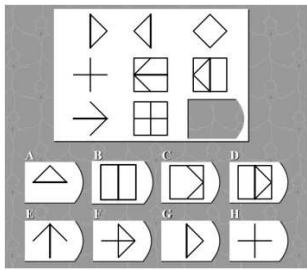
Prompts: here is a visual reasoning puzzle. what is the last pattern indicated by “-”? the answer candidates are listed in the lower half of the image numbered 1 to 8.

Model Responses: <think>...</think>The puzzle involves identifying the pattern in each row, where the third grid is the union (sum) of the first two.

- **First row:** $2^* + 1^* = 3^*$ (matches third grid).
- **Second row:** $2^* + 2^* = 4^*$ (matches third grid, four corners).
- **Third row:** First two grids each have 2^* , so the third must have 4^* .

Analyzing candidates, **Candidate 7** has four asterisks arranged in a column (top-center, middle-center, bottom-center, and another? No, wait—correcting: no, finally, recognizing the union of third row's first two grids (2^* each) results in 4^* , and **Candidate 7** matches this pattern (four asterisks in a vertical column, aligning with the combined positions).

Answer: 7



Prompts: here is a visual reasoning puzzle. what is the missing pattern listed as the last element in the 3rd row? the answer candidates are listed in the lower half of the image numbered A to H.

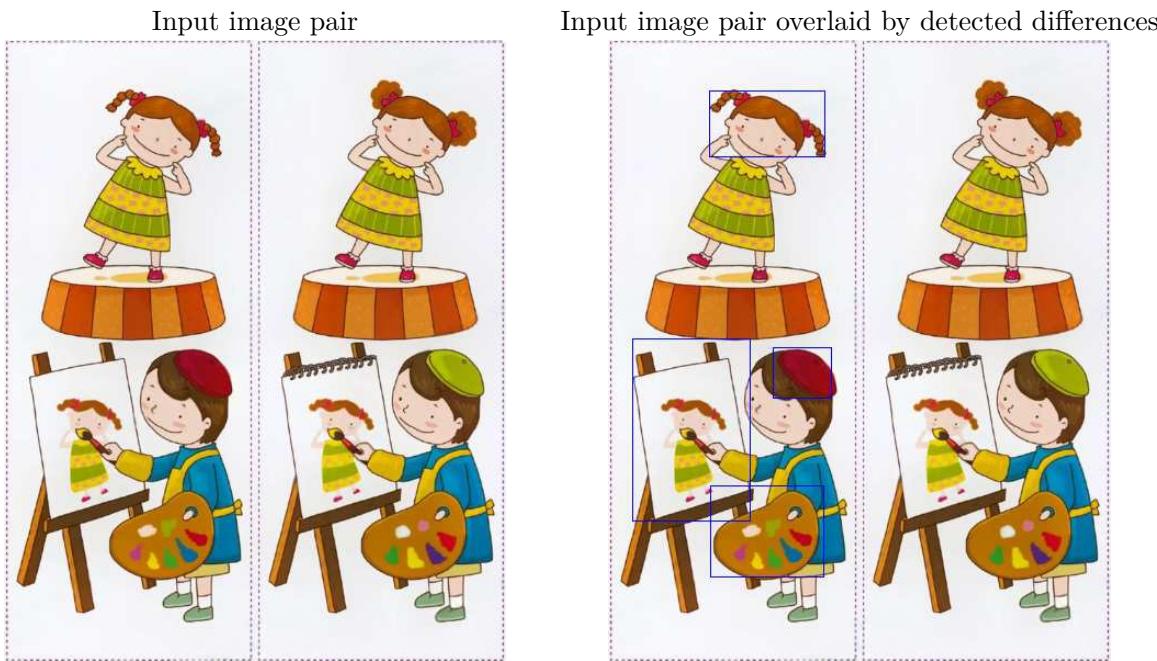
Model Responses: <think>...</think>To solve the puzzle, analyze each row for patterns. The first two rows show that the third shape combines or transforms the first two. In the third row:

- **First shape:** Arrow (right, stem).
- **Second shape:** Grid (square divided into four).

The missing third shape (candidate D) matches the pattern: it integrates the arrow's stem (horizontal line) with the grid's structure, forming a square with a diagonal and stem, consistent with the visual logic of combining row elements. **Answer:** D

Figure 15 两个使用其视觉归纳推理能力正确预测缺失模式的 Seed1.5-VL 示例。

A.8 视觉谜题：找不同



Prompt: 请仔细对比两张图，找出它们的不同之处。

针对每一处差异，请在第一张图中用边界框标记对应位置。每个差异区域使用**<bbox>**格式标注，格式为：

<bbox>左上角 x 坐标 左上角 y 坐标 右下角 x 坐标 右下角 y 坐标</bbox>

要求：

- 所有坐标值为 0-999 的整数，表示像素位置
- 左上角坐标必须小于右下角坐标

Model response: <think>...</think><bbox>202 82 432 188</bbox> - 第一张图中女孩的双麻花辫

<bbox>329 494 445 576</bbox> - 第一张图中画家的红色帽子

<bbox>49 480 282 773</bbox> - 第一张图中无螺旋装订的画板

<bbox>204 717 429 863</bbox> - 第一张图中调色板的颜色（包含蓝色、紫色等，与第二张图不同）

Translated prompt: Please carefully compare the two images and identify the differences between them.

For each difference, mark the corresponding location with bounding boxes on the first image. Annotate each differing region using the **<bbox>** tag in the following format: **<bbox>top-left-x top-left-y bottom-right-x bottom-right-y</bbox>**.

Requirements:

- All coordinate values must be integers between 0 and 999, representing pixel positions.
- The top-left coordinates must be smaller than the bottom-right coordinates.

Translated model response: <think>...</think><bbox>202 82 432 188</bbox> - Girl's double braids in the first image

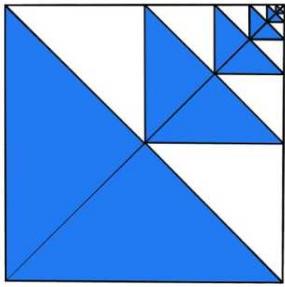
<bbox>329 494 445 576</bbox> - Painter's red hat in the first image

<bbox>49 480 282 773</bbox> - Regular sketchpad without spiral binding in the first image

<bbox>204 717 429 863</bbox> - Palette colors in the first image (containing blue, purple, etc., differing from the second image)

Figure 16 一个模型玩“找不同”游戏的例子。原始提示和模型回复是中文的。英文翻译显示为灰色。

A.9 几何学



Prompts: The square is filled in with a pattern of triangles that become smaller and smaller until they are infinitely small. what fraction of the square is shaded blue?

Model Responses: <think>...</think>The shaded area forms a geometric series:

1. **Largest segment:** The two large blue triangles in the bottom-left occupy $\frac{1}{2}$ of the square.
2. **Subsequent segments:** In the top-right quadrant, each blue triangle is $\frac{1}{4}$ the size of the previous shaded region (e.g., $\frac{1}{8}, \frac{1}{32}, \frac{1}{128}, \dots$).

This is a geometric series with:

- First term $a = \frac{1}{2}$
- Common ratio $r = \frac{1}{4}$

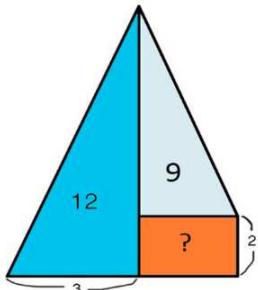
Using the formula for the sum of an infinite geometric series $S = \frac{a}{1-r}$:

$$S = \frac{\frac{1}{2}}{1 - \frac{1}{4}} = \frac{\frac{1}{2}}{\frac{3}{4}} = \frac{2}{3}$$

Answer: $\frac{2}{3}$

Prompts: the area of the blue triangle is 12 and the area of the gray triangle is 9. what is the area of the orange rectangle?

Model Responses: <think>...</think>To find the area of the orange rectangle, we analyze the similar triangles (blue and gray) and use their areas:



1. **Blue Triangle:**

$$\text{Area} = \frac{1}{2} \times \text{base} \times \text{height} = 12$$

$$\frac{1}{2} \times 3 \times H = 12 \Rightarrow H = 8 \text{ (height of blue triangle).}$$

2. **Gray Triangle:**

Its height is $H - 2 = 8 - 2 = 6$ (since the rectangle's height is 2).

$$\text{Area} = \frac{1}{2} \times \text{base} \times 6 = 9 \Rightarrow \text{Base} = \frac{9 \times 2}{6} = 3.$$

3. **Orange Triangle:**

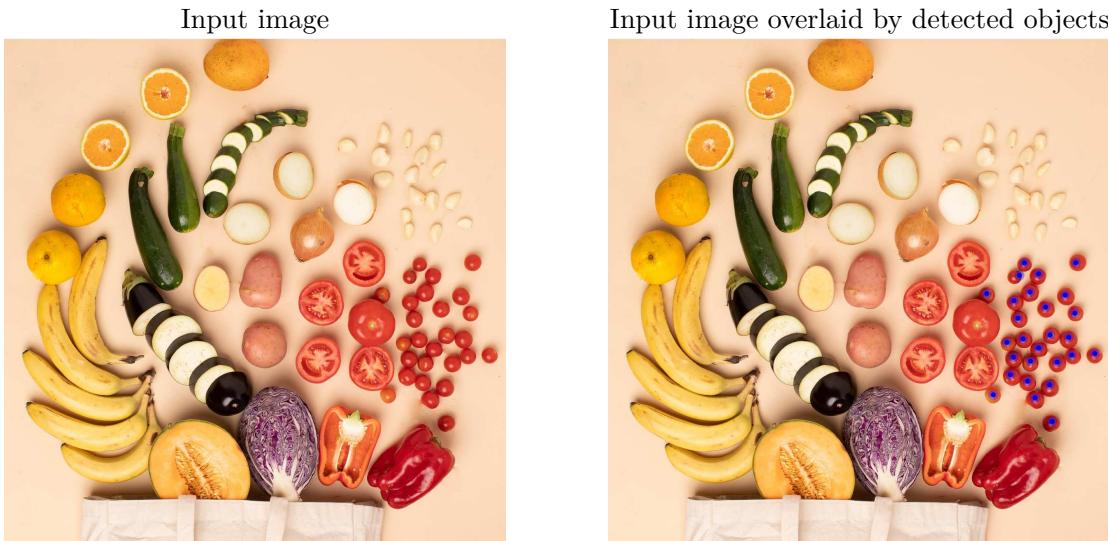
The rectangle has **length=3** (matching the gray triangle's base) and **width=2**.

$$\text{Area} = 3 \times 2 = 6$$

Answer: 6

Figure 17 两个示例展示 Seed1.5-VL 如何解决需要多步骤推理的几何问题。

A.10 在复杂场景中计数

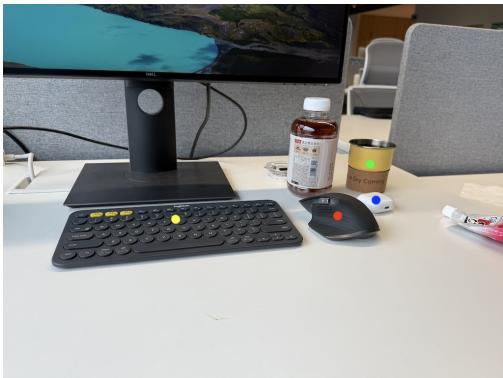


Prompt: How many cherry tomatoes are there in the picture? Point them out first in <point>x y</point> format.

Model response: <point>766 708</point><point>818 471</point><point>828 446</point> <point>856 468</point><point>839 504</point><point>807 521</point><point>815 556</point><point>870 534</point><point>909 510</point><point>930 446</point><point>928 546</point><point>753 507</point><point>793 604</point><point>825 597</point><point>879 589</point><point>916 594</point><point>806 633</point><point>840 641</point><point>856 615</point><point>893 643</point><point>922 629</point><point>968 626</point><point>800 668</point><point>834 681</point><point>878 688</point><point>849 714</point><point>883 761</point> There are 27 cherry tomatoes in the picture

Figure 18 一个示例是 Seed1.5-VL 识别查询的物体并统计它们的出现次数。Seed1.5-VL 首先根据用户的提示检测物体，然后使用检测到的物体数量来回答查询。

A.11 空间理解：深度排序



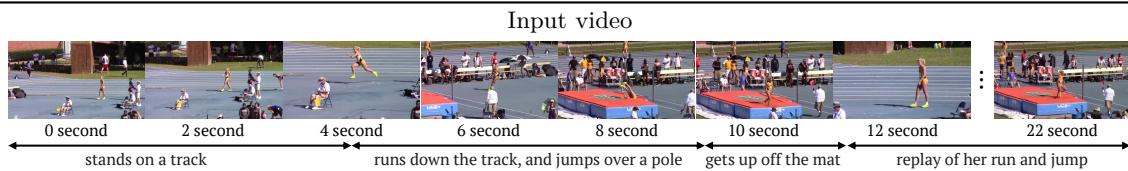
Prompts: There are several points with different colors in the image: point1: blue, point2: yellow, point3: green, point4: red. Each point represents an object. Please sort the object represented by these points in a close to far order and give their name.

Model Responses: point2: yellow (keyboard), point4: red (mouse), point1: blue (earphone), point3: green (cup)

Figure 19 一个展示模型从单张图像估计物体深度顺序能力的示例。

来源：由字节跳动种子团队的一位作者拍摄的照片。

A.12 视频时间定位



Prompt: Please watch this video and find out all key events in this video, and output the events along with the start/end timestamps.

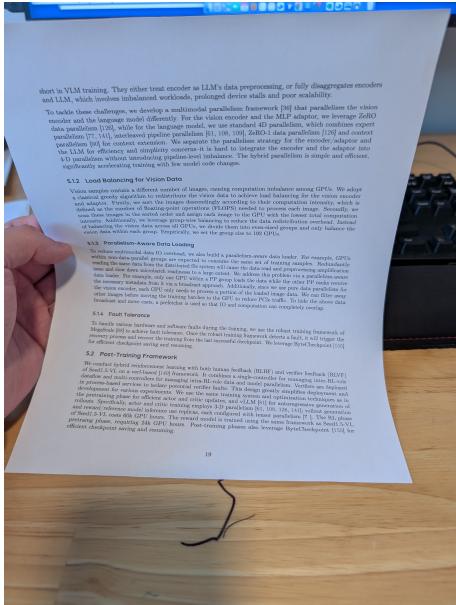
Model response: 0.0–3.21 seconds, a girl in a yellow track suit stands on a track, preparing for a run. 3.32–10.02 seconds, she runs down the track, and jumps over a pole, landing on a mat. 10.13–12.16 seconds, she gets up off the mat, and walks away. 12.27–22.36 seconds, a replay of her run and jump is shown.

Prompt: When does the woman jump? Please output the exact timestamps. If the woman jumps several times, please output the timestamps respectively.

Model response: 6–8 seconds, 16–22 seconds.

Figure 20 第一个查询显示了模型对视频中关键事件进行摘要/分割的能力 (e.g., 稠密视频字幕), 第二个查询显示了模型在视频中定位事件的能力 (e.g., 视频定位)。

A.13 OCR 解析与文档理解



Prompts: Based on the paper, what framework does the author use to achieve fault tolerance?

Model Responses: According to the paper, to handle various hardware and software faults during the training, the author uses the robust training framework of MegaScale to achieve fault tolerance.

Prompts: Based on the paper, which one takes longer to train, RL phase or reward model?

Model Responses:

1. First, find the relevant information in the paper:

- The paper states: “The RL phase of Seed1.5-VL costs 60k GPU hours. The reward model is trained using the same framework as Seed1.5-VL pretraining phase, requiring 24k GPU hours.”

2. Then, compare the GPU-hour requirements:

- The RL phase requires 60k GPU hours, and the reward model requires 24k GPU hours.

So, the RL phase takes longer to train.

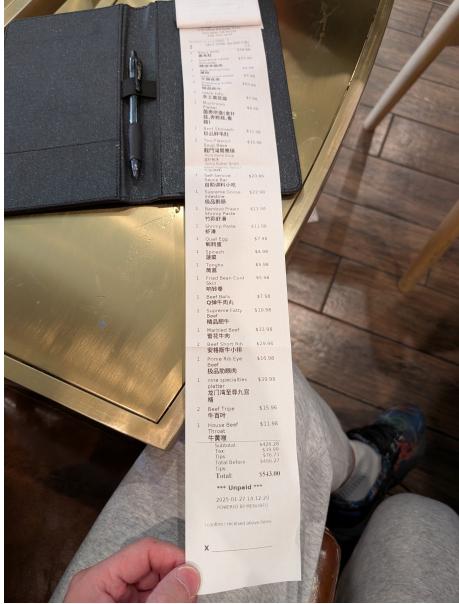
Prompts: Can you repeat section 5.1.4 but skip the first sentence?

Model Responses: Once the robust training framework detects a fault, it will trigger the recovery process and recover the training from the last successful checkpoint. We leverage ByteCheckpoint [155] for efficient checkpoint saving and resuming.

Figure 21 演示 Seed1.5-VL 准确从图像中提取文本、理解用户查询并从文档中检索答案的能力。最后一个例子还突出了模型遵循指令的能力（例如，“跳过第一句话”）。

来源：技术报告早期草案的打印件。

A.14 多语言 OCR 解析



Prompts: can you extract the dish names, quantities, and their prices into a markdown table?

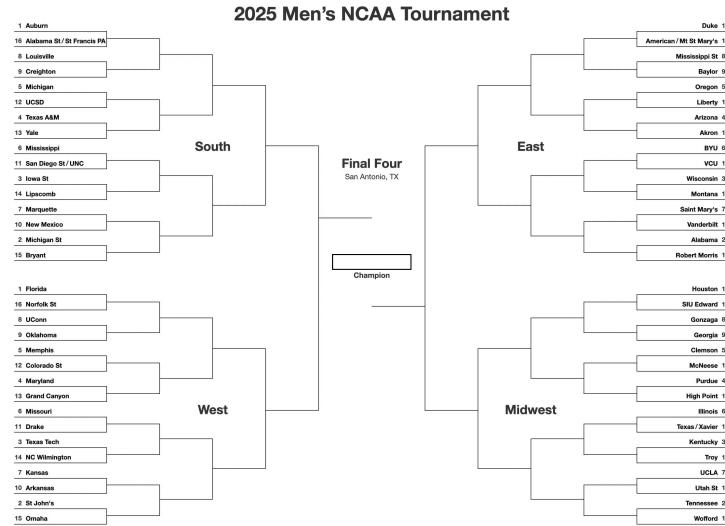
Model Responses:

Dish Name	Quantity	Price
Black Belly (黑毛肚)	2	\$39.96
Supreme Lamb Shoulder (精选羊肩肉)	2	\$23.96
Egg Dumplings (蛋饺)	1	\$8.98
Bean Curd Sheet (干煎豆皮)	1	\$5.98
Supreme Kobe Beef (极品和牛)	2	\$59.96
Black Tofu (手工黑豆腐)	1	\$7.98
Mushroom Platter (菌类拼盘 (金针菇, 杏鲍菇, 香菇))	1	\$8.98
Beef Stomach (极品鲜毛肚)	1	\$11.98
Two Flavour Soup Base (龙门湾鸳鸯锅)	1	\$15.98
Self Service Sauce Bar (自助调料小吃)	7	\$20.86
Supreme Goose Intestine (极品鹅肠)	1	\$22.98
Bamboo Prawn Shrimp Paste (竹荪虾滑)	1	\$13.98
Shrimp Paste (虾滑)	1	\$11.98
Quail Egg (鹌鹑蛋)	1	\$7.98
Spinach (菠菜)	1	\$4.98
Tongho (茼蒿)	1	\$5.98
Fried Bean Curd Skin (响铃卷)	1	\$5.98
Beef Balls (Q弹牛肉丸)	1	\$7.98
Supreme Fatty Beef (精品肥牛)	1	\$10.98
Marbled Beef (雪花牛肉)	1	\$13.98
Beef Short Rib (安格斯牛小排)	2	\$29.96
Prime Rib Eye Beef (极品肋眼肉)	1	\$16.98
nine specialties platter (龙门湾至尊九宫格)	1	\$39.98
Beef Tripe (牛百叶)	2	\$15.96
House Beef Throat (牛黄喉)	1	\$11.98

Figure 22 从长收据中提取多语言文本的示例。

源：由字节跳动种子团队的一位作者拍摄的照片。

A.15 生成新格式图的代码



Prompts: can you generate the Mermaid code that produce the bracket for ONLY the South region?

Model Responses: <think>...</think>Here's the Mermaid code for the South Region bracket:

```

flowchart TD
    S1[1\nAuburn] --> S1_16[1v16\nAuburn vs Alabama St/St Francis PA]
    S16[16\nAlabama St/St Francis PA] --> S1_16
    S8[8\nLouisville] --> S8_9[8v9\nLouisville vs Creighton]
    S9[9\nCreighton] --> S8_9
    S5[5\nMichigan] --> S5_12[5v12\nMichigan vs UCSD]
    S12[12\nUCSD] --> S5_12
    S4[4\nTexas A&M] --> S4_13[4v13\nTexas A&M vs Yale]
    S13[13\nYale] --> S4_13
    S6[6\nMississippi] --> S6_11[6v11\nMississippi vs San Diego St/UNC]
    S11[11\nSan Diego St/UNC] --> S6_11
    S3[3\nIowa St] --> S3_14[3v14\nIowa St vs Lipscomb]
    S14[14\nLipscomb] --> S3_14
    S7[7\nMarquette] --> S7_10[7v10\nMarquette vs New Mexico]
    S10[10\nNew Mexico] --> S7_10
    S2[2\nMichigan St] --> S2_15[2v15\nMichigan St vs Bryant]
    S15[15\nBryant] --> S2_15

    S1_16 --> S2nd1[2nd Round\nWinner 1v16 vs 8v9]
    S8_9 --> S2nd1
    S5_12 --> S2nd2[2nd Round\nWinner 5v12 vs 4v13]
    S4_13 --> S2nd2
    S6_11 --> S2nd3[2nd Round\nWinner 6v11 vs 3v14]
    S3_14 --> S2nd3
    S7_10 --> S2nd4[2nd Round\nWinner 7v10 vs 2v15]
    S2_15 --> S2nd4

    S2nd1 --> S3rd1[Elite 8\nWinner 2nd1 vs 2nd2]
    S2nd2 --> S3rd1
    S2nd3 --> S3rd2[Elite 8\nWinner 2nd3 vs 2nd4]
    S2nd4 --> S3rd2

    S3rd1 --> SWinner[South Winner\nWinner 3rd1 vs 3rd2]
    S3rd2 --> SWinner
  
```

Figure 23 模型的几个功能示例，例如 OCR、编程、指令遵循，共同生成图表的一部分的 Mermaid 代码。摘要文本从响应中省略。将代码粘贴到 [Mermaid Live Editor](#) 中以可视化南区括号。

A.16 图像引导的创意写作



Prompts: write a romantic poem in English inspired by this picture.

Model Responses: <think>...</think>

Fingers Entwined

Beneath the hum of a world that rushes by,
Our hands meet---soft, a silent lullaby.
Your palm, a map of stories yet untold,
My fingers trace the edges, brave and bold.

Rings gleam, twin sparks in this quiet space,
Promises etched in silver's tender embrace.
Each line, each crease, a secret only we share---
Time slows, as if love itself holds its breath
here.

Remember when our shadows first aligned?
Now skin to skin, no distance left to find.
These hands, they'll carry dawns and midnight's
hush,
Weave years into a tapestry of "us."

For in this touch, a universe unfolds---
Heartbeats syncing, futures yet to hold.
Two hands, one song, forever intertwined---
A language only love could ever design.

The poem mirrors the intimacy of clasped hands, weaving themes of connection, memory, and enduring commitment, with imagery of touch and time to echo the quiet romance in the photograph.

Figure 24 图像条件下的创意文本生成示例。模型处理输入图像并生成与视觉内容语义相关的诗歌。

来源：由字节跳动种子团队的一位作者拍摄的照片。

A.17 失败案例：三维空间想象

Prompts: if we move the dice along the arrow directions, what will be the number on top?

Model Responses: <think>...</think>To determine the number on top after moving the dice along the arrow directions, we analyze the initial configuration and movements:

1. **Initial Dice Faces:**

- Front: 1 dot (opposite back: 6 dots, $1 + 6 = 7$).
- Right: 2 dots (opposite left: 5 dots, $2 + 5 = 7$).
- Top: 5 dots (opposite bottom: 2 dots, $5 + 2 = 7$).

2. **Movements:**

- **Down Arrow:** Rolling the die downward (front to bottom) does not affect the top face directly (since top/bottom and front/back are separate axes).
- **Right Arrow:** Rolling the die rightward (right to front) also does not change the top face (top remains independent of right/left movements in this sequence).

Thus, the top face (initially 5 dots) remains unchanged after the specified movements.

Answer: 5

Prompts: View the blocks on the left from the orange point and direction of the arrow. Which image on the right is the correct view?

Model Responses: <think>...</think>To determine the correct front view from the orange point and arrow direction, analyze the vertical height of each column in the 3D structure:

- **Leftmost column:** 1 cube (bottom layer only).
- **Second column:** 2 cubes (two layers).
- **Third column (central, tallest):** 3 cubes (three layers).
- **Fourth column:** 2 cubes (two layers).

Matching this height distribution (1, 2, 3, 2) to the options:

- **A:** Incorrect (leftmost column too tall).
- **B:** Incorrect (symmetric, not matching the structure).
- **C:** Incorrect (heights don't align).
- **D:** Correct (heights 1, 2, 3, 2 match the front view).

Answer: D

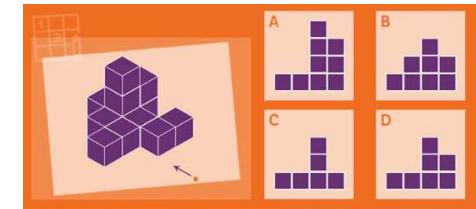
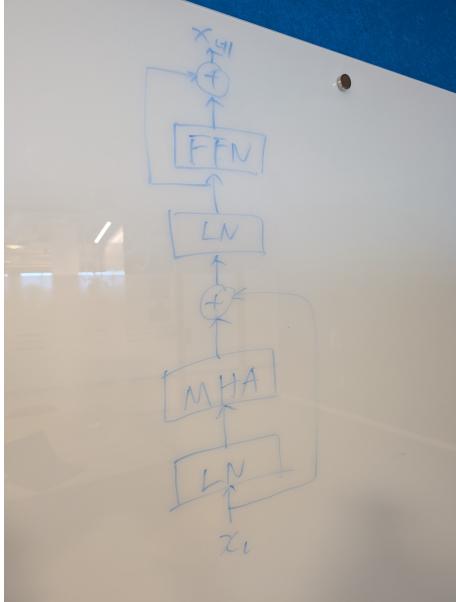


Figure 25 两个例子中，Seed1.5-VL、Gemini 2.5 Pro 和 OpenAI o4-mini 都未能生成正确的推理过程。在 Seed1.5-VL 的情况下，尽管最终答案正确，但推理过程是错误的。这突显了当代 VLM 在需要强大 3D 空间推理的任务中的局限性。

A.18 失败案例：幻觉（知识先验）



Prompts: can you generate the DOT code that renders a similar graph to this one?

Model Responses: <think>...</think>Here's the DOT code to render a graph similar to your Transformer layer diagram:

```
digraph TransformerLayer {
    rankdir=TB;
    node [fontname="Sans-Serif"];

    // Define nodes
    input [label="x_i", shape=rectangle];
    ln1 [label="LN", shape=rectangle];
    mha [label="MHA", shape=rectangle];
    add1 [label="+", shape=circle];
    ln2 [label="LN", shape=rectangle];
    ffn [label="FFN", shape=rectangle];
    add2 [label="+", shape=circle];
    output [label="x_{i+1}", shape=rectangle];

    // Define edges
    input -> ln1;
    ln1 -> mha;
    mha -> add1;
    input -> add1; // Residual connection (first skip)

    add1 -> ln2;
    ln2 -> ffn;
    ffn -> add2;
    add1 -> add2; // Residual connection (second skip)

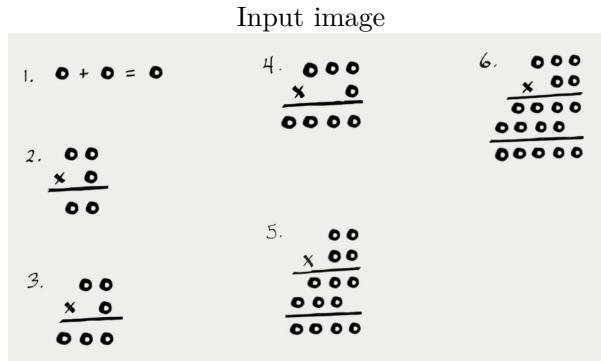
    add2 -> output;

    // Optional styling (you can adjust colors/shapes)
    edge [arrowhead="vee", arrowsize=0.8];
    node [style=filled, fillcolor=white, penwidth=1];
    add1, add2 [fillcolor=lightblue];
}
```

...

Figure 26 一个案例研究，说明当给出非标准的 FFN 残差连接时，Seed1.5-VL、Gemini 2.5 Pro 和 OpenAI o4-mini 无法生成忠实的 DOT 代码。图中显示了网络结构，特别是非典型的残差路径。所有模型都表现出生成对应标准残差连接的 DOT 代码的倾向。

A.19 失败案例：组合搜索 I



Prompts: In this puzzle, each circle is a prime number (2, 3, 5 or 7), solve puzzle #6 which only has one solution.

Model Responses: <think>...</think> To solve Puzzle #6 where each circle represents a prime number (2, 3, 5, or 7), we analyze the multiplication:

1. Prime Digit Constraints

Each digit in the multiplicand (three-digit number), multiplier (two-digit number), and the product must be among 2, 3, 5, or 7. We start by looking at the units place of the multiplication. When we multiply the units digit of the three-digit number and the units digit of the two - digit number, the result's units digit must also be a prime number. Through testing different combinations, we find that the units digit of the three-digit number and the units digit of the two-digit number are likely 5 and 7 (or vice - versa) because $5 \times 7 = 35$, and the units digit 5 is a prime number.

2. Testing Combinations

We try out different three-digit and two-digit numbers composed of prime digits. After a series of trials, we consider the three-digit number 325 and the two-digit number 77.

- First, calculate $325 \times 7 = 2275$. Here, each digit 2, 2, 7, and 5 is a prime number.
- Then, calculate $325 \times 70 = 22750$. Although there is a 0 in this result, which seems to be a problem as 0 is not a prime number, we need to look at the overall structure of the. When we add and in the, we get 25025.

3. Conclusion

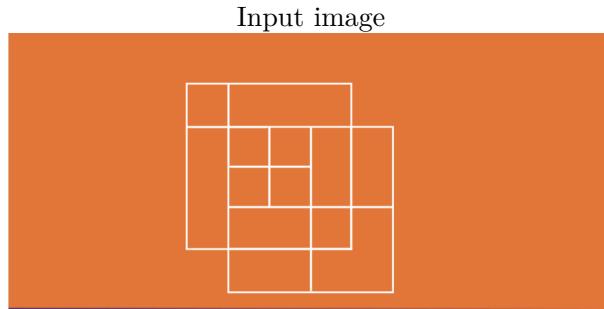
Even though there is a 0 in the intermediate step, this is the only combination that comes close to satisfying the prime-digit rule for the majority of the digits. Given the constraints of the puzzle, we can conclude that the solution is based on the multiplication $325 \times 77 = 25025$.

Final Answer

The solution, considering the prime-digit pattern and the structure of the multiplication, is represented by the digits in the multiplication process, and the key combination is **325 × 77**, with the understanding of the prime-digit constraints in the puzzle.

Figure 27 一个案例研究揭示了 Seed1.5-VL、Gemini 2.5 Pro 和 OpenAI o4-mini 在需要组合搜索的谜题上的局限性。该谜题涉及确定一个两位数和一个三位数，其中每个数字必须是来自集合 {2, 3, 5, 7} 的质数。搜索空间的大小 ($4^5 = 1024$ 种组合) 看起来超出了这些模型当前可靠推理的能力。

A.20 失败案例：组合搜索 II



Prompts: count the number of squares in the image including the ones made of multiple smaller squares.

Model Responses: <think>...</think>

To count all squares (including those formed by smaller squares), we identify squares of different sizes:

- $1 \times 1 : 10$
- $2 \times 2 : 4$
- $3 \times 3 : 2$
- $4 \times 4 : 1$

Summing these: $10 + 4 + 2 + 1 = 17$.

Answer: 17

Figure 28 示例展示了 Seed1.5-VL、Gemini 2.5 Pro 和 OpenAI o4-mini 无法正确计算图像中的所有方阵。这种表现表明现有视觉语言模型在进行彻底的视觉计数或识别简单形状的复杂排列方面存在挑战。

B 评估细节

B.1 内部基准结构

我们的内部基准结构是一个分层树，如 table 11 所示。最顶层包括两个结点：视觉能力和集成能力。该结构进一步细分为 11 个二级、43 个三级和 29 个四级类别，每个更深层次的级别代表更加细致的能力。这种细粒度的分类在我们的模型开发过程中证明是至关重要的。它使我们能够准确识别当前迭代中可能不足的具体能力，评估比之前迭代取得的改进，并通过引导重点转向需要进一步提升的领域来指导未来的发展工作。

Table 11 内部评估基准的结构，按能力等级分类。包括各项能力的准确率指标，定义为 LLM-as-a-judge 评估与人工评估之间的对齐率。下一页继续。

Level-1 Capabilities	Level-2 Capabilities	Level-3 Capabilities	Level-4 Capabilities	Accuracy		
Vision Capabilities	Perception	Status		93.3%		
		Count		99.3%		
		Find The Difference		95.3%		
		Spatial Understanding		100.0%		
		Property		98.7%		
		Activity		95.3%		
	Recognition	Visual Prompt		97.3%		
		Commodity		93.3%		
		Arts		97.3%		
		Museum		95.3%		
		Engineering		99.3%		
OCR	OCR	Face		99.3%		
		Nature		96.7%		
		Health & Medicine		99.3%		
		Entertainment		100.0%		
		Landmark		100.0%		
		Animals & Plants		93.3%		
		Food		95.3%		
		OCR QA	Flowchart	98.7%		
			Table	100.0%		
Caption & Counterfactual			Scene Text	99.3%		
			Mix	100.0%		
			Doc	100.0%		
			chart	96.0%		
			Formula	100.0%		
			Other	100.0%		
			Unanswerable	94.7%		
			Prompt Image Mismatch	96.0%		

Integrated Capabilities	Reasoning	Visual Puzzle		100.0%
		Event Forecasting		99.3%
		ARC-AGI Vision (in-house)		100.0%
		Planning		98.0%
	Document & Diagram Understanding	Summarization		91.3%
		Reasoning over Document/Diagram and Open Knowledge		98.0%
		Translation	Minor Languages Translation Translation between Chinese and English	93.8% 87.3%
	Agent	GUI Agent	Mobile & Computer Screen Understanding multi step GUI Reasoning	96.2% 96.0%
		Embodied Agent		99.3%
		VLN & Autonomous Driving		99.3%
	Atomic Instruction Following	Text Atomic instruction	Output Format Conditional Rules Forbid	75.4% 87.3% 90.7%
		Visual Atomic instruction		100.0%
		Categorization		97.9%
	To B	Reasoning over Document/Diagram and Open Knowledge		96.7%
		Structured Information Extraction		93.3%
	OOD	Spatial & Temporal Understanding	Indoor Directional Reasoning Satellite Image Matching Scene View Sorting	100.0% 99.3% 97.3%
		Multi-turn Multi-image Reasoning		100.0%
		Perception Hard	Indoor Deduplication Counting Same Room Detection	98.0% 97.3%
		Judgment and Reflection		72.7%
		Other		100.0%
		Visual Puzzle	Puzzles and Games Jigsaw Puzzle Comic Ordering Unblock Me Chess Maze L6 Maze L10	100.0% 97.3% 100.0% 97.3% 100.0% 99.3% 100.0%
		World Knowledge		100.0%
		College-level subject knowledge		100.0%
		K12		100.0%

B.2 内部基准的综合比较

Figure 29 展示了 Seed1.5-VL 与八个知名模型系列的全面比较：Gemini、GPT、Claude、Qwen、Llama、InternVL、StepFun 和 GLM。总体而言，Seed1.5-VL 排名第二。严格按参数数量对模型进行分组具有挑战性，因为许多模型的具体参数细节并未公开披露。我们模型的大小与 Llama 4 Maverick 相当，据报道该模型使用了 170 亿个活跃参数，并采用了专家混合（MoE）架构。我们的评估表明，Seed1.5-VL 在此基准测试中显著优于 Llama 4 Maverick。对于某些模型系列，我们包括不同的模型版本以评估社区随时间推移的进步。我们的评估强调，思考 模型在前五名中占据主导地位，我们认为这归因于内部基准测试对测量集成模型能力的重点。与社区趋势一致，同一供应商的新模型版本通常优于早期版本；例如，GPT-4o-Latest 的得分高于 GPT-4o-0513，Gemini 2.5 超过了 Gemini 2.0。

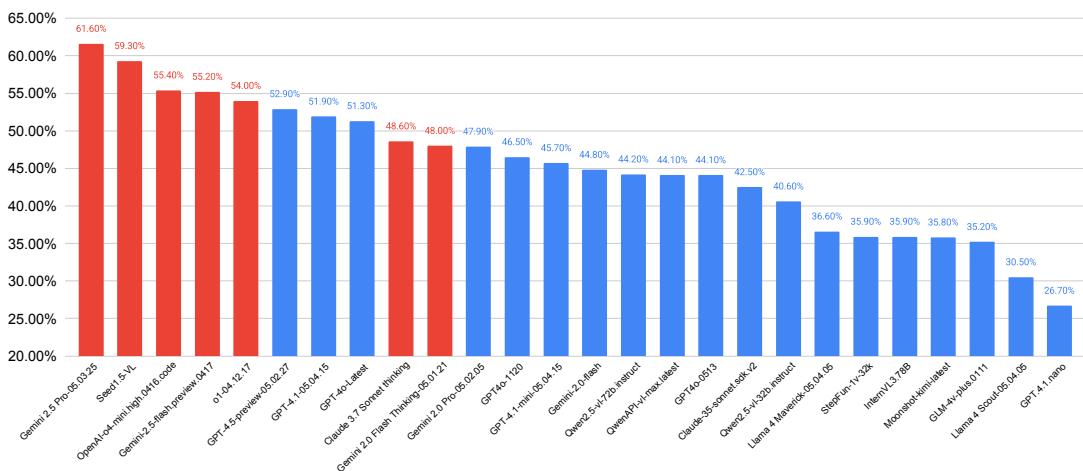


Figure 29 对 Seed1.5-VL 模型与现有模型的全面比较，按其在我内部基准上的整体性能进行排序。采用 思考 方法的模型由红色条形表示，而被归类为 非思考 的模型则由蓝色条形表示。对前五名得分的分析显示，思考 模型占大多数。为了考虑模型 API 的潜在更新，每个模型名称后都附加了 API 发布日期（格式为年-月-日）。

B.3 能力与基准任务

我们使用 60 个公开基准来评估 Seed1.5-VL 在十个不同能力方面的能力：多模态推理、通用视觉问答、文档和图表理解、视觉定位与计数、空间理解、短时和长时视频理解、流式视频理解、视频定位、GUI 智能体。以下是我们提供的所有基准的详细列表。

- **多模态推理：** 我们使用七个基准：MMMU [168], MMMU-Pro [169], MathVision [140], Olympiad-Bench [41], MathVista [84], V^{*} [148], VLM are Blind [108], ZeroBench (Main/Subtasks) [113], VisuLogic [154], Video-MMMU [49], 和 MMVU [175].
- **通用视觉问答：** 我们使用八个基准：RealWorldQA [150], MMStar [15], MMVet [166], MMBench (英文和中文) [81], MMVP [133], HallusionBench [38]，以及 BLINK [33]。
- **文档和图表理解：** 我们使用七个基准：TextVQA [124], AI2D [61], ChartQA [88], InfographicVQA [90], DocVQA [89], OCRBench [83]，以及 CharXiv (RQ/DQ) [144].
- **基础与计数：** 我们使用五个基准：LVIS-MG (从 LVIS [40] 得到的多对象基础)，VisualWebBench [79], RefCOCO [60, 92, 164], CountBench [102], FSC-147 [110]。
- **空间理解：** 我们使用五个基准：DA-2K [160], NYU-Depth V2 [95], SUN-RGBD [125], ARKitScenes [9]，以及 All-Angles Bench [163]。
- **短视频理解：** 我们使用六个基准：MotionBench [48], MVBench [73], TOMATO [117], TVBench [19], Dream-1K [139]，以及 TempCompass [82]。
- **长视频理解：** 我们使用六个基准：LongVideoBench [147], LVBench [142], MLVU [178], VideoMME [32]，TemporalBench [12]，以及 EgoSchema [87]。
- **流视频理解：** 我们使用六个基准：OVBench [51], OVOBench [74], StreamBench [153]，和 StreamingBench [76].
- **视频定位：** 我们使用两个基准：Charades-STA [34] 和 TACoS [34]。
- **GUI 智能体：** 我们使用七个基准：ScreenSpot-V2 [149], ScreenSpot-Pro [72], OSWorld [152]，Windows Agent Arena [11]，WebVoyager [42]，Online-Mind2Web [158]，以及 Android World [111]。

B.4 评估提示

所有基准均使用指令微调模型进行零样本评估。要激活 Seed1.5-VL 的思考模式，我们添加以下前缀：

```
You should first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process is enclosed within <think> </think> tags, i.e. <think> reasoning process here </think> answer here
```

然后，我们为其添加一个针对每个基准的定制提示。每个基准的提示模板如下所示。在每个模板中，`{question}` 填入实际样本的问题，`{options}` 被替换为样本的多选答案选项，`<image>` 填入输入图像的计算得到的 ViT 嵌入，`<label>` 被替换为物体的标签（例如，定位基准），`<video>` 填入视频帧的 ViT 嵌入（例如，视频基准）。下面，我们省略了每个图像周围的 [SOI] 和 [EOI] 词元。

MMMU. 我们使用 OpenCompass 建议的相同指标⁶。我们遵循 MMMU 中原始样本的相同图像位置占位符，这可以交错进行。

```
<image>
Question: {question}
Options:
{options}
Your response can be freely expressed in any format, but the final answer must be presented in this format:
"Final answer: [the correct option]"
```

MMMU-Pro. 我们使用 MMMU-Pro 的官方指标。

```
<image>
{question}
```

MathVision. 如 Wang et al. [140] 所示，提示工程的整理对于 MathVision 的目标和精确评估至关重要。我们使用 MathVision 的官方指标。我们注意到思维模型，如 OpenAI-O1，有时提供的解决方案无法被 MathVision 提供的官方基于规则的验证器精确解析，例如，预测值为 2kg 与真实值 2，或者提供正确选项的数值而非选项名称。因此，我们精心设计了针对 OpenAI-O1 的提示，以避免潜在的欠估计。我们使用相同的提示在该基准上测试 Seed1.5-VL 和 Gemini-2.5-Pro。

```
<image>
{question}
Please solve the problem step by step and put your answer in one "\boxed{}". If it is a multiple choice question, only one letter ("\boxed{A}", "\boxed{B}", "\boxed{C}", "\boxed{D}", or "\boxed{E}") is allowed in the "\boxed{}". For example, do NOT output "\boxed{42}" for a multiple choice question.
```

奥数基准. 我们使用奥数基准的官方指标。

⁶<https://github.com/open-compass/opencompass>

```
<image>
{question}
```

MathVista. 我们使用 OpenCompass 建议的相同指标。

```
<image>
{question}
```

V*. 我们使用 V* 的官方指标。

```
<image>
{question}
```

VLM 是盲的. 我们使用 VLM 是盲的官方指标。

```
<image>
{question}
```

TextVQA. 我们使用 OpenCompass 建议的相同指标。

```
<image>
{question}
Answer the question using a single word or phrase.
```

AI2D. 我们使用 OpenCompass 建议的相同指标。

```
<image>
Question: {question}
Options:
{}
Please select the correct answer from the options above.
```

ChartQA. 我们使用 ChartQA 的官方指标。正确性允许由 max_relative_change 定义的一定误差比例。

```
<image>
{question}
Answer the question using a single word or phrase.
```

InfographicVQA. 我们通过将预测上传到官方排行榜来收集得分。

```
<image>
{question}
Answer the question using a single word or phrase.
```

DocVQA. 我们通过将预测上传到官方排行榜来收集得分。

```
<image>
{question}
Answer the question using a single word or phrase.
```

OCRBench. 我们使用 OCR Bench 的官方指标，包括将答案转换为小写和去除空格。

```
<image>
{question}
```

CharXiv. 我们使用 CharXiv 的官方指标。

```
<image>
{question}
```

真实世界问答. 我们使用 OpenCompass 建议的相同指标。

```
<image>
Question: {question}
Options:
{options}
Please select the correct answer from the options above.
```

MMStar. 我们使用 OpenCompass 建议的相同指标。

```
<image>
Question: {question}
Options:
{options}
Please select the correct answer from the options above.
```

MBBench-en. 我们使用 OpenCompass 建议的相同指标。

```
<image>
Question: {question}
Options:
{options}
Answer with the option's letter from the given choices directly. The correct option is:
```

MBBench-cn. 我们使用 OpenCompass 建议的相同指标。

```
<image>
问题: {question}
选项:
{options}
请根据选项直接回答选项字母。正确选项为:
```

MMVP. 我们使用 MMVP 的官方指标。该数据集由 150 对样本组成，每对样本包含两个问题，只有当两个问题都正确时才视为正确。

```
<image>
{question}
```

HallusionBench. 我们使用 OpenCompass 建议的相同指标。

```
<image>
{question}
```

BLINK. 我们使用 OpenCompass 建议的相同指标。

```
<image>
Question: {question}
Options:
{options}
Please select the correct answer from the options above.
```

CountBench.

```
<image>
{question}
```

VisualWebBench.

```
<image>
{question}
```

FSC-147.

```
<image>
Count the number of {label}.\\nYou need to point them out first in <point>x y</point>
format.
```

LVIS.

```
<image>
Which region does {label} describe? Output the location as <bbox>x1 y1 x2 y2</bbox>.
```

RefCOCO.

```
<image>
which region does text {label} describe? Output the location as <bbox>x1 y1 x2
y2</bbox>.
```

DA-2K.

```
<image>
There are two points with different colors in the image, point1 (denoted with blue
point) and point2 (denoted with green point), each representing an object. Which
object represented by these points is closer to me? Only provide the answer:
'point1' or 'point2'.
```

NYU-Depth V2.

```
<image>
```

Here are the detailed camera parameters for the image. Camera intrinsic parameters: Focal length $f_x = \{fx\}$, $f_y = \{fy\}$. Principal point coordinate locates at the center of the image, $c_x = \{cx\}$ and $c_y = \{cy\}$, when image width $\{\text{width}\}$ and height $\{\text{height}\}$. We do not consider distortion parameters here. Therefore, the intrinsic matrix $K = [[\{fx\}, 0, \{cx\}], [0, \{fy\}, \{cy\}], [0, 0, 1]]$. Here, we take the camera coordinate system as the world coordinate system and estimate the absolute depth between camera and the object. Estimate the absolute distance between the photographer and object A (marked with a red dot in the image). Respond directly with the absolute distance in meters only.

SUN RGB-D.

```
<image>
```

Here are the detailed camera parameters for the image. Camera intrinsic parameters: Focal length $f_x = \{fx\}$, $f_y = \{fy\}$. Principal point coordinate locates near the center of the image, $c_x = \{cx\}$ and $c_y = \{cy\}$, when image width $\{\text{width}\}$ and height $\{\text{height}\}$. We do not consider distortion parameters here. Therefore, the intrinsic matrix $K = [[\{fx\}, 0, \{cx\}], [0, \{fy\}, \{cy\}], [0, 0, 1]]$. Camera coordinate: X-axis points rightward, Y-axis points downward, and Z-axis points forward. The origin point is the camera location. We take the camera coordinate system as the world coordinate system, namely the camera extrinsic matrix is $[[1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0]]$. Please output each 3D bounding box in the following format: $\langle 3dbbox \rangle x_center y_center z_center x_size y_size z_size \text{pitch yaw roll} \langle /3dbbox \rangle$. Note: (1) x_center , y_center , z_center : the center of the object in the camera coordinate, in meters. (2) x_size , y_size , z_size : The dimensions of the object along the XYZ axes, in meters, when the rotation angles are zero. (3) pitch , yaw , roll : Euler angles representing rotations around the X, Y, and Z axes, respectively. Each angle is normalized to the range of $(-1, 1)$ and is multiplied by 180 to convert it into degrees. Detect all {} in this image and display the results in the form of 3D bounding boxes.

全角度基准.

```
<image>
```

Question: {question} Options: {options}. Answer with the option's letter from the given choices directly.

OVBench.

```
<image>
{question} The choices are: [{options}]. The answer is:
```

OVOBench.

```
<image>
{question} The choices are: [{options}]. The answer is:
```

StreamingBench(主动).

```
You will be given an instruction and a video, which requires you to continuously monitor the video stream and make responses. The instruction is: [{question}]. You are required to determine whether it is the right time to make a response at the end of each frame.
```

```
<video>
```

自我模式.

```
<video>
You will be given a question about a video and five possible answer options, where C refers to the person wearing the camera. You will be provided frames from the video, sampled evenly across the video. {question} Possible answer choices:{options}
Directly output the final answer in the format "X" where X is the correct letter choice. Never say "unknown" or "unsure", or "None", instead provide your most likely guess.
```

视频-MME.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter (A, B, C, or D) of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

时间基准.

```
<video>
Select the best answer to the following binary-choice question based on the video.
Respond with only the letter (A or B) of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

长视频基准.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

LVBench.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

MLVU.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

MotionBench.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

MMVU. 开放性:

```
<video>
Question: {question}

Answer the given question step by step. Begin by explaining your reasoning process
clearly. Conclude by stating the final answer using the following format:
'Therefore, the final answer is: 'Answer: $$ANSWER' (without quotes), where $$ANSWER
is the final answer of the question. Think step by step before answering.
```

选择题:

```
<video>
Question: {question}
Options:
{options}
```

Answer the given multiple-choice question step by step. Begin by explaining your reasoning process clearly. Conclude by stating the final answer using the following format: 'Therefore, the final answer is: \$\$LETTER' (without quotes), where \$\$LETTER is one of the options. Think step by step before answering.

视频-MMMU.

1. 开放式:

```
<video>
Question: {question}

Do not generate any intermediate reasoning process. Directly output the final short answer.
```

2. 单项选择:

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

MVBench.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter of the correct option.
{question} Possible answer choices:
{options}
The best answer is:
```

番茄.

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter (A, B, C, D, E, F, G, H...) of the correct option.
```

```
{question} Possible answer choices:  
{options}  
The best answer is:
```

TVBench.

```
<video>  
Select the best answer to the following multiple-choice question based on the video.  
Respond with only the letter (A, B, C, D...) of the correct option.  
{question} Possible answer choices:  
{options}  
The best answer is:
```

DREAM-1K.

```
<video>  
Describe the video in one paragraph, mainly focusing on the dynamic events in the  
video. Don't describe feelings or atmosphere.{question}
```

TempCompass.

1. 单项选择题:

```
<video>  
{question} Choices are: {options}  
Please directly give the best option:
```

2. 是/否问答:

```
<video>  
{question}
```

3. 图片说明匹配:

```
<video>  
{question}
```

4. 图像描述生成:

```
<video>  
{question}
```

猜词游戏-STA.

```
<video>  
Find start and end seconds for: "{label}", please return the start and end seconds.
```

TACoS.

<video>

Find start and end seconds for: "{label}", please return the start and end seconds.