

DeepSeek-VL2: 面向高级多模态理解的混合专家视觉语言模型

Zhiyu Wu*, Xiaokang Chen*, Zizheng Pan*, Xingchao Liu*, Wen Liu*[†], Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, Chong Ruan[‡]

DeepSeek-AI

Abstract

我们提出 DeepSeek-VL2，这是一个先进的大型混合专家（MoE）视觉语言模型系列，通过两项关键重大升级显著改进了其前身 DeepSeek-VL。在视觉组件方面，我们引入了一种动态分块视觉编码策略，专为处理不同宽高比的高分辨率图像而设计。在语言组件方面，我们采用了具备多头潜在注意力（Multi-head Latent Attention）机制的 DeepSeekMoE 模型，该机制将键值（Key-Value）缓存压缩为潜在向量，从而实现高效推理和高吞吐量。在改进的视觉语言数据集上进行训练后，DeepSeek-VL2 在多项任务中展现出卓越的能力，包括但不限于视觉问答、光学字符识别、文档/表格/图表理解以及视觉定位。我们的模型系列包含三个变体：DeepSeek-VL2-Tiny、DeepSeek-VL2-Small 和 DeepSeek-VL2，激活参数分别为 1.0B、2.8B 和 4.5B。与现有的开源稠密模型和基于 MoE 的模型相比，DeepSeek-VL2 在激活参数数量相当或更少的情况下，取得了具有竞争力或最先进的性能。代码和预训练模型已公开于 <https://github.com/deepseek-ai/DeepSeek-VL2>。

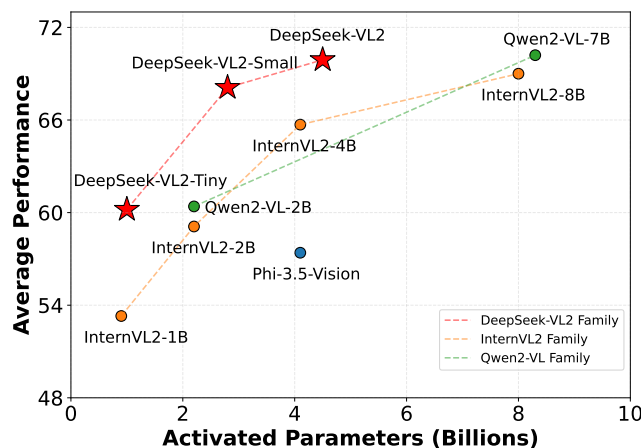


图 1 | 不同开源模型的平均性能与激活参数对比。我们对 MMBench v1.1、MMStar、MMMU (Val)、MathVista (TestMini)、AI2D (Test) 和 OCRBench 的准确率进行了平均。OCRBench 的得分除以 10 以将其缩放到 [0, 100] 区间。

*: 核心贡献者。†: 项目负责人。‡: 通讯作者。

目录

1	引言	3
2	模型架构	3
3	数据构建	5
3.1	视觉-语言对齐数据	6
3.2	视觉-语言预训练数据	6
3.3	监督微调数据	7
4	训练方法	9
4.1	训练流程	9
4.2	超参数与基础设施	10
5	评估	11
5.1	多模态性能	11
5.2	定性研究	12
6	结论	20

1. 引言

大型视觉-语言模型 (VLMs) 已成为人工智能领域的一项变革性力量 [15, 54, 59, 63, 83, 88, 94], 将大型语言模型 (LLMs) 的卓越能力扩展至无缝处理视觉与文本信息。这一进展极大地拓展了人工智能系统应对需要多模态理解的复杂现实世界应用的潜力。

在本技术报告中, 我们推出了 DeepSeek-VL2, 这是一系列开源的视觉-语言模型。该模型采用混合专家 (MoE) 架构, 相较于其前身 DeepSeek-VL [59], 在性能与效率方面均取得了显著提升。我们的改进主要围绕三个核心方面展开: (1) 一种动态高分辨率视觉编码策略, 以增强视觉理解能力; (2) 一种优化的语言模型架构, 显著提升了训练与推理效率; (3) 一套精细化的视觉-语言数据构建流程, 不仅提升了整体性能, 还将模型能力扩展至精确视觉定位等新领域。

在视觉组件方面, 我们引入了一种动态分块视觉编码策略, 能够高效处理具有不同宽高比的高分辨率图像。该方法改进了 DeepSeek-VL 的混合视觉编码器 (后者仅在两个固定分辨率 (384×384 和 1024×1024) 下提取图像特征)。我们的方法克服了旧版固定尺寸编码器的局限性, 在需要超高分辨率的任務中表现优异, 包括视觉定位、文档/表格/图表分析以及细粒度特征提取, 同时保持了可控的视觉 token 数量。借鉴成熟的切片分块方法, 我们的系统动态地将高分辨率输入分割为局部图块, 通过共享的视觉 Transformer 处理每个图块, 并在语言模型内无缝融合提取的特征。该设计保留了具有局部注意力机制的视觉 Transformer 的优势, 能够在不随图像分辨率增加而产生典型二次方计算复杂度增长的前提下, 实现丰富的特征提取。

在语言组件方面, 我们采用了 DeepSeek 语言模型 [20, 53], 其核心特性为多头潜在注意力 (MLA) 机制。MLA 通过将键值 (KV) 缓存压缩为潜在向量, 显著降低了计算成本, 从而实现了更快的推理速度与更高的吞吐量。我们进一步通过 DeepSeekMoE 框架 [20, 86] 提升效率, 该框架采用了稀疏计算技术。我们的模型系列包含三种 MoE 变体: 3B、16B 和 27B。这些大语言模型的激活参数分别为 0.57B、2.4B 和 4.1B。

此外, 我们在质量、数量和多样性方面大幅提升了视觉-语言训练数据。这一全面的数据集使模型在广泛的任務中实现了更好的泛化能力与性能表现, 涵盖视觉问答 (VQA)、光学字符识别 (OCR)、文档/表格/图表理解、视觉推理以及通用聊天机器人应用。改进后的训练数据还赋予了模型视觉定位和图形用户界面 (GUI) 感知等新能力。

综上所述, DeepSeek-VL2 标志着大规模混合专家视觉-语言建模领域的一次重大飞跃。通过全新的视觉处理策略与优化的语言模型, 我们开发出了一系列在性能与效率之间取得良好平衡的模型。通过开源这些预训练模型, 我们旨在加速该领域的技术进步, 并推动协同研究的深入发展。

2. 模型架构

DeepSeek-VL2 由三个核心模块组成: (1) 视觉编码器, (2) 视觉-语言适配器, 以及 (3) 混合专家 (Mixture-of-Experts) 语言模型。在其前代模型仅解码器 (decoder-only) 的 LLaVA 风格 [54] 架构基础上, DeepSeek-VL2 引入了两项主要改进: 动态分块策略 (dynamic tiling strategy) 以

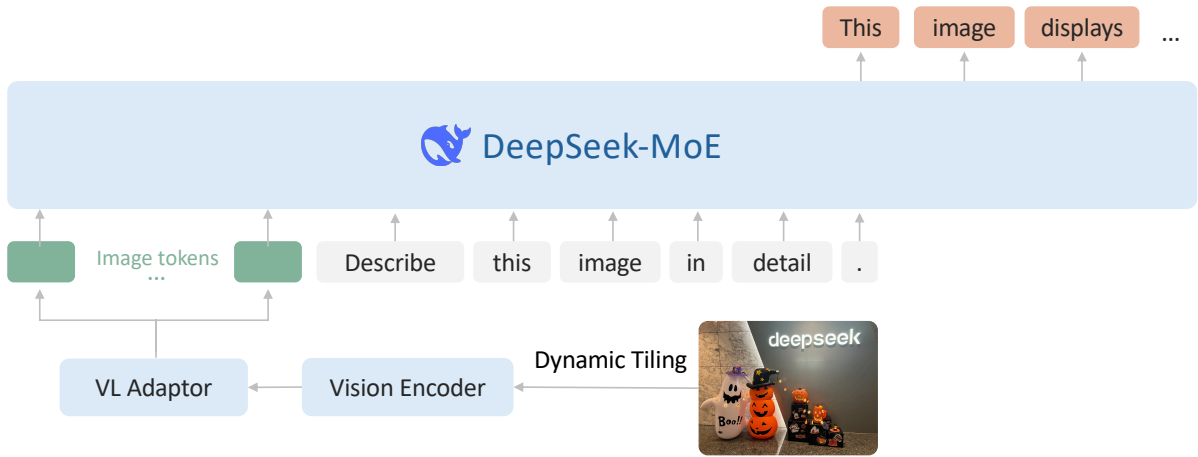


图 2 | **DeepSeek-VL2 概览**。整体结构采用 LLaVA 风格架构，包含视觉编码器、视觉-语言适配器（VL adaptor）以及基于混合专家（MoE）的大语言模型。

及采用多头潜在注意力（Multi-head Latent Attention） [53] 的 DeepSeekMOE [20, 86] 语言模型。这些创新使得模型能够更高效地处理高分辨率视觉输入与文本数据。

动态分块策略。原始 DeepSeek-VL 采用了一种混合视觉编码器，结合 SigLIP [106] 在 384×384 分辨率下进行粗粒度特征提取，以及 SAM-B [35] 在 1024×1024 分辨率下进行细粒度特征提取。尽管这种融合方法生成了适用于多种视觉-语言任务的丰富视觉表征，但其受限于固定的 1024×1024 分辨率约束。这一限制在处理更高分辨率及极端宽高比的图像时尤为突出，例如 InfographicVQA [67]、密集光学字符识别（OCR）以及细粒度视觉定位任务中的图像。

受近期视觉-语言模型（VLMs）进展的启发 [16, 21, 55]，我们通过将高分辨率图像分割为多个图块（tiles）来实现动态分块策略。该方法使得单个 SigLIP-SO400M-384 视觉编码器 [106] 能够高效处理具有不同宽高比的各类高分辨率图像。预训练的 SigLIP 基础分辨率为 384×384 。为适应不同的宽高比，我们定义了一组候选分辨率： $C_R = \{(m \cdot 384, n \cdot 384) \mid m \in \mathbb{N}, n \in \mathbb{N}, 1 \leq m, n, mn \leq 9\}$ ，其中 $m : n$ 表示宽高比。对于尺寸为 (H, W) 的输入图像，我们计算将其调整大小¹至 C_R 中每个候选分辨率所需的填充区域。我们选择使填充区域最小的分辨率 $(m_i \cdot 384, n_i \cdot 384)$ 。随后，调整大小后的图像被划分为 $m_i \times n_i$ 个尺寸为 384×384 像素的局部图块，以及一个全局缩略图图块。SigLIP-SO400M-384 视觉编码器对所有 $(1 + m_i \times n_i)$ 个图块进行处理，每个图块输出 $27 \times 27 = 729$ 个维度为 1152 的视觉嵌入。出于计算效率与上下文长度管理的考虑，在处理多张（ > 2 ）图像时，我们禁用了动态分块策略。

视觉-语言适配器。在完成视觉图块处理后，我们执行 2×2 像素洗牌（pixel shuffle）操作，将每个图块的视觉令牌（tokens）从 27×27 压缩至 $14 \times 14 = 196$ 个。随后，在处理 $(1 + m_i \times n_i)$ 个图块时，我们引入了三种特殊令牌。对于全局缩略图图块（ 14×14 ），我们在每一行末尾添加 14 个 `<tile_newline>` 令牌，使得总数达到 $14 \times 15 = 210$ 个。对于排列成形状为 $(m_i \cdot 14, n_i \cdot 14)$ 的二维网格的 $m_i \times n_i$ 个局部图块，我们在最后一列末尾追加 $m_i \cdot 14$ 个 `<tile_newline>` 令牌，以表示所

¹我们首先调整原始图像尺寸，直至其长边与目标分辨率匹配，随后在保持原始宽高比不变的情况下对另一维度进行填充。

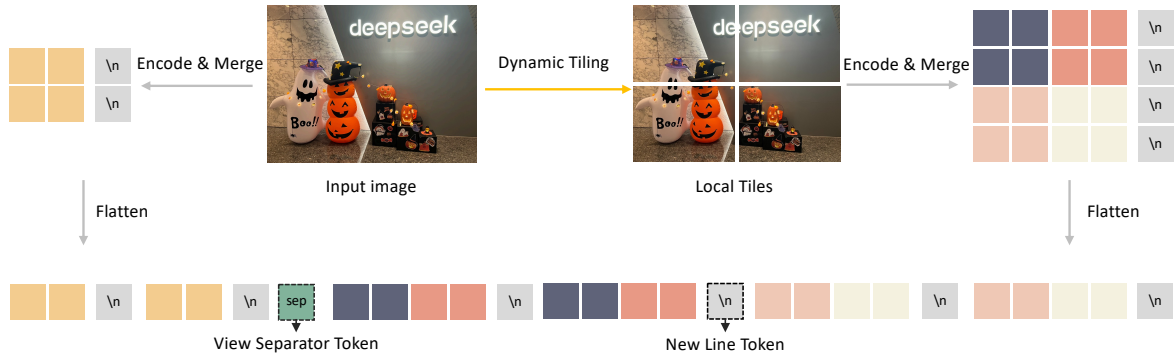


图 3 | DeepSeek-VL2 中的动态分块策略示意图。通过将图像划分为多个图块，DeepSeek-VL2 相较于 DeepSeek-VL 实现了更强的细粒度理解能力。

表 1 | DeepSeek-VL2 的架构配置。我们列出了该架构的超参数以及与混合专家训练相关的详细信息。

	DeepSeek-VL2-Tiny	DeepSeek-VL2-Small	DeepSeek-VL2
词表大小	129,280	102,400	129,280
嵌入维度	1,280	2,048	2,560
注意力头数	10	16	32
层数	12	27	30
注意力机制	多头注意力	MLA (rank=512)	MLA (rank=512)
路由专家数	64	64	72
共享专家数	2	2	2
专家选择 Top-K	6	6	6
路由函数	Softmax	Softmax	Sigmoid
专家校正偏置	×	×	✓

有局部图块的一行结束。此外，在全局缩略图图块与局部图块之间插入一个 `<view_separator>` 令牌。完整的视觉序列包含 $210 + 1 + m_i \cdot 14 \times (n_i \cdot 14 + 1)$ 个视觉令牌，随后通过一个两层多层感知机 (MLP) 将其投影至语言模型的嵌入空间。我们动态分块策略的可视化示意图如图 3 所示。

DeepSeekMoE 大语言模型。我们的语言模型基于 DeepSeekMoE [20, 86]，并集成了多头潜在注意力 (Multi-head Latent Attention, MLA) 机制 [53]。MLA 通过将键值 (Key-Value) 缓存压缩为潜在向量来提升推理效率，从而增加吞吐量。该模型还采用了混合专家 (MoE) 架构 [20]，通过稀疏计算实现高效推理。在 MoE 训练期间，我们为每个专家引入了一个全局偏置项 [86]，以低成本的方式改善专家间的负载均衡。DeepSeek-VL2 提供三种变体，模型参数量分别为：1.0B、2.8B 和 4.5B。完整的架构规格详见表 1。

3. 数据构建

我们为 DeepSeek-VL2 从多种来源构建了一个全面的视觉-语言 (Vision-Language) 数据集。训练过程分为三个不同的阶段：(1) 视觉-语言对齐 (VL alignment)，(2) 视觉-语言预训练 (VL pretraining)，以及 (3) 监督微调 (supervised fine-tuning, SFT)。在以下部分中，我们将详细介绍每个阶段所使用的数据。

3.1. 视觉-语言对齐数据

对齐阶段主要侧重于训练 MLP 连接器，以桥接预训练的视觉编码器与大语言模型 (LLM)。在此初始预热阶段，我们使用了 ShareGPT4V [12] 数据集，该数据集包含约 120 万条图像描述和对话样本。

3.2. 视觉-语言预训练数据

遵循 DeepSeek-VL [59] 的做法，我们的预训练数据结合了视觉-语言 (VL) 数据与纯文本数据，以在 VL 能力与纯文本性能之间保持平衡。对于 DeepSeek-VL2，我们保持约 70% 的 VL 数据与 30% 的纯文本数据比例，后者直接来源于我们基础 LLM 的预训练语料库。接下来，我们将 VL 数据分为几类并详细描述其构成。

交错图像-文本数据。 我们的数据收集始于多个开源数据集，包括 WIT [79]、WikiHow [38] 以及从 OBELICS [41] 中随机抽取的 30% 样本。这一特定的混合比例是通过在 DeepSeek-VL2-Tiny 上的初步实验确定的。为增强多语言能力，我们在以英语为主的数据集中补充了从 Wanjuan [29] 中提取的中文内容。此外，我们开发了内部数据集，以扩展对通用现实世界知识的覆盖范围。

图像描述数据。 图像描述是 VLM 训练中的基础数据，提供了视觉与文本信息之间的直接对齐。我们最初利用了多种开源数据集 [8, 25, 28, 36, 37, 39, 40, 48, 50, 51, 73, 78, 80, 82]。然而，我们的初步分析显示这些数据集之间存在严重的质量差异，从由先进 VLM 生成的密集、准确的描述，到描述简略、图文不匹配或存在明显幻觉的问题案例不等。为解决这些质量不一致问题，我们开发了一套综合的图像描述生成流水线，该流水线综合考虑了：(1) OCR 提示，(2) 元信息 (如地理位置、相机参数)，以及 (3) 相关的原始描述作为提示词。我们使用内部描述生成器，遵循与 PixelProse [78] 类似的提示策略对图像重新生成描述，并采用多样化的指令来引导 VLM 的描述生成过程。

尽管描述质量整体有所提升，但我们发现大规模标注流水线中存在重复问题。为缓解这一问题，我们实施了一套质量控制流水线，使用 DeepSeek Chat [53] 仅基于写作质量对所有描述进行评分。在实践中，该方法在过滤低质量描述方面既高效又有效。

光学字符识别 (OCR) 数据。 为开发 OCR 能力，我们使用了包括 LaTeX OCR [7] 和 1200 万条 RenderedText [93] 在内的开源数据集。我们将这些数据集与一个涵盖多种文档类型的庞大内部 OCR 数据集相结合。目前，我们的内部数据集主要侧重于英文和中文的字符识别。我们计划在未来的工作中扩展至其他语言。

视觉问答 (QA) 数据。 在早期的探索中，我们发现通用 QA 数据对模型预训练具有明显的益处。因此，我们构建了一个综合的视觉 QA 数据集，包含以下类别：

- **通用 VQA**。我们继承了来自 DeepSeek-VL 的通用 VQA 数据。更多详情请参阅 [59]。
- **表格、图表与文档理解**。我们采用 PubTabNet [112]、FinTabNet [111] 和 Docmatix [42] 以增强文档理解能力。
- **网页转代码与图表转 Python 生成**。遵循 DeepSeek-VL 的做法，我们利用 Websight [44] 提升网页转代码能力，并使用从公开 Jupyter 笔记本中获取的 Python 图表数据。我们通过使用 DeepSeek V2.5 复现部分 Websight 数据来增强该数据集。此外，我们还利用 DeepSeek V2.5 生成的 Python 绘图代码来减少图表转代码数据中的噪声。
- **带视觉提示的 QA**。我们遵循 [9] 的方法，通过将各种视觉指示器（箭头、方框、圆圈和涂鸦）叠加到来自 [9, 89, 90] 的图像上，构建视觉提示理解数据。随后，我们创建了聚焦于这些视觉提示所高亮对象的 QA 对。

视觉定位 (Visual Grounding) 数据。我们基于 [71, 75] 构建视觉定位数据集。针对每张图像的目标检测标注，我们将数据结构化如下：

- 提示词 (Prompt): `Locate <|ref|><query></ref|> in the given image.`
- 回复 (Response): `<|ref|><query></ref|><|det|>[[x1, y1, x2, y2],...]</det|>`

在训练过程中，问题提示词从候选池中随机采样。`<|ref|>`、`</ref|>`、`<|det|>` 和 `</det|>` 为特殊标记 (special tokens)。`<query>` 是类别名称 (如 “car”) 或物体描述 (如 “the leftmost person”) 的占位符。`[[x1, y1, x2, y2], ...]` 是一个边界框列表，其中每个边界框对应一个物体的位置。坐标 `x1`, `y1` 与 `x2`, `y2` 分别表示左上角与右下角，并根据图像分辨率归一化至 0 至 999 的范围内。此外，我们构建了负样本，其中查询目标在图像中故意缺失，以提升模型的鲁棒性。

视觉定位对话数据。我们的视觉定位对话数据集基于 [71] 构建，采用如下格式：

- 提示词: `<|grounding|> 请描述图像的内容？`
- 回复: `两只 <|ref|> 狗 </ref|><|det|>[[x1, y1, x2, y2],...]</det|> 正在草地上奔跑。`

与其他视觉定位数据类似，`<|grounding|>`、`<|ref|>`、`</ref|>`、`<|det|>`、`</det|>` 均为特殊标记，且 `x1`, `y1`, `x2`, `y2` 遵循相同的归一化方案。

3.3. 监督微调数据

我们的 SFT 数据结合了多样化的开源数据集与高质量的内部问答 (QA) 对。下文将详细说明我们在提升 SFT 数据集质量方面所做的努力。

通用视觉问答。 尽管公开的视觉问答数据集种类繁多 [9, 10, 27, 31, 43, 47, 74]，但它们通常存在三个主要局限性：(1) 回复简短，(2) OCR 质量较差，以及 (3) 存在幻觉内容。为解决这些问题，我们通过综合考虑原始问题、图像和 OCR 信息来重新生成回复。我们的实验表明，该方法能够生成更全面、准确的结果。在开发过程中，我们观察到 DeepSeek-VL2 的早期版本（尤其是 Tiny 变体）在中文回复中偶尔会不当插入英文单词。该问题在我们的较大模型中并未出现，这表明其源于模型容量有限以及视觉-语言预训练阶段中英数据的不平衡。为克服小模型的这一局限，我们开发了一个内部中文 QA 数据集，其中包含多样化的图像描述以及单轮/多轮对话。该数据集有助于缓解语言混杂问题。此外，我们额外构建了一个内部数据集，以补充现实世界与文化相关的视觉知识，涵盖动漫、网络迷因 (memes)、美食和艺术等领域。

OCR 与文档理解。 得益于我们先进的图像描述生成流水线，DeepSeek-VL2 已展现出优于其他最先进视觉语言模型 (VLM) 的 OCR 能力。因此，在 SFT 阶段，我们并未进一步追求 OCR 性能的提升，而是专注于清洗现有的开源数据集 [24, 31, 43, 66, 67, 77, 92, 104]，移除了 OCR 质量较差的样本。在文档理解方面，我们从内部数据中精选了一个多样化的文档页面子集。随后，我们生成了专门针对文档理解的多轮对话问答对。初步结果表明，该方法提升了基于文档的交互效果。

表格与图表理解。 我们通过基于原始问题重新生成回复，提升了基于表格的 QA 数据质量，涵盖了所有公开数据集 [14, 49]，Cauldron [43] 除外（该数据集本身已具备高质量）。与在视觉-语言预训练阶段发展的 OCR 能力类似，我们的模型在图表理解方面已展现出强劲性能，无需额外投入。

推理、逻辑与数学。 我们为面向推理的公开数据集 [17, 43, 61, 76, 102, 109] 补充了更详细的推理过程，并标准化了回复格式，将最终答案置于回复末尾。我们观察到，在训练较小的 VLM 时，过于详细的回复效果反而不佳。在我们的探索中，DeepSeek-VL2-Tiny 在使用更简洁的回复时表现更佳。

教材与学术问题。 我们基于文档库构建了一个专注于教材的内部数据集。该数据集主要侧重于涵盖多个学科领域的大学级别内容。

网页转代码与图表转 Python 代码生成。 我们扩展了用于网页代码和 Python 绘图代码的内部数据集，其规模超出了预训练阶段所使用的范围。对于开源数据集，我们通过重新生成答案来提升其质量。

视觉定位。 我们利用来自 [2, 23, 64, 85, 101, 110] 的数据构建了视觉定位数据集。为提升模型能力，我们将查询短语翻译为中文，并构建了额外的负样本。我们还添加了上下文视觉定位数

据，该任务要求在给定参考图像中由矩形或椭圆高亮标记的参考对象的情况下，在多张图像中定位同类对象。数据格式遵循以下结构：

- Prompt: `<|grounding|>The first image shows <object>.Please identify the object of the same category in the second image.`
- Response: `<|ref|><description></ref|><|det|>[[x1, y1, x2, y2]]</det|>`

在此格式中，`<|grounding|>`、`<|ref|>`、`</ref|>`、`<|det|>` 和 `</det|>` 为特殊标记 (tokens)。占位符 `<object>` 代表诸如“红色边界框内的对象”之类的短语，而 `<description>` 则是模型对检测到的对象的描述（例如“猫”）。

视觉定位对话。 我们利用 [62, 72] 构建了视觉定位对话数据，以进一步增强模型在预训练阶段已建立的能力。

纯文本数据集。 为保持模型的语言能力，我们在 SFT 阶段还使用了纯文本指令微调数据集 [4, 6, 18, 19, 68, 70, 84, 91, 98]。

4. 训练方法

4.1. 训练流程

DeepSeek-VL2 的训练采用三阶段流程：(1) 初始阶段，我们使用第 3.1 节详述的图文配对数据，在固定语言模型的同时训练视觉编码器和视觉-语言适配器 MLP；(2) 预训练阶段，我们使用第 3.2 节描述的数据进行视觉-语言预训练；(3) 微调阶段，我们使用第 3.3 节概述的数据进行监督微调。在预训练和微调阶段，所有模型参数（包括视觉编码器、视觉-语言适配器和语言模型）均被解冻并同时训练。在所有阶段中，我们均强调视觉理解能力，并且仅在文本 token 上计算下一个 token 预测损失。

视觉-语言对齐。 基于预训练语言模型 (DeepSeekMoE 3B/16B/27B)，我们的主要目标是在视觉特征与语言特征之间建立稳健的连接。该对齐过程使预训练语言模型能够有效处理视觉输入。与保持预训练视觉编码器和语言模型固定的先前方法 [54, 59] 不同，我们对固定分辨率的视觉编码器进行了适配，以支持动态高分辨率图像。在此阶段，我们在冻结语言模型的同时，对视觉编码器和视觉-语言适配器进行优化。

视觉-语言预训练。 在嵌入空间中建立视觉-语言对齐后，我们将大部分计算资源投入到视觉-语言预训练中。该阶段侧重于在多样化任务上构建全面的联合视觉-语言知识。我们解冻所有参数（包括视觉编码器、视觉-语言适配器 MLP 和 DeepSeekMoE 语言模型），以实现全模型优化。使用约 800B 图文 token（第 3.2 节），该阶段在保持模型大部分语言能力的同时，显著提升了其多模态理解能力。

表 2 | **DeepSeek-VL2 训练超参数**. Step LR 调度器在总训练步数的 50% 和 75% 处将学习率除以 $\sqrt{10}$ 。

	DeepSeek-VL2-Tiny			DeepSeek-VL2-Small			DeepSeek-VL2		
总参数量 (LLM)	3B			16B			27B		
激活参数量 (LLM)	0.57B			2.4B			4.1B		
视觉编码器	SigLIP-SO400M			SigLIP-SO400M			SigLIP-SO400M		
超参数	阶段 1	阶段 2	阶段 3	阶段 1	阶段 2	阶段 3	阶段 1	阶段 2	阶段 3
学习率	5.4×10^{-4}	5.4×10^{-4}	3.0×10^{-5}	4.2×10^{-4}	4.2×10^{-4}	1.4×10^{-5}	4.5×10^{-4}	4.5×10^{-4}	2×10^{-5}
视觉编码器学习率乘数	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
固定语言模型	✓	×	×	✓	×	×	✓	×	×
学习率调度器	Cosine	Step	Constant	Cosine	Step	Constant	Cosine	Step	Constant
权重衰减	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
梯度裁剪	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
优化器	AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)			AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)			AdamW($\beta_1 = 0.9, \beta_2 = 0.95$)		
BF16 优化器	×	×	×	×	×	×	✓	✓	✓
辅助损失权重	0.001	0.001	0.001	0.001	0.001	0.001	0.0001	0.0001	0.0001
专家偏置校正步数	-	-	-	-	-	-	0	0.001	0
训练 token 数	2.0B	798.5B	19.5B	2.0B	808.9B	20.0B	2.0B	796.5B	19.5B
批次大小	256	2304	64	256	2304	64	256	3360	64
序列长度	4096	4096	4096	4096	4096	4096	4096	4096	4096
序列打包	×	✓	✓	×	✓	✓	×	✓	✓
流水线并行	×	✓	✓	✓	✓	✓	✓	✓	✓

监督微调。在最后阶段，我们通过监督微调增强预训练模型的指令遵循与对话能力。使用我们自有的视觉-语言 SFT 数据，我们在优化所有参数的同时，仅对答案和特殊 token 进行监督，并对系统和用户提示进行掩码处理。为强化对话理解能力，我们将多模态数据与 DeepSeek-V2 [53] 的纯文本对话数据相结合。该方法确保了模型在多样化视觉-语言任务上的稳健表现，包括密集图像描述、通用 VQA、OCR、表格/图表/文档/图像理解、视觉转代码、视觉推理、视觉定位以及语言理解等。

4.2. 超参数与基础设施

DeepSeek-VL2 训练的详细超参数列于表 2 中。我们使用专为大模型设计的高效轻量级平台 HAILLM [30] 进行了训练与评估。在我们的流水线并行策略中，一个主要挑战源于视觉编码器相较于 LLM 模块所具有的独特计算特性。作为模型流水线的起始组件，视觉编码器需要在 GPU 之间进行精细的负载均衡，以避免流水线气泡并优化 GPU 利用率。为此，我们在流水线并行策略中实现了视觉编码器的细粒度层划分。此外，我们在前向与反向传播过程中，于不同数据并行 rank 之间执行图像分块负载均衡，以缓解动态分辨率策略导致的图像分块数量不均衡问题。我们的训练过程还融合了张量并行与专家并行技术，以实现最优训练效率。鉴于部分数据批次仅含文本数据，而另一些则包含图像数据，我们为不同类型的数据设计了两种不同的流水线策略，并可根据需求动态切换。DeepSeek-VL2 的训练分别在 7/10/14 天内完成，所用集群规模分别为 16/33/42 个节点，每个节点均配备 8 张 NVIDIA A100 GPU。

表 3 | 在 OCR 相关多模态基准测试上与最先进模型的对比. †: MoE 模型的激活参数量.

Model	#Params (LLM)	#Params (VE)	#Params (Activated)	DocVQA (测试集)	ChartQA (测试集)	InfoVQA (测试集)	TextVQA (验证集)	OCRBench
闭源模型								
GPT-4V [69]	-	-	-	87.2	78.1	75.1	78.0	645
GPT-4o [32]	-	-	-	92.8	85.7	79.2	77.4	736
Claude 3.5 Sonnet [5]	-	-	-	95.2	90.8	74.1	74.1	788
Gemini-1.5-Pro [81]	-	-	-	93.1	87.2	80.1	78.7	754
开源模型 (0.5B - 3B)								
LLaVA-OV 0.5B [45]	0.5B	0.4B	0.9B	70.0	61.4	41.8	-	-
InternVL2-1B [16]	-	-	0.9B	81.7	72.9	50.9	70.5	754
MM 1.5-1B [107]	-	-	1B	81.0	67.2	50.5	72.5	605
DeepSeek-VL2-Tiny	0.6B[†]	0.4B	1.0B[†]	88.9	81.0	66.1	80.7	809
MolmoE-1B [22]	1.2B [†]	0.3B	1.5B [†]	77.7	78.0	53.9	78.8	-
MiniCPM-V 2.0 [99]	2.4B	0.4B	2.8B	71.9	-	-	74.1	605
InternVL2-2B [16]	1.9B	0.3B	2.2B	86.9	76.2	58.9	73.4	784
Qwen2-VL-2B [88]	1.5B	0.7B	2.2B	90.1	73.5	65.5	79.7	794
MM 1.5-3B [107]	-	-	3B	87.7	74.2	58.5	76.5	657
DeepSeek-VL2-Small	2.4B[†]	0.4B	2.8B[†]	92.3	84.5	75.8	83.4	834
开源模型 (4B - 13B)								
Phi-3.5-Vision [1]	3.8B	0.3B	4.1B	69.3	81.8	36.6	72.0	599
InternVL2-4B [16]	3.8B	0.3B	4.1B	89.2	81.5	67.0	74.4	788
Aria-MoE [46]	3.9B [†]	0.4B	4.3B [†]	92.6	86.4	-	81.1	-
MM 1.5-7B [107]	-	-	7B	88.1	78.6	59.5	76.5	635
LLaVA-OV 7B [45]	7.6B	0.4B	8.0B	87.5	80.0	68.8	-	-
Molmo-7B-O [22]	7.3B	0.3B	7.6B	-	80.4	70.0	80.4	-
MiniCPM-V2.6 [99]	7.6B	0.4B	8.0B	90.8	82.4	-	80.1	852 (CoT)
InternVL2-8B [16]	7.7B	0.3B	8.0B	91.6	83.3	74.8	77.4	794
Qwen2-VL-7B [88]	7.6B	0.7B	8.3B	94.5	83.0	76.5	84.3	845
Pixtral-12B [3]	12.0B	0.4B	12.4B	90.7	81.8 (CoT)	50.8	75.7	-
DeepSeek-VL 7B [59]	6.9B	0.4B	7.3B	-	-	-	-	456
DeepSeek-VL2	4.1B[†]	0.4B	4.5B[†]	93.3	86.0	78.1	84.2	811

5. 评估

5.1. 多模态性能

基准测试 我们在一系列常用基准测试上对 DeepSeek-VL2 进行了全面评估,包括 DocVQA [66]、ChartQA [65]、InfoVQA² [67]、TextVQA [77]、RealWorldQA [95]、OCRBench [57]、AI2D [34]、MMMU [105]、MMStar [13]、MathVista [60]、MME [26]、MMBench、MMBench-V1.1 [58] 以及 MMT-Bench [100]。这些基准测试涵盖了从文档理解、图表解析到现实世界问题解决等多种任务,从而能够全面评估我们模型的能力。为了评估模型的定位能力,我们在 RefCOCO、RefCOCO+ 和 RefCOCOg 基准测试上对 DeepSeek-VL2 进行了测试 [33, 64]。

与最先进模型的对比 在多模态理解基准测试上,我们将 DeepSeek-VL2 与最先进模型进行了对比,包括 LLaVA-OV [45]、InternVL2 [15]、DeepSeek-VL [59]、Qwen2-VL [88]、Phi-3.5-Vision [1]、Molmo [22]、Pixtral [3]、MM1.5 [107] 和 Aria-MoE [46]。结果报告于表 3 和 4。得益于我们的 MoE 架构,DeepSeek-VL2 以更少的激活参数实现了相当或更优的性能。在定位基准测试上,我们将

²鉴于 InfoVQA 包含宽高比极端且尺寸过大的图像,我们在评估时将候选分辨率扩大为 $C_R = \{(m \cdot 384, n \cdot 384) \mid m \in \mathbb{N}, n \in \mathbb{N}, 1 \leq m, n, mn \leq 18\}$ 。

表 4 | 在通用问答与数学相关多模态基准测试上与最先进模型的对比。[†]: MoE 模型的激活参数量。*: 在不同设置下评估。

模型	参数量 (激活)	MMStar	A12D (测试集)	MMMU (验证集)	MME	MMBench (总分)	MMBench (英文测试集)	MMBench-V1.1 (中文测试集)	MMT-Bench	RealWorldQA	MathVista (测试子集)
闭源模型											
GPT-4V [69]	-	56.0	89.4	63.1	1,927	81	80.2	80	64.3	61.4	58.1
GPT-4o [32]	-	63.9	94.2	69.1	2,329	83.4	82.1	82.2	65.5	75.4	63.8
Claude 3.5 Sonnet [5]	-	62.2	94.7	68.3	1,920	79.7	80.7	78.5	-	60.1	67.7
Gemini-1.5-Pro [81]	-	-	94.4	62.2	-	-	-	-	64.5	70.4	63.9
开源模型 (0.5B - 3B)											
LLaVA-OV 0.5B [45]	0.9B	37.7	57.1	31.4	1,478	61.6	55.5	59.6	-	55.6	34.8
InternVL2-1B [16]	0.9B	45.7	64.1	35.4	1,794	65.4	60.7	61.6	49.5	50.3	37.7
MM 1.5-1B [107]	1B	-	59.3	35.8	1,611	-	-	-	-	53.3	37.2
DeepSeek-VL2-Tiny	1.0B [†]	45.9	71.6	40.7	1,915	73.3	69.2	68.3	53.2	64.2	53.6
MolmoE-1B [22]	1.5B [†]	-	86.4*	34.9	-	-	-	-	-	60.4	34
MiniCPM-V 2.0 [99]	2.8B	-	-	38.2	1,809	69.6	68.1	-	-	-	38.7
InternVL2-2B [16]	2.2B	49.8	74.1	36.3	1,877	73.2	70.9	69.6	50.4	57.3	46.3
Qwen2-VL-2B [88]	2.2B	48	74.4	41.1	1,872	74.9	73.5	72.2	54.5	62.9	47.8
MM 1.5-3B [107]	3B	-	65.7	37.1	1,798	-	-	-	-	56.9	44.4
DeepSeek-VL2-Small	2.8B [†]	57.0	80.0	48.0	2,123	82.3	80.3	79.3	62.9	65.4	60.7
开源模型 (4B - 13B)											
Phi-3.5-Vision [1]	4.1B	47.5	78.1	43	-	76	66.1	72.1	53.6	53.6	43.9
InternVL2-4B [16]	4.1B	54.3	78.9	47.9	2,060	78.6	73.9	75.8	55.7	60.7	58.6
Aria-MoE [46]	4.3B [†]	-	-	54.9	-	-	-	-	-	-	66.1
MM 1.5-7B [107]	7B	-	72.2	41.8	1,861	-	-	-	-	62.5	47.6
LLaVA-OV 7B [45]	8.0B	-	81.4	48.8	1,998	80.8	-	-	-	66.3	63.2
Molmo-7B-O [22]	7.6B [†]	-	90.7*	39.3	-	-	-	-	-	67.5	44.5
MiniCPM-V2.6 [99]	8.0B	57.5	82.1	49.8 (CoT)	2,348 (CoT)	81.5	79.3	78.0	60.8	65.0	60.6
InternVL2-8B [16]	8.0B	61.5	83.8	51.8	2,210	81.7	81.2	79.4	60.0	64.4	58.3
Qwen2-VL-7B [88]	8.3B	60.7	83	54.1	2,327	83	80.5	80.7	63.7	70.1	58.2
Pixtral-12B [3]	12.4B	-	-	52.5 (CoT)	-	-	-	-	-	65.4	58 (CoT)
DeepSeek-VL 7B [59]	7.3B	-	-	36.6	-	73.2	-	-	-	-	-
DeepSeek-VL2	4.5B [†]	61.3	81.4	51.1	2,253	83.1	79.6	79.2	63.6	68.4	62.8

DeepSeek-VL2 与 Groudning DINO [56]、UNINEXT [52]、ONE-PEACE [87]、mPLUG-2 [97]、Florence-2 [96]、InternVL2 [16]、Shikra [11]、TextHawk2 [103]、Ferret-v2 [108]、MM1.5 [107] 和 Qwen2 [88] 进行了对比。我们的模型在相似规模下优于其他视觉语言模型 (VLMs)。

5.2. 定性研究

在本节中，我们展示了 DeepSeek-VL2 的不同能力，涵盖从通用问答到视觉故事生成和视觉定位。

通用视觉问答。 得益于我们新的视觉-语言 (VL) 预训练数据集和多样化的监督微调 (SFT) 数据，DeepSeek-VL2 在通用视觉问答方面展现出显著提升的能力，如图 4 所示。总体而言，该模型在密集图像描述方面表现优异，能够识别常见地标、通用视觉知识以及中英文富文本。此外，它在图表理解方面表现良好，能够准确识别属性。进一步地，我们在图 5 中展示了 DeepSeek-VL2 在梗图 (meme) 理解方面的改进，它能够准确描述上下文，并结合有意义的文化背景解释其中的幽默之处。

多图对话。 如图 6 所示，DeepSeek-VL2 在多图对话方面展现出提升的能力。我们的模型能够分析多张图像之间的关联与差异，同时通过整合多张图像的内容实现简单的推理。例如，它能够根据特定食材的图片思考如何烹饪一道菜肴。

Image Description



Describe this scene in detail.

An expansive coastal landscape dominated by towering limestone stacks rising dramatically from turquoise waters of the ocean. These natural formations are part of The Twelve Apostles geological formation located on the Great Ocean Road in Victoria, Australia. In the foreground stands a prominent stack closest to the viewer's perspective; it is flanked by smaller stacks extending into the distance towards the horizon where they gradually diminish due to atmospheric haze or mist shrouding them partially. To the right side of the frame lies a sandy beach that curves gently around as if embracing the rugged cliffs which form the backdrop for this scene. The sky above appears overcast yet diffused sunlight illuminates parts of the cliff faces casting subtle shadows across some areas while highlighting others, adding depth and texture to the overall composition. Small waves break against both the base of the stacks and nearer shoreline creating white foam patterns contrasting sharply with the deep blue-green sea water below. Vegetation can also be seen at various points including atop certain stacks providing hints of greenery amidst predominantly earth-toned rock structures. This serene but powerful seascape captures nature's grandeur through its interplay between landforms, water bodies, light conditions, and vegetation all harmoniously coexisting within the same visual space.

Landmark Recognition




猜猜这是哪里?

这幅图描绘了一个湖边的日落景象,周围有山峦和一些停泊在水面的船只。特别显眼的是中间的一艘大船,船上有着类似中国传统建筑风格的屋檐。这种建筑风格非常像中国古代建筑,尤其是皇家或寺庙建筑。

这种场景可能位于中国的某个著名湖泊,例如杭州的西湖或南京的玄武湖等。特别是如果这艘船是仿古式的游船,那么它很可能位于西湖景区。西湖是中国最著名的湖泊之一,周围有着秀丽的山景和古建筑,常常成为游客观光和体验传统文化的地方。

因此,图片中展示的可能是中国的西湖,那个有着丰富的历史文化和自然美景的地方。

Poet Writing



请给这幅景象写一首诗

General Knowledge



Can you introduce the background story of this cake?

The cake in the image is themed around the iconic cartoon series "Tom and Jerry." This series, created by William Hanna and Joseph Barbera, first aired in 1940 and became one of the most famous animated shows of all time. The story revolves around the never-ending chase between the cunning Tom the Cat and Jerry the Mouse. Each episode typically features Tom making various plans to catch Jerry, only to be outsmarted or thwarted by Jerry's cleverness and cunning. The dynamic duo is a symbol of constant rivalry and humor, and their interactions have been a part of popular culture for decades. This cake captures the essence of their playful and chaotic relationship through its design and decoration.

Text Recognition

《望公山》

浮云出岫晓轻寒, 郁郁苍山展画屏。 几树孤影来绿野, 一泓深水抱金城。 鸿鸣远岫千山起, 鹤伴闲崖万物兴。 尽望烟霞山色秀, 却寻春色过江亭。

Chart Understanding



Can you explain this chart to me?

Comparison of Multilingual Datasets (ar, fr, ru, th, tr, zh)

The chart shows a comparison of multilingual datasets based on the average normalized score across six languages: Arabic (ar), French (fr), Russian (ru), Thai (th), Turkish (tr), and Chinese (zh). The x-axis represents the number of training tokens in billions, while the y-axis shows the average normalized score, which is a measure of model performance.

Key Points:

- Datasets Compared:**
 - FineWeb 2 (Ours)
 - CulturaX
 - HPLT-2
 - CC-100
 - CommonCrawl
 - mC4
 - HPLT
- Trend Analysis:**
 - FineWeb 2 consistently achieves the highest average normalized score, indicating superior performance compared to other datasets across the given range of training tokens.
 - The other datasets show varying levels of performance but tend to converge towards a similar score as the number of training tokens increases.
- Performance Insights:**
 - Initially, FineWeb 2 has a significant advantage over other datasets, which highlights the effectiveness of the data collected in this dataset.
 - As training tokens increase, the performance gap between datasets narrows, suggesting diminishing returns in terms of dataset size.
 - FineWeb 2 demonstrates robustness, maintaining a lead over the other datasets throughout the training process.

Interpretation:

The chart underscores the importance of dataset quality and size in achieving high model performance in multilingual tasks. The superior results of FineWeb 2 indicate its potential to enhance language model training for various applications across multiple languages.

Text Recognition



What's written on the large blackboard.

Serving Soul Since Twenty Twelve

图中在宣传什么东西?

图中展示的是与西湖醋鱼相关的食品和饮品。左边的广告牌上展示的是西湖醋鱼蛋糕 (West Lake Fish Cake), 它是一种以西湖醋鱼为主题食品。右边的广告牌则是在宣传西湖醋鱼咖啡, 这种咖啡被称为“铁狗咖啡”, 其特色是将西湖醋鱼的特色和杭州当地的经验结合在一起, 把西湖醋鱼的味道融入到咖啡中。西湖醋鱼是杭州的一道著名美食, 主要使用草鱼制作, 经过料酒、醋、糖等调料烹饪, 酸甜可口。现在通过将西湖醋鱼加入到其他食品和饮品中, 这一传统美食得到了全新的诠释。

图 4 | DeepSeek-VL2 的通用问答能力。我们的模型具备多方面的能力, 包括地标识别、看图作诗、常识问答、图表理解、文本识别等。

LIFE OF A PhD student



What does this image imply?

The image humorously depicts the **contrasting thoughts** of a PhD student. On the left, the student is **working at a desk**, surrounded by books and papers, **dreaming of a relaxing beach vacation**. On the right, the student is **actually at the beach**, but their mind is still consumed by thoughts of work, represented by stacks of **papers and a laptop**. This illustrates the common experience of PhD students who often **find it difficult to fully disconnect from their research**, even when they are physically away from their work environment.



Explain this meme to me.

This image depicts a humorous situation involving **a young child and a cake**. The child, appearing determined and mischievous, is depicted biting into a large chocolate cake on a table. The text at the top of the image reads: **"Mom told me not to touch it."**

Analysis and Meaning:

- Disobedience theme**: The text implies that the child **has intentionally ignored the instruction** or warning given by their mom. This reflects the **common tendency of children to follow rules reluctantly** and test boundaries when told otherwise.
- Humor in contrast**: The exaggerated act of biting the cake makes the situation absurd, adding to the humor. The over-the-top behavior contrasts with the seemingly simple instruction, creating an unexpected and funny outcome.
- Relatable scenario**: This image humorously captures a widely understood, everyday behavior—**children wanting things they are told they cannot have**, leading to playful and sometimes rebellious actions.
- Visual cue**: The child's action of taking a bite adds a dynamic and engaging element to the humor, amplifying the sense of mischief and delight in the moment.

In short, this image humorously captures the playful defiance of childhood, using the visual of a child **taking a bite of an untouchable cake** to amplify the humor and relatability.



详细解释一下这张梗图

这张梗图展示了一位老人在海边，似乎正在试图用锤子将一些东西打进沙滩的情景。上方配文是“试图挽留假期的我”，显然将老人的行为与“留不住”假期的情况联系在一起。

分析与解读:

- 图像情景与动作对比的幽默性**:
 - 老人的动作显得十分**徒劳**，因为沙滩与水边的沙子**难以固定**，他试图用锤子钉住什么，这显然是一件无法实现的事情。这形象地表现了一个人在**面对不可抗拒的力量（如假期结束）时的无奈和无助**。
- 文字与画面的结合**:
 - “试图挽留假期的我”与图片形成直接关联，画面展现了一种滑稽的、略带夸张的尝试挽留假期的情景。人们在**生活中也常常试图阻止即将结束的假期**，比如试图在假期最后几天拖延，推迟归程等。这个梗将这种心理状态以幽默的方式表现出来。
- 文化背景**:
 - 假期结束是**很多现代工作族的共同感受**，尤其是面临节后综合症，人们会感到非常焦虑、不适应，因此图中的夸张表现方式非常贴近人们真实的情感体验，具有较强的共鸣性。

总结:

这张梗图通过图像和文字的结合，利用一种幽默和夸张的手法，**表达了人们在假期结束时试图“挽留”假期的无奈和心理挣扎**。它反映了一种普遍的情感，尤其是在现代快节奏的生活中，假期是人们难得的放松时间，**当假期即将结束时，人们往往感到焦虑和不舍**。



这张图为什么好笑

这张图是一种常见的幽默梗图，表达了在考试时学生可能会感受到的**“无力感”或“打击感”**。

图中含义:

- 上方**: “当你为考试做足了功课”，表示学生在面对考试时自认为已经做好充分的准备，充满信心。
- 下方**: “第一题:”搭配图片中的骑士被射中眼睛的场景，象征考试开始后，**第一题的内容超出了自己准备的范围或难度，瞬间让人陷入迷茫**，也可能对考试信心造成重大打击。

背景和文化:

- 中世纪骑士与武士形象**: 骑士和武士通常被用来表现强大、无畏的形象，但在这里，骑士被箭射中，造成了**反差感**，增强了对考试的调侃效果。
- 考试的普遍性**: 考试对学生来说是生活中的重要部分，所以这种夸张的对比能引起大家**共鸣**。
- 调侃考试难度**: 第一题通常难度较高，成为考生在考试初期最可能受到冲击的部分，**这种“第一题必超纲”的既视感非常普遍**。

结论:

这张图通过**戏剧化的反差**，幽默地讽刺了在考试中面对挑战时的矛盾心理，很容易让经历过类似情景的人会心一笑。

图 5 | DeepSeek-VL2 的表情包理解能力。我们的模型能够理解表情包中的幽默元素并提供解释。

Multi-image conversation



I am having lunch now and I like sweet things. What would be a good drink to pair with my meal?



A good drink to pair with your meal might be the milkshake from **image 3**. The sweetness of the **milkshake** can complement the flavors of your **spaghetti dish**.



能帮我用这几个食材做一道菜吗?



当然可以! 这个食谱使用了**胡萝卜**、**玉米**和**猪肋骨**, 可以做出的味道非常丰富。你可以尝试这道胡萝卜玉米猪肋骨锅。

胡萝卜玉米猪肋骨锅

材料:

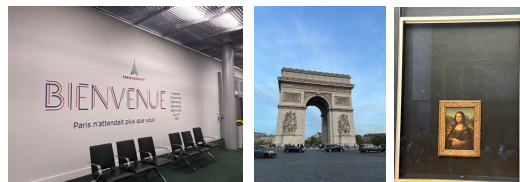
- 猪肋骨 500克
- 胡萝卜 3根
- 玉米 2根
- 葱 适量
- 姜 适量
- 蒜 适量
- 八角 1颗
- 料酒 2汤匙
- 酱油 1汤匙
- 糖 1汤匙
- 盐 适量
- 鸡精或味精 (可选)

步骤:

- **处理食材:****
 - 猪肋骨洗净后切块, 用清水泡30分钟**去血水**, 捞出备用。
 - **胡萝卜去皮**, 切成块; **玉米剥皮**, 切成段。
- **焯水:****
 - 猪肋骨放入锅中, 加入清水和姜片, 烧开**焯水**3分钟, **去除浮沫**, 捞出备用。
- **炒香料:****
 - 锅中留少量底油, 放入八角爆香, 然后加入葱姜蒜炒香。
- **炖煮:****
 - 倒入焯水后的猪肋骨, 翻炒均匀, 加入料酒、酱油、糖、盐, 倒入适量热水, 盖锅炖煮45分钟, 直到肉熟透。
- **加入配菜:****
 - **炖煮过程中**, 加入**胡萝卜**和**玉米**, 继续炖煮20分钟, 直到蔬菜软熟。
- **调味:****
 - 最后加入鸡精或味精, 根据个人口味调味, 搅拌均匀, 即可出锅享用。

这道胡萝卜玉米猪肋骨锅, **味道丰富**, **营养均衡**, 非常适合喜欢尝试不同口感的朋友。

图 6 | DeepSeek-VL2 的多图对话能力。我们的模型能够理解多张图像之间的关系, 并基于此进行推理。



Can you help me write a short travel story based on these photos?



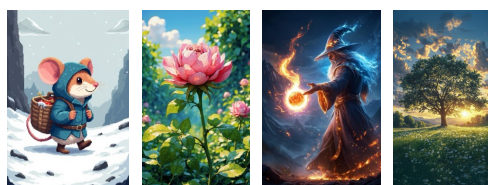
Title: "A Voyage Through Time and Light"

The wait was a symphony of anticipation, each note a silent prayer to the gods of travel. I sat in the Paris Airport terminal, the word "BIENVENUE" etched in bold, inviting letters above me. The Eiffel Tower symbol floated serenely, a promise of what was to come. I reached into my bag, feeling the cool metal of my suitcase handle, and bit my lip, trying not to sink further into the comfort of the grey seats. The ceiling lights buzzed softly, casting a sterile glow that neither cheered nor soothed my nerves.

I boarded the plane, the roar of engines a symphony of promise. As we soared above the clouds, the world below became a patchwork of life—houses, roads, rivers—each a story waiting to be told. Then, the familiar spire of the Arc de Triomphe appeared, its grandeur a testament to human ambition. I watched, entranced, as the plane descended, the iconic archway a beacon of hope and history. The sky, a canvas of blue, welcomed us with a smile.

Finally, we landed. The terminal buzzed with life, a cacophony of languages and laughter. I walked through the glass doors, the world outside a blur of green and grey. I found myself in a room, the Mona Lisa's eyes staring back at me, a mystery untouched by time. The golden frame shimmered, and I felt a connection, a fleeting moment of understanding. The reflection in the glass showed me a different perspective, a reminder that beauty is in the eye of the beholder.

I left the room, the Mona Lisa's enigmatic smile still etched in my mind. The journey had been a dance of light and shadow, a symphony of moments that painted a picture of a city where every corner held a story. As I stepped outside, the Parisian night embraced me, a welcome that felt both ancient and new, just like the city itself.



用这些图片讲一个故事吧



《雪夜的奇遇》

冬日的寒风在山谷中低语，雪花轻轻飘落在老鼠先生的蓝色连帽外套上。他肩上的柳条篮子里装满了红苹果，仿佛是这个寒冷世界中唯一的温暖。老鼠先生微笑着，尽管天气寒冷，他的心中却充满了期待。他知道，今晚将是一个不平凡夜晚。

夜幕降临，老鼠先生走进了一片神秘的玫瑰花园。中央那朵粉红色的玫瑰，花瓣在月光下闪烁着柔和的光芒。老鼠先生轻轻触摸花瓣，突然，花园中的一切都静止了，只有他心中那股莫名的力量在涌动。

就在这时，一位白发苍苍的巫师出现在他面前。巫师手中的火球闪烁着橙黄色的光芒，仿佛在诉说着古老的魔法。巫师告诉老鼠先生，这片花园隐藏着一个秘密，只有真正的勇士才能解开。

老鼠先生鼓起勇气，跟随巫师来到了一片沐浴在金色阳光中的田野。高大的树影在阳光下显得格外神秘。巫师将火球放在树下，瞬间，树干中涌出了一股温暖的力量。老鼠先生感到一股力量从心底涌出，他知道，这是大自然的馈赠。

当夜幕再次降临，老鼠先生带着满篮的红苹果回到了山谷。他微笑着，心中充满了满足与希望。这个冬天，他不仅找到了温暖，还发现了内心的力量。

图 7 | DeepSeek-VL2 的视觉叙事能力。我们的模型可以接受多张图像作为输入，并根据图像用中文或英文讲述一个故事。

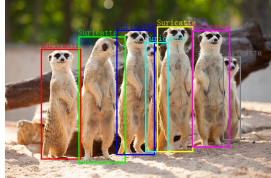
表 5 | 在视觉定位基准测试上与最先进模型的比较。我们不同规模的模型均在同等规模的 MLLM 中取得了最佳结果。

模型	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
视觉模型								
Grounding DINO-Tiny [56]	89.2	91.9	86.0	81.1	87.4	74.7	85.2	84.9
Grounding DINO-Largey [56]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
UNINEXT-H [52]	92.6	94.3	91.5	85.2	89.6	79.8	88.7	89.4
VLM + 任务特定微调								
ONE-PEACE [87]	92.6	94.2	89.3	88.8	92.2	83.2	89.2	89.3
mPLUG-2 [97]	90.1	92.8	86.1	-	-	86.1	84.7	85.1
Florence-2-B [96]	92.6	94.8	91.5	86.8	91.7	82.2	89.8	82.2
Florence-2-L [96]	93.4	95.3	92.0	88.3	92.9	83.6	91.2	91.7
开源 VLM (0.5B - 3B)								
InternVL2-1B [16]	83.6	88.7	79.8	76.0	83.6	67.7	80.2	79.9
DeepSeek-VL2-Tiny	84.7	87.8	78.4	75.9	83.9	67.4	73.8	83.9
InternVL2-2B [16]	82.3	88.2	75.9	73.5	82.8	63.3	77.6	78.3
DeepSeek-VL2-Small	93.9	95.3	91.3	89.4	92.9	84.8	92.6	92.6
开源 VLM (4B - 9B)								
Shikra-7B [11]	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
TextHawk2-7B [103]	91.9	93.0	87.6	86.2	90.0	80.4	88.2	88.1
Ferret-v2-7B [108]	92.8	94.7	88.7	87.4	92.8	79.3	89.4	89.3
InternVL2-8B [16]	87.1	91.1	80.7	79.8	87.9	71.4	82.7	82.7
MM1.5-7B [107]	-	92.5	86.7	-	88.7	77.8	-	87.1
Qwen2-VL-7B [88]	91.7	93.6	87.3	85.8	90.5	79.5	87.3	87.8
DeepSeek-VL2	95.1	96.7	92.7	91.2	94.9	87.4	92.8	92.9

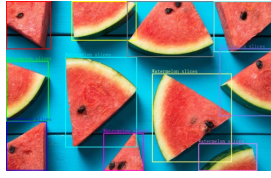
视觉故事生成。 在图 7 中，我们展示了 DeepSeek-VL2 能够根据少量图片创作故事。故事生成能力得益于其强大的通用视觉能力，如地标识别和光学字符识别 (OCR)，文中以绿色文本高亮显示。此外，由于故事生成能力源自纯文本的 DeepSeek Chat 模型，该模型已具备良好的安全对齐，因此在我们内部测试中未观察到 DeepSeek-VL2 产生明显的有害及 NSFW 输出。然而，值得注意的是，现实场景中的创意故事生成需要更多样化的体裁 (*e.g.* 恐怖、喜剧、动作) 和多样的情节类型 (*e.g.* 圆满或悲剧结局)，这可能与大型语言模型/视觉语言模型 (LLM/VLM) 研究中的安全要求存在内在冲突。我们旨在探索解决方案，在考虑这些挑战的同时拓宽故事生成的范围。

视觉定位。 视觉定位是我们为 DeepSeek-VL2 引入的一项新能力。在图 8 中，我们展示了 DeepSeek-VL2 的通用定位能力。有趣的是，尽管我们训练集中的大部分图像来自自然场景，且指代表达式多为物体类别名称或具体描述，但我们发现该模型能够泛化到其他场景 (如梗图和动漫)，并具备识别特定名人和抽象概念的能力。此外，我们在图 10 中展示了 DeepSeek-VL2 具备上下文视觉定位能力。给定第一张图像 (其中某个物体由视觉提示指代)，模型能够在第二张图像中定位同类物体。我们还观察到模型展现出涌现能力。给定一张图像和文本描述，模型能够结合图像与文本信息，在第二张图像中识别出对应的物体。相关示例列于图 10 的第二行和第三行。

Natural Scenes



<ref>Suricata suricatta</ref>



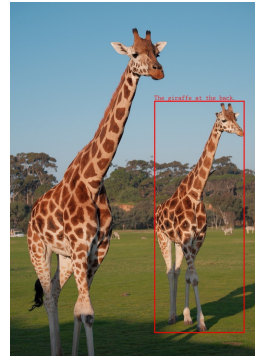
Find all the <ref>Watermelon slices</ref>



Locate the <ref>space needle</ref> in the given image.



Identify <ref>Albert Einstein</ref> in this picture.

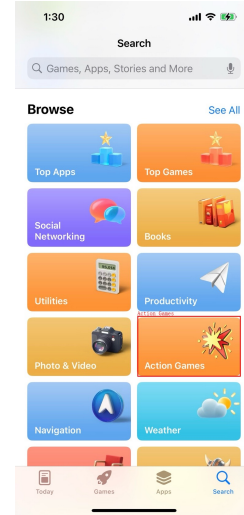


<ref>The giraffe at the back.</ref>

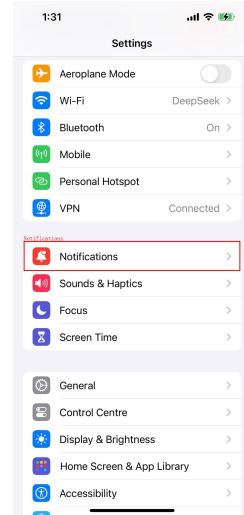


Identify all objects in the image and output them in bounding boxes.

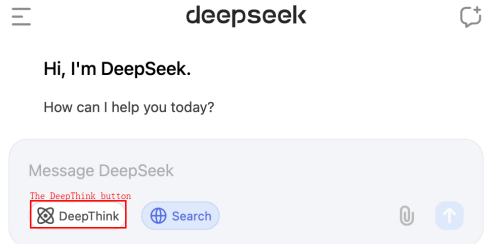
Web UI/Screenshot



<ref>Action Games</ref>



Pinpoint <ref>Notifications</ref> in the image with its coordinates.



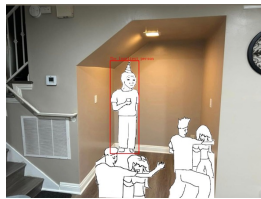
AI-generated, for reference only

Find <ref>The DeepThink button</ref>.

Meme/Anime



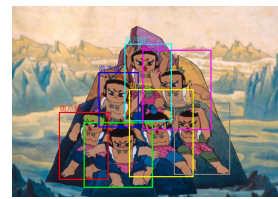
找到<ref>淡定姐</ref>。



<ref>The loneliest man</ref>



<ref>孙悟空</ref>



找到所有的<ref>葫芦娃</ref>。

图 8 | DeepSeek-VL2 的视觉定位能力。我们的模型能够根据类别名称、描述或某些抽象概念来定位物体。

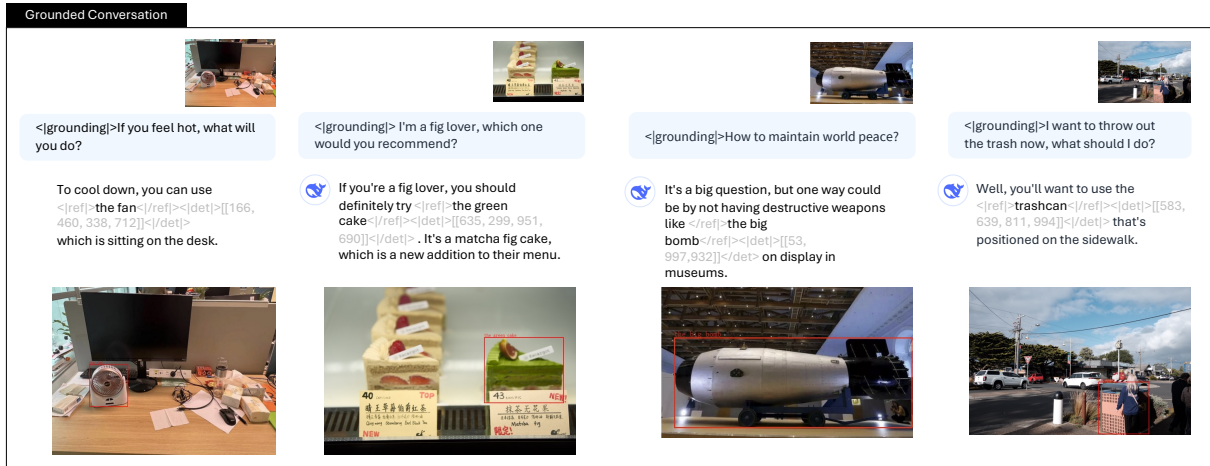


图 9 | 基于 DeepSeek-VL2 的定位对话。我们的模型能够在对图像进行推理的同时识别相关物体的位置，从而使其具备与现实世界交互的可能性。

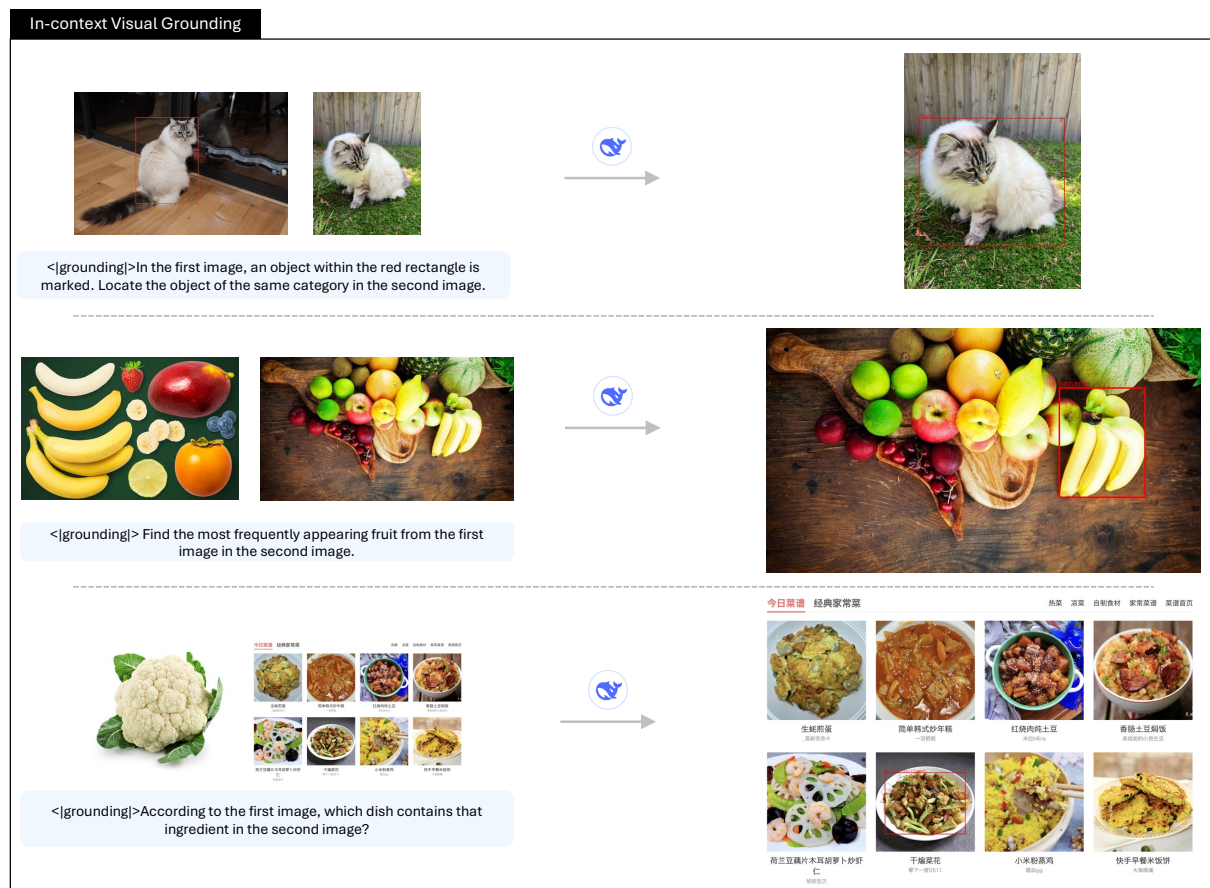


图 10 | 基于 DeepSeek-VL2 的上下文视觉定位。给定一张图像（无论是否包含视觉提示），DeepSeek-VL2 均能在另一张图像中找到相关物体。

定位对话。 借助特殊标记 `<|grounding|>`，DeepSeek-VL2 能够释放其定位对话能力，在回复中准确指出关键物体的位置，如图 9 所示。这使得模型能够更好地与现实世界交互，从而为其在具身智能（embodied AI）及电脑/手机智能体等领域发挥更大作用创造机遇。

6. 结论

在本技术报告中，我们介绍了 DeepSeek-VL2，这是一种基于混合专家（MoE）的视觉-语言模型的增强版本，提供总计 3B、16B 和 27B 参数的规模，对应的激活参数分别为 1.0B、2.8B 和 4.5B。该配置在训练与推理阶段均能实现高效的计算资源消耗。值得注意的是，我们的 3B、16B 和 27B 模型可分别部署在显存为 10 GB、40GB 和 80GB 的单张 GPU 上。我们采用动态分块视觉编码策略，以高效处理具有不同宽高比的高分辨率图像。通过公开代码与预训练模型，我们旨在推动视觉与语言交叉领域的进一步发展与应用。

局限性与未来工作 尽管 DeepSeek-VL2 在多项任务中展现出强大的能力，但仍存在若干可进一步改进的方面。目前，DeepSeek-VL2 的上下文窗口仅允许每个对话会话包含少量图像。我们计划在下一版本中扩展上下文窗口，以支持更丰富的多图像交互。此外，与其他现有的视觉-语言模型（VLMs）类似，该模型在处理模糊图像或未见过的物体时偶尔会面临挑战，这为未来版本提升模型鲁棒性提供了改进空间。最后，尽管 DeepSeek-VL2 在视觉感知与识别任务中表现优异，但我们仍致力于增强其推理能力。这些已明确的改进方向将指导我们后续的研究工作，以持续推动模型能力的提升。

参考文献

- [1] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. [arXiv preprint arXiv:2404.14219](#), 2024.
- [2] agentsea. Wave-ui 25k. <https://huggingface.co/datasets/agentsea/wave-ui-25k>, 2024.
- [3] P. Agrawal, S. Antoniak, E. B. Hanna, B. Bout, D. Chaplot, J. Chudnovsky, D. Costa, B. De Monicault, S. Garg, T. Gervet, et al. Pixtral 12b. [arXiv preprint arXiv:2410.07073](#), 2024.
- [4] A. Amini, S. Gabriel, P. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. [arXiv preprint arXiv:1905.13319](#), 2019.
- [5] Anthropic. Claude 3.5 sonnet. <https://www.anthropic.com/news/claude-3-5-sonnet>, 2024.
- [6] Y. Bai, X. Du, Y. Liang, Y. Jin, Z. Liu, J. Zhou, T. Zheng, X. Zhang, N. Ma, Z. Wang, et al. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. [arXiv preprint arXiv:2403.18058](#), 2024.
- [7] L. Blecher. Latex-ocr — a tool to convert images of latex equations into latex code. <https://github.com/lukas-blecher/LaTeX-OCR>, 2023. Accessed: 2023-10-17.
- [8] O. B. Bohan and H. Face. Megalith 10m dataset. <https://huggingface.co/datasets/madebyollin/megalith-10m>, 2024.
- [9] M. Cai, H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D. Park, and Y. J. Lee. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In [CVPR](#), pages 12914–12923. IEEE, 2024.
- [10] G. H. Chen, S. Chen, R. Zhang, J. Chen, X. Wu, Z. Zhang, Z. Chen, J. Li, X. Wan, and B. Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. [arXiv preprint arXiv:2402.11684](#), 2024.
- [11] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. [arXiv preprint arXiv:2306.15195](#), 2023.
- [12] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. Sharegpt4v: Improving large multi-modal models with better captions. [ECCV](#), 2023.

- [13] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al. Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330, 2024.
- [14] W. Chen, H. Wang, J. Chen, Y. Zhang, H. Wang, S. Li, X. Zhou, and W. Y. Wang. Tabfact: A large-scale dataset for table-based fact verification. In International Conference on Learning Representations.
- [15] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, B. Li, P. Luo, T. Lu, Y. Qiao, and J. Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238, 2023.
- [16] Z. Chen, W. Wang, H. Tian, S. Ye, Z. Gao, E. Cui, W. Tong, K. Hu, J. Luo, Z. Ma, et al. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024.
- [17] A. Cherian, K.-C. Peng, S. Lohit, K. Smith, and J. B. Tenenbaum. Are deep neural networks smarter than second graders? arXiv preprint arXiv:2212.09993, 2022.
- [18] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.
- [19] M. Conover, M. Hayes, A. Mathur, J. Xie, J. Wan, S. Shah, A. Ghodsi, P. Wendell, M. Zaharia, and R. Xin. Free dolly: Introducing the world’s first truly open instruction-tuned llm, 2023. URL <https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm>.
- [20] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. arXiv preprint arXiv:2401.06066, 2024.
- [21] W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, and W. Ping. Nvlm: Open frontier-class multimodal llms. arXiv preprint, 2024.
- [22] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. arXiv preprint arXiv:2409.17146, 2024.
- [23] X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, and Y. Su. Mind2web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems, 36, 2024.

- [24] M. Diem, S. Fiel, F. Kleber, R. Sablatnig, J. M. Saavedra, D. Contreras, J. M. Barrios, and L. S. Oliveira. Icfhr 2014 competition on handwritten digit string recognition in challenging datasets (hdsr 2014). In 2014 14th International Conference on Frontiers in Handwriting Recognition, pages 779–784. IEEE, 2014.
- [25] B. Egan, A. Redden, XWAVE, and SilentAntagonist. Dalle3 1 Million+ High Quality Captions, May 2024. URL <https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions>.
- [26] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. URL <https://arxiv.org/abs/2306.13394>.
- [27] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [28] J. Gu, X. Meng, G. Lu, L. Hou, N. Minzhe, X. Liang, L. Yao, R. Huang, W. Zhang, X. Jiang, C. Xu, and H. Xu. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In NeurIPS, 2022.
- [29] C. He, Z. Jin, C. Xu, J. Qiu, B. Wang, W. Li, H. Yan, J. Wang, and D. Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. arXiv preprint arXiv:2308.10755, 2023.
- [30] High-flyer. HAI-LLM: Efficient and lightweight training tool for large models, 2023. URL <https://www.high-flyer.cn/en/blog/hai-llm>.
- [31] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019.
- [32] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4v(ision) system card. 2023.
- [33] S. Kazemzadeh, V. Ordonez, M. Matten, and T. 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.
- [34] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. 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.

- [35] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023.
- [36] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023.
- [37] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In NeurIPS, 2023.
- [38] M. Koupaee and W. Y. Wang. Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305, 2018.
- [39] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallocci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV, 2020.
- [40] LAION. Laion-aesthetics, 2023. URL <https://laion.ai/blog/laion-aesthetics>. Accessed: 2023-10-27.
- [41] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In NeurIPS, 2023.
- [42] H. Laurençon, A. Marafioti, V. Sanh, and L. Tronchon. Building and better understanding vision-language models: insights and future directions., 2024.
- [43] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh. What matters when building vision-language models?, 2024.
- [44] H. Laurençon, L. Tronchon, and V. Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset, 2024.
- [45] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.
- [46] D. Li, Y. Liu, H. Wu, Y. Wang, Z. Shen, B. Qu, X. Niu, G. Wang, B. Chen, and J. Li. Aria: An open multimodal native mixture-of-experts model. arXiv preprint arXiv:2410.05993, 2024.
- [47] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024.

- [48] L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong, and Q. Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In ACL, 2024.
- [49] L. Li, Y. Wang, R. Xu, P. Wang, X. Feng, L. Kong, and Q. Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. arXiv preprint arXiv:2403.00231, 2024.
- [50] X. Li, F. Zhang, H. Diao, Y. Wang, X. Wang, and L.-Y. Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. arXiv preprint arXiv:2407.08303, 2024.
- [51] Z. Li, X. Yang, K. Choi, W. Zhu, R. Hsieh, H. Kim, J. H. Lim, S. Ji, B. Lee, X. Yan, et al. Mmsci: A dataset for graduate-level multi-discipline multimodal scientific understanding. arXiv preprint arXiv:2407.04903, 2024.
- [52] F. Lin, J. Yuan, S. Wu, F. Wang, and Z. Wang. Uninext: Exploring a unified architecture for vision recognition. In Proceedings of the 31st ACM International Conference on Multimedia, pages 3200–3208, 2023.
- [53] A. Liu, B. Feng, B. Wang, B. Wang, B. Liu, C. Zhao, C. Dengr, C. Ruan, D. Dai, D. Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. arXiv preprint arXiv:2405.04434, 2024.
- [54] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2023.
- [55] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [56] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. 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, 2025.
- [57] Y. Liu, Z. Li, B. Yang, C. Li, X. Yin, C.-l. Liu, L. Jin, and X. Bai. On the hidden mystery of ocr in large multimodal models. arXiv preprint arXiv:2305.07895, 2023.
- [58] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al. Mmbench: Is your multi-modal model an all-around player? In European Conference on Computer Vision, pages 216–233. Springer, 2025.
- [59] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang, et al. Deepseek-vl: towards real-world vision-language understanding. arXiv preprint arXiv:2403.05525, 2024.

- [60] P. Lu, H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In The Twelfth International Conference on Learning Representations.
- [61] P. Lu, L. Qiu, J. Chen, T. Xia, Y. Zhao, W. Zhang, Z. Yu, X. Liang, and S.-C. Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. arXiv preprint arXiv:2110.13214, 2021.
- [62] C. Ma, Y. Jiang, J. Wu, Z. Yuan, and X. Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In European Conference on Computer Vision, pages 417–435. Springer, 2025.
- [63] Y. Ma, X. Liu, X. Chen, W. Liu, C. Wu, Z. Wu, Z. Pan, Z. Xie, H. Zhang, L. Zhao, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. arXiv preprint arXiv:2411.07975, 2024.
- [64] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 11–20, 2016.
- [65] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244, 2022.
- [66] M. Mathew, D. Karatzas, and C. 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.
- [67] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706, 2022.
- [68] A. Mitra, H. Khanpour, C. Rosset, and A. Awadallah. Orca-math: Unlocking the potential of slms in grade school math, 2024.
- [69] OpenAI. Gpt-4v(ision) system card. <https://openai.com/research/gpt-4v-system-card>, 2023.
- [70] B. Peng, C. Li, P. He, M. Galley, and J. Gao. Instruction tuning with gpt-4. arXiv preprint arXiv:2304.03277, 2023.
- [71] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824, 2023.

- [72] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, pages 2641–2649, 2015.
- [73] B. Saleh and A. Elgammal. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855, 2015.
- [74] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar. Kvqa: Knowledge-aware visual question answering. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 8876–8884, 2019.
- [75] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8430–8439, 2019.
- [76] W. Shi, Z. Hu, Y. Bin, J. Liu, Y. Yang, S.-K. Ng, L. Bing, and R. K.-W. Lee. Mathllava: Bootstrapping mathematical reasoning for multimodal large language models. arXiv preprint arXiv:2406.17294, 2024.
- [77] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317–8326, 2019.
- [78] V. Singla, K. Yue, S. Paul, R. Shirkavand, M. Jayawardhana, A. Ganjdanesh, H. Huang, A. Bhatele, G. Somepalli, and T. Goldstein. From pixels to prose: A large dataset of dense image captions. CoRR, abs/2406.10328, 2024.
- [79] K. Srinivasan, K. Raman, J. Chen, M. Bendersky, and M. Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In SIGIR, page 2443–2449, 2021.
- [80] K. Sun, J. Pan, Y. Ge, H. Li, H. Duan, X. Wu, R. Zhang, A. Zhou, Z. Qin, Y. Wang, J. Dai, Y. Qiao, L. Wang, and H. Li. Journeydb: A benchmark for generative image understanding. In NeurIPS, 2023.
- [81] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.
- [82] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. Yfcc100m: The new data in multimedia research. Communications of the ACM, 59(2):64–73, 2016.

- [83] S. Tong, E. L. Brown II, P. Wu, S. Woo, A. J. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In The Thirty-eighth Annual Conference on Neural Information Processing Systems.
- [84] S. Toshniwal, W. Du, I. Moshkov, B. Kisacanin, A. Ayrapetyan, and I. Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. arXiv preprint arXiv:2410.01560, 2024.
- [85] J. Wang, P. Zhang, T. Chu, Y. Cao, Y. Zhou, T. Wu, B. Wang, C. He, and D. Lin. V3det: Vast vocabulary visual detection dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 19844–19854, 2023.
- [86] L. Wang, H. Gao, C. Zhao, X. Sun, and D. Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. CoRR, abs/2408.15664, 2024. doi: 10.48550/ARXIV.2408.15664. URL <https://doi.org/10.48550/arXiv.2408.15664>.
- [87] P. Wang, S. Wang, J. Lin, S. Bai, X. Zhou, J. Zhou, X. Wang, and C. Zhou. One-peace: Exploring one general representation model toward unlimited modalities. arXiv preprint arXiv:2305.11172, 2023.
- [88] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.
- [89] W. Wang, M. Shi, Q. Li, W. Wang, Z. Huang, L. Xing, Z. Chen, H. Li, X. Zhu, Z. Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. arXiv preprint arXiv:2308.01907, 2023.
- [90] W. Wang, Y. Ren, H. Luo, T. Li, C. Yan, Z. Chen, W. Wang, Q. Li, L. Lu, X. Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In European Conference on Computer Vision, pages 471–490. Springer, 2025.
- [91] Y. Wei, Z. Wang, J. Liu, Y. Ding, and L. Zhang. Magicoder: Empowering code generation with OSS-instruct. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 52632–52657. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/wei24h.html>.
- [92] C. Wendler. wendlerc/renderedtext.
- [93] C. Wendler. Renderedtext dataset. <https://huggingface.co/datasets/wendlerc/RenderedText>, 2023. Accessed: 2023-10-17.
- [94] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. arXiv preprint arXiv:2410.13848, 2024.

- [95] xAI. Grok-1.5 vision preview. 2024.
- [96] B. Xiao, H. Wu, W. Xu, X. Dai, H. Hu, Y. Lu, M. Zeng, C. Liu, and L. Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4818–4829, 2024.
- [97] H. Xu, Q. Ye, M. Yan, Y. Shi, J. Ye, Y. Xu, C. Li, B. Bi, Q. Qian, W. Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In International Conference on Machine Learning, pages 38728–38748. PMLR, 2023.
- [98] Z. Xu, F. Jiang, L. Niu, Y. Deng, R. Poovendran, Y. Choi, and B. Y. Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. arXiv preprint arXiv:2406.08464, 2024.
- [99] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024.
- [100] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. In Forty-first International Conference on Machine Learning.
- [101] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. 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.
- [102] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models. arXiv preprint arXiv:2309.12284, 2023.
- [103] Y.-Q. Yu, M. Liao, J. Zhang, and J. Wu. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. arXiv preprint arXiv:2410.05261, 2024.
- [104] Y. Yuan, X. Liu, W. Dikubab, H. Liu, Z. Ji, Z. Wu, and X. Bai. Syntax-aware network for handwritten mathematical expression recognition. arXiv preprint arXiv:2203.01601, 2022.
- [105] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. 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.

- [106] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 11975–11986, 2023.
- [107] H. Zhang, M. Gao, Z. Gan, P. Dufter, N. Wenzel, F. Huang, D. Shah, X. Du, B. Zhang, Y. Li, et al. Mm1.5: Methods, analysis & insights from multimodal llm fine-tuning. arXiv preprint arXiv:2409.20566, 2024.
- [108] H. Zhang, H. You, P. Dufter, B. Zhang, C. Chen, H.-Y. Chen, T.-J. Fu, W. Y. Wang, S.-F. Chang, Z. Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv preprint arXiv:2404.07973, 2024.
- [109] R. Zhang, X. Wei, D. Jiang, Y. Zhang, Z. Guo, C. Tong, J. Liu, A. Zhou, B. Wei, S. Zhang, P. Gao, and H. Li. Mavis: Mathematical visual instruction tuning, 2024. URL <https://arxiv.org/abs/2407.08739>.
- [110] B. Zheng, B. Gou, J. Kil, H. Sun, and Y. Su. Gpt-4v(ision) is a generalist web agent, if grounded. 2024. URL <https://openreview.net/forum?id=piecKJ2D1B>.
- [111] X. Zheng, D. Burdick, L. Popa, P. Zhong, and N. X. R. Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. Winter Conference for Applications in Computer Vision (WACV), 2021.
- [112] X. Zhong, E. ShafieiBavani, and A. Jimeno-Yepes. Image-based table recognition: Data, model, and evaluation. In ECCV, volume 12366, pages 564–580, 2020.