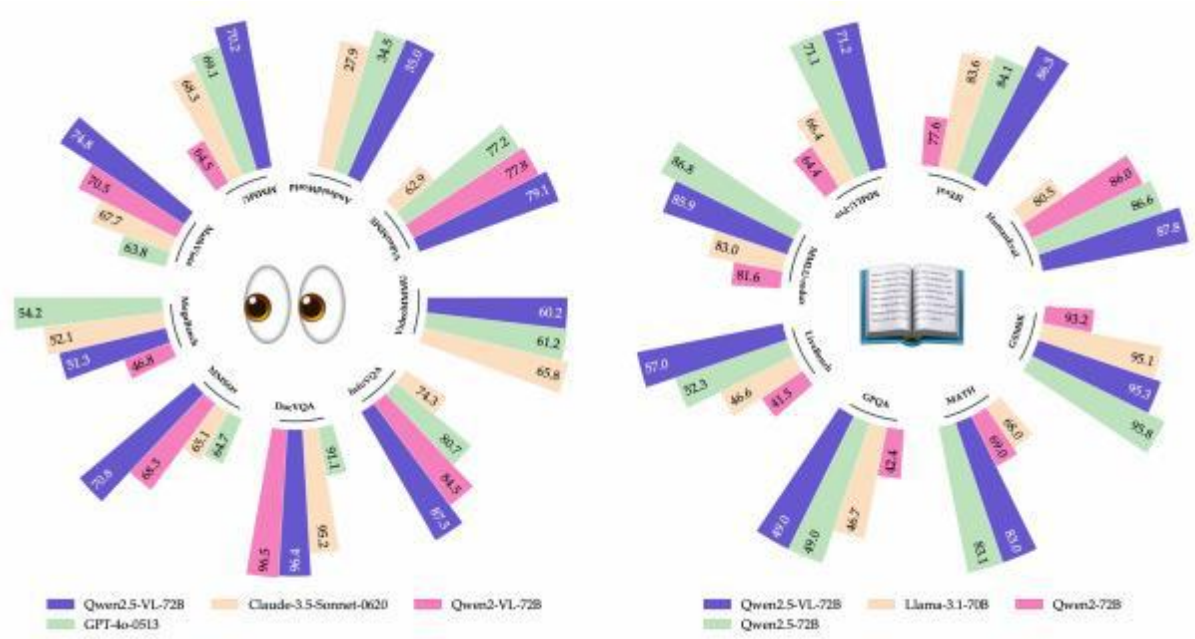
Qwen2.5-VL技术报告

阿里巴巴集团Qwen团队

https://chat.qwenlm.ai
https://huggingface.co/Qwen
https://modelscope.cn/organization/qw恩
https:// github。com/ QwenLM/ Qwen2.5- VL

摘要

我们介绍了Qwen2.5-VL，Qwen视觉语言系列的最新旗舰模型，它展示了在基础功能和创新功能方面的重要广告改进。Qwen2.5- VL通过增强的视觉识别、精确的对象局部化、鲁棒的文档解析和长视频理解，实现了在理解和与世界交互方面的重大飞跃。Qwen2.5- VL的一个sta ndout特性是它能够使用边界框或点精确地定位对象。它提供了来自发票、表单和表格的健壮的结构化数据扩展操作，以及对图表、图表和布局的详细分析。为了处理复杂的输入，Qwen2.5- VL引入了动态分辨处理和绝对时间编码，使其能够通过二级事件定位处理不同时间（长达数小时）的视频。这允许模型原生地感知痉挛的尺度和时间动态，而不依赖于传统的归一化技术。通过训练一个本地动态分辨率视觉变压器（ViT）和不良窗口注意，我们显著减少了补偿在保持本地分辨率的同时增加所有开销。因此，Qwen2.5-VL不仅在静态图像和文档理解方面表现出色，而且作为一个交互式视觉代理，能够在操作计算机和移动设备等现实场景中进行推理、工具使用和任务执行。该模型实现了跨领域的强泛化。Qwen2.5- VL有三种尺寸，解决了从边缘A智能到高性能计算的各种用例。旗舰的Qwen2.5- VL-72B模型匹配了最先进的模型，如GPT-4o和Claude 3.5十四行诗，特别是在文档和图表理解方面表现出色。规模较小的Qwen2.5-VL-7B和Qwen2.5- VL-3 B模型的性能优于同类竞争对手，即使在资源有限的环境中也提供了强大的能力。此外，Qwen2.5-VL保持了强大的语言性能，保留了Qwen 2.5 LLM的核心语言能力。



1介绍

大型视觉语言模型（LVLMs）(OpenAI， 2024;人类的，2024a;团队等人， 20 23;Wang等人， 2024f)代表了人工智能的一个关键突破，标志了多模式理解交互的转换方法。通过将视觉感知与自然语言处理无缝集成，这些高级模型从根本上重塑了机器解释和分析跨不同领域的复杂信息的方式。尽管多模态大型语言模型取得了重大进步，但这些模型目前的功能可以比作亚桑德威奇公司在各种任务中的中间层，但没有达到出色的性能。细粒度的视觉任务是这个类比的基础层。在Qwen2.5-VL的这次迭代中，我们致力于探索细粒度的感知能力，旨在为lvlm建立一个强大的基础，并为现实世界的应用创建一个代理放大器。该框架的顶层是多模态推理方法，它通过利用最新的Qwen2.5 LLM和使用多模态QA数据构建而得到增强。

一系列的作品促进了多模态大型模型的发展，其特点是建筑设计、视觉输入处理和数据管理。lvlm进步的主要驱动力之一是体系结构的持续创新。在(Alayrac等人， 2022; 李等人， 2022a;2023b;Liu等人，20 23b;a;Wang等人， 2024i;张等人， 2024b;Wang等人， 2023)已经逐步形成了当前的范式，它通常由一个视觉编码器、一个跨模态投影仪和LLM组成。细粒度感知模型已经成为另一个关键领域。模型，如(Xiao et al.， 2023;Liu等人， 2023c;Ren等人， 2024;张等人， 2024a;d;彭等人，2023;Deitke等人， 2024)已经突破了在详细的视觉理解方面有可能实现的界限。Omni的架构(Li et al.，2024g;2025b;叶氏等人， 2024)和MoE (Riquelme等人， 2021;Lee等人， 2024;李等人，2024h;c;吴等人， 2024b)也激发了lvlm的未来发展。视觉编码器的增强(Chen等人，2023;Liu等人， 2024b;梁等人， 2025)和分辨率缩放(Li等人， 2023c;Ye等人，2023;李等人，2023a)在提高实际视觉理解的质量方面发挥了关键作用。管理更多样化的场景和更高质量的数据是培训高级lvlm的必要步骤。在(Guo et al.， 2024;陈等人， 2024d;Liu等人， 2024a;陈等人，2024a;Tong等人， 2024;李氏等人， 2024a)是对这一努力的非常有价值的贡献。

然而，尽管视觉语言模型取得了显著的进展，但它们目前面临着发展瓶颈，包括计算复杂性、有限的上下文理解、糟糕的细粒度视觉感知，以及在不同序列长度上不一致的性能。

在本报告中，我们介绍了最新的工作Qwen2.5-VL，它延续了Qwen系列的开源哲学，在各种基准测试上实现甚至超越了顶级的闭源代码模型。从技术上讲，我们的贡献有四方面： (1)在视觉编码器中实现窗口注意以优化推理效率；(2)引入动态FPS采样，将动态分辨率扩展到时间维度，并在不同采样率下实现全面的视频理解；

(3)通过调整绝对时间，在时间域升级MRoPE，从而促进最复杂的时间序列学习；(4)我们在管理训练前和监督微调的高质量数据方面做出了重大努力，进一步将训练前语料库从1.2万亿代币扩展到4.1万亿代币。

Qwen2.5- VL的发光特点如下：

- 强大的文档解析功能： Qwen2.5-VL将文本识别升级到全文档解析，擅长处理多场景、多语言和各种内置（手写、表格、图表、化学公式和音乐表）文档。
- 跨格式的精确对象接地： Qwen2.5- VL解锁提高了检测、指向和计数对象的精度，适应了绝对坐标和JSON格式的高级空间推理。
- 超长视频理解和细粒度视频接地：我们的模型将本地动态分辨率扩展到时间维度，增强了在数秒内理解vi演示的能力，同时提取事件片段。
- 针对计算机和移动设备的增强代理功能：利用先进的基础、推理和决策能力，在智能手机和计算机上使用优越的代理功能来增强模型。

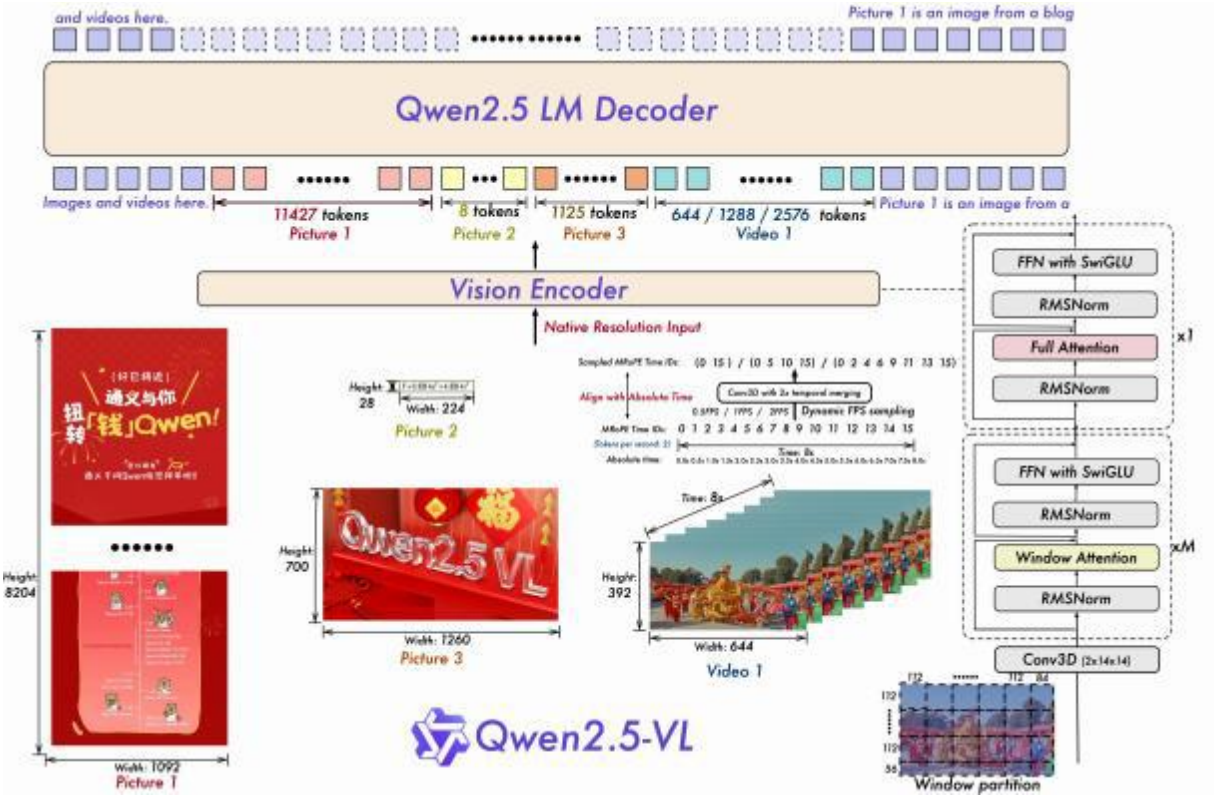


图1：Qwen2.5-VL框架演示了一个视觉编码器和一个语言模型解码器的集成，以处理多模态输入，包括图像和视频。视觉编码器被设计用来处理输入t的原生分辨率，并支持动态FPS采样。不同大小的图像和具有不同FPS速率的视频帧被动态地映射到不同长度的令牌序列上。值得注意的是，MRoPE将时间id与时间维度上的绝对时间对齐，使模型能够更好地理解时间动态，如事件的p ace和精确的矩定位。处理后的视觉数据随后被输入Qwen2.5 LM解码器。我们重新设计了视觉变压器（ViT）架构，结合了先进的组件，如具有SwiGLU激活的FFN、用于标准化的RMSNorm和基于窗口的注意机制来提高性能和效率。

2方法

在本节中，我们首先概述了Qwen2.5- VL系列模型的架构更新，并提供了数据和培训细节的概述。

2.1模型架构

Qwen2.5-VL的整体模型架构由三个组成部分组成：

大型语言模型： TheQwen2.5-VL系列采用大型语言模型作为其基本组件。该模型用Qwen2.5 LLM中预先训练的权重进行初始化。为了更好地满足多模态理解的要求，我们将一维RoPE（旋转位置嵌入）修改为多模态旋转位置嵌入。

视觉编码器：Qwen2.5-VL的视觉编码器采用了重新设计的视觉变压器（ViT）架构。在结构上，我们结合了2D-RoPE和窗口注意来支持本机输入分辨率，同时加速了整个视觉编码器的计算。在训练和推理过程中，输入图像的高度和宽度被调整到28的倍数，然后被输入ViT。视觉编码器通过将图像分割成14步的补丁来处理图像，生成一组图像特征。我们在第2.1.1节中提供了对视觉编码器的更详细的介绍。

基于MLP的视觉语言合并：为了解决长图像特征序列带来的效率挑战，我们采用了一种简单而有效的方法来压缩特征序列，然后将其输入大型语言模型（LLM）。具体来说，而不是直接使用原始补丁

利用视觉变换器（ViT）提取的特征，我们首先对四个斑块特征的空间相邻集进行分组。然后将这些分组的特征连接起来，并通过两个多层感知器（MLP）将它们投影到与中使用的文本嵌入数据对齐的维度中局部白细胞运动该方法不仅计算了计算代价，而且为动态压缩不同长度的图像特征序列提供了一种灵活的方法。

在表1中 ，详细介绍了Qwen2.5- VL的架构和配置。

配置	Qwen2.5- VL-3B	Qwen2.5- VL-7B	Qwen2.5- VL-72B
视觉变压器（V iT）			
隐藏的大小	1280	1280	1280
#层	32	32	32
# Num头	16	16	16
中间尺寸	3456	3456	3456
修补程序大小	14	14	14
窗口大小	112	112	112
全注意块索引	{7,15, 23, 31}	{7,15, 23, 31}	{7,15, 23, 31}
视觉语言合并			
在通道中	1280	1280	1280
输出通道	2048	3584	8192
大型语言模型（LLM）			
隐藏的大小	2048	3,584	8192
#层	36	28	80
# KV头	2	4	8
头部尺寸	128	128	128
中间尺寸	4864	18944	29568
嵌入接合	√	X	X
词汇量	151646	151646	151646
#培训的令牌	4.1T	4.1T	4.1T

表1：Qwen2.5-VL的配置。

2.1.1快速和高效的视觉编码器

视觉编码器在多模态大型语言模型（MLLMs）中起着关键的作用。为了解决在训练和推理过程中由于本地分辨率输入而产生的计算负荷计算所带来的挑战，我们重新设计了视觉变压器（ViT）架构。一个关键的问题来自于与处理不同大小的图像相关的二次计算复杂度。为了缓解这种情况，我们在大多数布局中引入了窗口注意，这确保了计算成本与斑块的数量呈线性比例，而不是二次比例。在我们的体系结构中，只有四层具有充分的自我注意，而其余的层利用窗口注意，最大窗口大小为112×112（对应8个×8补丁）。对小于112×112的区域进行不填充的处理，保持其原始分辨率。这种设计允许模型在输入分辨率下原生运行，避免了不必要的任意缩放或失真。

在位置编码方面，我们采用二维旋转波位置嵌入（RoPE）来有效地捕获二维空间中的空间关系。此外，为了更好地处理视频输入，我们将我们的方法扩展到3D补丁分区。具体来说，我们使用14个×14个图像补丁作为基本单元，与传统的静态图像ViTs相一致。对于视频数据，两个连续的帧被分组在一起，显著减少了输入到语言模型中的令牌的数量。这种设计不仅保持了与现有架构的兼容性，而且提高了处理顺序视频数据的效率。

为了简化整体的网络结构，我们将ViT体系结构与大型语言模型（llm）的设计原则更紧密地对齐。具体来说，我们采用RMSNorm(张&Sennrich， 2019) 对于标准化和SwiGLU (Dauphin等人，2017)作为函数上的激活者。这些选择提高了计算效率和模型的视觉和语言组件之间的兼容性。

在训练方面，我们从头开始训练重新设计的ViT。培训过程包括多个阶段，包括CLIP预训练、视觉-语言对齐和端到端微调。为了确保不同输入分辨率之间的鲁棒性，我们在原生分辨率上采用动态采样

训练图像根据其原始的纵横比进行随机采样，使模型能够有效地推广到不同分辨率的输入。该方法不仅提高了模型的适应性，而且保证了不同大小的视觉数据的训练。

2.1.2本地动态分辨率和帧率

Qwen2.5-VL引入了空间和时间维度的进步，以处理不同的多模态输入。

在空间域内，Qwen2.5-VL动态地将不同大小的图像转换为具有相应长度的标记序列。与传统的坐标规格化方法不同，我们的模型直接使用输入图像的实际尺寸来表示边界框、点和其他空间特征。这允许模型固有地学习比例信息，提高其处理不同分辨率图像的能力。

对于视频输入，Qwen2.5- VL结合了动态帧率（FPS）训练和绝对时间编码。通过适应可变帧率，该模型可以更好地捕捉视频内容的时间动态。与其他合并文本时间戳或使用额外的头来实现时间接地的方法不同，我们引入了一种新颖而有效的策略，将MroPEid直接与时间戳对齐。这种方法允许模型通过时间维度id之间的间隔来理解时间的节奏，而不需要任何额外的计算开销。

2.1.3多模态旋转位置嵌入与绝对时间对齐

位置嵌入对于在视觉和语言模式中建模顺序数据都是至关重要的。在Qwen2-VL中引入的多模态旋转位置嵌入（MRoPE）的基础上，我们扩展了其功能，以更好地处理视频中的入口信息。

Qwen2-VL中的MRoPE将位置嵌入分解为三个不同的组件：时间I、高度和宽度，以有效地建模多模态输入。对于文本输入，所有三个组件都使用相同的位置id，使得MRoPE在功能上等同于传统的1D RoPE (Su等人， 2024).对于图像，时间ID在视觉标记之间保持不变，而唯一的ID则根据每个标记在图像中的空间位置分配给高度和宽度组件。当处理被视为帧序列的视频时，每一帧的时间ID都有增量，而高度和宽度组件遵循与静态图像相同的分配模式。

然而，在Qwen2-VL中，MRoPE中的主要位置id与输入帧数有关，这并没有解释内容变化的速度或视频内事件的绝对时间。为了解决这一限制，Qwen2.5-VL引入了一个关键的改进：将MRoPE的时间组件与绝对时间对齐。如图1所示 ，通过利用时间id之间的间隔，该模型能够在具有不同FPS采样率的视频中学习一致的时间对齐。

2.2培训前

在本节中，我们首先描述训练前数据集的构建，然后概述整个训练的p空间和配置。

2.2.1Pre-Training数据

与Qwen2-VL相比，我们显著增加了训练前数据的数量，从1.2万亿代币增加到大约4万亿代币。我们的预训练数据集是通过多种方法的组合来构建的，包括清理原始的网络数据，合成数据等。该数据集包括各种各样的多模态数据，如图像标题、交错图像-文本数据、光学字符识别（OCR）数据、视觉知识（如名人、地标、植物和动物识别）、多模态学术问题、定位数据、文档解析数据、视频描述、视频定位和基于代理的交互数据。在整个训练过程中，我们仔细调整了这些数据类型的比例和比例，以优化学习结果。

交错的图像-文本数据对于多模态学习是必要的，提供了三个关键的好处：(1)使上下文中的I获得同时的视觉和文本线索(Alayrac 以及其他人 2022)，(2)在图像缺失时保持强大的纯文本功能(Lin等人，20 24)和(3)包含广泛的一般信息。然而，许多可用的交错数据

缺乏有意义的文本-图像关联，而且我经常很嘈杂，限制了它对复杂推理和创造性生成的有用性。

为了解决这些挑战，我们开发了一个用于评分和清理数据的管道，以确保只使用高质量的、相关的交错数据。我们的过程包括两个步骤：标准数据清理(Li 以及其他人2024e)，随后采用了一个使用内部评估模型的四阶段评分系统。评分标准包括：(1)纯文本质量，(2)图像文本相关性，(3)图像文本互补性，(4)信息密度平衡。这种细致的方法提高了模型执行复杂推理和生成连贯的多模态内容的能力。

以下是对这些图像-文本评分标准的描述：

图像-文本相关性：得分越高，表示图像和文本之间的联系越强，图像会有意义地补充、解释或扩展文本，而不仅仅是装饰它。

信息互补性：得分越高，表示图像和文本之间的互补性信息越大。每一个都应该提供独特的细节，共同创造一个完整的叙述。

信息密度的平衡：得分越高，意味着图像和文本之间的信息分布越平衡，避免过多的文本或图像信息，确保两者之间的适当平衡。

我们采用具有绝对位置坐标的原生分辨率训练，目的是实现更准确的感知世界。相反，相对坐标不能有效地表示图像中物体的原始大小和位置。为了解决这个限制，Qwen2.5-VL在训练过程中使用基于输入图像的实际尺寸的坐标值来表示边界框和点。该方法确保了模型能够更好地捕捉目标的真实尺度和空间关系，从而提高了目标检测和定位等任务的性能。

为了提高接地能力的通用化，我们开发了一个全面的数据集，包含包含引用表达式的边界框和点，利用公开的数据集和专有数据。我们的方法包括将数据合成为各种格式，包括-XML、JSON和自定义格式，并使用诸如复制-粘贴增强等技术(Ghiasi 以及其他人2021)和现成模型的合成，如接地DINO (Liu等，2023c)和SAM (Kirillov等人， 2023).这一点有助于对接地能力进行更可靠的评估和提高。

为了提高模型在开放词汇表检测方面的性能，我们将训练数据集扩展到包括超过10,000个对象类别。此外，为了提高模型在极端目标检测场景中的有效性，我们在查询中合成了不存在的对象类别，并为每个对象构建了包含多个实例s的图像数据。

为了确保优越的基于点的对象接地能力，我们构建了一个包含公开可用数据和合成数据的综合指向数据集。具体来说，数据源包括公共点g和来自PixMo的计数数据(Deitke等人， 2024)，公开可访问的目标接地数据（来自目标检测和实例分割任务），以及由自动管道合成的数据，用于生成指向某些图像细节的精确数据。

当文档全解析数据到trainQwen2.5-VL时，我们合成了大量的文档数据语料库。解析文档内容的传统方法通常依赖于单独的模型来处理布局分析、文本提取、图表解释和插图处理。相比之下，Qwen2.5- VL旨在使通用目的模型具有解析、理解和转换文档格式的全面功能。具体来说，我们在文档中加入了多种多样的元素，如表格、图表、方程式、自然或合成图像、音乐效果表和化学公式。这些元素统一采用HTML格式，将布局框信息和插图描述集成到HTML标签结构中。我们还根据典型的阅读序列s丰富了文档布局，并在基于html的地面真相中包含了每个模块对应的坐标，如段落和图表。这种创新的方法允许任何文档的完整信息，包括其布局、文本、图表和插图，以一种标准化和统一的方式来表示。因此，Qwen2.5-VL实现了se没有对多模态文档元素的集成，从而促进了更高效和准确的文档理解和信息转换。

下面是QwenVL HTML格式：

QwenVL HTML格式

```
< html>< body>
#段
<p数据-bbox="x1y1x2y2">内容</p>
#表
<样式>表{id}样式</样式><表数据-bbox="x1y1x2y2"类="表{id}">表内容</表>
#图表
< div类="图表"数据-b箱="x1y1x2y2"><img数据-b箱="x1y1x2y2"/><表>图表内容</表></div>
公式
< div类="公式"数据-bbox="x1y1x2y2"><img数据-bbox="x1y1x2y2"/><<>公式内容
</<></div>
#图像标题
< div类="图像标题"数据-bbox="x1y1x2y2"><img数据-bbox="x1y1x2y2"/><p>图像标题
</p></div>
#图像ocr
< div类="图像ocr"数据-bbox="x1y1x2y2"><img数据-bbox="x1y1x2y2"/><p>图像ocr </p></div>
音乐工作表
< div类="音乐页"格式="abc符号"数据-bbox="x1y1x2y2"><img数据-bbox="x1y1x2y2"/> <版本>
音乐页内容</版本></div>
化学配方含量
< div类="化学公式"格式="微笑"数据-bbox="x1y1x2y2"><img数据-bbox="x1y1x2y2"/><数据
>化学公式内容</<></div>
</html></body>
```

这种格式确保所有文档元素都以结构化和可访问的方式表示，使Qwen2.5-VL能够高效地处理和理解。

OCR数据从不同的来源收集和整理数据，以提高OCR的性能，包括合成数据、开源数据和内部收集的数据。合成数据是通过一个可视化的文本生成引擎进行生成的，从而在野外生成高质量的文本图像。为了支持更广泛的语言并增强多语言功能，我们合并了一个大规模的多语言OCR数据集。该数据集包括对不同语言的支持，如法语、德语、意大利语、西班牙语、葡萄牙语、阿拉伯语、俄语、日语、韩语和越南语。该数据集是精心策划的，以确保多样性和质量，利用了高质量的合成图像和真实的自然场景图像。这种组合确保了在不同语言上下文下的鲁棒性能，并提高了模型对不同文本外观和环境条件的适应性。格式图类型的数据，我们合成了100万个样本和可视化库，包括矩阵库、海运库和程序库，包括图表类别，如柱状图、关系图和热图。关于表格数据，我们处理了600万真实世界 通过离线的端到端选项卡le识别模型进行采样，随后过滤出低置信度的表、重叠的表和单元格密度不足的表。

视频数据为了确保增强理解每秒变化帧（FPS）的鲁棒性，我们在训练期间对FPS进行动态采样，以实现训练数据集中更均匀的FPS分布表示。此外，对于长度超过半小时的视频，我们通过目标合成管道专门构建了一组长视频字幕。对于视频接地数据，我们以基于第二格式的格式和小时-分钟-秒帧（hmsf）格式制定了时间戳，确保模型能够准确理解和输出各种格式的时间。

代理数据我们增强了感知和决策能力，以构建Qwen2.5- VL的代理能力。为了进行感知，我们在移动、网络和桌面平台上收集屏幕截图。一个合成数据引擎用于生成屏幕截图标题和UI元素接地注释。上限任务帮助Qwen2.5-VL理解图形界面，而接地任务帮助它对齐元素的外观和功能。对于决策矩阵ng，我们首先将跨移动、web和桌面平台的操作统一为具有共享操作空间的函数调用格式。从开源数据中收集的注释多步骤轨迹，由agent框架合成(Wang et al.， 2025;2024b;c)在虚拟环境中被重新格式化为一种函数格式。我们进一步生成一个

通过人类和模型注释器对每一步的推理过程(Xu et al.， 2024).具体来说，给定一个地面真实操作，我们将在屏幕截图中突出显示它。然后，我们向注释者提供全局que ry，以及此操作之前和之后的屏幕截图，并要求他们编写推理内容来解释此操作背后的意图。采用基于模型的筛选筛选低质量的推理内容。这样的推理内容可以了Qwen2.5-VL对地面真实操作的过拟合，并使其在现实场景中更健壮。

阶段	视觉预训练	多模态预训练	长上下文预训练
数据	图像标题 知识 OCR	+ 纯文本 交错数据 VQA、视频 接地、代理	+ 长视频 长期代理 长文件
代币	1.5T	2T	0.6T
序列长度	8192	8192	32768
训练	ViT	ViT & LLM	ViT & LLM

表2：跨不同阶段的训练数据量和组成。

2.2.2培训配方

我们使用DataComp从头开始训练一个视觉变压器（ViT）(Gadre等人， 2023)和一些内部数据集作为视觉编码器的初始化，同时利用预先训练过的Qwen2.5大型语言模型（LLM）(Yang等人，2024a)作为LLM组件的初始化。如表2所示 ，将训练的过程分为三个不同的阶段，每个阶段采用不同的数据配置和训练策略来逐步提高模型的能力。

在第一阶段，只有视觉变换器（ViT）被训练，以改善其与语言模型的对齐，为多模态理解奠定了坚实的基础。该阶段的主要数据源包括图像标题、视觉知识和OCR数据。这些数据集是精心挑选的，以培养ViT提取有意义的视觉表示的能力，可以有效地与文本信息集成。

在第二阶段，将所有模型参数解冻，模型在一组不同的多峰图像数据上进行训练，以增强其处理复杂视觉信息的能力。这个阶段引入了更复杂和推理密集型的数据集，如交错数据、多任务学习数据集、视觉问题回答（VQA）、多模态数学、基于代理的任务、视频理解和纯文本数据集。这些数据集加强了该模型在视觉和语言模式之间建立更深层次联系的能力，使其能够处理越来越复杂的任务。

在第三阶段，为了进一步增强模型在较长序列上的推理能力，我们合并了视频和基于代理的数据，同时还增加了序列长度。这使得该模型能够更精确地处理更高级和更复杂的多模态任务。通过扩展序列长度，该模型获得了处理扩展上下文的能力，这对需要长期依赖和复杂推理的任务特别有利。

为了解决不同的图像大小和文本长度可能导致在训练过程中计算负荷不平衡的挑战，我们采用了一种优化训练效率的策略。主要的计算成本来自于LLM和视觉编码器。考虑到视觉编码器的参数相对较少，并且我们引入了窗口注意以进一步减少其计算需求，我们专注于平衡跨不同gpu的LLM的计算负载。具体来说，我们根据数据样本对应的输入序列长度将其动态打包到LLM中，以确保一致的计算负载。在第一和第二阶段，数据被均匀地打包到8192的序列长度，而在第三阶段，序列长度被增加到32768，以适应模型处理更长的序列的高级能力。

2.3培训后

Qwen2.5-VL的训练后对齐框架采用了一种双阶段优化范式，包括监督精细化（SFT）和直接偏好优化（DPO）(Rafailov等人， 2023).这种层次对齐策略将参数高效的领域适应与人类偏好蒸馏协同起来，通过不同的优化目标来解决表示基础和行为细化问题。

监督微调（SFT）旨在通过有针对性的指令优化来弥补图前d表示和下游任务需求之间的差距。在此阶段，我们使用Cha tML格式(Openai， 2024)来构建指令跟踪数据，在使用whQwen2-VL保持架构一致性的同时，故意偏离训练前的数据模式(Wang et al.， 2024e).这种格式转换可以实现三种关键的调整：1)显式对话角色标记的多模式转换，2)与文本指令相结合的结构化视觉嵌入的结构化注入，以及3)通过格式感知的p打包保存跨模态位置关系。通过将模型暴露给这种增强模式下的管理多模态指令-响应对，SFT在保持高效的网络信息传输的同时实现了预训练特性的完整性。

2.3.1指令数据

监督微调（SFT）阶段采用了一个精心策划的数据集，旨在增强模型跨不同模式的指令跟踪能力。该数据集由大约200万个条目组成，均匀分布在纯文本数据（50%）和多模态数据（50%）之间，其中包括图像-文本和视频-文本组合。多模态数据的加入使模型能够有效地处理复杂的问题。值得注意的是，尽管纯文本和多模态条目的表示是相同的，但由于嵌入的视觉和时间信息，多模态条目在训练过程中消耗了明显更多的标记和计算资源。该数据集主要由中文和英语数据组成，并补充了多语言条目，以支持更广泛的语言多样性。

该数据集的结构反映了不同层次的对话复杂性，包括双单回合和多回合交互。这些交互作用通过从单图像输入到多图像序列的场景进一步情境化，从而模拟真实的会话动态。查询源主要来自开源存储库，另外还有来自精心购买的数据集和在线查询数据。这种组合确保了广泛的覆盖范围，并增强了数据集的代表性。

为了解决广泛的应用程序场景，该数据集包括用于通用视觉问题回答（VQA）、图像字幕、数学问题解决、编码任务和安全相关查询的专门子集。此外，还构建了用于文档和光学字符识别（Doc和OCR）、接地、视频分析和代理交互的专用数据集，以增强特定领域的熟练程度。有关这些数据的详细信息可以在本文的相关部分中找到。这种结构化和多样化的组合确保了SFT阶段有效地将预先训练过的表示与下行多模态任务的微妙需求对齐，促进了健壮和实际感知的模型性能。

2.3.2数据过滤管道

训练数据的质量是影响视觉语言模型性能的关键因素。开源和合成数据集通常表现出显著的可变性，通常包含噪声、重复或低质量的样本。因此，严格的数据清理和过滤过程是解决这些问题的关键。低质量的数据可能导致预训练的代表和下游任务需求之间的次优对齐，从而降低了模型有效地手处理多模态任务的能力。因此，确保高质量的数据对于实现稳健和可靠的模型性能至关重要。

为了解决这些挑战，我们实现了一个两阶段的数据文件环管道，旨在系统地提高监督微调（SFT）数据集的质量。该管道包括以下两个翼台：

第一阶段：领域特异性分类在初始阶段，我们使用Qwen2-VL-Instag，一个来自Qwen2-VL-72B的专门分类模型，对问答（QA）对进行分层分类。该模型将QA对组织为8个主要领域，如编码和规划，它们被进一步划分为30个细粒度的子类别。例如，主域编码被细分为子类别，包括代码调试、代码生成、代码翻译、andCode_Understanding。这种层次结构促进了领域感知和子域感知的过滤策略，使管道能够针对每个类别的特定特征优化数据清理过程。因此，这提高了监督微调（SFT）数据集的质量和相关性。

第二阶段涉及领域定制过滤，该过滤包括基于规则和基于模型的方法，以全面提高数据质量。指定的

由于文档处理、光学字符重新识别（OCR）和视觉接地等领域的多样性，每个领域都可能需要独特的过滤策略。下面，我们将概述在这些领域中应用的一般过滤策略。

基于规则的过滤使用pre定义的启发式方法来消除低质量或有问题的条目。具体来说，对于与文档处理、OCR和视觉基础任务相关的数据集，重复模式将被识别和删除，以防止模型的学习过程失真，并确保最佳性能。另外，还排除了包含不完整、截断或格式不当的响应的条目——这在合成数据集和多模态上下文中很常见。为了保持相关性和维护道德标准，不相关或可能导致有害产出的查询和答案也会被丢弃。这种结构化的方法可以确保了数据集遵循道德准则，并满足特定于任务的需求。

基于模型的过滤通过利用在Qwen2.5- VL系列上训练的奖励模型，进一步细化了数据集。这些模型评估了跨多个维度的多模态QA对。对查询的复杂性和相关性进行评估，只保留那些具有适当挑战性和上下文相关性的示例。对swers的评估是基于正确性、完整性、清晰度、与查询的相关性和帮助性。在基于视觉的任务中，特别注意验证视觉信息的准确解释和利用。这种多维评分确保了只有高质量的数据才能进展到SFT阶段。

基于推理的2.3.3拒绝抽样

为了补充我们的结构化数据过滤方法，我们采用拒绝抽样作为一种策略来细化数据集，并增强视觉语言模型（VLM）的推理能力。这种方法对于需要复杂推理的任务尤其重要，比如进行数学问题解决、代码生成和特定于领域的视觉任务回答（VQA）。先前的研究表明，结合思维链（CoT）Wei等人。(2022)的推理显著提高了amodel的推理性能。(DeepSeek-AI等人， 2024)我们的训练后实验证实了这一点，强调了结构化推理过程对于实现高质量结果的重要性。

拒绝采样过程从富含地面真实注释的数据集开始。这些数据集经过精心设计，包括需要多步骤推理的任务，如数学问题解决、代码生成和特定于领域的VQA。使用Qwen2.5- VL模型的中间版本，我们评估生成的响应。只有模型输出与预期答案匹配的样本被保留，以确保数据集只包含高质量、有效的例子。

为了进一步提高数据质量，我们应用了额外的约束来过滤掉不需要的输出。具体来说，我们排除了显示语码转换、注释长度或重复模式的响应。这些标准确保了CoT推理过程的清晰度和一致性，这对下游应用程序是至关重要的。

将CoT推理应用于视觉语言模型中的一个关键挑战是它们对文本和视觉模式的依赖。中间推理步骤可能无法充分整合视觉信息，要么忽略了相关的视觉线索，要么误解了它们。为了解决这个问题，我们开发了基于规则和模型的iven博士过滤策略来验证中间推理步骤的准确性。这些机制确保了CoT过程中的每一步都有效地集成了视觉和文本模式。尽管有这些努力，实现最佳的模式对准器t仍然是一个持续的挑战，需要进一步的进步。

通过拒绝抽样生成的数据显著提高了模型的推理能力。通过迭代地细化数据集并去除低质量或错误的样本，我们使模型能够从强调精确和连贯推理的高保真例子中学习。这种方法不仅增强了模型处理复杂任务的能力，而且为视觉语言建模的未来改进奠定了基础。

2.3.4培训配方

Qwen2.5-VL的训练后过程包括两个阶段：监督微调（SFT）和直接偏好优化（DPO），这两个阶段都冻结了视觉变压器（ViT）参数。在sft阶段，模型对不同的多模态数据进行微调，包括图像-文本对、视频和纯文本，来自一般的VQA、拒绝采样，以及专门的数据集，如文档和OCR、接地、视频和代理相关的任务。DPO阶段只关注图像-文本和纯文本数据，利用偏好数据将模型与人类的偏好对齐，每个样本只处理一次，以确保有效的优化。这种简化的过程增强了模型的性能

跨模态推理和特定于任务的性能，同时与用户意图保持一致。

3实验

在本节中，我们首先介绍整个模型，并将其与当前最先进的（SoTA）模型进行比较。然后，我们评估了模型跨各种子功能的性能。

3.1与SOTA模型的比较

表3：Qwen2.5-VL的性能和最先进的性能。

数据集	上一个开源SoTA	克劳德-3.5十四行诗-0620	GPT-4o 0513	InternVL 2.5 78B	Qwen2-VL 72B	Qwen2.5- VL 72B	Qwen2.5- VL 7B	Qwen2.5- VL 3B
大学层面的问题								
MMMU val (Yue等人， 2023)	70.1陈氏等人。 （2024d ）	68.3	69.1	70.1	64.5	**70.2**	58.6	53.1
MMMU- Pro整体版(Yue等人， 2024)	48.6陈等人。 （2024d ）	51.5	**51.9**	48.6	46.2	51.1	38.3	31.56
数学								
迷你版(Lu等人， 2024)	72.3 Chen等人。 （2024d ）	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
数学-视觉完整(Wang等人， 2024d)	32.2 Chen等人。 （2024d ）	-	30.4	32.2	25.9	**38.1**	25.1	21.2
数学诗歌迷你(Zhang等人， 2024c)	51.7陈氏等人。 （2024d ）	-	50.2	51.7	-	**57.6**	49.2	47.6
一般的视觉问题回答								
医学(Chen等人， 2024b)	47.4 MiniMax等人。 （2025 ）	52.1	**54.2**	45.6	46.8	51.3	36.8	28.9
MMBench-EN测试(Liu等， 2023d)	88.3陈氏等人。 （2024d ）	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
MMBench- CN测试(Liu等， 2023d)	88.5陈等人。 （2024d ）	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
MMBench-V1.1- EN测试(Liu等， 2023d)	87.4陈等人。 （2024d ）	80.9	83.1	87.4	86.1	**88.4**	82.6	77.4
MMStar (Chen等人， 2024c)	69.5陈等人。 （2024d ）	65.1	64.7	69.5	68.3	**70.8**	63.9	55.9
MME总和(Fu等人， 2023)	**2494陈等人。** （2024d ）	1920	2328	**2494**	2483	2448	2347	2157
老师老师(Wang等人， 2024a)	63.5陈等人。 （2024d ）	-	68.0	63.5	-	**70.7**	59.6	47.7
眨眼时间(Fu等人， 2024c)	63.8陈等人。 （2024d ）	-	**68. 0**	63.8	-	64.4	56.4	47.6
CRPE关系(Wang等人， 2024h)	78.8陈等人。 （2024d ）	-	76.6	78.8	-	**79.2**	76.4	73.6
(Huanal.等， 2023)	**58.1 Wang等人。** （2024f ）	55.5	55.0	57.4	**58.1**	55.2	52.9	46.3
MTVQA (Tang等人，2024)	**31.9陈等人。** （2024d ）	25.7	27.8	31.9	30.9	31.7	29.2	24.8
RealWorldQA平均值(X. AI， 2024)	78.7陈等人。 （2024d ）	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
MME-RealWorlde(Zhang等人， 2024f)	62.9陈等人。 （2024d ）	51.6	45.2	62.9	-	**63.2**	57.4	53.1
MMVet涡轮增压器(Yu等人， 2024)	74.0 Wang等人。 （2024f ）	70.1	69.1	72.3	74.0	**76.2**	67.1	61.8
(Agrawal等人， 2024)	7.4阿格拉瓦尔等人。 （2024 ）	7.5	**7.72**	-	6.59	7.6	6.3	5.7

实验部分评估了Qwen2.5- VL在各种数据集上的性能，并与最先进的模型，如Claude-3.5-十四行诗-0620(人类学， 2024a), GPT-4o-0513 (Ope- nAI ,2024)，InterVL2.5(Chen等， 2024d)，和不同大小的Qwen2- VL (Wang et al.， 2024e).在大学水平的问题中，Qwen2.5- VL-72 B在MMMU上获得了70.2分(Yue等人， 2023).(Yue等人， 2024)，Qwen2.5- VL-72 B得分为51.1分，超过了之前的开源的最先进的模型，并实现了可与GPT-4o相媲美的性能。

在与数学相关的任务中，Qwen2.5- VL-72 B展示了强大的能力。(Lu等人， 2024)，它获得了74.8分，超过了之前的开源的最先进技术水平的72.3分。对于数学视觉(Wang et al.， 2024d)，Qwen2.5- VL-72 B得分为38.1分，而MathVerse (Zhang等人， 202 4c) 达到57.6，与其他领先模型相比都具有竞争结果。

对于一般的视觉问题回答，Qwen2.5- VL-72将跨越多个基准测试。(Liu等人， 2023d)，它达到了88.6分，略高于之前的最佳分88.3分。该模型在MuirBench模型中也表现良好(Wang等人，2024a)，得分为70.7分，然后是眨眼(Fu et al.， 2024c)与64.4年。在MTVQA的多语言能力评估中(Tang等， 2024)，Qwen2.5- VL-72B获得了31.7分，显示了其强大的m种多语言文本识别能力。在主观评价中，如MMVet (Yu等人， 2024)和MM- MT- Bench (Agrawal等人， 2024)，Qwen2.5- VL-72B得分分别为76.2分和7.6分，表现出良好的自然对话体验和用户满意度。

3.2对纯文本任务的性能

来批判性地评估指令调优模型在纯文本任务上的性能，如表4所示 ，我们选择了几个具有代表性的基准测试来评估模型在多个领域的能力，包括一般任务(Wang et al.， 2024j;Gema等人， 2024;怀特等人， 2024)、数学和科学任务(Rein等人， 2023;亨德里克斯等人， 2021;科比等人， 2021)、编码任务(Chen等人， 2021; 卡萨诺等人， 2023)和对准任务(Zhou et al.， 2023).我们将

Qwen2.5- VL与几个相似大小的大型语言模型（llm）进行了比较。结果表明，Qwen2.5- VL不仅在多模态任务上取得了最先进的（SoTA）性能，而且在纯文本任务上表现出领先的性能，展示了其在不同评价标准上的通用性和鲁棒性。

表4：70B+指令模型和Qwen2.5-VL的纯文本任务的性能。

数据集	Llama-3.1-70 B	Llama-3.1-405 B	Qwen2-72 B	Qwen2.5-72 B	Qwen2.5- VL-72B
常规任务					
MMLU-专业	66.4	**73.3**	64.4	71.1	71.2
MMLU-复用	83.0	86.2	81.6	**86.8**	85.9
LiveBench-0831	46.6	53.2	41.5	52.3	**57.0**
数学与科学任务					
格帕	46.7	**51.1**	42.4	49.0	49.0
数学	68.0	73.8	69.0	**83.1**	83.0
GSM 8K	95.1	**96.8**	93.2	95.8	95.3
编码任务					
人类生存期	80.5	**89.0**	86.0	86.6	87.8
MultiPL-E	68.2	73.5	69.2	75.1	**79.5**
对齐任务					
IFEval	83.6	86.0	77.6	84.1	**86.3**

3.3定量结果

3.3.1通用视觉问题回答

为了全面评估该模型在一般视觉问题回答（VQA）和对话中的能力，我们对不同范围的数据集进行了广泛的实验。如表3所示 ，Qwen2.5-VL演示了在各种VQA任务、主观评价、多语言场景和多图像问题中的rt表现状态。具体来说，它擅长于基准数据集，如MMBench系列(Liu等人，2023d)、MMStar (Chen等人， 2024c)、MME (Fu等人， 2023)，MuirBench (Wang等人， 2024a)，眨眼(Fu等人， 2024c)、CRPE (Wang等人， 2024h)，HallBench(关 以及其他人 2023)、MTVQA (Tang等人， 2024)，MME- RealWorld (Zhang等人， 2024f)、MMVet (Yu等人， 2024)和MM- MT- Bench (Agrawal等人， 2024).

在视觉细节理解和推理领域，Qwen2.5- VL-72 B在MMBench- EN-V1.1数据集上达到了88.4%的准确率，超过了以前的最先进的模型，如IntrarVL2.5（78B）和Claude-3.5十四行诗-0620。类似地，在MMStar数据集上，Qwen2.5-VL获得了70.8%的分数，在该基准测试中优于其他领先模型。这些结果强调了该模型在不同语言语境中的稳健性和适应性。

此外，在高分辨率的现实世界场景中，特别是在MME-RealWorld基准测试上，Qwen2.5- VL以63.2分的分数展示了最先进的性能，展示了其对现实环境的广泛适应能力。此外，在在MuirBench数据集上评估的多图像理解任务中，Qwen2.5-VL获得了70.7分的领先分数，进一步突出了其优越的泛化能力。总的来说，这些结果说明了Qwen2.5-VL在解决不同场景下的通用视觉问题回答（VQA）任务方面具有强大的多功能性和有效性。

值得注意的是，即使是Qwen的Qwen2.5- VL版本，特别是Qwen2.5- VL-7 B和Qwen2.5-VL-3B，也表现出极高竞争力的性能。例如，在MMStar数据集上，Qwen2.5-VL-7B achieves63.9%，而Qwen2.5- VL-3 B的得分为55.9%。这表明Qwen2.5-VL的架构不仅功能强大，而且可扩展，即使使用更少的对数表，也能保持强大的性能。

3.3.2文档理解和OCR

我们通过不同的OCR、图表和文档理解基准来评估了我们的模型。表5 演示了betweenQwen2.5-VL模型和顶级模型在以下与ocr相关的基准测试： AI2D (Kembhavi等人， 2016)，textVQA（Singh） 以及其他人 2019)、DocVQA (Mathew等人， 2021b)，InfoVQA (Mathew等人，2021a)，ChartQA (Masry等人， 2022)、CharXiv (Wang等人， 2024k(Li等人， 2024b)，OCRBench (Liu等人， 2023e)、OCRBench_v2 (Fu等人， 2024b)、CC- OCR (Yang等人，2024b(欧阳等人， 2024)、VCR (Zhang et al.， 2024e).

对于与ocr相关的元素解析的解析基准，公式场景、多语言、各种内置（手写、表格、图表、化学公式和数学表达式）文档，

作为CC-OCR和OmniDocBench，Qwen2.5-VL-72B模型凭借精心策划的训练数据和LLM模型的卓越能力而了先进。

对于与ocr相关的场景文本、图表、图表和文档的理解基准，Qwen2.5-VL模型以良好的理解能力取得了令人印象深刻的性能。值得注意的是，在与ocr相关的理解基准上，如OCRBench，专注于信息图形的InfoVQA，以及种子-板凳2+覆盖文本丰富的场景，包括图表、地图和网络，Qwen2.5-VL-72Bachieves显著的结果，显著优于强大的竞争对手，如InternVL2.5-78B。此外，对于与ocr相关的全面基准，如OCRBench_ v2，包括广泛与ocr相关的解析和理解任务，Qwen2.5-VL也实现了最佳性能型号，基本上比最佳车型双子座1.5- Pro分别多了9.6%和20.6%。

表5：Qwen2.5-VL和其他模型在OCR、图表和文档理解基准上的性能。

数据集	克劳德-3.5十四行诗	双子座1.5专业人员	GPT 4o	InternVL 2.5 78B	Qwen2.5- VL 72B	Qwen2.5- VL 7B	Qwen2.5- VL 3B
与OCR相关的解析任务							
CC- OCR	62.5	73.0	66.9	64.7	**79.8**	77.8	74.5
OmniDocBenchedit en/ zh↓	0.330/0.381	0.230/ 0.281	0.265/0.435	0.275/0.324	**0.226/0.324**	0.308/0.398	0.409/0.543
与OCR相关的理解任务							
AI 2 D w。M.	81.2	88.4	84.6	**89.1**	88.7	83.9	81.6
TextVQA val	76.5	78.8	77.4	83.4	83.5	**84.9**	79.3
DocVQA测试	95.2	93.1	91.1	95.1	**96.4**	95.7	93.9
信息VQA测试	74.3	81.0	80.7	84.1	**87.3**	82.6	77.1
ChartQA测试平均。	**90.8**	87.2	86.7	88.3	89.5	87.3	84.0
CharXivRQ/DQ	**60.2/84.3**	43.3/72.0	47.1/84.5	42.4/82.3	49.7/ 87.4	42.5/73.9	31.3/58.6
SEED-Bench-2+	71.7	70.8	72.0	71.3	**73.0**	70.4	67.6
OCR工作台	788	754	736	854	**885**	864	797
VCREn-硬电子显微镜	41.7	28.1	73.2	-	79.8	**80.5**	37.5
OCR相关的综合任务							
OCRBench_v2 en/ zh	45.2/39.6	51.9/43.1	46.5/32.2	49.8/5 2.1	61.5/ 63.7	56.3/57.2	54.3/52.1

3.3.3空间理解

理解臀部的空间关系对于开发可以像人类一样解释和与世界互动的人工智能模型至关重要。在大型视觉语言模型中，视觉接地允许基于自然语言查询或描述对图像中的特定对象、区域或元素进行精确的定位或识别。这种能力超越了传统的对象检测，建立了视觉内容和语言上下文之间的语义关系，促进了更微妙和上下文感知的视觉推理。我们在参考表达式理解基准上评估了Qwen2.5-VL的接地能力。 2014;毛等人， 2016)，在野外的物体检测(Li et al.， 2022b)，自我管理的点接地基准，以及柜台工作台上(Paiss等人， 2023).

我们比较了Qwen2.5-VL的视觉接地能力与其他领先的lvlm，包括双子座、接地-dino(Liu等人， 2023c)，Molmo (Deitke等人， 202 4)和InirVL2.5。

Qwen2.5-VL在从盒接地、点接地到计数的不同基准上实现了领先的性能。通过为Qwen2.5-VL配备盒接地和点接地能力，它能够理解、定位和推理图像的某些部分的细节。对于开放词汇表对象检测，Qwen2.5-VL在ODinW-13上实现了43.1 mAP的良好性能，超过了大多数lvlm，并迅速缩小了多面手模型和专家模型之间的差距。此外，Qwen2.5-VL还解锁了基于点的接地能力，从而可以精确地定位某个物体的每个细节，这在过去很难用边界框来表示。Qwen2.5-VL的计数能力也取得了很大的进展，通过Qwen2.5- VL-72 B使用“检测然后计数”式提示，达到了93.6的领先精度。

3.3.4视频理解和接地

我们在不同的视频理解和基础任务中评估了我们的模型，利用了包括长度从几秒钟到几个小时不等的视频的标准化基准。表8 在以下视频基准测试上，Qwen2.5- VL模型和顶级专有模型之间的性能比较： Video-MME (Fu等人， 2024a)、Video- MMMU (Hu等人， 2025)，MMVU(赵

表6：Qwen2.5- VL等型号的接地性能。

数据集	双子座1.5接地Pro	Molmo内部DINO	VL2.5 72B	Qwen2.5-VL 78B	Qwen2.5-VL 72B	Qwen2.5- VL 7B	Qwen2.5- VL 3B
Refcoco val	73.2	90.6	-	93.7	92.7	90.0	89.1
瑞可可测试器	72.9	93.2	-	95.6	94.6	92.5	91.7
Refcoco testB	74.6	88.2	-	92.5	89.7	85.4	84.0
Refcoco+ val	62.5	88.2	-	90.4	88.9	84.2	82.4
雷可可+测试A	63.9	89.0	-	94.7	92.2	89.1	88.0
Refcoco+ testB	65.0	75.9	-	86.9	83.7	76.9	74.1
Refcocog val	75.2	86.1	-	92.7	89.9	87.2	85.2
Refcolg测试	76.2	87.0	-	92.2	90.3	87.2	85.7
奥丁	36.7	55.0	-	31.7	43.1	37.3	37.5
点接地	-	-	69.2	-	67.5	67.3	58.3

表7：Qwen2.5-VL及其他模型的计数性能。

数据集	双子座1.5-专业版	GPT-4o	克劳德-3.5十四行诗	Molmo-72b	InternVL2.5-78B	Qwen2.5-VL-72B
伯爵的板凳	85.5	87.9	89.7	91.2	72.1	93.6

以及其他人 2025)、MVBench (Li等人， 2024d(Fang等人， 2024(Wu等人， 2024a)，EgoSchema (Mangalam等人， 2023)、感知测试(Patraucean等人， 2024)、MLVU (Zhou等人， 2024)、LVBench (Wang等人，2024g(Liu等人，2024c)和字字鱼(Gao等人，2017).值得注意的是，在LVBench和MLVU上，通过评估长形式的视频理解能力，Qwen2.5-VL-72B取得了显著的结果，显著优于强大的竞争对手，如GPT-4o。

通过利用所提出的同步MRoPE，Qwen2.5-VL增强了其在时间敏感的视频理解方面的能力，具有改进的d时间戳引用、时间接地、密集的字幕和额外的功能。在Charades-STA数据集上，它评估了使用精确的时间戳快速定位事件或活动的能力，Qwen2.5-VL-72B获得了令人印象深刻的mIoU分数50.9，从而超过了GPT-4o的性能。对于所有评估的基准测试，我们将每个视频分析的最大帧数限制在768帧，而视频令牌的总数不超过24,576帧。

表8：Qwen2.5-VL和其他模型在视频基准测试上的性能。

数据集	双子座1.5-专业版	GPT-4o	Qwen2.5- VL-72B	Qwen2.5- VL-7B	Qwen2.5- VL-3B
视频理解任务					
视频-MMEw/o子。	**75.0**	71.9	73.3	65.1	61.5
视频-MME w子。	**81.3**	77.2	79.1	71.6	67.6
视频-MMMU	53.9	**61.2**	60.2	47.4	-
MMVU val	65.4	**67.4**	62.9	50.1	-
MV工作台	60.5	64.6	**70.4**	69.6	67.0
MMBench-视频	1.30	1.63	**2.02**	1.79	1.63
LongVideoBench val	64.0	**66.7**	60.7	56.0	54.2
LV工作台	33.1	30.8	**47.3**	45.3	43.3
EgoSchema测试	71.2	72.2	**76.2**	65.0	64.8
百分比测试	-	-	**73.2**	70.5	66.9
MLVUM-Avg	-	64.6	**74.6**	70.2	68.2
TempCompass Avg	67.1	73.8	**74.8**	71.7	64.4
视频接地任务					
任务-STAmIoU	-	35.7	**50.9**	43.6	38.8

3.3.5代理

多模态模型中的代理能力对于使这些模型能够有效地与真实世界的设备进行交互至关重要。我们通过各个方面评估了Qwen2.5- VL的代理能力。用户界面

元素接地由ScreenSpot (Cheng等， 2024)和屏幕(Li等人， 2025a).离线评估是在安卓控制系统上进行的(Li et al.， 2 024f)，而在线评估是在包括机器人世界在内的平台上进行的(Rawles等人，2024)，移动迷你++++(Rawles等人， 2024)和OSWorld (Xie等人， 2025).我们比较了Qwen2.5- VL-72 B与其他著名模型，如GPT-4o (OpenAI， 2024双子座2.0，20 24），克劳德公司（人类学公司， 2024b)，Aguvis-72B (Xu等人，2024)和Qwen2-VL-72B (Wang et al.，2024e).结果如表9所示。

表9：Qwen2.5-VL和其他模型在GUI代理基准测试上的性能。

基准测试	GPT-4o	双子座2.0	克劳德	Aguvis-72 B	Qwen2- VL-72 B	Qwen2.5- VL-72B
屏幕位置	18.1	84.0	83.0	**89.2- 87.1**		
屏幕点Pro	-	-	17.1	23.6 1.6	43.6	
安卓控制高EM	20.8	28.5	12.5	66.4	59.1	67.36
安卓控制低EM	19.4	60.2	19.4	84.4	59.2	93.7
AndroidWorld SR	34.5% (SoM)	26% (SoM)	27.9%	26.1%6% (SoM)	35%	
MobileMiniWob++ SR	61%	42% (SoM)	61% (SoM)	66% 50% (SoM)	68%	
OSWorld	5.03	4.70	**14.90**	10.26	2.42	8.83

Qwen2.5-VL-72B模型的性能在整个GUI接地基准测试上取得了非凡的进步。它在屏幕点上达到了87.1%的准确率，与双子座2.0（84.0%）和克劳德（83.0%）激烈竞争，同时值得注意的是以43.6%的准确率在屏幕spotpro上设置了一个新标准——远远超过agu维斯-72B（23.6%）和基础Qwen2- VL-72B（1.6%）。利用这些优越的接地能力，Qwen2.5-VL-72B si在所有离线评估基准上都显著优于基线，但差距很大。在在线评估中，一些基线由于接地能力有限，难以完成任务。因此，我们将标记集（SoM）应用于这些模型的输入。结果表明，Qwen2.5-VL-72B在机器人世界和移动系统++上的基线，并且在OSWorld上取得了相当的性能。这一观察结果表明，Qwen2.5-VL-72B能够在真实和动态的环境中发挥作用。

4结论

我们提出了Qwen2.5- VL，一个最先进的视觉-语言模型系列，在多模态理解和交互方面取得了显著的重大进步。Qwen2.5-V L在视觉识别、对象定位、文档解析和长视频理解方面的能力不断增强，它在静态和动态任务方面都表现出色。它的原生动态分辨率处理和绝对时间编码使鲁棒处理不同的输入，而窗口注意减少了计算开销，而不牺牲分辨率保真度。Qwen2.5- VL满足了广泛的应用范围，从边缘人工智能到高性能计算。旗舰产品Qwen2.5- VL-72 B匹配或超过GPT-4o和Claude 3.5十四行诗等领先机型，特别是在文档和图表理解方面，同时在纯文本任务上保持强大的性能。较小的Qwen2.5- VL-7 B和Qwen2.5- VL-3 B变体的性能优于类似规模的竞争对手，提供了效率和多功能性。Qwen2.5-VL为视觉语言模型设置了一个新的基准测试，展示了跨领域的特殊泛化和任务执行。它的创新为更智能和交互式的系统铺平了道路动态感知和现实世界的应用。

5位作者

核心发起人：白帅、陈克勤、刘学静、王家林、葛文斌、宋思博、党凯、王鹏、王世杰、唐军、润记、朱元志、杨明坤、李赵海、万建强、王鹏飞、王伟、富义和、徐家宝、张祥、谢成、成、张贤、杨博、徐海阳、林俊洋

贡献者1： 杨、宾元慧、余宝文、陈成、刘大恒、范红、黄飞、刘佳伟、金旭、曾建元、张杰、张金凯、张建伟、张建仁、周可心、杨可心、李明、明艳、倪、瑞人、宋江、邓晓东、黄晓明、周希明、任兴章、杨范、宜昌张、朱一凯、刘玉琼、郭志芳

1字母顺序。

参考文献

阿格拉瓦尔、西蒙·安东尼亚克、艾玛·博汉纳、巴普蒂斯特·布特、德特德拉·查普洛特、杰西卡·丘德诺夫斯基、迪奥戈·科斯塔、鲍杜因·德莫尼科、索拉布·格格、格尔维特等。像素12b.arXiv预印本arXiv：2410.07073,2024年。

让-巴蒂斯特·阿莱拉克、杰夫·多纳休、波林·卢克、安托万·米埃赫、艾恩巴尔、亚纳·哈森、卡雷尔·伦克、亚瑟·门施、凯瑟琳·米利肯、马尔科姆·雷诺兹等。火烈鸟：一种少镜头学习的视觉语言模型。在NeurIPS，2022年。

人类的。克劳德，3.5十四行诗，202 4年a。URLhttps://www.anthropic.com/news/claude-3-5-sonnet .

人类的。介绍计算机使用，一个新的克劳德3.5十四行诗，和克劳德3.5俳句，2024b。URLhttps://www.anthropic.com/news/3-5-models-and-computer-use .

费德里科卡萨诺，约翰古瓦尔，丹尼尔，阮悉尼阮，卢娜菲普斯-科斯汀，唐纳德平克尼，明何叶，杨田子，卡罗琳简安德森，莫莉Q。费尔德曼、阿尔琼·古哈、迈克尔·格林伯格和阿比纳夫·扬达。MultiPL-E：一种可扩展的和具有多种语言的方法来基准测试神经代码
一代人IEEE跨。软件工程。，49(7)：3675-3691年，2023年。

陈桂明、陈秀敏、张瑞飞、陈俊英、吴相波、张志毅、陈志宏、李建全、项万、王本。利用gpt4v合成的数据
精简视觉语言模型。arXiv预印本arXiv：2402.11684,202 4a。

陈嘉成、梁天浩、萧谢尔曼、王正清、王凯、王玉波、倪元生、王朱、江紫彦、吕博翰等。大型板凳：将多模式评估扩展到超过500个
在现实世界中的任务。arXiv预印本arXiv：2410.10563,2024b。

陈林、李锦江、董晓义、张全、藏宇航、陈泽辉、段浩东、王佳琪、于乔、林大华等。我们是在正确的方式评估慷慨的语言models？arXiv：2403.20330,2024c。

陈、杰瑞特华克、海宇俊、元铭、平托、爱德华、爱德华、华、布尔达、布罗克曼、布罗克曼、普雷、普尔、克鲁格、凯撒、巴伐利亚、克莱斯、温特、彼得罗斯基、卡明斯、尚齐斯、巴恩斯、巴恩斯、古斯、高克斯、尼克尔、佩诺、特扎克、唐、巴布斯金、巴拉吉、贾因，威廉·桑德斯、克里斯托弗·黑塞、安德鲁·卡尔、扬·莱克、约书亚·亚齐姆、编辑米斯拉、埃文·莫里川、雷德福、马修·奈特、迈尔斯·布伦戴奇、米拉·穆拉特、凯蒂·梅耶、彼得·韦林德、鲍勃·麦克格鲁、达里奥·阿莫迪、山姆·麦坎德利什、伊利亚·苏斯克弗和沃克·扎雷姆巴。评估经过代码训练的大型语言模型。CoRR，abs/2107.03374,2021年。

陈哲、吴剑南、王文海、苏伟杰、郭陈、森兴、钟勇、张青龙、朱喜洲、吕乐伟、李斌、罗平、童路、于乔、戴继峰。内部设计：扩展视觉基础模型和对齐通用的视觉-语言tasks.arXiv预印本arXiv：2312.14238,2023年。

陈哲、王伟云、曹越、刘扬州、高张伟、崔二飞、朱金国、叶胜龙、郝浩、刘昭阳等。通过模型、数据和测试时间的缩放来扩展开源多模态模型的性能边界。arXiv预印本arXiv：2412.05271,2024年d。

程健智、孙秋实、朱友刚、徐芳智、李燕涛、张建兵、吴志勇。利用高级gui接地的视觉gui代理。arXiv预印本arXiv：2401.10935,2024年。

卡尔·科布，科萨拉朱，穆罕默德·巴伐利亚，陈、宇宇军、凯泽、马提亚斯·普拉普特、杰瑞·托沃克、雅克布·希尔顿、中野良一郎、克里斯托弗·赫斯泽和约翰·舒尔曼。训练验证者来解决数学单词问题。CoRR，abs/2110.14168,2021年。

扬N.多芬，安吉拉范，迈克尔奥利，和大卫格兰奇尔。使用通用的卷积网络进行语言建模。ICML，机器学习研究论文集第70卷，页。933–941.PMLR，2017.

谷歌的深层思想。介绍双子座2.0：我们在2024年的农业时代推出的新ai模型。URLhttps://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/ .

深爱、刘爱新、北峰、冰峰、冰峰、王炳轩、吴宝超、成达、卢成达、赵成刚、邓成奇、陈宇、张阮、戴迈、郭大亚、杨德健、陈德力、李东杰、李二亨、林芳云、富聪、戴福、郝广博、陈官厅、李国伟、张伟、汉宝、徐汉伟、王浩诚、张浩伟、丁鸿伟、丁华辉、高华硕、李慧、屈。蔡、梁、郭忠、倪家琪、李嘉师、王佳伟、金晨、陈景昌、景阳、邱俊杰、李俊龙、容晓松、开东、胡凯、高凯歌、康管、黄可欣、快玉、林、张乐从、雷磊、乐霞、赵梁亮、王立通、张立月、李、王妙君、张英川、张明华、唐明华、李明明、宁田、潘潘、王培义、王彭毅、王千成、朱齐浩、秦玉、陈秋实、陈秋实、陈玲。金、葛瑞奇、张瑞松、潘瑞哲、王龙吉、徐龙信、张若愚、陈如意、李、陆上浩、周上岩、陈山皇、吴少卿、叶胜丰、叶胜丰、世荣、马世余、王双周、水平、周丰、周、潘亭、王 陶云、天培、孙天宇、肖、曾王丁。Deepseal-v3技术报告。CoRR，abs/2412.19437,2024。doi：10.48550/ ARXIV .2412.19437 。URL https:// doi .org/10.48550/ arXiv .2412.19437 .

马特代克、克里斯托弗·克拉克、李桑浩、罗亨特里帕蒂、杨悦、朴在成、萨利希、尼尼霍夫、凯尔、卢卡·索尔达尼等。Molmo和pixmo：打开重量和打开
为最先进的多模态模型提供的数据。arXiv预印本arXiv：2409.17146,2024年。

方、毛康瑞、段浩东、赵象宇、李伊宁、林大华、陈凯。mmbench视频：一个为整体视频理解的长形式的多镜头基准。arXiv预印本
arXiv:2406.14515, 2024.

傅朝佑、陈弦、沈云亨、秦仁、张孟丹、徐林、邱振宇、魏林、杨金瑞、郑下武等。夫人：一个针对多模态大型企业的综合评估基准
语言models.arXiv：2306.13394,2023年。

傅超宇、戴玉涵、罗永东、李磊、任淑怀、张仁瑞、王和、周陈宇、沈云亨、张孟丹等。视频短信：第一次进行综合评估
视频分析中多模态llms的基准测试。arXiv:2405.21075, 202 4a.

凌福、杨彪、陈斌、宋家俊、李玉哲、朱、林浩、罗、新余万、陆、黄明信、张力、唐、山斌、林春、刘齐、吴炳洪、吴浩峰、刘浩、黄健、京群、陈魏、金连文、刘玉良、项白。Ocrbencv2：一个改进的基准，以评估大型多模态模型的视觉文本定位和
推理，2024b。URL https:// arxiv .org/ abs/2501.00321 .

傅兴宇、胡于时、李邦正、于峰、王浩宇、林旭东、丹罗斯、诺亚阿史密斯、马伟楚、兰杰奎师那。眨眼：多模态的大型语言模型可以看到，但不能感知。在欧洲计算机视觉会议上，第3页。148–166.2024年c年的春天。

萨米尔·伊扎克·加德雷、加布里埃尔·伊尔哈科、亚历克斯·方、乔治斯·尼斯、阮、瑞安·马滕、米切尔·沃斯曼、德鲁巴、高希、杰玉鸿等。数据组：正在寻找下一个数据组
生成多模态数据集。arXiv:2304.14108, 2023.

高济阳、陈孙、杨振环、内华达兰等。高：通过语言查询进行时间活动定位。《IEEE计算机视觉国际会议论文集》，第3页。5267–5275, 2017.

格马、俊梁、洪、王、曼西诺、何宣礼、余昭、杜晓堂、马尼等。我们用mmlu了吗？CoRR，abs/2406.04127,2024年。

陈琪、崔银、陈倩、陈毅、林志健、李、佐柏。简单的复制粘贴是一种强大的实例分割数据增强方法。IEEE/CVF计算机视觉与模式识别会议论文集，第3页。2918–2928, 2021.

关天瑞、刘富晓、吴西洋阳、西安瑞奇、李宗霞、刘小雨、王锡君、陈厂、黄芙蓉、雅卓、马诺查、周天一。幻觉台：一种用于纠缠语言幻觉和视觉幻觉的大型视觉语言models.arXiv：2310.14566,2023。

郭、郑尼、白岳林、李波、王玉波、朱王、李一智、纽博、陈文虎、项悦。通过大规模的指令调优来激发多模态推理。
arXiv预印本arXiv：2412.05237,2024年。

丹亨德里克斯，科林伯恩斯，索拉夫卡达瓦斯，阿库尔阿罗拉，史蒂文巴萨特，埃里克唐，D芒宋，和雅各布斯坦哈特。用数学数据集测量如何解决数学问题。在NeurIPS数据集和基准测试中，2021年。

胡凯瑞、吴鹏浩、蒲一、王晓、张元汉、项悦、李波、刘紫薇。评估从多学科的专业视频中获取的知识。arXiv预印本
arXiv:2501.13826, 2025.

萨哈尔·卡泽姆扎德，维森特·奥多涅斯，马克·马滕和塔玛拉·伯格。参考游戏：指自然场景照片中的物体。在EMNLP，2014年。

阿尼鲁达肯巴维，迈克萨尔瓦托，埃里克科尔夫，明琼秀，汉南嗯哈吉希尔齐，和阿里法哈迪。一个图表值十几张图片。在ECCV，2016年。

亚历山大·基里洛夫、埃里克·明顿、尼希拉·拉维、毛、克洛伊·罗兰、劳拉·古斯塔夫森、泰特肖、斯宾塞·怀特黑德、亚历山大·伯格、万仁罗等。段任何东西。在ICCV，2023年。

李炳坤、朴博元、蔡胜金、曼罗勇。Moai：大型语言和视觉的所有智能的混合。在欧洲计算机视觉会议上，第3页。273–302.施普林格，2024年。

李波、张培元、杨景康、张元汉、法尼浦、刘紫薇。一种高分辨率的多模态模型。arXiv:2311.04219, 2023a.

李波、张元汉、董国、张仁瑞、李峰、张浩、张臣、张培元、李彦伟、刘紫薇等。简单的视觉任务转移。arXiv预印本arXiv：2408.03326,2024a。

李博浩、葛育英、陈毅、葛一晓、张雷茂、英山。种子板凳2+：具有文本丰富的视觉理解的多模态大型语言模型。arXiv preprintarXiv:2404.16790, 2024b.

李东旭、刘玉东、吴浩宁、王岳、沈智、曲文、牛新耀、王国因、陈备、李俊南。Aria：一个开放的多模态本地专家混合模型。arXiv预印本
arXiv:2410.05993, 2024c.

李俊南、李东旭、熊凯明、何志伟。引导语言-图像预训练，以实现统一的视觉-语言理解和生成。在ICML，2022a。

李俊南、李东旭、萨瓦泽、何。blip2：使用冻结图像编码器和大型语言模型进行引导语言图像预训练。arXiv:2301.12597, 2023b.

李开心、资阳孟、林红展、罗资阳、田宇晨、马静、黄志勇、蔡达成。专业：专业高分辨率计算机专用接地，2025a。紫外线
https://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf .预印本。

李昆昌、王雅丽、何一南、李依火、王毅、刘毅、王尊、徐吉兰、郭陈、罗平等。一个全面的多模态视频理解基准测试。在CVPR，2024d。

李连日安·哈罗德、张鹏川、张浩天、杨建伟、李春元、钟义乌、王丽娟、陆元、张雷、黄仁能等。地面语言-图像预训练。在IEEE/CVF计算机Visio n和模式识别会议论文集上，页。10965–10975, 2022b.

李云、陈哲、王伟云、王文海、叶胜龙、金镇江、陈兰州、何一南、高张伟、崔崔等。一个包含100亿个水平的统一的多模态语料库
与文本交织的图像。arXiv预印本arXiv：2406.08418,2024 e。

李伟、威廉毕晓普、李爱丽丝、克里斯罗尔斯、坎贝尔-阿贾拉、泰玛贡鲁和奥里安娜里瓦。论数据尺度对计算机控制代理的影响。arXiv预印本arXiv：2406.03679,2024年f。

李亚东、孙浩泽、林明安、李天鹏、董国胜、张涛、丁博文、宋伟子、郑正林、霍雨琪等。百川-全方位的技术报告。arXiv预印本arXiv：2410.08565,3 (7)，2024g。

李亚东、刘俊军、张涛、宋晨、李天鹏、李泽环、刘立军、明灵峰、董国胜、大潘等。百川-1.5技术报告。arXiv预印本arXiv：2501.15368,2025b。

李云新、姜深、胡宝田、王龙岳、钟万奇、罗文涵、马林等、张敏。统一模式：用混合的专家来扩展统一的多模态模式。arXiv预印本
arXiv:2405.11273, 2024 h.

张李、杨彪、刘强、“马马、张硕、杨静旭、孙亚宝、刘玉良、香白。猴子：图像分辨率和文本标签是大型模式模型的重要因素。
arXiv:2311.06607, 2023c.

梁宇轩、徐丽、陈晓雷、陈浩天、郑毅、赖成红、李斌、薛襄阳。高分辨率大视觉语言模型中的全局语义引导子图像特征权值分配。arXiv预印本arXiv：2501.14276,2025年。

纪林、尹虹、魏平、莫尔查诺夫、肖碧、宋汉。关于视觉语言模型的预训练。在IEEE/CVF计算机视觉和模式识别会议论文集上，第页。26689–26699, 2024.

刘浩天、李春元、李裕衡、李永杰。改进的基线与视觉指令tuning.arXiv：2310.03744,2023a。

刘浩天、李春元、吴庆阳、李永在。视觉指令调优。arXiv:2304.0848 5, 2023b.

刘石龙、曾朝阳、天河人、冯丽、张浩、杨杰、李岳、杨建伟、杭州、朱俊娟、张磊。接地恐龙：在训练前结合恐龙
对于开放集对象detection.arXiv：2303.05499,2023c。

刘扬州、曹越、高张伟、王伟云、陈哲、王文海、天浩海、陆乐伟、朱西洲、童陆等。一个高质量的多节点指令调优数据集
广泛的多样性。《中国信息科学》，67（12）：1-16,2024a。

刘元、段浩东、张博李元汉、张松雅、赵王博、元一可、王佳琪、何同盟会、刘紫薇、陈凯、林大华。Mmbench：你的多模式模式是一个全能的球员吗？arXiv:2307.06281, 2023d.

刘元、赵中银、庄资源、乐田、小周、周杰。要点：用可负担得起的策略来改善你的视觉语言模式。arXiv预印本arXiv：2409.04828,2024b。

刘元信、李世诚、刘易、王玉祥、任淑怀、李磊、陈西神、徐孙、陆侯。视频llms真的理解视频吗？arXiv预印本arXiv：2403.00476,202 4c。

刘玉良、张力、黄明信、杨彪、余文文、李春元、尹徐成、刘成林、金连文、向白。奥克本奇：关于大型多模态模型中ocr隐藏的秘密。
arXiv:2305.07895, 2023e.

陆、班萨、夏托尼、刘佳成、李春元、哈尔滨、郝诚、张开伟、高丽、高剑。在视觉上下文下评估基础模型的数学推理。在ICLR，2024年。

曼加拉姆，阿克舒拉科夫和吉滕德拉·马利克。自我模式：用于长时间视频语言理解的诊断基准。在NeurIPS，2023年。

毛俊华、华、亚历山大托舍夫、坎布鲁、尤尔和墨菲。生成和理解明确的对象描述。在CVPR，2016年。

艾哈迈德·马斯里、多玄龙、贾庆谭、乔蒂和名男。Chartqa：用视觉和逻辑推理回答关于图表的问题的基准。arXiv:2203.10244, 2022.

马修，巴加尔，铁托，卡拉扎斯，欧内斯特·瓦尔维尼，和C .V。贾瓦哈尔。Infographicvqa .2022年IEEE/CVF冬季计算机视觉应用会议（WACV），第2582-2591页，2021a页。

马修，卡拉扎斯和贾瓦哈尔。Docvqa：一个用于文档图像上的vqa的数据集。在WACV，2021b。

迷你、宏、李、龚、邦伟、杨、博子、陈子、陈子、程珠、张春浩、郭刚超、大陈、董力、娇心、耿心、张军、孙浩海、后东、朱嘉井、庄家佳、宋祖、金珠、汉、李景阳、谢俊斌、徐俊浩、燕俊杰、张凯顺、张柯诚、柯西康、乐河、乐阳、王、余莲飞、冯立恒、林正、林宝柴、龙兴、梅居、米源、张茂智、黄培凯、彭城

牛、李鹏飞、曹鹏宇、杨鹏宇、徐贵地、王哲香、王秦、李秋辉、李瑞涛冷、石盛敏、余书琪、李世臣、松泉、黄涛、梁天润、孙伟高、伟轩、孙伟宇、李文凯、宋晓秀、韩晓东、张新捷、新竹、徐民、迅津、沈阳、旭阳、严刚、朱鹏、周一鹏、钟永然、胡永义、范元宇、杨裕宇、李宇浩、云南、李云基、黄云鹏、徐宇新、李泽汉、李子康、陶泽文英、朝阳阳、振秦、范振华、余志华、卓智昂、吴子佳。最小极大值-01：使用闪电来缩放基础模型
注意，2025年。URL https:// arxiv .org/ abs/2501.08313 .

奥普奈。Chatml文档，2024年。URLhttps://github.com/openai/openai-python/blob/main/chatml .md.

OpenAI。你好，gpt-4o，2024年。URLhttps://openai.com/index/hello-gpt-4o .

欧阳林克、袁屈、周宏斌、朱家伟、张瑞、林群书、王斌、赵志远、曼江、赵小蒙、金石、范武、裴楚、刘明浩、李振祥、赵超、张博、施博田、图中英、何同盟会。全功能平台：使用全面的注释对各种pdf文档解析进行基准测试，2024年。URLhttps://arxiv.org/abs/2412.07626 .

罗尼·派斯、阿里尔·埃弗拉特、奥默尔·托夫、扎达、莫塞里、伊拉尼和塔利·德克尔。教学剪辑，以计数到十。《IEEE/ CVF计算机视觉国际会议集》，第3170-3180页，2023年。

维奥丽卡·帕特劳西安、卢卡斯·斯迈拉、安什古普塔、阿德里亚·里卡森斯、拉里莎·马基瓦、迪伦·巴纳斯、斯坎达·科普拉、马图斯·马林诺夫斯基、伊杨、卡尔·杜尔施等。感知测试：一个多模态视频模型的诊断基准。在NeurIPS，2024年。

彭志亮、王文辉、李东、雅浩、黄少汉、马漱明、福伟。科斯莫斯-2: 为世界建立多模态大型语言模型。arXiv:2306.14824, 2023.

拉斐尔·拉法洛夫，沙尔玛，埃里克·米切尔，克里斯托弗·D。曼宁，埃尔蒙和切尔西·芬恩。直接偏好优化：你的语言模型是一个秘密的奖励模型。在《爱丽丝哦》、特里斯坦·诺曼、阿米尔·格洛伯森、凯特·森科、莫里茨·哈特和谢尔盖·莱文（eds.）中，神经信息处理系统的进展36：神经信息处理系统年会2023，NeurIPS 2023，新奥尔良，美国，2023,2023年12月10 - 16日。URLhttp://papers.nips.cc/论文/论文/2023/散希，摘要-会议。html .

克里斯托弗·罗尔斯、萨拉·克林克迈利、张一凡、乔纳森·华尔兹、刘布里埃尔、玛丽贝丝·费尔、李爱丽丝、威廉·毕晓普、魏李、福拉维约·坎贝尔-阿贾拉等。仙女座世界：一个动态的长凳
针对自主代理的标记环境。arXiv:2405.14573, 2024.

雷因、李侯、陈兰、佩蒂、庞、迪拉尼、迈克尔、鲍曼。GPQA：研究生水平的谷歌证明的问答基准。CoRR，abs/2311.12022,202 3年。

天河人、江清、刘石龙、曾昭阳、刘文、韩高宏、黄宏杰、马正宇、姜小渴、陈一号等。接地dino1.5：向前推进开放集对象的“边缘”
察觉arXiv预印本arXiv：2405.10300,2024年。

卡洛斯·里克尔姆、琼·普格塞弗、巴兹尔·穆斯塔法、马克西姆·诺伊曼、鲁杜夫·杰纳顿、安德烈苏-萨诺·平托、丹尼尔·凯瑟斯和尼尔·霍尔斯比。用稀疏的专家混合物来扩展视觉。神经信息处理系统的进展，34：8583-8595,2021。

辛格、纳塔拉扬，遇见沙阿、于江、陈欣蕾、巴德拉、帕里克和罗尔巴赫。面向可以阅读的vqa模型。在CVPR，2019年。

苏健林、艾哈迈德、吕玉隆、潘胜丰、文博、刘云峰。旋转变压器：旋转式位置嵌入式增强型变压器。Neurocomputing, 568:127063, 2024.

唐景群、刘齐、叶永杰、吕景辉、舒伟、林春、李万清、马哈茂德、郝冯、赵曾、王彦杰、刘玉良、刘浩、项白、黄灿。Mtvqa：对多语言以文本为中心的可视化问题回答进行基准测试。arXiv:2405.11985, 2024.

双子座团队、安尼、博格、吴永辉、王力、余家慧、郭、戴、豪等。双子座：一个非常有能力的家庭多模态模型。arXiv预印本arXiv：2312.11805,2023年。

童、小龙、吴鹏浩、吴相贤、木兰、西花、杨翰、杨树生、杨、潘西臣等。寒武纪-1：一个完全开放的，以视觉为中心的地方
多模态llms的探索。arXiv预印本arXiv：2406.16860,20 24年。

王飞、傅兴宇、黄玉玉、李泽坤、刘秦、刘小公、马明宇、徐南、周文轩、张凯等。Muirbench：一个鲁棒的多图像的综合基准
了解arXiv预印本arXiv：2406.09411,2024a。

王俊阳、徐海阳、贾哈涛、张喜、严明、沈伟洲、张吉、黄飞、桑日涛。移动代理-v2：具有有效导航导航代理的移动设备操作助手
协作arXiv预印本arXiv：2406.01014,2024b。

王俊阳、徐海洋、叶家宝、严明国、沈伟洲、张智、黄飞、桑季涛等。移动代理：具有视觉感知功能的自主多模态移动设备代理。arXiv预印本
arXiv:2401.16158, 2024c.

王柯、潘俊亭、石伟康、路齐木、詹明杰、李宏生。用数学视觉数据集测量多模态数学推理。
arXiv:2402.14804, 202 4d.

王鹏、白帅、谭思南、王石杰、范志浩、白晋泽、陈克勤、刘学静、王家林、葛文斌、杨范、凯党、杜孟飞、宣城任、瑞人、刘大恒、长周、周景仁、林俊阳。增强视觉语言模型的感知
世界在任何决议下。arXiv:2409.12191, 202 4e.

王鹏、白帅、谭思南、王石杰、范志浩、白晋泽、陈克勤、刘学晶、王嘉林、葛文斌等。增强视觉语言模型对世界的感知
决心arXiv预印本arXiv：2409.12191,2024年f。

王伟涵、何泽海、洪文义、程仁、张小涵、吉启、顾晓涛、黄士余、徐斌、优晓东等。一个非常长的视频理解基准。arXiv预印本
arXiv:2406.08035, 2024g.

王伟云、任一鸣、罗浩文、李天童、陈艳、陈哲、王文海、李云、陆乐伟、朱喜洲等。全视项目v2：向一般关系的理解
开放的世界。arXiv预印本arXiv：2402.19474,2024小时。

王文海、戴继峰、陈哲、黄振环、李志奇、朱喜洲、胡小伟、陆童伟、陆乐伟、李宏生等。内部图像：探索具有可变形卷积的大规模视觉基础模型。在IEEE/CVF计算机信息识别与模式识别会议论文集上，第页。14408–14419, 2023.

王新龙、张晓松、罗正雄、孙权、崔玉峰、吴金生、张范、王月泽、李真、余齐英等。Emu3：下一个令牌预测是你所需要的。arXiv prepr int
arXiv:2409.18869, 2024i.

王玉波、马学光、张葛、倪元生、阿伯拉尼尔·钱德拉、郭世光、任、阿兰、玄和、江紫彦、李天乐、库、王凯、庄、范、项悦、陈文虎。MMLU-Pro：一个更健壮和更具挑战性的多任务语言理解基准测试。CoRR，abs/2406.01574,2024j。

王振隆、徐海洋、王军阳、张西、明燕、张吉、黄飞、恒智。移动代理-e：完成复杂任务的自我进化的移动助手。arXiv预印本arXiv：2501.11733,2025年。

王智瑞、夏孟州、何鲁西、陈霍华德、刘一涛、朱开曲、梁、吴新地、刘浩天、马拉迪、阿列克谢的骑士、桑吉夫·阿罗拉、陈丹奇。在多模态llms中绘制现实图表理解中的差距。arXiv预印本arXiv：2406.18521,2024k。

魏杰生、王学志、舒尔曼、博斯马、志志、乐、周丹尼。思维链促使人们在大型语言模型中引发推理。CoRR，abs/2201.11903,2022年。URLhttps://arxiv.org/abs/2201 .11903 .

科林·怀特、塞缪尔杜利、曼利罗伯茨、阿尔卡帕尔、本杰明福尔、西达尔塔贾因、拉维德什瓦兹齐夫、尼尔贾因、哈立德赛富拉、西达尔塔奈杜、钦梅赫格德、扬勒昆、汤姆戈尔茨坦、威利尼斯万格和弥卡戈尔德布卢姆。直播台：一个具有挑战性的，无污染的LLM基准。CoRR，abs/2406.19314,2024年。

吴鸿、李东旭、陈北、李俊南。长视频注释：长上下文交织视频语言理解的基准，2024a。
URLhttps://arxiv.org/abs/2407.15754 .

吴志宇、陈小康、潘子正、刘行超、刘文家、戴大泰、高华藏、马益阳、吴承月、王炳轩等。深度vl2：专家的混合视觉语言模型
先进的多模态的理解。arXiv预印本arXiv：2412.10302,2024b。X.AI.Grok-1.5视

觉版本。https://x.ai/blog/grok-1.5v ，2024.

萧斌、吴海平、徐伟建、戴西阳、胡东、陆玉茂、曾迈、刘策、陆源。佛罗伦萨-2：推进各种视觉任务的统一表示（2023年）。URL
https://arxiv.org/abs/2311.06 242, 2023.

谢天、张丹阳、陈季选、李小川、赵思恒、曹瑞生、华华、程周军、陈东强、方宇磊等。Osworld：基准测试真实计算机环境中的开放式任务的多模态代理。神经信息处理系统的进展，37：52040-52094,2025。

徐义亨、王泽坤、王俊礼、吕敦杰、谢图腾、阿姆利塔萨哈、渡萨虎、于涛、熊廷。Aguvis：用于自主gui交互的统一纯视觉代理。一个rXiv预印本
arXiv:2412.04454, 2024.

杨、杨宝钢、张元、许斌元、波正、余博文、李成元、刘大恒、黄飞等。Qwen 2.5技术报告。
arXiv:2412.15115, 2024a.

杨志波、唐俊君、李兆海、王鹏飞、万建强、钟门、刘学静、杨明坤、王鹏、白帅、金连文、林俊阳。Cc-ocr：一个评估大型多元数字扫盲模型的全面和具有挑战性的ocr基准，2024b。URLhttps://arxiv.org/abs/2412.02210 .

叶韩荣、黄德安、姚陆、余志定、魏萍、涛、简考茨、宋汉、徐丹、帕夫罗·莫尔查诺夫等。用于大型语言模型的跨模态对齐。arXiv预印本
arXiv:2405.19335, 2024.

叶庆浩、徐海洋、叶亚博、燕明、刘浩伟、齐谦、张智、黄飞、京仁Zhou.mplug-owl2：革新多模式大型语言模式。
arXiv:2311.04257, 2023.

余伟浩、杨正元、李林杰、王健、林凯文、刘自成、王新超、王丽娟。兽医：评估大型多模态模型的集成能力。在ICML，2024年。

项悦、倪元生、张凯、郑天宇、刘若琪、张葛、史蒂文斯、姜东福、任伟明、孙宇轩等。一个庞大的多学科的多模式理解
以及专家敏捷的推理基准。arXiv:2311.16502, 2023.

项悦、郑天宇、倪元生、王玉波、张家一、张圣邦童、孙宇轩、尹明涛、余博涛、张葛等。Mmmu-pro：一个更健壮的多学科的多模态理解
基准arXiv预印本arXiv：2409.02813,2024年。

张彪和里科·森恩里希。均方根层归一化。在NeurIPS，2019年。

张浩天、你郝轩、杜夫德、张宝文、陈晨、陈洪友、徐久富、杨伟王、张世富、干哲、杨银飞。雪貂-v2：一个改进的基线
用大型语言模型进行参考和接地。arXiv:2404.07973 , 2024 a.

张潘、董孝义、曹宇航、藏宇航、瑞倩、西林魏、陈林、李亦飞、牛俊波、丁双瑞等。Internlm-xcomposer2.5项：一个全面的多模式系统的长期流媒体视频和音频交互。arXiv预印本arXiv：2412.09596,2024 b。

张仁瑞、江东治、张一一、林浩坤、郭子玉、邱鹏朔、周奥军、陆、张开伟、于乔等。你的多模态llm真的看到了视觉数学问题中的图表吗？在欧洲计算机视觉会议上，第3页。169–186.弹簧r，2024年c。

张涛、李祥泰、郝飞、袁浩博、吴生琼、季顺平、车变乐、水城燕。连接图像级、对象级、像素级的推理和理解。
arXiv预印本arXiv：2406.19389,2024年d。

张天宇、王素臣、卢李宇、张葛、塔斯拉基、赛拉杰斯瓦、傅杰、刘邦、本耀。视觉标题恢复。
arXiv:2406.06462, 2024e.

张一凡、张环宇、天浩臣、傅超友、张双清、吴俊飞、李峰、王昆、文优化、张张等。现实世界模式：你的多模态llm能挑战那些对人类来说很困难的高分辨率现实世界场景吗？arXiv预印本arXiv：2408.13257,2024年f。

赵一伦、谢路静、张浩伟、郭甘、龙一涛、胡志远、胡通彦、陈威远、李楚汉、宋俊杨、徐志坚、王承业、潘伟峰、上官子耀、唐相如、梁振文、刘义新、陈赵、德阿曼可汉。Mmvu：测量专家级别多学科的视频理解，2025年。URLhttps://arxiv.org/abs/2501.123 80 .

周、陆天健、梵天、巴苏、易一、周丹尼、勒侯。对大型语言模型的指令-后续评估。CoRR，abs/2311.07911,2023年。

周俊杰、严淑、赵波、吴博雅、肖石涛、杨西涛、熊勇、张波、黄窃贼、刘郑等。Mlvu：一个全面的基准，为多视频要求长视频理解。arXiv
预印arXiv：2406.04264,2024年。