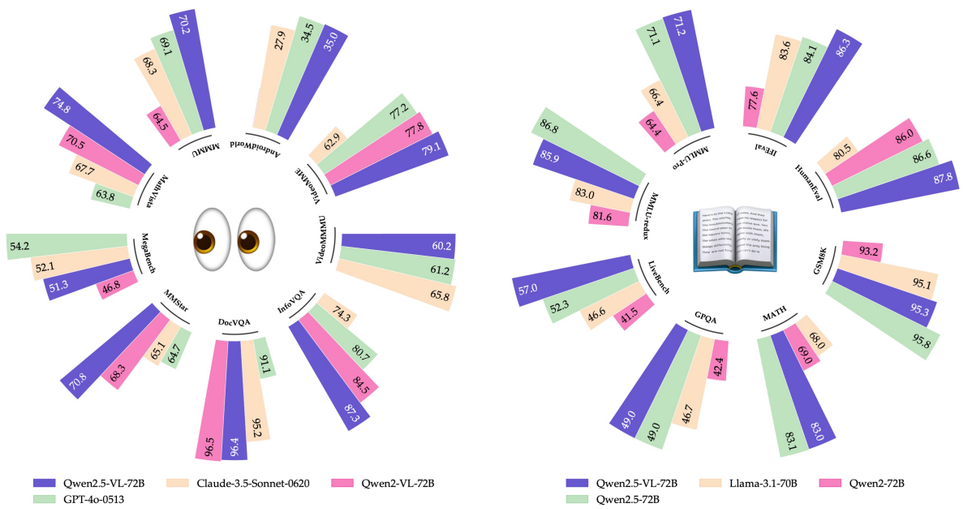
Qwen2.5-VL Technical Report

Qwen Team, Alibaba Group

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

Abstract

We introduce Qwen2.5-VL, the latest flagship model of Qwen vision-language series, which demonstrates significant advancements in both foundational capabilities and innovative functionalities. Qwen2.5-VL achieves a major leap forward in understanding and interacting with the world through enhanced visual recognition, precise object localization, robust document parsing, and long-video comprehension. A standout feature of Qwen2.5-VL is its ability to localize objects using bounding boxes or points accurately. It provides robust structured data extraction from invoices, forms, and tables, as well as detailed analysis of charts, diagrams, and layouts. To handle complex inputs, Qwen2.5-VL introduces dynamic resolution processing and absolute time encoding, enabling it to process images of varying sizes and videos of extended durations (up to hours) with second-level event localization. This allows the model to natively perceive spatial scales and temporal dynamics without relying on traditional normalization techniques. By training a native dynamic-resolution Vision Transformer (ViT) from scratch and incorporating Window Attention, we have significantly reduced computational overhead while maintaining native resolution. As a result, Qwen2.5-VL excels not only in static image and document understanding but also as an interactive visual agent capable of reasoning, tool usage, and task execution in real-world scenarios such as operating computers and mobile devices. The model achieves strong generalization across domains without requiring task-specific fine-tuning. Qwen2.5-VL is available in three sizes, addressing diverse use cases from edge AI to high-performance computing. The flagship Qwen2.5-VL-72B model matches state-of-the-art models like GPT-4o and Claude 3.5 Sonnet, particularly excelling in document and diagram understanding. The smaller Qwen2.5-VL-7B and Qwen2.5-VL-3B models outperform comparable competitors, offering strong capabilities even in resource-constrained environments. Additionally, Qwen2.5-VL maintains robust linguistic performance, preserving the core language competencies of the Qwen2.5 LLM.



[cs.CV] 19 Feb 2025

arXiv:2502.13923v1

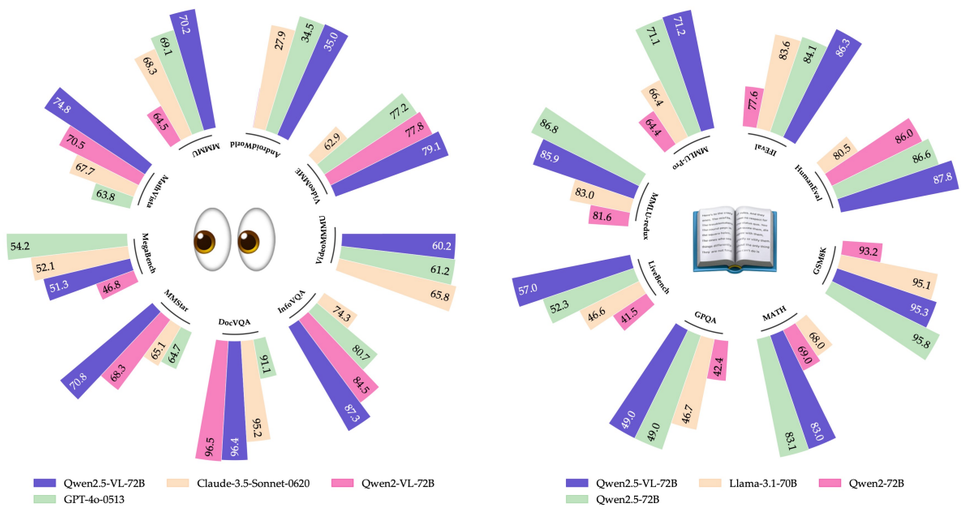
Qwen2.5-VL 技术报告

Qwen 团队，阿里巴巴集团

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

摘要

我们介绍了Qwen2.5-VL，这是Qwen视觉语言系列的最新旗舰模型，它在基础能力和创新功能方面都取得了显著进步。Qwen2.5-VL通过增强的视觉识别、精确的物体定位、稳健的文档解析和长视频理解，在理解和与世界互动方面实现了重大飞跃。
Qwen2.5-VL的一个突出特点是能够使用边界框或点精确地定位物体。它可以从发票、表格和表格中提供稳健的结构化数据提取，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引入了动态分辨率处理和绝对时间编码，使其能够处理不同尺寸的图像和长达数小时的视频，并实现二级事件定位。这使得模型能够原生感知空间尺度和时间动态，而无需依赖传统的归一化技术。通过从头开始训练原生动态分辨率视觉Transformer（ViT）并整合窗口注意力机制，我们显著降低了计算开销，同时保持了原生分辨率。因此，Qwen2.5-VL不仅在静态图像和文档理解方面表现出色，还作为一个能够进行推理、工具使用和任务执行的可交互视觉代理，在操作计算机和移动设备等实际场景中表现出色。该模型在跨领域实现了强大的泛化能力，而无需进行特定任务的微调。Qwen2.5-VL有三种尺寸，涵盖了从边缘人工智能到高性能计算的多样化用例。旗舰级的Qwen2.5-VL-72B模型与GPT-4o和Claude 3.5 Sonnet等当前最先进的模型相当，特别是在文档和图表理解方面表现出色。较小的Qwen2.5-VL-7B和Qwen2.5-VL-3B模型优于可比竞争对手，即使在资源受限的环境中也能提供强大的功能。
此外，Qwen2.5-VL保持了稳健的语言性能，保留了Qwen2.5大语言模型的核心语言能力。



1 Introduction

Large vision-language models (LVLMs) (OpenAI, 2024; Anthropic, 2024a; Team et al., 2023; Wang et al., 2024f) represent a pivotal breakthrough in artificial intelligence, signaling a transformative approach to multimodal understanding and interaction. By seamlessly integrating visual perception with natural language processing, these advanced models are fundamentally reshaping how machines interpret and analyze complex information across diverse domains. Despite significant advancements in multimodal large language models, the current capabilities of these models can be likened to the middle layer of a sandwich cookie—competent across various tasks but falling short of exceptional performance. Fine-grained visual tasks form the foundational layer of this analogy. In this iteration of Qwen2.5-VL, we are committed to exploring fine-grained perception capabilities, aiming to establish a robust foundation for LVLMs and create an agentic amplifier for real-world applications. The top layer of this framework is multi-modal reasoning, which is enhanced by leveraging the latest Qwen2.5 LLM and employing multi-modal QA data construction.

A spectrum of works have promoted the development of multimodal large models, characterized by architectural design, visual input processing, and data curation. One of the primary drivers of progress in LVLMs is the continuous innovation in architecture. The studies presented in (Alayrac et al., 2022; Li et al., 2022a; 2023b; Liu et al., 2023b;a; Wang et al., 2024i; Zhang et al., 2024b; Wang et al., 2023) have incrementally shaped the current paradigm, which typically consists of a visual encoder, a cross-modal projector, and LLM. Fine-grained perception models have emerged as another crucial area. Models like (Xiao et al., 2023; Liu et al., 2023c; Ren et al., 2024; Zhang et al., 2024a;d; Peng et al., 2023; Deitke et al., 2024) have pushed the boundaries of what is possible in terms of detailed visual understanding. The architectures of Omni (Li et al., 2024g; 2025b; Ye et al., 2024) and MoE (Riquelme et al., 2021; Lee et al., 2024; Li et al., 2024h;c; Wu et al., 2024b) also inspire the future evolution of LVLMs. Enhancements in visual encoders (Chen et al., 2023; Liu et al., 2024b; Liang et al., 2025) and resolution scaling (Li et al., 2023c; Ye et al., 2023; Li et al., 2023a) have played a pivotal role in improving the quality of practical visual understanding. Curating data with more diverse scenarios and higher-quality is an essential step in training advanced LVLMs. The efforts proposed in (Guo et al., 2024; Chen et al., 2024d; Liu et al., 2024a; Chen et al., 2024a; Tong et al., 2024; Li et al., 2024a) are highly valuable contributions to this endeavor.

However, despite their remarkable progress, vision-language models currently face developmental bottlenecks, including computational complexity, limited contextual understanding, poor fine-grained visual perception, and inconsistent performance across varied sequence length.

In this report, we introduce the latest work Qwen2.5-VL, which continues the open-source philosophy of the Qwen series, achieving and even surpassing top-tier closed-source models on various benchmarks. Technically, our contributions are four-folds: (1) We implement window attention in the visual encoder to optimize inference efficiency; (2) We introduce dynamic FPS sampling, extending dynamic resolution to the temporal dimension and enabling comprehensive video understanding across varied sampling rates; (3) We upgrade MRoPE in the temporal domain by aligning to absolute time, thereby facilitating more sophisticated temporal sequence learning; (4) We make significant efforts in curating high-quality data for both pre-training and supervised fine-tuning, further scaling the pre-training corpus from 1.2 trillion tokens to 4.1 trillion tokens.

The sparkling characteristics of Qwen2.5-VL are as follows:

- **Powerful document parsing capabilities:** Qwen2.5-VL upgrades text recognition to omni-document parsing, excelling in processing multi-scene, multilingual, and various built-in (handwriting, tables, charts, chemical formulas, and music sheets) documents.
- **Precise object grounding across formats:** Qwen2.5-VL unlocks improved accuracy in detecting, pointing, and counting objects, accommodating absolute coordinate and JSON formats for advanced spatial reasoning.
- **Ultra-long video understanding and fine-grained video grounding:** Our model extends native dynamic resolution to the temporal dimension, enhancing the ability to understand videos lasting hours while extracting event segments in seconds.
- **Enhanced agent Functionality for computer and mobile devices:** Leverage advanced grounding, reasoning, and decision-making abilities, boosting the model with superior agent functionality on smartphones and computers.

1 简介

大视觉语言模型（LVLMs）（OpenAI，2024；Anthropic，2024a；团队等，2023；王等人，2024f）代表了人工智能领域的一项关键突破，标志着多模态理解和交互方法的变革性进展。通过将视觉感知与自然语言处理无缝集成，这些高级模型正在从根本上改变机器如何跨不同领域解释和分析复杂信息的方式。尽管多模态大语言模型取得了显著进展，但这些模型的当前能力可以类比为三明治饼干的中层——在各种任务中表现合格，但未能达到卓越性能。细粒度视觉任务构成了这个类比的基础层。在Qwen2.5-VL的这次迭代中，我们致力于探索细粒度感知能力，旨在为LVLMs建立坚实的基础，并为现实世界应用创建一个代理式放大器。这个框架的顶层是多模态推理，它通过利用最新的Qwen2.5大语言模型并采用多模态QA数据构建而得到增强。

一系列研究推动了多模态大模型的发展，其特点在于架构设计、视觉输入处理和数据管理。LVLMs进展的主要驱动力之一是架构的持续创新。在(Alayrac等, 2022; 李等人, 2022a; 2023b; 刘等人, 2023b;a; 王等人, 2024i; 张等人, 2024b; 王等人, 2023)中提出的研究逐步塑造了当前范式，通常由视觉编码器、跨模态投影器和LLM组成。细粒度感知模型是另一个关键领域。像(Xiao等人, 2023; 刘等人, 2023c; Ren等人, 2024; 张等人, 2024a;d; Peng等人, 2023; Deitke等人, 2024)这样的模型在详细视觉理解方面拓展了可能性边界。Omni(李等人, 2024g; 2025b; Ye等人, 2024)和MoE(Riquelme等, 2021; Lee等人, 2024; 李等人 , 2024h;c; Wu 等人 , 2024b) 的架构也启发了 LVLMs 的未来演进。视觉编码器 (陈等人 , 2023; 刘等人 , 2024b; Liang 等人 , 2025) 和分辨率缩放 (李等人 , 2023c; Ye 等人 , 2023; 李等人 , 2023a)的改进在提升实际视觉理解质量方面发挥了关键作用。用更多样化的场景和更高质量的场景管理数据是训练高级LVLMs的重要步骤。在(Guo等人, 2024; Chen等人, 2024d; Liu等人, 2024a; Chen等人, 2024a; Tong等人, 2024; 李等人, 2024a)中提出的努力对这个事业具有高度价值的贡献。

然而，尽管取得了显著进展，当前视觉语言模型仍面临发展瓶颈，包括计算复杂度高、上下文理解有限、细粒度视觉感知能力差以及在不同序列长度下性能不稳定等问题。

在本报告中，我们介绍了最新成果Qwen2.5-VL，该模型延续了Qwen系列的开源理念，在多个基准测试中超越甚至达到了顶尖闭源模型。从技术角度来看，我们的贡献主要有四个方面：(1) 我们在视觉编码器中实现了窗口注意力机制，以优化推理效率；(2) 我们引入了动态帧率采样技术，将动态分辨率扩展到时间维度，从而支持跨不同采样率的全面视频理解；(3) 我们通过将MRoPE对齐到绝对时间，对时间域进行了升级，从而促进更复杂的时序序列学习；(4) 我们在构建高质量预训练和监督微调数据集方面付出了巨大努力，进一步将预训练语料库规模从120万亿token扩展至410万亿token。

Qwen2.5-VL 的闪亮特性如下：

- **强大的文档解析能力：** Qwen2.5-VL 将文字识别升级为全文档解析，擅长处理多场景、多语言以及各种内置（手写、表格、图表、化学公式和乐谱）文档。
- **跨格式的精准目标 grounding：** Qwen2.5-VL 解放了检测、指向和计数目标的准确性，支持绝对坐标和 JSON 格式以进行高级空间推理。
- **超长视频理解和细粒度视频 grounding：** 我们的模型将原生动态分辨率扩展到时间维度，增强了理解长达数小时视频的能力，同时提取秒级的事件片段。
- **增强 Agent 功能，适用于电脑和移动设备：** 利用先进的 grounding、推理和决策能力，提升模型在智能手机和电脑上的 Agent 功能表现。

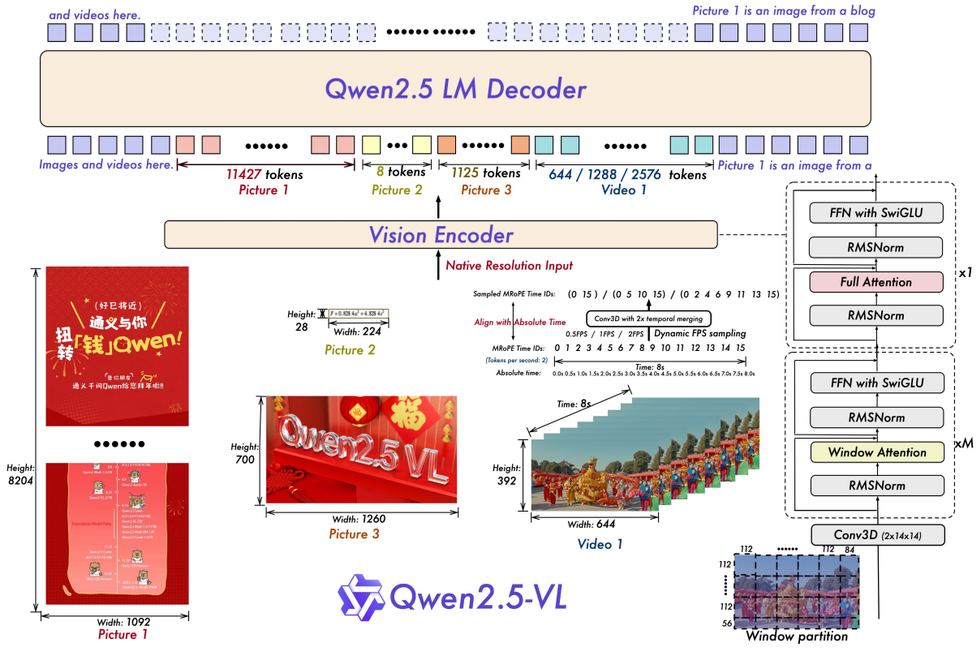


Figure 1: The Qwen2.5-VL framework demonstrates the integration of a vision encoder and a language model decoder to process multimodal inputs, including images and videos. The vision encoder is designed to handle inputs at their native resolution and supports dynamic FPS sampling. Images of varying sizes and video frames with different FPS rates are dynamically mapped to token sequences of varying lengths. Notably, MRoPE aligns time IDs with absolute time along the temporal dimension, enabling the model to better comprehend temporal dynamics, such as the pace of events and precise moment localization. The processed visual data is subsequently fed into the Qwen2.5 LM Decoder. We have re-engineered the vision transformer (ViT) architecture, incorporating advanced components such as FFN with SwiGLU activation, RMSNorm for normalization, and window-based attention mechanisms to enhance performance and efficiency.

2 Approach

In this section, we first outline the architectural updates of the Qwen2.5-VL series models and provide an overview of the data and training details.

2.1 Model Architecture

The overall model architecture of Qwen2.5-VL consists of three components:

Large Language Model: The Qwen2.5-VL series adopts large language models as its foundational component. The model is initialized with pre-trained weights from the Qwen2.5 LLM. To better meet the demands of multimodal understanding, we have modified the 1D RoPE (Rotary Position Embedding) to our Multimodal Rotary Position Embedding Aligned to Absolute Time.

Vision Encoder: The vision encoder of Qwen2.5-VL employs a redesigned Vision Transformer (ViT) architecture. Structurally, we incorporate 2D-RoPE and window attention to support native input resolutions while accelerating the computation of the entire visual encoder. During both training and inference, the height and width of the input images are resized to multiples of 28 before being fed into the ViT. The vision encoder processes images by splitting them into patches with a stride of 14, generating a set of image features. We provide a more detailed introduction to the vision encoder in Section 2.1.1.

MLP-based Vision-Language Merger: To address the efficiency challenges posed by long sequences of image features, we adopt a simple yet effective approach to compress the feature sequences before feeding them into the large language model (LLM). Specifically, instead of directly using the raw patch

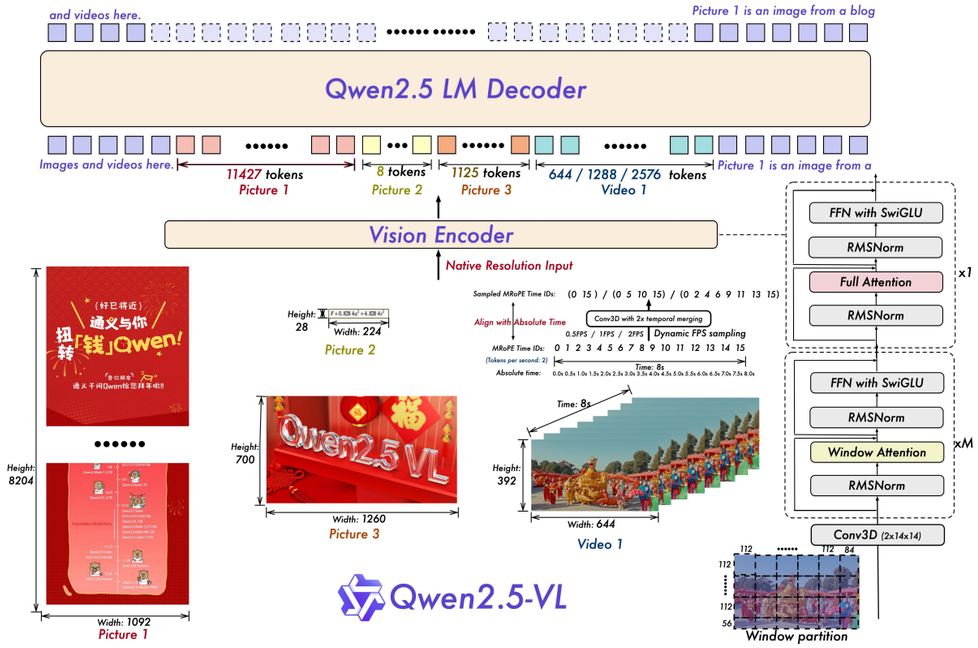


图1：Qwen2.5-VL框架展示了视觉编码器和语言模型解码器的集成，用于处理多模态输入，包括图像和视频。视觉编码器设计用于处理原始分辨率输入，并支持动态帧率采样。不同尺寸的图像和具有不同帧率的视频帧被动态映射到不同长度的标记序列。值得注意的是，MRoPE将时间ID与时间维度上的绝对时间对齐，使模型能够更好地理解时间动态，例如事件的速度和精确时刻定位。处理后的视觉数据随后被输入到Qwen2.5 LM解码器。我们对视觉Transformer（ViT）架构进行了重新设计，加入了高级组件，如带有SwiGLU激活的FFN、用于归一化的RMSNorm以及基于窗口的注意力机制，以提升性能和效率。

2 方法

在本节中，我们首先概述 Qwen2.5-VL 系列模型的架构更新，并提供数据和训练细节的概览。

2.1 模型架构

Qwen2.5-VL的整体模型架构由三个组件构成：

大语言模型：Qwen2.5-VL系列采用大语言模型作为其基础组件。模型使用Qwen2.5大语言模型的预训练权重进行初始化。为了更好地满足多模态理解的需求，我们将1D RoPE（旋转位置编码）修改为与绝对时间对齐的多模态旋转位置编码。

视觉编码器：Qwen2.5-VL的视觉编码器采用了一种重新设计的视觉Transformer（ViT）架构。在结构上，我们集成了2D-RoPE和窗口注意力机制，以支持原生输入分辨率，同时加速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都会被调整为28的倍数，然后再输入到ViT中。视觉编码器通过以14的步长将图像分割成块来处理图像，生成一组图像特征。我们将在第2.1.1节中更详细地介绍视觉编码器。

基于MLP的视觉语言合并器：为了应对长序列图像特征带来的效率挑战，我们采用一种简单而有效的方法，在将特征序列输入大语言模型（LLM）之前对其进行压缩。具体来说，我们没有直接使用原始的图像块

features extracted by the Vision Transformer (ViT), we first group spatially adjacent sets of four patch features. These grouped features are then concatenated and passed through a two-layer multi-layer perceptron (MLP) to project them into a dimension that aligns with the text embeddings used in the LLM. This method not only reduces computational costs but also provides a flexible way to dynamically compress image feature sequences of varying lengths.

In Table 1, the architecture and configuration of Qwen2.5-VL are detailed.

Configuration	Qwen2.5-VL-3B	Qwen2.5-VL-7B	Qwen2.5-VL-72B
Vision Transformer (ViT)			
Hidden Size	1280	1280	1280
# Layers	32	32	32
# Num Heads	16	16	16
Intermediate Size	3456	3456	3456
Patch Size	14	14	14
Window Size	112	112	112
Full Attention Block Indexes	{7, 15, 23, 31}	{7, 15, 23, 31}	{7, 15, 23, 31}
Vision-Language Merger			
In Channel	1280	1280	1280
Out Channel	2048	3584	8192
Large Language Model (LLM)			
Hidden Size	2048	3,584	8192
# Layers	36	28	80
# KV Heads	2	4	8
Head Size	128	128	128
Intermediate Size	4864	18944	29568
Embedding Tying	✓	✗	✗
Vocabulary Size	151646	151646	151646
# Trained Tokens	4.1T	4.1T	4.1T

Table 1: Configuration of Qwen2.5-VL.

2.1.1 Fast and Efficient Vision Encoder

The vision encoder plays a pivotal role in multimodal large language models (MLLMs). To address the challenges posed by computational load imbalances during training and inference due to native resolution inputs, we have redesigned the Vision Transformer (ViT) architecture. A key issue arises from the quadratic computational complexity associated with processing images of varying sizes. To mitigate this, we introduce windowed attention in most layers, which ensures that computational cost scales linearly with the number of patches rather than quadratically. In our architecture, only four layers employ full self-attention, while the remaining layers utilize windowed attention with a maximum window size of 112×112 (corresponding to 8×8 patches). Regions smaller than 112×112 are processed without padding, preserving their original resolution. This design allows the model to operate natively at the input resolution, avoiding unnecessary scaling or distortion.

For positional encoding, we adopt 2D Rotary Positional Embedding (RoPE) to effectively capture spatial relationships in 2D space. Furthermore, to better handle video inputs, we extend our approach to 3D patch partitioning. Specifically, we use 14×14 image patches as the basic unit, consistent with traditional ViTs for static images. For video data, two consecutive frames are grouped together, significantly reducing the number of tokens fed into the language model. This design not only maintains compatibility with existing architectures but also enhances efficiency when processing sequential video data.

To streamline the overall network structure, we align the ViT architecture more closely with the design principles of large language models (LLMs). Specifically, we adopt RMSNorm (Zhang &Sennrich, 2019) for normalization and SwiGLU (Dauphin et al., 2017) as the activation function. These choices enhance both computational efficiency and compatibility between the vision and language components of the model.

In terms of training, we train the redesigned ViT from scratch. The training process consists of several stages, including CLIP pre-training, vision-language alignment, and end-to-end fine-tuning. To ensure robustness across varying input resolutions, we employ dynamic sampling at native resolutions during

通过视觉Transformer（ViT）提取的特征，我们首先将空间上相邻的四个块特征分组。然后，这些分组后的特征被连接起来，并通过一个两层多层感知机（MLP）将其投影到与LLM中使用的文本嵌入相匹配的维度。这种方法不仅降低了计算成本，还提供了一种灵活的方式来动态压缩不同长度的图像特征序列。

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Transformer			
隐藏大小	1280	1280	1280
层数	32	32	32
头数	16	16	16
中间大小	3456	3456	3456
块大小	14	14	14
窗口大小	112	112	112
全注意力块索引	{7, 15, 23, 31}	{7, 15, 23, 31}	{7, 15, 23, 31}
视觉语言合并器			
输入通道	1280	1280	1280
输出通道	2048	3584	8192
大语言模型			
隐藏大小	2048	3,584	8192
# 层数	36	28	80
# KV 头	2	4	8
头大小	128	128	128
中等尺寸	4864	18944	29568
嵌入绑定	✓	✗	✗
词汇量大小	151646	151646	151646
# 训练的token	4.1T	4.1T	4.1T

表1：Qwen2.5-VL的配置。

2.1.1 快速高效的VisionEncoder

视觉编码器在多模态大语言模型（MLLMs）中扮演着关键角色。为了应对由于原生分辨率输入在训练和推理过程中导致的计算负载不平衡问题，我们重新设计了视觉Transformer（ViT）架构。一个关键问题是处理不同尺寸图像时相关的二次计算复杂度。为了缓解这一问题，我们在大多数层中引入了窗口注意力机制，确保计算成本与块数呈线性关系而非二次方关系。在我们的架构中，只有四层采用全自注意力机制，其余层则使用窗口注意力机制，其最大窗口大小为 112×112 (对应于 8×8 个块)。小于 112×112 的区域无需填充即可处理，从而保留了其原始分辨率。这种设计使得模型能够原生地在输入分辨率下运行，避免了不必要的缩放或失真。

对于位置编码，我们采用二维旋转位置嵌入（RoPE）来有效地捕捉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三维分块。具体来说，我们使用 14×14 图像块作为基本单元，这与传统的用于静态图像的 ViTs 保持一致。对于视频数据，两个连续的帧被组合在一起，显著减少了输入到语言模型中的 token 数量。这种设计不仅保持了与现有架构的兼容性，而且在处理序列视频数据时也提高了效率。

为了简化整体网络结构，我们将 ViT 架构更紧密地与大型语言模型（LLM）的设计原则对齐。具体来说，我们采用 RMSNorm（Zhang &Sennrich， 2019）进行归一化，并使用 SwiGLU（Dauphin et al.，2017）作为激活函数。这些选择提高了计算效率，并增强了模型中视觉和语言组件之间的兼容性。

在训练方面，我们从零开始训练重新设计的 ViT。训练过程包括多个阶段，包括 CLIP 预训练、视觉-语言对齐和端到端微调。为了确保在不同输入分辨率下的鲁棒性，我们在原生分辨率下采用动态采样，

training. Images are randomly sampled according to their original aspect ratios, enabling the model to generalize effectively to inputs of diverse resolutions. This approach not only improves the model's adaptability but also ensures stable and efficient training across different sizes of visual data.

2.1.2 Native Dynamic Resolution and Frame Rate

Qwen2.5-VL introduces advancements in both spatial and temporal dimensions to handle diverse multimodal inputs effectively.

In the spatial domain, Qwen2.5-VL dynamically converts images of varying sizes into sequences of tokens with corresponding lengths. Unlike traditional approaches that normalize coordinates, our model directly uses the actual dimensions of the input image to represent bounding boxes, points, and other spatial features. This allows the model to learn scale information inherently, improving its ability to process images across different resolutions.

For video inputs, Qwen2.5-VL incorporates dynamic frame rate (FPS) training and absolute time encoding. By adapting to variable frame rates, the model can better capture the temporal dynamics of video content. Unlike other approaches that incorporate textual timestamps or utilize additional heads to enable temporal grounding, we introduce a novel and efficient strategy that aligns MRoPE IDs directly with the timestamps. This approach allows the model to understand the tempo of time through the intervals between temporal dimension IDs, without necessitating any additional computational overhead.

2.1.3 Multimodal Rotary Position Embedding Aligned to Absolute Time

Positional embeddings are crucial for modeling sequential data in both vision and language modalities. Building upon the Multimodal Rotary Position Embedding (MRoPE) introduced in Qwen2-VL, we extend its capabilities to better handle temporal information in videos.

The MRoPE in Qwen2-VL decomposes the position embedding into three distinct components: temporal, height, and width to effectively model multimodal inputs. For textual inputs, all three components use identical position IDs, making MRoPE functionally equivalent to traditional 1D RoPE (Su et al., 2024). For images, the temporal ID remains constant across visual tokens, while unique IDs are assigned to the height and width components based on each token's spatial position within the image. When processing videos, which are treated as sequences of frames, the temporal ID increments for each frame, while the height and width components follow the same assignment pattern as for static images.

However, in Qwen2-VL, the temporal position IDs in MRoPE were tied to the number of input frames, which did not account for the speed of content changes or the absolute timing of events within the video. To address this limitation, Qwen2.5-VL introduces a key improvement: aligning the temporal component of MRoPE with absolute time. As shown in Figure 1, by leveraging the intervals between temporal IDs, the model is able to learn consistent temporal alignment across videos with different FPS sampling rates.

2.2 Pre-Training

In this section, we first describe the construction of the pre-training dataset, followed by an overview of the overall training pipeline and configuration.

2.2.1 Pre-Training Data

Compared to Qwen2-VL, we have significantly expanded the volume of our pre-training data, increasing it from 1.2 trillion tokens to approximately 4 trillion tokens. Our pre-training dataset was constructed through a combination of methods, including cleaning raw web data, synthesizing data, etc. The dataset encompasses a wide variety of multimodal data, such as image captions, interleaved image-text data, optical character recognition (OCR) data, visual knowledge (e.g., celebrity, landmark, flora, and fauna identification), multi-modal academic questions, localization data, document parsing data, video descriptions, video localization, and agent-based interaction data. Throughout the training process, we carefully adjusted the composition and proportions of these data types at different stages to optimize learning outcomes.

Interleaved Image-Text Data Interleaved image-text data is essential for multimodal learning, offering three key benefits: (1) enabling in-context learning with simultaneous visual and textual cues (Alayrac et al., 2022), (2) maintaining strong text-only capabilities when images are missing (Lin et al., 2024), and (3) containing a wide range of general information. However, much of the available interleaved data

在训练过程中。图像根据其原始宽高比随机采样，使模型能够有效地泛化到不同分辨率的输入。这种方法不仅提高了模型的适应性，还确保了在不同尺寸的视觉数据上稳定且高效的训练。

2.1.2 原生动态分辨率与帧率

Qwen2.5-VL 在空间和时间维度上引入了进步，以有效处理多样化的多模态输入。

在空间域中，Qwen2.5-VL 会动态地将不同尺寸的图像转换为具有相应长度的标记序列。与需要归一化坐标的传统方法不同，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。这使得模型能够内建地学习比例信息，从而提高其处理不同分辨率图像的能力。

对于视频输入，Qwen2.5-VL 包含动态帧率（FPS）训练和绝对时间编码。通过适应可变的帧率，模型能够更好地捕捉视频内容的时序动态。与其他方法（如结合文本时间戳或使用额外的头部来实现时序定位）不同，我们引入了一种新颖且高效的策略，将 MRoPE ID 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，而无需任何额外的计算开销。

2.1.3 多模态旋转位置嵌入对齐到绝对时间

位置嵌入对于建模视觉和语言模态中的序列数据至关重要。基于Qwen2-VL中引入的多模态旋转位置嵌入(MRoPE)，我们扩展了其功能，以更好地处理视频中时间信息。

Qwen2-VL中的MRoPE将位置嵌入分解为三个不同组件：时间、高度和宽度，以有效建模多模态输入。对于文本输入，所有三个组件使用相同的位ID，使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，而高度和宽度组件根据每个标记在图像中的空间位置分配唯一ID。在处理视频时，视频被视为帧序列，时间ID为每帧递增，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的时间位置ID与输入帧的数量绑定，这并未考虑内容变化的速度或视频内事件发生的绝对时间。为了解决这一局限性，Qwen2.5-VL引入了一个关键改进：将MRoPE的时间分量与绝对时间对齐。如图1所示，通过利用时间ID之间的间隔，模型能够学习到跨不同FPS采样率视频的一致时间对齐。

2.2 预训练

在本节中，我们首先描述预训练数据集的构建，然后概述整体训练流程和配置。

2.2.1 预训练数据

与Qwen2-VL相比，我们将预训练数据的规模显著扩大，从120万亿token增加到约400万亿token。我们的预训练数据集是通过多种方法构建的，包括清理原始网络数据、合成数据等。该数据集涵盖了多种多模态数据，如图像描述、交错图像-文本数据、光学字符识别（OCR）数据、视觉知识（例如，名人、地标、动植物识别）、多模态学术问题、定位数据、文档解析数据、视频描述、视频定位以及基于Agent的交互数据。在训练过程中，我们仔细调整了不同阶段这些数据类型的组成和比例，以优化学习效果。

交错图像-文本数据交错图像-文本数据对于多模态学习至关重要，它提供了三大关键优势：(1) 通过同时提供视觉和文本提示实现情境学习（Alayrac等，2022），(2) 当图像缺失时保持强大的纯文本能力（Lin等，2024），以及(3) 包含广泛的一般信息。然而，目前可用的交错数据

lacks meaningful text-image associations and is often noisy, limiting its usefulness for complex reasoning and creative generation.

To address these challenges, we developed a pipeline for scoring and cleaning data, ensuring only high-quality, relevant interleaved data is used. Our process involves two steps: standard data cleaning (Li et al., 2024e) followed by a four-stage scoring system using an internal evaluation model. The scoring criteria include: (1) text-only quality, (2) image-text relevance, (3) image-text complementarity, and (4) information density balance. This meticulous approach improves the model's ability to perform complex reasoning and generate coherent multimodal content.

The following is a description of these image-text scoring criteria:

Image-text Relevance: A higher score indicates a stronger connection between the image and text, where the image meaningfully supplements, explains or expands on the text rather than just decorating it.

Information Complementarity: A higher score reflects greater complementary information between the image and text. Each should provide unique details that together create a complete narrative.

Balance of Information Density: A higher score means a more balanced distribution of information between the image and text, avoiding excessive text or image information, and ensuring an appropriate balance between the two.

Grounding Data with Absolute Position Coordinates We adopt native resolution training with the aim of achieving a more accurate perception of the world. In contrast, relative coordinates fail to effectively represent the original size and position of objects within images. To address this limitation, Qwen2.5-VL uses coordinate values based on the actual dimensions of the input images during training to represent bounding boxes and points. This approach ensures that the model can better capture the real-world scale and spatial relationships of objects, leading to improved performance in tasks such as object detection and localization.

To improve the generalizability of grounding capabilities, we have developed a comprehensive dataset encompassing bounding boxes and points with referring expressions, leveraging both publicly available datasets and proprietary data. Our methodology involves synthesizing data into various formats, including XML, JSON, and custom formats, employing techniques such as copy-paste augmentation (Ghiasi et al., 2021) and synthesis with off-the-shelf models such as Grounding DINO (Liu et al., 2023c) and SAM (Kirillov et al., 2023). This approach facilitates a more robust evaluation and advancement of grounding abilities.

To enhance the model's performance on open-vocabulary detection, we expanded the training dataset to include over 10,000 object categories. Additionally, to improve the model's effectiveness in extreme object detection scenarios, we synthesized non-existent object categories within the queries and constructed image data containing multiple instances for each object.

To ensure superior point-based object grounding capabilities, we have constructed a comprehensive pointing dataset comprising both publicly available and synthetic data. Specifically, the data source includes public pointing and counting data from PixMo (Deitke et al., 2024), publicly accessible object grounding data (from both object detection and instance segmentation tasks), and data synthesized by an automated pipeline for generating precise pointing data towards certain image details.

Document Omni-Parsing Data To train Qwen2.5-VL, we synthesized a large corpus of document data. Traditional methods for parsing document content typically rely on separate models to handle layout analysis, text extraction, chart interpretation, and illustration processing. In contrast, Qwen2.5-VL is designed to empower a general-purpose model with comprehensive capabilities for parsing, understanding, and converting document formats. Specifically, we incorporated a diverse array of elements into the documents, such as tables, charts, equations, natural or synthetic images, music sheets, and chemical formulas. These elements were uniformly formatted in HTML, which integrates layout box information and descriptions of illustrations into HTML tag structures. We also enriched the document layouts according to typical reading sequences and included the coordinates corresponding to each module, such as paragraphs and charts, in the HTML-based ground truth. This innovative approach allows the complete information of any document, including its layout, text, charts, and illustrations, to be represented in a standardized and unified manner. As a result, Qwen2.5-VL achieves seamless integration of multimodal document elements, thereby facilitating more efficient and accurate document understanding and transformation.

Below is the QwenVL HTML format:

缺乏有意义的文本-图像关联，且通常存在噪声，限制其用于复杂推理和创意生成的作用。

为应对这些挑战，我们开发了一套用于评分和清洗数据的流程，确保仅使用高质量、相关的交错数据。我们的流程包含两个步骤：标准数据清洗（李等， 2024e）随后是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) 文本质量、(2) 图文相关性、(3) 图文互补性，以及(4) 信息密度平衡。这种严谨的方法提升了模型进行复杂推理和生成连贯多模态内容的能力。

以下是对这些图文评分标准的描述：

图文相关性：高分表示图像与文本之间联系更强，图像能对文本进行有意义的补充、解释或扩展，而不仅仅是装饰。
信息互补性：高分反映图像与文本之间互补信息更多。每部分应提供独特细节，共同构成完整叙事。

信息密度平衡：高分意味着图文信息分布更均衡，避免文本或图像信息过载，确保两者间适当平衡。

使用绝对位置坐标进行Grounding数据标注我们采用原生分辨率训练，旨在更准确地感知世界。相比之下，相对坐标无法有效地表示图像中物体的原始大小和位置。为了解决这个问题，Qwen2.5-VL在训练过程中使用基于输入图像实际尺寸的坐标值来表示边界框和点。这种方法确保模型能够更好地捕捉物体的现实世界尺度和空间关系，从而在目标检测和定位等任务中提高性能。

为提升 grounding 能力的泛化性，我们开发了一个包含边界框和带指代表达式的点的综合数据集，该数据集利用了公开数据集和专有数据。我们的方法涉及将数据合成多种格式，包括 XML、JSON 和自定义格式，采用了诸如复制粘贴增强（Ghiasi 等人, 2021) 以及使用现成模型（如 Grounding DINO (刘等人,2023c) 和 SAM (Kirillov 等人, 2023) 进行合成）等技术。这种方法促进了 grounding 能力的更稳健评估和进步。

为提升模型在开放词汇检测上的性能，我们将训练数据集扩展至包含超过10,000个物体类别。此外，为提高模型在极端物体检测场景下的有效性，我们在查询中合成了不存在的物体类别，并为每个物体构建了包含多个实例的图像数据。

为确保卓越的基于点的物体 grounding 能力，我们构建了一个包含公开可用和合成数据的综合指向数据集。具体而言，数据来源包括 PixMo（Deitke 等人， 2024）的公开指向和计数数据，以及公开可用的物体 grounding 数据（来自物体检测和实例分割任务），还有通过自动化流程合成、用于生成精确指向特定图像细节的数据。

文档全场景解析数据 为了训练 Qwen2.5-VL，我们合成了一个大规模的文档数据集。传统的文档内容解析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 旨在赋予通用模型全面的解析、理解和转换文档格式的功能。具体而言，我们将多种元素融入文档中，例如表格、图表、公式、自然或合成图像、乐谱和化学公式。这些元素以 HTML 格式统一格式化，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顺序丰富文档布局，并在基于 HTML 的真实标签中包含每个模块（如段落和图表）的坐标。这种创新方法使得任何文档的完整信息，包括其布局、文本、图表和插图，都能以标准化和统一的方式表示。因此，Qwen2.5-VL 实现了多模态文档元素的无缝集成，从而促进更高效和准确的文档理解和转换。

以下是 QwenVL HTML 格式：

QwenVL HTML Format

```
<html><body>
# paragraph
<p data-bbox="x1 y1 x2 y2"> content </p>
# table
<style>table{id} style</style><table data-bbox="x1 y1 x2 y2" class="table{id}"> table content
</table>
# chart
<div class="chart" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><table> chart content
</table></div>
# formula
<div class="formula" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /> <div> formula
content </div></div>
# image caption
<div class="image caption" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> image
caption </p></div>
# image ocr
<div class="image ocr" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> image ocr
</p></div>
# music sheet
<div class="music sheet" format="abc notation" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> music sheet content </div></div>
# chemical formula content
<div class="chemical formula" format="smile" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> chemical formula content </div></div>
</html></body>
```

This format ensures that all document elements are represented in a structured and accessible manner, enabling efficient processing and understanding by Qwen2.5-VL.

OCR Data Data from different sources are gathered and curated to enhance the OCR performance, including synthetic data, open-sourced data and in-house collected data. Synthetic data is generated through a visual text generation engine to produce high-quality text images in the wild. To support a wider range of languages and enhance multilingual capabilities, we have incorporated a large-scale multilingual OCR dataset. This dataset includes support for diverse languages such as French, German, Italian, Spanish, Portuguese, Arabic, Russian, Japanese, Korean, and Vietnamese. The dataset is carefully curated to ensure diversity and quality, utilizing both high-quality synthetic images and real-world natural scene images. This combination ensures robust performance across various linguistic contexts and improves the model's adaptability to different text appearances and environmental conditions. For chart-type data, we synthesized 1 million samples using visualization libraries including matplotlib, seaborn, and plotly, encompassing chart categories such as bar charts, relational diagrams, and heatmaps. Regarding tabular data, we processed 6 million real-world samples through an offline end-to-end table recognition model, subsequently filtering out low-confidence tables, overlapping tables, and tables with insufficient cell density.

Video Data To ensure enhanced robustness in understanding video data with varying frames per second (FPS), we dynamically sampled FPS during training to achieve a more evenly distributed representation of FPS within the training dataset. Additionally, for videos exceeding half an hour in length, we specifically constructed a set of long video captions by synthesizing multi-frame captions through a targeted synthesis pipeline. Regarding video grounding data, we formulated timestamps in both second-based formats and hour-minute-second-frame (hmsf) formats, ensuring that the model can accurately understand and output time in various formats.

Agent Data We enhance the perception and decision-making abilities to build the agent capabilities of Qwen2.5-VL. For perception, we collect screenshots on mobile, web, and desktop platforms. A synthetic data engine is used to generate screenshot captions and UI element grounding annotations. The caption task helps Qwen2.5-VL understand the graphic interface, while the grounding task helps it align the appearance and function of elements. For decision-making, we first unify the operations across mobile, web, and desktop platforms into a function call format with a shared action space. A set of annotated multi-step trajectories collected from open-source data and synthesized by agent framework (Wang et al., 2025; 2024b;c) on virtual environments are reformatted into a function format. We further generate a

QwenVL HTML 格式

```
<html><body># 段落<p data-bbox="x1 y1 x2 y2">内容 </p># 表格<style>table{id} style</
style><table data-bbox="x1 y1 x2 y2" class="table{id}"> table 内容</table># 图表<div
class="图表" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1y1 x2 y2" /><table> 图表 内容</
table></div># 公式<div class="公式" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2
y2" /> <div>公式 内容 </div></div># 图像描述<div class="图像描述" data-bbox="x1 y1 x2
y2"> <img data-bbox="x1 y1 x2 y2" /><p> 图像描述 </p></div># 图像 OCR<div class="图像
OCR" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> 图像 OCR</p></div>#
乐谱<div class="乐谱" 格式="abc 记谱法" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1
x2 y2" /> <div> 乐谱 内容 </div></div># 化学公式 内容<div class="化学公式" 格式="smile"
data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /> <div>化学公式 内容 </div></
div></html></body>
```

此格式 确保所有文档元素以结构化和可访问的方式表示 r,
启用 ef 高效处理和理解

OCR 数据 来自不同来源的数据被收集和整理，以提升 OCR 性能，包括合成数据、开源数据和内部收集的数据。合成数据通过视觉文本生成引擎生成，以在真实场景中产生高质量的文本图像。为了支持更广泛的语言并增强多语言能力，我们整合了一个大规模的多语言 OCR 数据集。该数据集支持多种语言，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。该数据集经过精心整理，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。这种组合确保了在各种语言环境下的鲁棒性能，并提高了模型对不同文本外观和环境条件的适应性。对于图表类型数据，我们使用 matplotlib、seaborn 和 plotly 等可视化库合成了 100 万个样本，涵盖柱状图、关系图和热力图等图表类别。关于表格数据，我们通过离线的端到端表格识别模型处理了 600 万个真实世界样本，随后过滤掉了低置信度表格、重叠表格和单元格密度不足的表格。

视频数据为确保在理解不同帧率（FPS）的视频数据时具有更强的鲁棒性，我们在训练过程中动态采样 FPS，以实现训练数据集中FPS分布的更均匀表示。此外，对于时长超过半小时的视频，我们通过一个有针对性的合成流程，专门构建了一套长视频字幕，通过合成多帧字幕生成。关于视频定位数据，我们制定了基于秒的格式和小时-分钟-秒-帧（hmsf）格式的时戳，确保模型能够准确理解和输出各种格式的时间。

Agent 数据我们增强了感知和决策能力，以构建 Qwen2.5-VL 的 Agent 功能。在感知方面，我们在移动端、网页端和桌面端收集截图。使用合成数据引擎生成截图描述和 UI 元素 Grounding 标注。描述任务帮助 Qwen2.5-VL 理解图形界面，而 Grounding 任务帮助它对齐元素的形态和功能。在决策方面，我们首先将移动端、网页端和桌面端的操作统一为具有共享动作空间的函数调用格式。一组从开源数据中收集并由 Agent 框架（王等人,2025; 2024b;c）在虚拟环境中合成的多步轨迹被重新格式化为函数格式。我们进一步生成一个

reasoning process for each step through human and model annotators (Xu et al., 2024). Specifically, given a ground-truth operation, we highlight it on the screenshot. Then, we provide the global query, along with screenshots from before and after this operation, to the annotators and require them to write reasoning content to explain the intention behind this operation. A model-based filter is used to screen out low-quality reasoning content. Such reasoning content prevents Qwen2.5-VL from overfitting to the ground-truth operations and makes it more robust in real-world scenarios.

Stages	Visual Pre-Training	Multimodal Pre-Training	Long-Context Pre-Training
Data	Image Caption Knowledge OCR	+ Pure text Interleaved Data VQA, Video Grounding, Agent	+ Long Video Long Agent Long Document
Tokens	1.5T	2T	0.6T
Sequence length	8192	8192	32768
Training	ViT	ViT & LLM	ViT & LLM

Table 2: Training data volume and composition across different stages.

2.2.2 Training Recipe

We trained a Vision Transformer (ViT) from scratch using DataComp (Gadre et al., 2023) and some in-house datasets as the initialization for the vision encoder, while leveraging the pre-trained Qwen2.5 large language model (LLM) (Yang et al., 2024a) as the initialization for the LLM component. As shown in Table 2, the pre-training process is divided into three distinct phases, each employing different data configurations and training strategies to progressively enhance the model's capabilities.

In the first phase, only the Vision Transformer (ViT) is trained to improve its alignment with the language model, laying a solid foundation for multimodal understanding. The primary data sources during this phase include image captions, visual knowledge, and OCR data. These datasets are carefully selected to foster ViT's ability to extract meaningful visual representations that can be effectively integrated with textual information.

In the second phase, all model parameters are unfrozen, and the model is trained on a diverse set of multimodal image data to enhance its capacity to process complex visual information. This phase introduces more intricate and reasoning-intensive datasets, such as interleaved data, multi-task learning datasets, visual question answering (VQA), multimodal mathematics, agent-based tasks, video understanding, and pure-text datasets. These datasets strengthen the model's ability to establish deeper connections between visual and linguistic modalities, enabling it to handle increasingly sophisticated tasks.

In the third phase, to further enhance the model's reasoning capabilities over longer sequences, video, and agent-based data are incorporated, alongside an increase in sequence length. This allows the model to tackle more advanced and intricate multimodal tasks with greater precision. By extending the sequence length, the model gains the ability to process extended contexts, which is particularly beneficial for tasks requiring long-range dependencies and complex reasoning.

To address the challenges posed by varying image sizes and text lengths, which can lead to imbalanced computational loads during training, we adopted a strategy to optimize training efficiency. The primary computational costs arise from the LLM and the vision encoder. Given that the vision encoder has relatively fewer parameters and that we introduced window attention to further reduce its computational demands, we focused on balancing the computational load of the LLM across different GPUs. Specifically, we dynamically packed data samples based on their corresponding input sequence lengths to the LLM, ensuring consistent computational loads. In the first and second phases, data were uniformly packed to a sequence length of 8,192, while in the third phase, the sequence length was increased to 32,768 to accommodate the model's enhanced capacity for handling longer sequences.

2.3 Post-training

The post-training alignment framework of Qwen2.5-VL employs a dual-stage optimization paradigm comprising Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO) (Rafailov et al., 2023). This hierarchical alignment strategy synergizes parameter-efficient domain adaptation with human preference distillation, addressing both representational grounding and behavioral refinement through distinct optimization objectives.

每一步的推理过程通过人工和模型标注者（徐等人, 2024）。具体来说，给定一个真实操作，我们在截图上高亮显示它。然后，我们将全局查询以及该操作前后的截图提供给标注者，要求他们编写推理内容来解释该操作的意图。使用基于模型的过滤器筛选出低质量的推理内容。这样的推理内容防止 Qwen2.5-VL 过拟合真实操作，并使其在实际场景中更加鲁棒。

阶段	视觉预训练	多模态预训练	长上下文预训练
Data	图像描述 知识 OCR	+ 纯文本 交错数据 VQA，视频 Grounding，Agent	+ 长视频 长Agent 长文档
标记	1.5T	2T	0.6T
序列长度	8192	8192	32768
训练	ViT	视觉 Transformer & 大语言模型	视觉 Transformer & 大语言模型

表2：不同阶段训练数据量及构成。

2.2.2 训练配方

我们使用DataComp（Gadre等人。， 2023）和部分内部数据集作为视觉编码器的初始化，从头训练了一个视觉Transformer（ViT），同时利用预训练的Qwen2.5大语言模型（LLM）（Yang等人。，2024a）作为LLM组件的初始化。如表2所示，预训练过程分为三个不同阶段，每个阶段采用不同的数据配置和训练策略，以逐步提升模型的能力。

在第一阶段，仅训练视觉Transformer（ViT），以提升其与语言模型的对齐度，为多模态理解奠定坚实基础。此阶段的主要数据来源包括图像描述、视觉知识和OCR数据。这些数据集经过精心选择，旨在培养ViT提取有意义的视觉表示的能力，以便有效整合文本信息。

在第二阶段，所有模型参数都被解冻，模型在多样化的多模态图像数据上进行训练，以增强其处理复杂视觉信息的能力。这一阶段引入了更复杂、推理更密集的数据集，例如交错数据、多任务学习数据集、视觉问答（VQA）、多模态数学、基于Agent的任务、视频理解和纯文本数据集。这些数据集增强了模型在视觉和语言模态之间建立更深层联系的能力，使其能够处理日益复杂的任务。

在第三阶段，为了进一步增强模型在较长序列、视频和基于Agent数据上的推理能力，同时增加了序列长度。这使得模型能够更精确地处理更高级和复杂的多模态任务。通过延长序列长度，模型获得了处理扩展上下文的能力，这对于需要长距离依赖和复杂推理的任务尤其有益。

为了应对图像大小和文本长度变化带来的挑战，这些变化可能导致训练期间计算负载不平衡，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大语言模型（LLM）和视觉编码器。鉴于视觉编码器的参数相对较少，并且我们引入了窗口注意力机制来进一步减少其计算需求，我们专注于平衡LLM在不同GPU上的计算负载。具体来说，我们根据数据样本对应输入序列长度动态打包数据，以确保计算负载一致。在第一阶段和第二阶段，数据被均匀打包到8,192的序列长度，而在第三阶段，序列长度增加到32,768，以适应模型处理更长序列的增强能力。

2.3 迁移学习

Qwen2.5-VL 的 post-training 对齐框架采用双阶段优化范式，包括监督微调 (SFT) 和直接偏好优化 (DPO) (Rafailov 等人,2023)。这种分层对齐策略将参数高效的领域适应与人类偏好蒸馏相结合，通过不同的优化目标解决表征接地和行为优化问题。

Supervised Fine-Tuning (SFT) aims to bridge the gap between pretrained representations and downstream task requirements through targeted instruction optimization. During this phase, we employ the ChatML format (Openai, 2024) to structure instruction-following data, deliberately diverging from the pretraining data schema while maintaining architectural consistency with Qwen2-VL (Wang et al., 2024e). This format transition enables three critical adaptations: 1) Explicit dialogue role tagging for multimodal turn-taking, 2) Structured injection of visual embeddings alongside textual instructions, and 3) Preservation of cross-modal positional relationships through format-aware packing. By exposing the model to curated multimodal instruction-response pairs under this enhanced schema, SFT enables efficient knowledge transfer while maintaining the integrity of pre-trained features.

2.3.1 Instruction Data

The Supervised Fine-Tuning (SFT) phase employs a meticulously curated dataset designed to enhance the model's instruction-following capabilities across diverse modalities. This dataset comprises approximately 2 million entries, evenly distributed between pure text data (50%) and multimodal data (50%), which includes image-text and video-text combinations. The inclusion of multimodal data enables the model to process complex inputs effectively. Notably, although pure text and multimodal entries are equally represented, multimodal entries consume significantly more tokens and computational resources during training due to the embedded visual and temporal information. The dataset is primarily composed of Chinese and English data, with supplementary multilingual entries to support broader linguistic diversity.

The dataset is structured to reflect varying levels of dialogue complexity, including both single-turn and multi-turn interactions. These interactions are further contextualized by scenarios ranging from single-image inputs to multi-image sequences, thereby simulating realistic conversational dynamics. The query sources are primarily drawn from open-source repositories, with additional contributions from curated purchased datasets and online query data. This combination ensures broad coverage and enhances the representativeness of the dataset.

To address a wide range of application scenarios, the dataset includes specialized subsets for General Visual Question Answering (VQA), image captioning, mathematical problem-solving, coding tasks, and security-related queries. Additionally, dedicated datasets for Document and Optical Character Recognition (Doc and OCR), Grounding, Video Analysis, and Agent Interactions are constructed to enhance domain-specific proficiency. Detailed information regarding the data can be found in the relevant sections of the paper. This structured and diverse composition ensures that the SFT phase effectively aligns pre-trained representations with the nuanced demands of downstream multimodal tasks, fostering robust and contextually aware model performance.

2.3.2 Data Filtering Pipeline

The quality of training data is a critical factor influencing the performance of vision-language models. Open-source and synthetic datasets typically exhibit significant variability, often containing noisy, redundant, or low-quality samples. Therefore, rigorous data cleaning and filtering processes are essential to address these issues. Low-quality data can lead to suboptimal alignment between pretrained representations and downstream task requirements, thereby diminishing the model's ability to effectively handle complex multimodal tasks. Consequently, ensuring high-quality data is paramount for achieving robust and reliable model performance.

To address these challenges, we implement a two-stage data filtering pipeline designed to systematically enhance the quality of the Supervised Fine-Tuning (SFT) dataset. This pipeline comprises the following stages:

Stage 1: Domain-Specific Categorization In the initial stage, we employ *Qwen2-VL-Instag*, a specialized classification model derived from Qwen2-VL-72B, to perform hierarchical categorization of question-answer (QA) pairs. This model organizes QA pairs into eight primary domains, such as *Coding* and *Planning*, which are further divided into 30 fine-grained subcategories. For example, the primary domain *Coding* is subdivided into subcategories including *Code_Debugging*, *Code_Generation*, *Code_Translation*, and *Code_Understanding*. This hierarchical structure facilitates domain-aware and subdomain-aware filtering strategies, enabling the pipeline to optimize data-cleaning processes tailored to each category's specific characteristics. Consequently, this enhances the quality and relevance of the supervised fine-tuning (SFT) dataset.

Stage 2: Domain-Tailored Filtering The second stage involves domain-tailored filtering, which integrates both rule-based and model-based approaches to comprehensively enhance data quality. Given

监督微调 (SFT) 旨在通过目标指令优化弥合预训练表征与下游任务需求之间的差距。在此阶段，我们采用 ChatML 格式 (OpenAI, 2024) 来组织指令跟随数据，在有意偏离预训练数据模式的同时保持与 Qwen2-VL 的架构一致性 (王等人, 2024e)。这种格式转换使三种关键适应成为可能：1) 多模态轮流中的显式对话角色标记，2) 在文本指令旁边结构化注入视觉嵌入，3) 通过格式感知打包保留跨模态位置关系。通过在这种增强模式下向模型展示精心策划的多模态指令-响应对，SFT 能够实现高效的知识迁移，同时保持预训练特征的完整性。

2.3.1 指令数据

监督微调（SFT）阶段采用一个精心策划的数据集，旨在提升模型跨多种模态的指令跟随能力。该数据集包含约200万个条目，纯文本数据（50%）和多模态数据（50%）均等分布，其中多模态数据包括图像-文本和视频-文本组合。多模态数据的加入使模型能有效处理复杂输入。值得注意的是，尽管纯文本和多模态条目数量相同，但由于嵌入的视觉和时序信息，多模态条目在训练过程中消耗的token和计算资源显著更多。该数据集主要由中文和英文数据组成，并辅以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级别的对话复杂度，包括单轮和多轮交互。这些交互通过从单图像输入到多图像序列的场景进一步进行情境化，从而模拟真实的对话动态。查询来源主要来自开源仓库，并辅以精选购买数据集和在线查询数据。这种组合确保了广泛覆盖，并提升了数据集的代表性。

为应对广泛的应用场景，数据集包含针对通用视觉问答（VQA）、图像描述、数学问题解决、编程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）、Grounding、视频分析和Agent交互的专用数据集，以提升特定领域的专业能力。有关数据的详细信息，请参阅论文的相关章节。这种结构化和多样化的组成确保了SFT阶段能够有效地将预训练表示与下游多模态任务的细微需求对齐，从而促进模型稳健且具有上下文感知能力的性能。

2.3.2 数据过滤流程

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变异性，经常包含噪声、冗余或低质量样本。因此，严格的数据清理和过滤过程对于解决这些问题至关重要。低质量数据会导致预训练表示与下游任务需求之间的对齐不佳，从而削弱模型有效处理复杂多模态任务的能力。因此，确保高质量数据对于实现稳健可靠的模型性能至关重要。

为应对这些挑战，我们实现了一个两阶段数据过滤流程，旨在系统性地提升Supervised Fine-Tuning (SFT)数据集的质量。该流程包含以下阶段：

阶段 1：领域特定分类 在初始阶段，我们采用*Qwen2-VL-Instag*，一个基于Qwen2-VL-72B衍生的专用分类模型，对问答（QA）对进行分层分类。该模型将QA对组织成八个主要领域，例如编程 和规划，这些领域进一步细分为 30 个细粒度子类别。例如，主要领域编程 被细分为包括代码_调试、代码_生成、代码_翻译和代码_理解等子类别。这种分层结构促进了领域感知和子领域感知的过滤策略，使流程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。

阶段 2：领域定制过滤 第二阶段涉及领域定制过滤，该过程整合了基于规则和基于模型的方法，以全面提升数据质量。给定

the diverse nature of domains such as Document Processing, Optical Character Recognition (OCR), and Visual Grounding, each may necessitate unique filtering strategies. Below, we provide an overview of the general filtering strategies applied across these domains.

Rule-Based Filtering employs predefined heuristics to eliminate low-quality or problematic entries. Specifically, for datasets related to Document Processing, OCR, and Visual Grounding tasks, repetitive patterns are identified and removed to prevent distortion of the model's learning process and ensure optimal performance. Additionally, entries containing incomplete, truncated, or improperly formatted responses—common in synthetic datasets and multimodal contexts—are excluded. To maintain relevance and uphold ethical standards, queries and answers that are unrelated or could potentially lead to harmful outputs are also discarded. This structured approach ensures that the dataset adheres to ethical guidelines and meets task-specific requirements.

Model-Based Filtering further refines the dataset by leveraging reward models trained on the Qwen2.5-VL series. These models evaluate multimodal QA pairs across multiple dimensions. Queries are assessed for complexity and relevance, retaining only those examples that are appropriately challenging and contextually pertinent. Answers are evaluated based on correctness, completeness, clarity, relevance to the query, and helpfulness. In visual-grounded tasks, particular attention is given to verifying the accurate interpretation and utilization of visual information. This multi-dimensional scoring ensures that only high-quality data progresses to the SFT phase.

2.3.3 Rejection Sampling for Enhanced Reasoning

To complement our structured data filtering pipeline, we employ rejection sampling as a strategy to refine the dataset and enhance the reasoning capabilities of the vision-language model (VLM). This approach is particularly critical for tasks requiring complex inference, such as mathematical problem-solving, code generation, and domain-specific visual question answering (VQA). Prior research has shown that incorporating Chain-of-Thought (CoT) Wei et al. (2022) reasoning significantly improves a model's inferential performance. (DeepSeek-AI et al., 2024) Our post-training experiments confirm this, underscoring the importance of structured reasoning processes for achieving high-quality outcomes.

The rejection sampling process begins with datasets enriched with ground truth annotations. These datasets are carefully curated to include tasks that demand multi-step reasoning, such as mathematical problem-solving, code generation, and domain-specific VQA. Using an intermediate version of the Qwen2.5-VL model, we evaluate the generated responses against the ground truth. Only samples where the model's output matches the expected answers are retained, ensuring the dataset consists solely of high-quality, accurate examples.

To further improve data quality, we apply additional constraints to filter out undesirable outputs. Specifically, we exclude responses that exhibit code-switching, excessive length, or repetitive patterns. These criteria ensure clarity and coherence in the CoT reasoning process, which is crucial for downstream applications.

A key challenge in applying CoT reasoning to vision-language models is their reliance on both textual and visual modalities. Intermediate reasoning steps may fail to adequately integrate visual information, either by ignoring relevant visual cues or misinterpreting them. To address this, we have developed rule-based and model-driven filtering strategies to validate the accuracy of intermediate reasoning steps. These mechanisms ensure that each step in the CoT process effectively integrates visual and textual modalities. Despite these efforts, achieving optimal modality alignment remains an ongoing challenge that requires further advancements.

The data generated through rejection sampling significantly enhances the model's reasoning proficiency. By iteratively refining the dataset and removing low-quality or erroneous samples, we enable the model to learn from high-fidelity examples that emphasize accurate and coherent reasoning. This methodology not only strengthens the model's ability to handle complex tasks but also lays the groundwork for future improvements in vision-language modeling.

2.3.4 Training Recipe

The post-training process for Qwen2.5-VL consists of two phases: Supervised Fine-Tuning (SFT) and Direct Preference Optimization (DPO), both with the Vision Transformer (ViT) parameters frozen. In the SFT phase, the model is fine-tuned on diverse multimodal data, including image-text pairs, video, and pure text, sourced from general VQA, Rejection Sampling, and specialized datasets such as Document and OCR, Grounding, Video, and Agent-related tasks. The DPO phase focuses exclusively on image-text and pure text data, utilizing preference data to align the model with human preferences, with each sample processed only once to ensure efficient optimization. This streamlined process enhances the model's

由于文档处理、光学字符识别 (OCR) 和视觉接地等领域的多样性，每个领域可能需要独特的过滤策略。以下，我们概述了在这些领域中应用的通用过滤策略。

基于规则的过滤采用预定义的启发式规则来消除低质量或存在问题的条目。具体而言，对于与文档处理、OCR 和视觉接地任务相关的数据集，会识别并移除重复模式，以防止扭曲模型的训练过程并确保最佳性能。此外，包含不完整、截断或不正确格式化响应的条目——常见于合成数据集和多模态上下文中——也会被排除。为了保持相关性并遵守道德标准，与问题无关或可能产生有害输出的查询和答案也会被丢弃。这种结构化方法确保数据集符合道德规范并满足特定任务要求。

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优化数据集。这些模型从多个维度评估多模态问答对。查询会根据复杂性和相关性进行评估，仅保留那些具有适当挑战性和上下文相关性的示例。答案会根据正确性、完整性、清晰度、与查询的相关性以及有用性进行评估。在视觉接地任务中，特别关注验证对视觉信息的准确解释和利用。这种多维评分确保只有高质量的数据才会进入 SFT 阶段。

2.3.3 拒绝采样增强推理

为补充我们的结构化数据过滤流程，我们采用拒绝采样作为一种策略来优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为关键，例如数学问题求解、代码生成和特定领域的视觉问答（VQA）。先前研究表明，结合思维链（CoT）魏等人(2022) 推理可以显著提升模型的推理性能。（DeepSeek-AI 等人， 2024) 我们的离线训练实验证实了这一点，强调了结构化推理过程对于实现高质量结果的重要性。

拒绝采样流程始于带有真实标注的数据集。这些数据集经过精心筛选，包含需要多步推理的任务，例如数学问题求解、代码生成和特定领域的VQA。使用Qwen2.5-VL模型的中间版本，我们评估生成响应与真实标注的匹配程度。只有模型输出与预期答案一致的样本才会被保留，确保数据集仅包含高质量、准确的示例。

为了进一步提升数据质量，我们应用额外的约束条件来过滤掉不理想的输出。具体来说，我们排除了那些存在代码转换、过长或重复模式的响应。这些标准确保了CoT推理过程的清晰性和连贯性，这对于下游应用至关重要。

将CoT推理应用于视觉语言模型的一个关键挑战在于它们同时依赖文本和视觉模态。中间推理步骤可能无法充分整合视觉信息，要么忽略相关的视觉线索，要么误读它们。为了解决这个问题，我们开发了基于规则和模型驱动的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过程中的每一步都能有效整合视觉和文本模态。尽管付出了这些努力，但要实现最优的模态对齐仍然是一个需要进一步发展的持续挑战。

通过拒绝采样生成的数据显著提升了模型的推理能力。通过迭代优化数据集并移除低质量或错误的样本，我们使模型能够从强调准确和连贯推理的高保真示例中学习。这种方法不仅增强了模型处理复杂任务的能力，也为未来视觉语言模型的改进奠定了基础。

2.3.4 训练配方

Qwen2.5-VL的post-training过程包含两个阶段：监督微调（SFT）和直接偏好优化（DPO），这两个阶段都冻结了视觉Transformer（ViT）参数。在SFT阶段，模型在多样化的多模态数据上进行微调，包括图像-文本对、视频和纯文本，数据来源包括通用VQA、拒绝采样以及专业数据集，如文档、OCR、Grounding、视频和Agent相关任务。在DPO阶段，模型专注于图像-文本和纯文本数据，利用偏好数据使模型与人类偏好保持一致，每个样本仅处理一次以确保高效优化。这个简化的流程提升了模型的

cross-modal reasoning and task-specific performance while maintaining alignment with user intent.

3 Experiments

In this section, we first introduce the overall model and compare it with the current state-of-the-art (SoTA) models. Then, we evaluate the model's performance across various sub-capabilities.

3.1 Comparison with the SOTA Models

Table 3: **Performance of Qwen2.5-VL and State-of-the-art.**

Datasets	Previous Open-source SoTA	Claude-3.5 Sonnet-0620	GPT-4o 0513	InternVL2.5 78B	Qwen2-VL 72B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
College-level Problems								
$MMMU_{val}$ (Yue et al., 2023)	70.1 Chen et al. (2024d)	68.3	69.1	70.1	64.5	**70.2**	58.6	53.1
$MMMU\text{-}Pro_{overall}$ (Yue et al., 2024)	48.6 Chen et al. (2024d)	51.5	**51.9**	48.6	46.2	51.1	38.3	31.56
Math								
$MathVista_{mini}$ (Lu et al., 2024)	72.3 Chen et al. (2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$MATH\text{-}Vision_{full}$ (Wang et al., 2024d)	32.2 Chen et al. (2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$MathVerse_{mini}$ (Zhang et al., 2024c)	51.7 Chen et al. (2024d)	-	50.2	51.7	-	**57.6**	49.2	47.6
General Visual Question Answering								
MegaBench (Chen et al., 2024b)	47.4 MiniMax et al. (2025)	52.1	**54.2**	45.6	46.8	51.3	36.8	28.9
$MMBench\text{-}EN_{test}$ (Liu et al., 2023d)	88.3 Chen et al. (2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$MMBench\text{-}CN_{test}$ (Liu et al., 2023d)	88.5 Chen et al. (2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$MMBench\text{-}V1.1\text{-}EN_{test}$ (Liu et al., 2023d)	87.4 Chen et al. (2024d)	80.9	83.1	87.4	86.1	**88.4**	82.6	77.4
MMStar (Chen et al., 2024c)	69.5 Chen et al. (2024d)	65.1	64.7	69.5	68.3	**70.8**	63.9	55.9
MME_{sum} (Fu et al., 2023)	**2494** Chen et al. (2024d)	1920	2328	**2494**	2483	2448	2347	2157
MuirBench (Wang et al., 2024a)	63.5 Chen et al. (2024d)	-	68.0	63.5	-	**70.7**	59.6	47.7
$BLINK_{val}$ (Fu et al., 2024c)	63.8 Chen et al. (2024d)	-	**68.**0	63.8	-	64.4	56.4	47.6
$CRPE_{relation}$ (Wang et al., 2024h)	78.8 Chen et al. (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$HallBench_{avg}$ (Guan et al., 2023)	**58.1** Wang et al. (2024f)	55.5	55.0	57.4	**58.1**	55.2	52.9	46.3
MTVQA (Tang et al., 2024)	**31.9** Chen et al. (2024d)	25.7	27.8	31.9	30.9	31.7	29.2	24.8
$RealWorldQA_{avg}$ (X.AI, 2024)	78.7 Chen et al. (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$MME\text{-}RealWorld_{en}$ (Zhang et al., 2024f)	62.9 Chen et al. (2024d)	51.6	45.2	62.9	-	**63.2**	57.4	53.1
$MMVet_{turbo}$ (Yu et al., 2024)	74.0 Wang et al. (2024f)	70.1	69.1	72.3	74.0	**76.2**	67.1	61.8
MM-MT-Bench (Agrawal et al., 2024)	7.4 Agrawal et al. (2024)	7.5	**7.72**	-	6.59	7.6	6.3	5.7

The experimental section evaluates the performance of Qwen2.5-VL across a variety of datasets, comparing it with state-of-the-art models such as Claude-3.5-Sonnet-0620 (Anthropic, 2024a), GPT-4o-0513 (OpenAI, 2024), InternVL2.5 (Chen et al., 2024d), and different sizes of Qwen2-VL (Wang et al., 2024e). In college-level problems, Qwen2.5-VL-72B achieves a score of 70.2 on MMMU (Yue et al., 2023). For MMMU-Pro (Yue et al., 2024), Qwen2.5-VL-72B scores 51.1, surpassing the previous open-source state-of-the-art models and achieving performance comparable to GPT-4o.

In math-related tasks, Qwen2.5-VL-72B demonstrates strong capabilities. On MathVista (Lu et al., 2024), it achieves a score of 74.8, outperforming the previous open-source state-of-the-art score of 72.3. For MATH-Vision (Wang et al., 2024d), Qwen2.5-VL-72B scores 38.1, while MathVerse (Zhang et al., 2024c) achieves 57.6, both showing competitive results compared to other leading models.

For general visual question answering, Qwen2.5-VL-72B excels across multiple benchmarks. On MMbench-EN (Liu et al., 2023d), it achieves a score of 88.6, slightly surpassing the previous best score of 88.3. The model also performs well in MuirBench (Wang et al., 2024a) with a score of 70.7 and BLINK (Fu et al., 2024c) with 64.4. In the multilingual capability evaluation of MTVQA (Tang et al., 2024), Qwen2.5-VL-72B achieves a score of 31.7, showcasing its powerful multilingual text recognition abilities. In subjective evaluations such as MMVet (Yu et al., 2024) and MM-MT-Bench (Agrawal et al., 2024), Qwen2.5-VL-72B scores 76.2 and 7.6, respectively, demonstrating excellent natural conversational experience and user satisfaction.

3.2 Performance on Pure Text Tasks

To critically evaluate the performance of instruction-tuned models on pure text tasks, as illustrated in Table 4, we selected several representative benchmarks to assess the model's capabilities across a variety of domains, including general tasks (Wang et al., 2024j; Gema et al., 2024; White et al., 2024), mathematics and science tasks (Rein et al., 2023; Hendrycks et al., 2021; Cobbe et al., 2021), coding tasks (Chen et al., 2021; Cassano et al., 2023), and alignment task (Zhou et al., 2023). We compared Qwen2.5-VL with several large language models (LLMs) of similar size. The results demonstrate that Qwen2.5-VL not only achieves state-of-the-art (SoTA) performance on multimodal tasks but also exhibits leading performance on pure text tasks, showcasing its versatility and robustness across diverse evaluation criteria.

交叉测试 推理能力和任务特定性能，同时保持与用户意图的一致性 t.

3 实验部分

在本节中，我们首先介绍整体模型，并将其与当前最先进技术（SoTA）模型进行比较。然后，我们评估模型在各个子能力上的性能。

3.1 与当前最先进技术（SoTA）模型比较

表3： **Qwen2.5-VL与当前最先进技术的性能对比。**

数据集	上一个开源SOTA	Claude-3.5 Sonnet-0620	GPT-4o 0513	InternVL2.5 78B	Qwen2-VL 72B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
大学水平问题								
$MMMU_{val}$ (岳等人, 2023)	70.1 陈等人 (2024d)	68.3	69.1	70.1	64.5	**70.2**	58.6	53.1
$MMMU\text{-}Pro_{overall}$ (岳等人,2024)	48.6 陈等人(2024d)	51.5	**51.9**	48.6	46.2	51.1	38.3	31.56
Math								
$MathVista_{mini}$ (陆等人, 2024)	72.3 陈等人(2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$MATH\text{-}Vision_{full}$ (王等人,2024d)	32.2 陈等人(2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$MathVerse_{mini}$ (张等人,2024c)	51.7 陈等人(2024d)	-	50.2	51.7	-	**57.6**	49.2	47.6
通用视觉问答								
MegaBench(陈等人,2024b)	47.4MiniMax 等人 (2025)	52.1	**54.2**	45.6	46.8	51.3	36.8	28.9
$MMBench\text{-}EN_{test}$ (刘等人, 2023d)	88.3 陈等人(2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$MMBench\text{-}CN_{test}$ (刘等人, 2023d)	88.5 陈等人(2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$MMBench\text{-}V1.1\text{-}EN_{test}$ (刘等人, 2023d)	87.4 陈等人(2024d)	80.9	83.1	87.4	86.1	**88.4**	82.6	77.4
MMStar (陈等人, 2024c)	69.5 陈等人 (2024d)	65.1	64.7	69.5	68.3	**70.8**	63.9	55.9
MME_{sum} (付等人,2023)	**2494** 陈等人 (2024d)	1920	2328	**2494**	2483	2448	2347	2157
MuirBench (王等人, 2024a)	63.5 陈等人 (2024d)	-	68.0	63.5	-	**70.7**	59.6	47.7
$BLINK_{val}$ (付等人,2024c)	63.8 陈等人 (2024d)	-	**68.**0	63.8	-	64.4	56.4	47.6
$CRPE_{relation}$ (王等人,2024h)	78.8 陈等人 (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$HallBench_{avg}$ (关等人, 2023)	**58.1** 王等人(2024f)	55.5	55.0	57.4	**58.1**	55.2	52.9	46.3
MTVQA (唐等人,2024)	**31.9**陈等人 (2024d)	25.7	27.8	31.9	30.9	31.7	29.2	24.8
$RealWorldQA_{avg}$ (X.AI, 2024)	78.7 陈等人 (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$MME\text{-}RealWorld_{en}$ (张等人, 2024f)	62.9 陈等人 (2024d)	51.6	45.2	62.9	-	**63.2**	57.4	53.1
$MMVet_{turbo}$ (余等人,2024)	74.0 王等人(2024f)	70.1	69.1	72.3	74.0	**76.2**	67.1	61.8
MM-MT-Bench(阿格拉瓦尔等人,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，将其与当前最先进技术模型（如 Claude-3.5-Sonnet-0620（Anthropic， 2024a），GPT-4o-0513（OpenAI， 2024），InternVL2.5（陈等人， 2024d），以及不同大小的 Qwen2-VL（王等人， 2024e）进行比较）。在大学水平问题中，Qwen2.5-VL-72B 在 MMMU 上获得 70.2 分（岳等人， 2023）。对于 MMMU- Pro（岳等人， 2024），Qwen2.5-VL-72B 得分为 51.1，超越了之前的开源当前最先进技术模型，并达到了 GPT-4o 的性能水平。

在数学相关任务中，Qwen2.5-VL-72B展现出强大的能力。在MathVista（陆等人， 2024）上，它获得了74.8分，超过了之前开源的当前最先进技术的72.3分。对于MATH-Vision（王等人，2024d），Qwen2.5-VL-72B得分为38.1，而MathVerse（张等人， 2024c）得分为57.6，两者与其他领先模型相比都显示出竞争性结果。

对于通用视觉问答，Qwen2.5-VL-72B在多个基准测试中表现出色。在MMbench- EN（刘等人，2023d）上，它获得了88.6分，略微超过了之前的最佳分数88.3。该模型在MuirBench（王等人，2024a）上的得分为70.7，在BLINK（付等人，2024c）上的得分为64.4。在MTVQA（唐等人， 2024）的多语言能力评估中，Qwen2.5-VL-72B获得了31.7分，展示了其强大的多语言文本识别能力。在主观评估如MMVet（余等人， 2024）和MM-MT-Bench（Agrawal等人， 2024）中，Qwen2.5-VL-72B分别得分为76.2和7.6，展示了出色的自然对话体验和用户满意度。

3.2 纯文本任务性能

为了批判性地评估指令微调模型在纯文本任务上的性能，如表4所示，我们选择了几个有代表性的基准测试来评估模型在不同领域的能力，包括一般任务（王等人,2024j;葛玛等人， 2024; 怀特等人， 2024），数学与科学任务（莱恩等人， 2023; 亨德里克斯等人， 2021; 科布等人， 2021），编程任务（陈等人，2021; 卡萨诺等人， 2023），以及对齐任务（周等人， 2023）。我们将Qwen2.5-VL与几个规模相似的大语言模型（LLMs）进行了比较。结果表明，Qwen2.5-VL不仅在多模态任务上实现了最先进（SoTA）的性能，而且在纯文本任务上也表现出领先性能，展示了其在不同评估标准下的通用性和鲁棒性。

Table 4: **Performance on pure text tasks of the 70B+ Instruct models and Qwen2.5-VL.**

Datasets	Llama-3.1-70B	Llama-3.1-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-VL-72B
General Tasks					
MMLU-Pro	66.4	**73.3**	64.4	71.1	71.2
MMLU-redux	83.0	86.2	81.6	**86.8**	85.9
LiveBench-0831	46.6	53.2	41.5	52.3	**57.0**
Mathematics & Science Tasks					
GPQA	46.7	**51.1**	42.4	49.0	49.0
MATH	68.0	73.8	69.0	**83.1**	83.0
GSM8K	95.1	**96.8**	93.2	95.8	95.3
Coding Tasks					
HumanEval	80.5	**89.0**	86.0	86.6	87.8
MultiPL-E	68.2	73.5	69.2	75.1	**79.5**
Alignment Tasks					
IFEval	83.6	86.0	77.6	84.1	**86.3**

3.3 Quantitative Results

3.3.1 General Visual Question Answering

To comprehensively evaluate the model's capabilities in general visual question answering (VQA) and dialogue, we conducted extensive experiments across a diverse range of datasets. As illustrated in Table 3, Qwen2.5-VL demonstrates state-of-the-art performance in various VQA tasks, subjective evaluations, multilingual scenarios, and multi-image questions. Specifically, it excels on benchmark datasets such as MMBench series (Liu et al., 2023d), MMStar (Chen et al., 2024c), MME (Fu et al., 2023), MuirBench (Wang et al., 2024a), BLINK(Fu et al., 2024c), CRPE (Wang et al., 2024h), HallBench (Guan et al., 2023), MTVQA (Tang et al., 2024), MME-RealWorld (Zhang et al., 2024f), MMVet (Yu et al., 2024), and MM-MT-Bench (Agrawal et al., 2024).

In the domain of visual detail comprehension and reasoning, Qwen2.5-VL-72B achieves an accuracy of 88.4% on the MMBench-EN-V1.1 dataset, surpassing previous state-of-the-art models such as InternVL2.5 (78B) and Claude-3.5 Sonnet-0620. Similarly, on the MMStar dataset, Qwen2.5-VL attains a score of 70.8%, outperforming other leading models in this benchmark. These results underscore the model's robustness and adaptability across diverse linguistic contexts.

Furthermore, in high-resolution real-world scenarios, specifically on the MME-RealWorld benchmark, Qwen2.5-VL demonstrates state-of-the-art performance with a score of 63.2, showcasing its broad adaptability to realistic environments. Additionally, in multi-image understanding tasks evaluated on the MuirBench dataset, Qwen2.5-VL achieves a leading score of 70.7, further highlighting its superior generalization capabilities. Collectively, these results illustrate the strong versatility and effectiveness of Qwen2.5-VL in addressing general-purpose visual question answering (VQA) tasks across various scenarios.

Notably, even the smaller-scale versions of Qwen2.5-VL, specifically Qwen2.5-VL-7B and Qwen2.5-VL-3B, exhibit highly competitive performance. For instance, on the MMStar dataset, Qwen2.5-VL-7B achieves 63.9%, while Qwen2.5-VL-3B scores 55.9%. This demonstrates that Qwen2.5-VL's architecture is not only powerful but also scalable, maintaining strong performance even with fewer parameters.

3.3.2 Document Understanding and OCR

We evaluated our models across a diverse range of OCR, chart, and document understanding benchmarks. Table 5 demonstrates the performance comparison between Qwen2.5-VL models and top-tier models on following OCR-related benchmarks: AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2021b), InfoVQA (Mathew et al., 2021a), ChartQA (Masry et al., 2022), CharXiv (Wang et al., 2024k), SEED-Bench-2-Plus (Li et al., 2024b), OCRBench (Liu et al., 2023e), OCRBench_v2 (Fu et al., 2024b), CC-OCR (Yang et al., 2024b), OmniDocBench (Ouyang et al., 2024), VCR (Zhang et al., 2024e).

For OCR-related parsing benchmarks on element parsing for multi-scene, multilingual, and various built-in (handwriting, tables, charts, chemical formulas, and mathematical expressions) documents,

表4： **70B+Instruct模型和Qwen2.5-VL在纯文本任务上的性能** .

数据集	Llama-3.1-70B	Llama-3.1-405B	Qwen2-72B	Qwen2.5-72B	Qwen2.5-VL-72B
一般任务					
MMLU-Pro	66.4	**73.3**	64.4	71.1	71.2
MMLU-redux	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					
GPQA	46.7	**51.1**	42.4	49.0	49.0
MATH	68.0	73.8	69.0	**83.1**	83.0
GSM8K	95.1	**96.8**	93.2	95.8	95.3
编程任务					
HumanEval	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 一般视觉问答

为了全面评估模型在通用视觉问答（VQA）和对话方面的能力，我们在多种数据集上进行了广泛的实验。如表3所示，Qwen2.5-VL在各种VQA任务、主观评估、多语言场景和多图像问题中均表现出最先进性能。具体而言，它在MMBench系列（刘等人，2023d）、MMStar（陈等人，2024c）、MME（付等人，2023）、MuirBench（王等人，2024a）、BLINK（付等人，2024c）、CRPE（王等人，2024h）、HallBench（关等人，2023）、MTVQA（唐等人，2024）、MME-RealWorld（张等人，2024f）、MMVet（余等人，2024）以及MM-MT-Bench（阿格拉瓦尔等人，2024）等基准数据集上表现优异。

在视觉细节理解和推理领域，Qwen2.5-VL-72B在MMBench-EN-V1.1数据集上达到88.4%的准确率，超越了之前的当前最先进模型，如InternVL2.5（78B）和Claude-3.5 Sonnet-0620。同样，在MMStar数据集上，Qwen2.5-VL获得70.8分，优于该基准中的其他领先模型。这些结果突显了该模型的鲁棒性和跨不同语言环境的适应性。

此外，在高分辨率真实场景中，特别是在MME-RealWorld基准上，Qwen2.5-VL展现出当前最先进性能，得分为63.2，展示了其对现实环境的广泛适应性。此外，在MuirBench数据集上评估的多图像理解任务中，Qwen2.5-VL获得领先分数70.7，进一步突显了其卓越的泛化能力。综合来看，这些结果说明了Qwen2.5-VL在处理各种场景下的通用视觉问答（VQA）任务时，具有强大的通用性和有效性。

值得注意的是，即使是 Qwen2.5-VL 的小规模版本，特别是 Qwen2.5-VL-7B 和 Qwen2.5-VL-3B，也表现出极具竞争力的性能。例如，在 MMStar 数据集上，Qwen2.5-VL-7B 达到了 63.9%，而 Qwen2.5-VL-3B 得分为 55.9%。这表明 Qwen2.5-VL 的架构不仅强大，而且可扩展，即使参数更少也能保持强劲的性能。

3.3.2 文档理解与OCR

我们对模型在多种 OCR、图表和文档理解基准测试中进行了评估。表 5 展示了 Qwen2.5-VL 模型与顶级模型在以下 OCR 相关基准测试上的性能比较：AI2D（Kembhavi 等人，2016 年）、TextVQA（Singh 等人，2019 年）、DocVQA（Mathew 等人，2021b）、InfoVQA（Mathew 等人，2021a）、ChartQA（Masry 等人，2022 年）、CharXiv（Wang 等人，2024k）、SEED-Bench-2-Plus（Li 等人，2024b）、OCRBench（Liu 等人，2023e）、OCRBench_v2（付等人，2024b）、CC-OCR（Yang 等人，2024b）、OmniDocBench（Ouyang 等人，2024 年）、VCR（张等人，2024e）。

针对多场景、多语言以及各种内置（手写、表格、图表、化学公式和数学表达式）文档的元素解析OCR相关解析基准，

as CC-OCR and OmniDocBench, Qwen2.5-VL-72B model sets the new state-of-the-art due to curated training data and excellent capability of LLM models.

For OCR-related understanding benchmarks for scene text, chart, diagram and document, Qwen2.5-VL models achieve impressive performance with good understanding abilities. Notably, on composite OCR-related understanding benchmarks as OCRBench, InfoVQA which focusing on infographics, and SEED-Bench-2-Plus covering text-rich scenarios including charts, maps, and webs, Qwen2.5-VL-72B achieves remarkable results, significantly outperforming strong competitors such as InternVL2.5-78B.

Furthermore, for OCR-related comprehensive benchmarks as OCRBench_v2 including a wide range of OCR-related parsing and understanding tasks, top performance is also achieved by Qwen2.5-VL models, largely exceeding best model Gemini 1.5-Pro by 9.6% and 20.6% for English and Chinese track respectively.

Table 5: **Performance of Qwen2.5-VL and other models on OCR, chart, and document understanding benchmarks.**

Datasets	Claude-3.5 Sonnet	Gemini 1.5 Pro	GPT 4o	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
OCR-related Parsing Tasks							
CC-OCR	62.5	73.0	66.9	64.7	**79.8**	77.8	74.5
$\text{OmniDocBench}_{\text{edit en/zh}}\downarrow$	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
OCR-related Understanding Tasks							
$\text{AI2D}_{\text{w. M.}}$	81.2	88.4	84.6	**89.1**	88.7	83.9	81.6
$\text{TextVQA}_{\text{val}}$	76.5	78.8	77.4	83.4	83.5	**84.9**	79.3
$\text{DocVQA}_{\text{test}}$	95.2	93.1	91.1	95.1	**96.4**	95.7	93.9
$\text{InfoVQA}_{\text{test}}$	74.3	81.0	80.7	84.1	**87.3**	82.6	77.1
$\text{ChartQA}_{\text{test Avg.}}$	**90.8**	87.2	86.7	88.3	89.5	87.3	84.0
$\text{CharXiv}_{\text{RQ/DQ}}$	**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-Plus	71.7	70.8	72.0	71.3	**73.0**	70.4	67.6
OCRBench	788	754	736	854	**885**	864	797
$\text{VCR}_{\text{En-Hard-EM}}$	41.7	28.1	73.2	-	79.8	**80.5**	37.5
OCR-related Comprehensive Tasks							
$\text{OCRBench\_v2}_{\text{en/zh}}$	45.2/39.6	51.9/43.1	46.5/32.2	49.8/52.1	**61.5/63.7**	56.3/57.2	54.3/52.1

3.3.3 Spatial Understanding

Understanding spatial relationships is crucial for developing AI models that can interpret and interact with the world as humans do. In Large Vision-Language Models, visual grounding allows for the precise localization and identification of specific objects, regions, or elements within an image based on natural language queries or descriptions. This capability transcends traditional object detection by establishing a semantic relationship between visual content and linguistic context, facilitating more nuanced and contextually aware visual reasoning. We evaluated Qwen2.5-VL's grounding capabilities on the referring expression comprehension benchmarks (Kazemzadeh et al., 2014; Mao et al., 2016), object detection in the wild (Li et al., 2022b), self-curated point grounding benchmark, and CountBench (Paiss et al., 2023).

We compare Qwen2.5-VL's visual grounding capabilities with other leading LVLMs including Gemini, Grounding-DINO (Liu et al., 2023c), Molmo (Deitke et al., 2024), and InternVL2.5.

Qwen2.5-VL achieves leading performance across different benchmarks from box-grounding, and point-grounding to counting. By equipping Qwen2.5-VL with both box and point-grounding capability, it is able to understand, locate, and reason on the very details of certain parts of an image. For open-vocabulary object detection, Qwen2.5-VL achieves a good performance of 43.1 mAP on ODinW-13, surpassing most LVLMs and quickly narrowing the gap between generalist models and specialist models. In addition, Qwen2.5-VL unlocks the point-based grounding ability so that it could precisely locate the very details of a certain object, which was difficult to represent by a bounding box in the past. Qwen2.5-VL's counting ability also makes great progress, achieving a leading accuracy of 93.6 on CountBench with Qwen2.5-VL-72B using a "detect then count"-style prompt.

3.3.4 Video Understanding and Grounding

We assessed our models across a diverse range of video understanding and grounding tasks, utilizing benchmarks that include videos ranging from a few seconds to several hours in length. Table 8 demonstrates the performance comparison between Qwen2.5-VL models and top-tier proprietary models on the following video benchmarks: Video-MME (Fu et al., 2024a), Video-MMMU (Hu et al., 2025), MMVU (Zhao

如CC-OCR和OmniDocBench，Qwen2.5-VL-72B模型集因精选的训练数据和LLM模型的出色能力，确立了当前最先进技术。

对于场景文本、图表、图表和文档的OCR相关理解基准，Qwen2.5-VL模型凭借良好的理解能力取得了令人印象深刻的性能表现。值得注意的是，在OCR相关理解基准如OCRbench、专注于信息图的InfoVQA以及涵盖图表、地图和网页等文本丰富场景的SEED-Bench-2-Plus上，Qwen2.5-VL-72B取得了显著成果，显著优于InternVL2.5-78B等强劲对手。

此外，在OCR相关综合基准OCRbench_v2（包含广泛的OCR相关解析和理解任务）上，Qwen2.5-VL模型同样取得了顶尖性能，英语和中文赛道分别大幅超越了最佳模型Gemini 1.5-Pro 9.6%和20.6%。

表5： **Qwen2.5-VL及其他模型在OCR、图表和文档理解基准上的性能表现**。

数据集	Claude-3.5 十四行诗	Gemini 1.5 Pro	GPT 4o	InternVL2.5 78B	Qwen2.5-VL 72B	Qwen2.5-VL 7B	Qwen2.5-VL 3B
*OCR*相关解析任务							
CC-OCR	62.5	73.0	66.9	64.7	**79.8**	77.8	74.5
$\text{OmniDocBench}_{\text{edit en/zh}}\downarrow$	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
*OCR*相关理解任务							
$\text{AI2D}_{\text{w. M。}}$	81.2	88.4	84.6	**89.1**	88.7	83.9	81.6
$\text{TextVQA}_{\text{val}}$	76.5	78.8	77.4	83.4	83.5	**84.9**	79.3
$\text{DocVQA}_{\text{test}}$	95.2	93.1	91.1	95.1	**96.4**	95.7	93.9
$\text{InfoVQA}_{\text{test}}$	74.3	81.0	80.7	84.1	**87.3**	82.6	77.1
$\text{图表QA}_{\text{test Avg。}}$	**90.8**	87.2	86.7	88.3	89.5	87.3	84.0
$\text{CharXiv}_{\text{RQ/DQ}}$	**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-Plus	71.7	70.8	72.0	71.3	**73.0**	70.4	67.6
OCRBench	788	754	736	854	**885**	864	797
$\text{VCR}_{\text{En-Hard-EM}}$	41.7	28.1	73.2	-	79.8	**80.5**	37.5
*OCR*相关综合任务							
$\text{OCRBench\_v2}_{\text{en/zh}}$	45.2/39.6	51.9/43.1	46.5/32.2	49.8/52.1	**61.5/63.7**	56.3/57.2	54.3/52.1

3.3.3 空间理解

理解空间关系对于开发能够像人类一样解释和与世界互动的AI模型至关重要。在大视觉语言模型中，视觉接地允许根据自然语言查询或描述，在图像中精确定位和识别特定对象、区域或元素。这项能力超越了传统的目标检测，通过在视觉内容和语言上下文之间建立语义关系，促进更细致和具有上下文感知能力的视觉推理。我们在指代表达理解基准（Kazemzadeh等， 2014； Mao等， 2016），野外目标检测（Li等， 2022b），自建点接地基准，以及计数基准（Paiss等， 2023）上评估了Qwen2.5-VL的接地能力。

We compa 将Qwen2.5-VL的视觉接地能力与其他领先的LVLMs（包括Gemin）进行比较 i, Grounding -DINO (刘等人, 2023c), Molmo (戴特克等人,2024), 和 InternVL2.5.

Qwen2.5-VL 在框接地、点接地和计数等不同基准测试中均取得领先性能。通过为 Qwen2.5-VL 配备框接地和点接地的双重能力，它能够理解、定位并对图像中某些部分的细节进行推理。在开放词汇目标检测方面，Qwen2.5-VL 在 ODinW-13 上取得了 43.1 mAP 的良好性能，超越了大多数大视觉语言模型，并迅速缩小了通用模型与专业模型之间的差距。此外，Qwen2.5-VL 解锁了基于点的接地能力，使其能够精确定位某些目标的细节，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显著进展，使用“先检测后计数”式提示的 Qwen2.5-VL-72B 在计数基准上达到了 93.6 的领先准确率。

3.3.4 视频理解与Grounding

我们对我们的模型在多种视频理解和 grounding 任务上进行了评估，使用了包含从几秒到数小时不等长度的视频的基准测试。表8 展示了 Qwen2.5-VL 模型与顶级专有模型在以下视频基准测试上的性能比较：Video-MME (付等人， 2024a)，Video-MMMU (胡等人， 2025)，MMVU (赵

Table 6: **Performance of Qwen2.5-VL and other models on grounding.**

Datasets	Gemini 1.5 Pro	Grounding DINO	Molmo 72B	InternVL2.5 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
Refcoco_{testA}	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
Refcoco+_{testA}	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
Refcocog_{test}	76.2	87.0	-	92.2	90.3	87.2	85.7
ODinW	36.7	55.0	-	31.7	43.1	37.3	37.5
PointGrounding	-	-	69.2	-	67.5	67.3	58.3

Table 7: **Performance of Qwen2.5-VL and other models on counting.**

Datasets	Gemini 1.5-Pro	GPT-4o	Claude-3.5 Sonnet	Molmo-72b	InternVL2.5-78B	Qwen2.5-VL-72B
CountBench	85.5	87.9	89.7	91.2	72.1	93.6

et al., 2025), MVBench (Li et al., 2024d), MMBench-Video (Fang et al., 2024), LongVideoBench (Wu et al., 2024a), EgoSchema (Mangalam et al., 2023), PerceptionTest (Patraucean et al., 2024), MLVU (Zhou et al., 2024), LVBench (Wang et al., 2024g), TempCompass (Liu et al., 2024c) and Charades-STA (Gao et al., 2017). Notably, on LVBench and MLVU, which evaluate long-form video understanding capabilities through question-answering tasks, Qwen2.5-VL-72B achieves remarkable results, significantly outperforming strong competitors such as GPT-4o.

By utilizing the proposed synchronized MRoPE, Qwen2.5-VL enhances its capabilities in time-sensitive video understanding, featuring improved timestamp referencing, temporal grounding, dense captioning, and additional functionalities. On the Charades-STA dataset, which assesses the capability to accurately localize events or activities with precise timestamps, Qwen2.5-VL-72B achieves an impressive mIoU score of 50.9, thereby surpassing the performance of GPT-4o. For all evaluated benchmarks, we capped the maximum number of frames analyzed per video at 768, with the total number of video tokens not exceeding 24,576.

Table 8: **Performance of Qwen2.5-VL and other models on video benchmarks.**

Datasets	Gemini 1.5-Pro	GPT-4o	Qwen2.5-VL-72B	Qwen2.5-VL-7B	Qwen2.5-VL-3B
Video Understanding Tasks					
$\text{Video-MME}_{\text{w/o sub.}}$	**75.0**	71.9	73.3	65.1	61.5
$\text{Video-MME}_{\text{w sub.}}$	**81.3**	77.2	79.1	71.6	67.6
Video-MMMU	53.9	**61.2**	60.2	47.4	-
MMVU_{val}	65.4	**67.4**	62.9	50.1	-
MVBench	60.5	64.6	**70.4**	69.6	67.0
MMBench-Video	1.30	1.63	**2.02**	1.79	1.63
$\text{LongVideoBench}_{\text{val}}$	64.0	**66.7**	60.7	56.0	54.2
LVBench	33.1	30.8	**47.3**	45.3	43.3
$\text{EgoSchema}_{\text{test}}$	71.2	72.2	**76.2**	65.0	64.8
$\text{PerceptionTest}_{\text{test}}$	-	-	**73.2**	70.5	66.9
$\text{MLVU}_{\text{M-Avg}}$	-	64.6	**74.6**	70.2	68.2
$\text{TempCompass}_{\text{Avg}}$	67.1	73.8	**74.8**	71.7	64.4
Video Grounding Tasks					
$\text{Charades-STA}_{\text{mIoU}}$	-	35.7	**50.9**	43.6	38.8

3.3.5 Agent

Agent capabilities within multimodal models are crucial for enabling these models to effectively interact with real-world devices. We assess the agent capabilities of Qwen2.5-VL through various aspects. The UI

表 6： **Qwen2.5-VL 和其他模型在 grounding 上的性能表现。**

数据集	Gemini 1.5 Pro	Grounding DINO	Molmo 72B	InternVL2.5 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
Refcoco_{testA}	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
Refcoco+_{testA}	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
Refcocog_{test}	76.2	87.0	-	92.2	90.3	87.2	85.7
ODinW	36.7	55.0	-	31.7	43.1	37.3	37.5
点标注	-	-	69.2	-	67.5	67.3	58.3

表 7： **Qwen2.5-VL 和其他模型在计数任务上的性能表现。**

数据集	Gemin 1.5-Pro	GPT-4o	Claude-3.5 Sonnet	Molmo-72B	InternVL2.5-78B	Qwen2.5-VL-72B
计数基准	85.5	87.9	89.7	91.2	72.1	93.6

等人， 2025)，MVBench (李等人， 2024d)，MMBench-Video (方等人， 2024)，LongVideoBench (吴等人,2024a)，EgoSchema (马加拉姆等人， 2023)，PerceptionTest (帕特鲁塞安等人， 2024)，MLVU (周等人,2024)，LVBench (王等人， 2024g)，TempCompass(刘等人， 2024c) 和 Charades-STA (高等人，2017)。值得注意的是，在 LVBench 和 MLVU 上，这两个基准测试通过问答任务评估长视频理解能力，Qwen2.5-VL-72B 取得了显著成果，显著优于 GPT-4o 等强劲竞争对手。

通过利用提出的同步 MRoPE，Qwen2.5-VL 增强了其在时间敏感视频理解方面的能力，具备改进的时间戳引用、时间 grounding、密集描述以及其他功能。在评估准确定位事件或活动并给出精确时间戳能力的 Charades-STA 数据集上，Qwen2.5-VL-72B 实现了令人印象深刻的 50.9 mIoU 分数，从而超越了 GPT-4o 的性能。对于所有评估基准，我们限制了每个视频分析的最大帧数为 768，视频 token 总数不超过 24,576。

表8：**Qwen2.5-VL和其他模型在视频基准测试上的性能。**

数据集	Gemini 1.5-Pro	GPT-4o	Qwen2.5-VL-72B	Qwen2.5-VL-7B	Qwen2.5-VL-3B
视频理解任务					
视频-$\text{MME}_{\text{w/o 子。}}$	**75.0**	71.9	73.3	65.1	61.5
视频-$\text{MME}_{\text{w 子。}}$	**81.3**	77.2	79.1	71.6	67.6
Video-MMMU	53.9	**61.2**	60.2	47.4	-
MMVU_{val}	65.4	**67.4**	62.9	50.1	-
MVBench	60.5	64.6	**70.4**	69.6	67.0
MMBench-Video	1.30	1.63	**2.02**	1.79	1.63
长视频基准测试$_{\text{val}}$	64.0	**66.7**	60.7	56.0	54.2
LVBench	33.1	30.8	**47.3**	45.3	43.3
$\text{EgoSchema}_{\text{test}}$	71.2	72.2	**76.2**	65.0	64.8
$\text{PerceptionTest}_{\text{test}}$	-	-	**73.2**	70.5	66.9
$\text{MLVU}^{\text{M-Avg}}$	-	64.6	**74.6**	70.2	68.2
$\text{TempCompass}^{\text{Avg}}$	67.1	73.8	**74.8**	71.7	64.4
视频*Grounding*任务					
Charades-STA mIoU	-	35.7	**50.9**	43.6	38.8

3.3.5 Agent

Agent capa 多模态模型中的 Agent 能力对于使这些模型能够有效地交互至关重要 ct
与真实- 世界设备。我们通过多个方面评估 Qwen2.5-VL 的 Agent 能力。U I

elements grounding is evaluated by ScreenSpot (Cheng et al., 2024) and ScreenSpot Pro (Li et al., 2025a). Offline evaluations are conducted on Android Control (Li et al., 2024f), while online evaluations are performed on platforms including AndroidWorld (Rawles et al., 2024), MobileMiniWob++ (Rawles et al., 2024), and OSWorld (Xie et al., 2025). We compare the performance of Qwen2.5-VL-72B againsts other prominent models, such as GPT-4o (OpenAI, 2024), Gemini 2.0 (Deepmind, 2024), Claude (Anthropic, 2024b), Aguvis-72B (Xu et al., 2024), and Qwen2-VL-72B (Wang et al., 2024e). The results are demonstrated in Table 9.

Table 9: **Performance of Qwen2.5-VL and other models on GUI Agent benchmarks.**

Benchmarks	GPT-4o	Gemini 2.0	Claude	Aguvis-72B	Qwen2-VL-72B	Qwen2.5-VL-72B
ScreenSpot	18.1	84.0	83.0	**89.2**	-	87.1
ScreenSpot Pro	-	-	17.1	23.6	1.6	**43.6**
Android Control High$_{EM}$	20.8	28.5	12.5	66.4	59.1	**67.36**
Android Control Low$_{EM}$	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

The performance of Qwen2.5-VL-72B demonstrates exceptional advancements across GUI grounding benchmarks. It achieves 87.1% accuracy on ScreenSpot, competing strongly with Gemini 2.0 (84.0%) and Claude (83.0%), while notably setting a new standard on ScreenSpot Pro with 43.6% accuracy - far surpassing both Aguvis-72B (23.6%) and its foundation Qwen2-VL-72B (1.6%). Leveraging these superior grounding capabilities, Qwen2.5-VL-72B significantly outperforms baselines across all offline evaluation benchmarks with a large gap. In online evaluation, some baselines have difficulty completing tasks due to limited grounding capabilities. Thus, we apply the Set-of-Mark (SoM) to the inputs of these models. The results show that Qwen2.5-VL-72B can outperform the baselines on AndroidWorld and MobileMiniWob++ and achieve comparable performance on OSWorld in online evaluation without auxiliary marks. This observation suggests that Qwen2.5-VL-72B is able to function as an agent in real and dynamic environments.

4 Conclusion

We present Qwen2.5-VL, a state-of-the-art vision-language model series that achieves significant advancements in multimodal understanding and interaction. With enhanced capabilities in visual recognition, object localization, document parsing, and long-video comprehension, Qwen2.5-VL excels in both static and dynamic tasks. Its native dynamic-resolution processing and absolute time encoding enable robust handling of diverse inputs, while Window Attention reduces computational overhead without sacrificing resolution fidelity. Qwen2.5-VL caters to a wide range of applications, from edge AI to high-performance computing. The flagship Qwen2.5-VL-72B matches or surpasses leading models like GPT-4o, and Claude 3.5 Sonnet, particularly in document and diagram understanding, while maintaining strong performance on pure text tasks. The smaller Qwen2.5-VL-7B and Qwen2.5-VL-3B variants outperform similarly sized competitors, offering efficiency and versatility. Qwen2.5-VL sets a new benchmark for vision-language models, demonstrating exceptional generalization and task execution across domains. Its innovations pave the way for more intelligent and interactive systems, bridging perception and real-world application.

5 Authors

Core Contributors: Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, Junyang Lin

Contributors[1]**:** An Yang, Binyuan Hui, Bowen Yu, Chen Cheng, Dayiheng Liu, Fan Hong, Fei Huang, Jiawei Liu, Jin Xu, Jianhong Tu, Jianyuan Zeng, Jie Zhang, Jinkai Wang, Jianwei Zhang, Jingren Zhou, Kexin Yang, Mei Li, Ming Yan, Na Ni, Rui Men, Songtao Jiang, Xiaodong Deng, Xiaoming Huang, Ximing Zhou, Xingzhang Ren, Yang Fan, Yichang Zhang, Yikai Zhu, Yuqiong Liu, Zhifang Guo

[1]Alphabetical order.

元素接地由 ScreenSpot（Cheng 等人，2024）和 ScreenSpot Pro（李等人，2025a）进行评估。离线评估在 Android Control（李等人，2024f）上进行，而在线评估则在包括 AndroidWorld（Rawles 等人，2024）、MobileMiniWob++ （Rawles 等人，2024）和 OSWorld（Xie 等人，2025）等平台进行。我们比较 Qwen2.5-VL-72B 与其他突出模型的性能，例如 GPT-4o（OpenAI，2024）、Gemini 2.0（Deepmind，2024）、Claude（Anthropic，2024b）、Aguvis-72B（Xu 等人，2024）和 Qwen2-VL-72B（Wang 等人，2024e）。结果展示在表 9 中。

表 9： **Qwen2.5-VL 和其他模型在 GUI Agent 基准测试上的性能表现。**

基准测试	GPT-4o	Gemini 2.0	Claude	Aguvis-72B	Qwen2-VL-72B	Qwen2.5-VL-72B
屏幕截图	18.1	84.0	83.0	**89.2**	-	87.1
屏幕截图 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**
安卓世界$_{SR}$	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 的性能在 GUI 基准测试中展现了卓越的进步。它在 ScreenSpot 上达到 87.1% 的准确率，与 Gemini 2.0 (84.0%) 和 Claude (83.0%) 竞争激烈，同时在 ScreenSpot Pro 上以 43.6% 的准确率创下了新标准——远超 Aguvis-72B (23.6%) 及其基础模型 Qwen2-VL-72B (1.6%)。凭借这些优越的 grounding 能力，Qwen2.5-VL-72B 在所有离线评估基准测试中均显著优于基线，差距巨大。在在线评估中，由于 grounding 能力有限，一些基线难以完成任务。因此，我们将 Set-of-Mark (SoM) 应用于这些模型的输入。结果显示，Qwen2.5-VL-72B 在 AndroidWorld 和 MobileMiniWob++ 上可以优于基线，并在在线评估中达到与 OSWorld 相当的性能，无需辅助标记。这一观察表明 Qwen2.5-VL-72B 能够在真实和动态环境中作为 agent 运行。

4 结论

我们介绍了 Qwen2.5-VL，这是一系列当前最先进的视觉语言模型，在多模态理解和交互方面取得了显著进步。凭借在视觉识别、物体定位、文档解析和长视频理解方面的增强能力，Qwen2.5-VL 在静态和动态任务中均表现出色。其原生动态分辨率处理和绝对时间编码能够稳健地处理多样化输入，而窗口注意力机制在降低计算开销的同时不牺牲分辨率保真度。Qwen2.5-VL 适用于从边缘人工智能到高性能计算的广泛应用。旗舰型号 Qwen2.5-VL-72B 与 GPT-4o、Claude 3.5 Sonnet 等领先模型相当或超越，尤其在文档和图表理解方面，同时在纯文本任务上保持强劲性能。较小的 Qwen2.5-VL-7B 和 Qwen2.5-VL-3B 变体在同等规模的竞争中表现更优，兼具效率和灵活性。Qwen2.5-VL 为视觉语言模型树立了新标杆，展示了跨领域的卓越泛化能力和任务执行能力。其创新成果为构建更智能、更交互的系统铺平道路，连接感知与现实应用。

5 作者

核心贡献者： 白帅, 陈克勤, 刘雪晶, 王嘉林, 郭文斌, 宋思博, 唐凯, 王鹏, 王时杰, 唐俊, 钟惠敏, 朱元智, 杨明坤, 李兆海, 万建强, 王鹏飞, 丁伟, 傅哲仁, 许毅恒, 叶嘉柏, 张希, 谢天宝, 成则森, 张航, 杨志博, 许海阳, 林俊阳

贡献者[1]**：** 杨安，胡斌元，余博文，陈成，刘大恒，黄帆，黄飞，刘佳伟，徐金，涂建红，曾建元，张杰，王金凯，张建伟，周景仁，杨克欣，李梅，闫明，倪娜，门瑞，蒋松涛，邓晓东，黄小明，周小明，任兴章，杨帆，张一畅，朱一凯，刘雨琼，郭志芳

[1]按字母顺序排列。

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

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

Anthropic. Claude 3.5 sonnet, 2024a. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. Introducing computer use, a new claude 3.5 sonnet, and claude 3.5 haiku, 2024b. URL https://www.anthropic.com/news/3-5-models-and-computer-use.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Trans. Software Eng.*, 49(7):3675–3691, 2023.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv preprint arXiv:2402.11684*, 2024a.

Jiacheng Chen, Tianhao Liang, Sherman Siu, Zhengqing Wang, Kai Wang, Yubo Wang, Yuansheng Ni, Wang Zhu, Ziyan Jiang, Bohan Lyu, et al. Mega-bench: Scaling multimodal evaluation to over 500 real-world tasks. *arXiv preprint arXiv:2410.10563*, 2024b.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv:2403.20330*, 2024c.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024d.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 933–941. PMLR, 2017.

Google Deepmind. Introducing gemini 2.0: our new ai model for the agentic era, 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

参考文献

Pravesh Agrawal，Szymon Antoniak，Emma Bou Hanna，Baptiste Bout，Devendra Chaplot，Jessica Chudnovsky，Diogo Costa，Baudouin De Monicault，Saurabh Garg，Theophile Gervet，等。Pixtral 12b。*arXiv* 预印本 *arXiv:2410.07073*，2024。

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, 等. Flamingo: 一种用于少样本学习的视觉语言模型. 在 *NeurIPS* 会议上, 2022.

Anthrop Anthropic. Claude 3.5 sonnet, 2024a. URLhttps://www.anthropic.com/news/claude-3-5-sonnet.

Anthropic. 推出计算机使用、新的 claude 3.5 sonnet 和 claude 3.5 haiku, 2024b. URLhttps://www.anthropic.com/news/3-5-models-and-computer-use.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, 和 Abhinav Jangda. MultiPL-E: 一种可扩展的多语言神经网络代码生成基准方法. *IEEE* 软件工程汇刊, 49(7):3675–3691, 2023.

陈贵明·哈迪·陈，陈顺年，张瑞飞，陈军英，吴祥波，张志毅，陈志宏，李建权，万翔，王本由。Allava：利用gpt4v合成数据进行轻量级视觉语言模型研究。*arXiv*预印本*arXiv:2402.11684*，2024a。

陈家成，梁天浩，苏尔曼，王正庆，王凯，王宇博，倪元生，朱王，蒋子岩，吕博文，等。Mega-bench：将多模态评估扩展到500多个真实世界任务。*arXiv*预印本*arXiv:2410.10563*，2024b。

陈林，李进松，董晓怡，张盘，臧宇航，陈泽辉，段浩东，王佳琪，乔宇，林大华，等。我们是否正在正确评估大视觉语言模型？*arXiv:2403.20330*，2024c。

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 评估在代码上训练的大型语言模型。 *CoRR*, abs/2107.03374, 2021。

陈哲，吴建南，王文海，苏伟杰，陈国，行森，钟沐言，张清龙，朱锡舟，陆磊伟，李斌，罗平，陆通，乔宇，以及戴继锋。Internvl：扩展视觉基础模型并针对通用视觉语言任务进行对齐。*arXiv*预印本*arXiv:2312.14238*，2023。

陈哲，王伟云，曹越，刘阳州，高张伟，崔尔飞，朱金国，叶胜龙，田浩，刘兆阳，等。通过模型、数据和测试时扩展扩展开源多模态模型的性能边界。*arXiv*预印本*arXiv:2412.05271*，2024d。

程侃之，孙启舒，褚友刚，徐方智，李言涛，张建兵，以及吴志勇。Seeclick：利用guigrounding为高级视觉gui代理赋能。*arXiv*预印本*arXiv:2401.10935*，2024。

Karl Cobbe，Vineet Kosaraju，Mohammad Bavarian，Mark Chen，Heewoo Jun，Lukasz Kaiser，Matthias Plappert，Jerry Tworek，Jacob Hilton，Reiichiro Nakano，Christopher Hesse，以及 John Schulman。训练验证者解决数学应用题。*CoRR*,abs/2110.14168,2021。

Yann N. Dauphin，Angela Fan，Michael Auli 和 David Grangier。使用门控卷积网络进行语言建模。在 *ICML*，机器学习研究论文集 的第 70 卷，第 933–941 页。PMLR，2017。

Google Deepmind。介绍gemini 2.0：我们为代理时代推出的新AI模型，2024年。URLhttps://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/。

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, and Wangding Zeng. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/ARXIV.2412.19437. URL https://doi.org/10.48550/arXiv.2412.19437.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv:2405.21075*, 2024a.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024b. URL https://arxiv.org/abs/2501.00321.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024c.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv:2304.14108*, 2023.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? *CoRR*, abs/2406.04127, 2024.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2918–2928, 2021.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models. *arXiv:2310.14566*, 2023.

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

DeepSeek-AI，刘爱欣，北风，薛冰，王冰轩，吴博文，陆成达，赵成刚，邓成奇，张晨宇，阮崇，戴大迈，郭大雅，杨德建，陈德立，季东杰，李尔航，林方云，戴福丛，罗福丽，郝光波，陈冠廷，李国伟，张汉，鲍汉，徐汉伟，王浩程，张浩伟，丁红会，辛华健，高华尊，李辉，李慧，蔡建良，梁建中，郭建忠，倪家奇，李家石，王继伟，陈金，陈景长，袁景阳，邱俊杰，李俊龙，宋俊晓，董凯，胡凯，高凯歌，关康，黄克欣，于快，王亮，张乐丛，徐雷，夏雷毅，赵亮，王立通，张立月，李梦，王妙军，张明川，张明华，张明辉，李明明，田宁，黄盼盼，王培毅，张鹏，王前程，朱启浩，陈启宇，杜启舒，陈启源，陈瑞，陈瑞泽，潘瑞哲，王润基，徐润欣，张若宇，陈若怡，李思思，陆尚豪，周尚岩，陈山黄，陈少庆，吴少庆，叶胜风，叶胜风，马世荣，王世宇，周双，余水萍，周顺峰，潘淑婷，王天，云涛，孙天宇，肖伟良，曾旺定。Deepseek-v3技术报告。*CoRR*，abs/2412.19437，2024年。doi: 10.48550/ARXIV.2412.19437。URLhttps://doi.org/10.48550/arXiv.2412.19437。

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, 等. Molmo 和 pixmo：用于当前最先进多模态模型的开放权重和开放数据. ***arXiv*** 预印本 ***arXiv:2409.17146***, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, 以及 Kai Chen. Mmbench-video：一个用于整体视频理解的长格式多镜头基准. ***arXiv*** 预印本 ***arXiv:2406.14515***, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, 等. Mme：一个用于多模态大型语言模型的综合评估基准. *arXiv:2306.13394*, 2023.

傅超友, 戴宇涵, 罗永东, 李雷, 任树会, 张仁睿, 王子涵, 周晨宇, 沈云航, 张梦丹, 等. Video-MME：首个视频分析中多模态大语言模型的全面评估基准. *arXiv:2405.21075*, 2024a.

付玲, 杨彪, 库壮斌, 宋嘉俊, 李宇哲, 朱令浩, 罗启迪, 王新宇, 陆浩, 黄明新, 李张, 唐国志, 山斌, 林春辉, 刘奇, 吴炳红, 冯浩, 刘浩, 黄灿, 唐景群, 陈伟, 金连文, 刘宇亮, 白翔. OCRBench v2：用于评估大型多模态模型在视觉文本定位和推理上的基准, 2024b. URLhttps://arxiv.org/abs/2501.00321.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In欧洲计算机视觉会议, pp. 148–166. Springer, 2024c.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, 等. Datacomp: 寻找下一代多模态数据集. *arXiv:2304.14108*, 2023.

Jiyang Gao，Chen Sun，Zhenheng Yang和Ram Nevatia。Tall：通过语言查询进行时序活动定位。在***IEEE***国际计算机视觉会议论文集，第5267-5275页，2017年。

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, 等. 我们完成 mmlu 吗？*CoRR*, abs/2406.04127, 2024.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, 和 Barret Zoph. 简单的复制粘贴是一种强大的实例分割数据增强方法。在***IEEE/CVF*** 计算机视觉与模式识别会议论文集，第 2918–2928 页, 2021年。

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, 和 Tianyi Zhou. Hallusionbench：一个用于大视觉语言模型中纠缠语言幻觉与视觉错觉的高级诊断套件。*arXiv:2310.14566*, 2023年。

郭嘉宇, 郑天宇, 白越林, 李波, 王宇博, 朱京, 李一知, Graham Neubig, 陈文虎, 和薛阳. Mammoth-vl: 大规模指令微调下多模态推理的生成. ***arXiv*** 预印本 ***arXiv:2412.05237***, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *ECCV*, 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023.

Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. Moai: Mixture of all intelligence for large language and vision models. In *European Conference on Computer Vision*, pp. 273–302. Springer, 2024.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model. *arXiv:2311.04219*, 2023a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024b.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024c.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023b.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use, 2025a. URL https://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf. Preprint.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024d.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

Qingyun Li, Zhe Chen, Weiyun Wang, Wenhai Wang, Shenglong Ye, Zhenjiang Jin, Guanzhou Chen, Yinan He, Zhangwei Gao, Erfei Cui, et al. Omnicorpus: An unified multimodal corpus of 10 billion-level images interleaved with text. *arXiv preprint arXiv:2406.08418*, 2024e.

Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the effects of data scale on computer control agents. *arXiv preprint arXiv:2406.03679*, 2024f.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 3(7), 2024g.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Da wn Song, 和 Jacob Steinhardt. 使用 MATH 数据集衡量数学问题解决能力. 在 ***NeurIPS*** 数据集与基准中, 2021. 胡凯瑞, 吴鹏浩, 浦繁艺, 王晓, 张远航, 薛阳, 李波, 和刘梓玮. Video- mmmu: 从多学科专业视频中评估知识获取能力. ***arXiv*** 预印本 ***arXiv:2501.13826***, 2025. Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, 和 Tamara Berg. Referitgame: 在自然场景照片中指代物体. 在 *EMNLP* 中, 2014. Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, 和 Ali Farhadi. 一图胜千言. 在 *ECCV* 中, 2016. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, 等. Segment anything. 在 *ICCV* 中, 2023. 李秉宽, 朴丙灿, 金彩温, 和罗永曼. Moai: 大型语言与视觉模型的混合智能. 在 欧洲计算机视觉会议 中, 第 273–302 页. Springer, 2024. 李波, 张培元, 杨景康, 张远航, 浦繁艺, 和刘梓玮. Otterhd: 高分辨率多模态模型. *arXiv:2311.04219*, 2023a. 李波, 张远航, 郭冬, 张仁瑞, 李峰, 张昊, 张凯辰, 张培元, 李岩伟, 刘梓玮, 等. Llava-onevision: 轻松实现视觉任务迁移. ***arXiv*** 预印本 ***arXiv:2408.03326***, 2024a. 李博文, 葛宇莹, 陈怡, 葛奕晓, 张瑞茂, 和山英. Seed-bench-2-plus: 使用富含文本的视觉理解基准测试多模态大语言模型. ***arXiv*** 预印本 ***arXiv:2404.16790***, 2024b.

董旭, 刘宇东, 吴昊宁, 王越, 沈志奇, 邱博文, 牛新瑶, 王国寅, 陈北, 李俊楠. Aria：一个开放的多模态原生专家混合模型. ***arXiv*** 预印本 ***arXiv:2410.05993***, 2024c.

李俊南, 李东旭, 熊才明, 和 Steven C. H. Hoi. Blip: 自举语言-图像预训练以实现统一的视觉-语言理解和生成. 收录于*ICML*, 2022a. 李俊南, 李东旭, Silvio Savarese, 和 Steven Hoi. Blip-2: 带冻结图像编码器和大型语言模型的自举语言-图像预训练. *arXiv:2301.12597*, 2023b. 李凯欣, 孟子洋, 林红赞, 罗子阳, 田宇晨, 马静, 黄志勇, 和 Tat-Seng Chua. Screenspot-pro: 专业高分辨率计算机使用的图形 grounding, 2025a. URLhttps://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf. 预印本. 李坤昌, 王亚丽, 何一男, 李奕卓, 王奕, 刘奕, 王尊, 许继兰, 陈国, 罗平, 等. Mvbench: 一个全面的多模态视频理解基准. 收录于*CVPR*, 2024d. 李伦年 Harold, 张鹏川, 张浩天, 杨建伟, 李春元, 中怡武, 王丽娟, 袁路, 张雷, 黄建宁, 等. Grounded 语言-图像预训练. 收录于***IEEE/CVF*** 计算机视觉与模式识别会议论文集, pp. 10965–10975, 2022b. 李庆云, 陈哲, 王伟云, 王文海, 叶胜龙, 金振江, 陈冠洲, 何一男, 高张伟, 崔尔飞, 等. Omni: 一个包含 10 亿级图像与文本交织的统一多模态语料库. *arXiv preprint arXiv:2406.08418*, 2024e. 李伟, William Bishop, 李爱丽丝, Chris Rawles, Campbell-Ajala Folawiyo, Tyamagundlu Divya, 和 Oriana Riva. 数据规模对计算机控制代理的影响. *arXivpreprintarXiv:2406.03679*, 2024f. 李亚东, 孙浩泽, 林明安, 李天鹏, 董国胜, 张涛, 丁 Bowen, 宋伟, 程正林, 胡宇琪, 等. Baichuan-omni 技术报告. *arXiv preprint arXiv:2410.08565*, 3(7), 2024g. 李亚东, 刘俊, 张涛, 陈宋, 李天鹏, 李泽桓, 刘丽君, 明灵峰, 董国胜, 潘达, 等. Baichuan-omni-1.5 技术报告. *arXiv preprint arXiv:2501.15368*,2025b.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *arXiv preprint arXiv:2405.11273*, 2024h.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv:2311.06607*, 2023c.

Yuxuan Liang, Xu Li, Xiaolei Chen, Haotian Chen, Yi Zheng, Chenghang Lai, Bin Li, and Xiangyang Xue. Global semantic-guided sub-image feature weight allocation in high-resolution large vision-language models. *arXiv preprint arXiv:2501.14276*, 2025.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv:2303.05499*, 2023c.

Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *Science China Information Sciences*, 67(12):1–16, 2024a.

Yuan Liu, Haodong Duan, Bo Li Yuanhan Zhang, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023d.

Yuan Liu, Zhongyin Zhao, Ziyuan Zhuang, Le Tian, Xiao Zhou, and Jie Zhou. Points: Improving your vision-language model with affordable strategies. *arXiv preprint arXiv:2409.04828*, 2024b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024c.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*, 2023e.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv:2203.10244*, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2582–2591, 2021a.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021b.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng

李云欣, 蒋申元, 胡宝天, 王龙越, 中钟魁, 罗文汉, 马林, 和 张敏. Uni-moe: 基于专家混合的统一多模态 LLM 扩展. *arXiv preprintarXiv:2405.11273*, 2024h. 张丽, 杨彪, 刘强, 马志寅, 张硕, 杨景旭, 孙亚博, 刘玉良, 和 白翔. Monkey: 大型多模态模型的图像分辨率和文本标签很重要.*arXiv:2311.06607*, 2023c. 梁宇轩, 李旭, 陈晓蕾, 陈浩天, 郑毅, 赖承航, 李斌, 和 薛向阳. 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重分配. *arXiv preprintarXiv:2501.14276*, 2025. 林继, 殷红旭, 平伟, Pavlo Molchanov, Mohammad Shoeybi, 和 韩松. Vila: 视觉语言模型的预训练研究. 收录于 ***IEEE/CVF*** 计算机视觉与模式识别会议论文集, pp. 26689–26699, 2024. 刘浩天, 李春元, 李宇恒, 和 Yong Jae Lee. 基于视觉指令微调的改进基线. *arXiv:2310.03744*, 2023a. 刘浩天, 李春元, 吴庆阳, 和 Yong Jae Lee. 视觉指令微调. *arXiv:2304.08485*, 2023b. 刘世龙, 曾昭阳, 任天鹤, 李峰, 张浩, 杨杰, 李春越, 杨建伟, 苏杭, 朱俊娟, 和 张雷. Grounding dino: 将 DINO 与 grounding 预训练结合以实现开放集目标检测. *arXiv:2303.05499*, 2023c. 刘阳州, 曹越, 高张伟, 王伟云, 陈哲, 王文海, 天天, 陆磊伟, 朱锡舟, 陆通, 等. Mminstruct: 一个具有广泛多样性的高质量 multimodal 指令微调数据集. 科学中国 信息科学, 67(12):1–16, 2024a. 刘元, 段好东, 李媛媛, 张张汉, 张宋阳, 赵王波, 袁一科, 王家齐, 何聪辉, 刘子伟, 和 陈凯. Mmbench: 你的多模态模型是全能选手吗? *arXiv:2307.06281*, 2023d. 刘元, 赵中银, 庄子源, 田乐, 周晓, 和 周杰. Points: 用经济策略提升你的视觉语言模型. *arXiv preprint arXiv:2409.04828*, 2024b. 刘云欣, 李世成, 刘毅, 王宇翔, 任树会, 李雷, 陈思硕, 孙旭, 和 侯路. Tempcompass: 视频 LLM 真的懂视频吗? *arXiv preprint arXiv:2403.00476*, 2024c. 刘玉良, 张丽, 黄明欣, 杨彪, 余文文, 李春元, 阴旭程, 刘成林, 金连文, 和 白翔. Ocrbench: 大型多模态模型中 OCR 的隐藏之谜.*arXiv:2305.07895*, 2023e. 陆磐, Bansal Hritik, Tony Xia, 刘家成, 李春元, Hannaneh Hajishirzi, Cheng Hao, Chang Kai-Wei, Galley Michel, 和 高建峰. Mathvista: 在视觉环境中评估基础模型在数学推理方面的能力 . 收录于 *ICLR*, 2024. Mangalam Karttikeya, Akshulakov Raiymbek, 和 Jitendra Malik. Egoschema: 一个用于超长视频语言理解的诊断基准. 收录于*NeurIPS*, 2023. 毛军华, Jonathan Huang, Alexander Toshev, Camburu Oana, Yuille Alan L, 和 Kevin Murphy. 不明确对象描述的生成与理解. 收录于*CVPR*, 2016. Ahmed Masry, 龙长隆, 谭家庆, Joty Shafiq, 和 Enamul Hoque. Chartqa: 关于图表的问答基准， 具有视觉和逻辑推理能力. *arXiv:2203.10244*, 2022. Mathew Minesh, Viraj Bagal, Rubèn Pérez Tito, Karatzas Dimosthenis, Valveny Ernest, 和 C.V. Jawahar. Infographicvqa. ***2022 IEEE/CVF*** 冬季计算机视觉应用会议 ***(WACV)***, pp. 2582–2591, 2021a. Mathew Minesh, Karatzas Dimosthenis, 和 CV Jawahar. Docvqa: 用于文档图像问答的数据集. 收录于*WACV*, 2021b. MiniMax, 李奥, 龚邦伟, 杨波, 阮博基, 刘畅, 朱成, 张春浩, 郭聪超, 陈达, 李东, 赵恩伟, 李庚新, 张国俊, 孙浩海, 董厚泽, 朱嘉戴, 朱嘉奇, 宋嘉源 , 朱金 , 韩景涛 , 李景阳 , 谢俊斌 , 许俊豪 , 徐俊豪 , 闫俊杰 , 张凯顺 , 小肖克诚 , 康克西 , 韩雷 , 王乐阳, 余连飞, 冯立恒, 郑林, 蔡林波, 邢龙, 蒋美芝, 张茂之, 黄培凯, 彭程

Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. Minimax-01: Scaling foundation models with lightning attention, 2025. URL https://arxiv.org/abs/2501.08313.

Openai. Chatml documents, 2024. URL https://github.com/openai/openai-python/blob/main/chatml.md.

OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations, 2024. URL https://arxiv.org/abs/2412.07626.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3170–3180, 2023.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2024.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, et al. Androidworld: A dynamic benchmarking environment for autonomous agents. *arXiv:2405.14573*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *CoRR*, abs/2311.12022, 2023.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, et al. Grounding dino 1.5: Advance the" edge" of open-set object detection. *arXiv preprint arXiv:2405.10300*, 2024.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, Yanjie Wang, Yuliang Liu, Hao Liu, Xiang Bai, and Can Huang. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv:2405.11985*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

牛，彭飞，李鹏宇，赵鹏宇，杨琪，徐启迪，王奇祥，王秦，李秋会，冷瑞涛，石胜民，余书琪，李思晨，朱松泉，黄涛，梁天润，孙伟高，孙伟轩，程伟宇，李文凯，宋祥军，苏晓，韩晓东，张新杰，侯新珠，徐鑫，邹勋，沈宇阳，龚岩，朱英杰，周奕鹏，钟一然，胡永毅，范元祥，余越，杨宇峰，李宇浩，黄云南，李云继，黄云鹏，徐云智，毛宇欣，李泽涵，李泽康，陶泽巍， Yingying Ying，陈振，陈振华，余志航，蒋卓，吴子嘉。MiniMax-01：闪电注意力下基础模型扩展，2025。URLhttps://arxiv.org/abs/2501.08313。

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

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

林凯阳，袁泉，周红斌，朱佳伟，张瑞，林群舒，王斌，赵志远，蒋曼，赵晓萌，石金，吴帆，褚培，刘明浩，李振祥，徐超，张博，石博文，涂中颖，何聪辉。OmniDocBench：对具有全面标注的多样化 PDF 文档解析进行基准测试，2024。URLhttps://arxiv.org/abs/2412.07626。

Roni Paiss，Ariel Ephrat，Omer Tov，Shiran Zada，Inbar Mosseri，Michal Irani，和 Tali Dekel。教 CLIP 数到十。在 ***IEEE/CVF*** 国际计算机视觉会议论文集，第 3170–3180 页，2023。

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, 等. 感知测试：多模态视频模型的诊断基准。在 NeurIPS，2024.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, 和 Furu Wei. Kosmos-2：将多模态大语言模型与世界关联起来。arXiv:2306.14824，2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, 和 Chelsea Finn. 直接偏好优化：你的语言模型本质上是一个奖励模型。在 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, 和 Sergey Levine (编), 神经信息处理系统进展 *36*：年度神经信息处理系统会议 *2023*，***NeurIPS 2023***，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，*2023*。*URL* http://papers.nips.cc/ paper_文件/论文/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html。

Christopher Rawles, Sarah Clinckemaillie, Yifan Chang, Jonathan Waltz, Gabrielle Lau, Marybeth Fair, Alice Li, William Bishop, Wei Li, Folawiyo Campbell-Ajala, 等. Androidworld：一个用于自主智能体的动态基准测试环境. *arXiv:2405.14573*, 2024.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, 和 Samuel R. Bowman. GPQA：一个研究生级别的 Google-免疫问答基准测试. *CoRR*, abs/2311.12022, 2023.

Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, 等. Grounding dino 1.5：推进开放集目标检测的“边缘”. ***arXiv*** 预印本 ***arXiv:2405.10300***, 2024.

Carlos Riquelme、Joan Puigcerver、Basil Mustafa、Maxim Neumann、Rodolphe Jenatton、André Su-sano Pinto、Daniel Keysers 和 Neil Houlsby。用稀疏专家混合模型扩展视觉能力。神经信息处理系统进展，34:8583–8595，2021年。

Amanpreet Singh、Vivek Natarajan、Meet Shah、Yu Jiang、Xinlei Chen、Dhruv Batra、Devi Parikh 和 Marcus Rohrbach。迈向能阅读的 VQA 模型。在*CVPR*，2019年。

Jianlin Su、Murtadha H. M. Ahmed、Yu Lu、Shengfeng Pan、Wen Bo 和 Yunfeng Liu。Roformer：带有旋转位置嵌入的增强型 Transformer。神经计算，568:127063，2024年。

Jingqun Tang、Qi Liu、YongjieYe、Jinghui Lu、Shu Wei、Chunhui Lin、Wanqing Li、Mohamad Fitri Faiz Bin Mahmood、Hao Feng、Zhen Zhao、Yanjie Wang、Yuliang Liu、Hao Liu、Xiang Bai 和 Can Huang。MTVQA：多语言文本中心视觉问答的基准测试。*arXiv:2405.11985*，2024年。

Gemini 团队，Rohan Anil，Sebastian Borgeaud，Yonghui Wu，Jean-Baptiste Alayrac，Jiahui Yu，Radu Soricut，Johan Schalkwyk，Andrew M Dai，Anja Hauth 等。Gemini：一个高度强大的多模态模型家族。***arXiv*** 预印本 ***arXiv:2312.11805***，2023。

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

Junyang Wang, Haiyang Xu, Haitao Jia, Xi Zhang, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent-v2: Mobile device operation assistant with effective navigation via multi-agent collaboration. *arXiv preprint arXiv:2406.01014*, 2024b.

Junyang Wang, Haiyang Xu, Jiabo Ye, Ming Yan, Weizhou Shen, Ji Zhang, Fei Huang, and Jitao Sang. Mobile-agent: Autonomous multi-modal mobile device agent with visual perception. *arXiv preprint arXiv:2401.16158*, 2024c.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024d.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024e.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024f.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024g.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024h.

Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14408–14419, 2023.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024i.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *CoRR*, abs/2406.01574, 2024j.

Zhenhailong Wang, Haiyang Xu, Junyang Wang, Xi Zhang, Ming Yan, Ji Zhang, Fei Huang, and Heng Ji. Mobile-agent-e: Self-evolving mobile assistant for complex tasks. *arXiv preprint arXiv:2501.11733*, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024k.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-free LLM benchmark. *CoRR*, abs/2406.19314, 2024.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024a. URL https://arxiv.org/abs/2407.15754.

Shengbang Tong，Ellis Brown，Penghao Wu，Sanghyun Woo，Manoj Middepogu，Sai Charitha Akula，Jihan Yang，Shusheng Yang，Adithya Iyer，Xichen Pan 等。Cambrian-1：一个完全开放、以视觉为中心的多模态 LLM 探索。*arXiv* 预印本 *arXiv:2406.16860*，2024。

王飞, 傅兴宇, 黄詹姆斯, 李泽坤, 刘秦, 刘晓耿, 马明宇 Derek, 许楠, 周文轩, 张凯, 等等. Muirbench: 一个用于鲁棒多图像理解的综合性基准. *arXiv* 预印本 *arXiv:2406.09411*, 2024a.

王俊阳, 许海阳, 贾海涛, 张希, 闫明, 沈伟周, 张继, 黄飞, 和桑继涛. Mobile-agent-v2: 通过多智能体协作实现有效导航的移动设备操作助手. *arXiv* 预印本 *arXiv:2406.01014*, 2024b.

王俊阳, 许海阳, 叶嘉宝, 闫明, 沈伟周, 张继, 黄飞, 和桑继涛. Mobile-agent: 具有视觉感知的自主多模态移动设备智能体. *arXiv* 预印本 *arXiv:2401.16158*, 2024c.

王克, 潘俊廷, 石伟康, 陆子木, 战明杰, 和李红生. 使用 math-vision 数据集测量多模态数学推理. *arXiv:2402.14804*, 2024d.

王鹏, 白帅, 谭思南, 王世杰, 范志浩, 白金泽, 陈克勤, 刘雪晶, 王嘉林, 葛文斌, 杨帆, 董凯, 杜梦飞, 任宣程, 门瑞, 刘大恒, 周畅, 周景仁, 和林俊阳. Qwen2-vl: 增强视觉语言模型在任何分辨率下对世界的感知. *arXiv:2409.12191*, 2024e.

王鹏, 白帅, 谭思南, 王世杰, 范志浩, 白金泽, 陈克勤, 刘雪晶, 王嘉林, 葛文斌, 等人. Qwen2-vl: 增强视觉语言模型在任何分辨率下对世界的感知. *arXiv preprint arXiv:2409.12191*, 2024f.

王伟汉，何泽海，洪文怡，程燕，张晓寒，齐吉，顾晓涛，黄诗宇，徐斌，董宇晓，等。LVBench：一个极端长视频理解基准。*arXiv*预印本 *arXiv:2406.08035*，2024年。

王伟云，任一鸣，罗浩文，李天童，闫晨祥，陈哲，王文海，李清云，陆磊伟，朱锡舟，等。全视项目 v2：迈向开放世界的通用关系理解。*arXiv*预印本*arXiv:2402.19474*, 2024h.

王文海，戴继锋，陈哲，黄振航，李志奇，朱锡舟，胡晓伟，陆通，陆磊伟，李红生，等. Internimage: 基于可变形卷积探索大规模视觉基础模型. In*IEEE/CVF*计算机视觉与模式识别会议论文集, pp. 14408–14419, 2023.

王新龙, 张晓松, 罗正雄, 孙权, 崔宇峰, 王金生, 张帆, 王越泽, 李振, 余启颖, 等. Emu3: 下一个词预测就够了. *arXiv* 预印本 *arXiv:2409.18869*, 2024i.

王宇博, 马学广, 张格, 牛元胜, Chandra Abhranil, 郭世光, 任伟明, Arulraj Aaran, 何轩, 姜子岩, 李天乐, Ku Max, 王凯, 朱翔, 樊荣奇, 岳翔, 和文虎. MMLU-Pro: 一个更鲁棒和具有挑战性的多任务语言理解基准. *CoRR*, abs/2406.01574, 2024j.

王振海龙, 许海阳, 王俊阳, 张希, 闫明, 张继, 黄飞, 和季恒. Mobile-agent-e: 用于复杂任务的自我进化移动助手. *arXiv* 预印本 *arXiv:2501.11733*, 2025.

王梓瑞, 夏梦舟, 何陆曦, 陈 Howard, 刘奕涛, 朱 Richard, 梁凯泉, 吴欣迪, 刘浩天, Malladi Sadhika, Chevalier Alexis, Arora Sanjeev, 以及 陈 Danqi. Charxiv: 多模态大语言模型在真实图表理解中的差距分析. *arXiv preprint arXiv:2406.18521*, 2024k.

魏建勋, 王学之, Schuurmans Dale, Bosma Maarten, Chi Ed H., Le Quoc, 以及 周Denny. 思维链提示激发大语言模型的推理能力. *CoRR*, abs/2201.11903, 2022.URL https://arxiv.org/abs/2201.11903.

Colin White, Samuel Dooley, Roberts Manley, Pal Arka, Feuer Benjamin, Jain Siddhartha, Shwartz-Ziv Ravid, Jain Neel, Saifullah Khalid, Naidu Siddartha, Hegde Chinmay, LeCun Yann, Goldstein Tom, Neiswanger Willie, 以及 Goldblum Micah. LiveBench: 一个具有挑战性、无污染的大语言模型基准.*CoRR*, abs/2406.19314, 2024.

吴昊宁，李东旭，陈北，和李俊楠。Longvideobench：一个用于长上下文视频语言理解的基准，2024a。URLhttps://arxiv.org/abs/2407.15754。

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024b.

X.AI. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks (2023). *URL https://arxiv. org/abs/2311.06242*, 2023.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Jing Hua Toh, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37: 52040–52094, 2025.

Yiheng Xu, Zekun Wang, Junli Wang, Dunjie Lu, Tianbao Xie, Amrita Saha, Doyen Sahoo, Tao Yu, and Caiming Xiong. Aguvis: Unified pure vision agents for autonomous gui interaction. *arXiv preprint arXiv:2412.04454*, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv:2412.15115*, 2024a.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy, 2024b. URL https://arxiv.org/abs/2412.02210.

Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv:2311.04257*, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv:2311.16502*, 2023.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS*, 2019.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv:2404.07973*, 2024a.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, et al. Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. *arXiv preprint arXiv:2412.09596*, 2024b.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024c.

Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024d.

Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. Vcr: Visual caption restoration. *arXiv:2406.06462*, 2024e.

吴智宇，陈晓康，潘子正，刘兴超，刘文，戴大迈，高华祖，马奕阳，吴成越，王冰璇等。Deepseek-vl2：用于高级多模态理解的专家混合视觉语言模型。*arXiv*预印本 *arXiv:2412.10302*，2024b。

X.AI. Grok-1.5 视觉预览。 https://x.ai/blog/grok-1.5v，2024。

Bin Xiao，Haiping Wu，Weijian Xu，Xiyang Dai，Houdong Hu，Yumao Lu，Michael Zeng，Ce Liu，以及 Lu Yuan。Florence-2：为多种视觉任务推进统一表示（2023）。*URLhttps://arxiv. org/abs/2311.06242*，2023。

Tianbao Xie，Danyang Zhang，Jixuan Chen，Xiaochuan Li，Siheng Zhao，Ruisheng Cao，Jing Hua Toh，Zhoujun Cheng，Dongchan Shin，Fangyu Lei，等。Osworld：在真实计算机环境中对开放式任务的多模态代理基准测试。神经信息处理系统进展，37：52040–52094，2025。

Yiheng Xu，Zekun Wang，Junli Wang，Dunjie Lu，Tianbao Xie，Amrita Saha，Doyen Sahoo，Tao Yu，以及 Caiming Xiong。Aguvis：用于自主GUI交互的纯视觉代理。*arXiv* 预印本 *arXiv:2412.04454*，2024。

杨安，杨保松，张北辰，胡斌元，郑波，余博文，李成元，刘大恒，黄飞，等. Qwen2.5技术报告. *arXiv:2412.15115*, 2024a.

杨志波，唐军，李兆海，王鹏飞，万建强，钟虎门，刘雪晶，杨明坤，王鹏，白帅，金连文，以及林俊阳. CC-OCR：一个全面且具有挑战性的OCR基准，用于评估大型多模态模型在识字方面的能力，2024b. URL https://arxiv.org/abs/2412.02210.

叶汉荣，黄德安，陆瑶，余志定，平伟，Andrew Tao，Jan Kautz，韩松，徐丹，Pavlo Molchanov，等. X-vila：大型语言模型的跨模态对齐. *arXivpreprintarXiv:2405.19335*, 2024.

叶清浩, 许海阳, 叶嘉宝, 闫明, 刘浩伟, 钱奇, 张继, 黄飞, 和周景仁. mplug-owl2: 通过模态协作革新多模态大语言模型。*arXiv:2311.04257*, 2023.

余伟豪, 杨正源, 李林杰, 王建峰, 林凯文, 刘子成, 王新超, 和王丽娟. Mm-vet: 评估集成能力的大型多模态模型。在 *ICML*, 2024.

岳翔, 倪元生, 张凯, 郑天宇, 刘若琪, 张格, Samuel Stevens, 姜东福, 任伟明, 孙宇轩, 等人. Mmmu: 一个面向专家级AGI的大规模多学科多模态理解和推理基准。*arXiv:2311.16502*, 2023.

向跃, 郑天宇, 倪元生, 王宇博, 张凯, 唐胜邦, 孙雨轩, 尹明, 余波涛, 张格, 等. Mmmu-pro: 一个更鲁棒的多学科多模态理解基准. *arXiv* 预印本 *arXiv:2409.02813*, 2024.

张标和 Rico Sennrich. 均方根层归一化. 在*NeurIPS*中, 2019.

张浩天, 尹浩轩, Philipp Dufter, 张博文, 陈晨, 陈宏宇, Tsu-Jui Fu, 王威廉, 张世福, 甘哲, 杨银飞. Ferret-v2: 一个用于指代和 grounding 的大型语言模型的改进基线. *arXiv:2404.07973*, 2024a.

张盘, 董晓怡, 曹宇航, 张宇航, 钱瑞, 魏锡林, 陈林, 李一飞, 牛俊波, 丁双瑞, 等. Internlm-xcomposer2.5-omnilive: 一个用于长期流式视频和音频交互的综合多模态系统. *arXiv* 预印本 *arXiv:2412.09596*, 2024b.

张仁睿, 姜东芝, 张奕驰, 林浩坤, 郭子宇, 邱鹏硕, 周澳君, 陆磐, 张凯伟, 乔宇, 等. Mathverse：你的多模态大语言模型真的能看懂视觉数学问题中的图表吗？在欧洲计算机视觉会议， 第169-186页. Springer, 2024c.

张涛, 李祥泰, 费浩, 袁浩波, 吴胜琼, 贾顺平, Loy Chen Change, 和严水成. Omg-llava：连接图像级、对象级、像素级推理和理解.*arXiv*预印本 *arXiv:2406.19389*, 2024d.

张天宇, 王素纯, 李路, 张格, Perouz Taslakian, Sai Rajeswar, 傅杰, 刘邦, 和 Yoshua Bengio. Vcr: 视觉标题恢复. *arXiv:2406.06462*, 2024e.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024f.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, Weifeng Pan, Ziyao Shangguan, Xiangru Tang, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. URL https://arxiv.org/abs/2501.12380.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024.

张一帆, 张焕宇, 田浩辰, 傅超宇, 张双庆, 吴俊飞, 李峰, 王坤, 温清松, 张张, 等. Mme-realworld: 你的多模态大语言模型能否挑战人类难以处理的高分辨率真实场景? *arXiv preprint arXiv:2408.13257*, 2024f.

赵一伦, 谢路景, 张浩伟, 甘果, 龙一涛, 胡志远, 胡通岩, 陈伟源, 李楚涵, 宋俊阳, 许志坚, 王成业, 潘伟峰, 尚子瑶 , 唐祥儒 , 梁振文 , 刘奕欣 , 赵晨 , 和阿曼 · 科汉 . MMVU: 测量专家级多学科视频理解 , 2025. URLhttps://arxiv.org/abs/2501.12380.

周建, 陆天健, 米什拉·斯瓦鲁普, 布拉马·西德哈拉, 巴苏·苏乔伊, 刘一澜, 周登, 和侯乐. 大型语言模型的指令跟随评估. *CoRR*, abs/2311.07911, 2023.

周俊杰, 舒岩, 赵波, 吴伯雅, 肖时涛, 杨希, 邢永平, 张波, 黄铁军, 和刘铮. MLVU: 多任务长视频理解的综合基准. *arXivpreprintarXiv:2406.04264*, 2024.