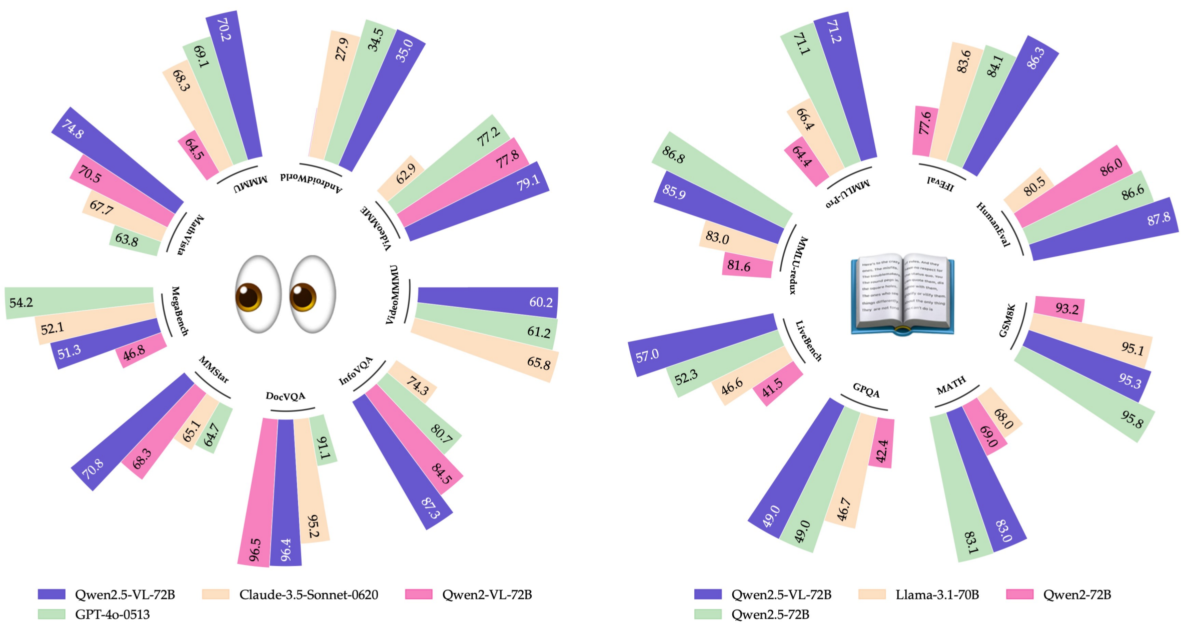
Qwen2.5-VL Technical Report

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.



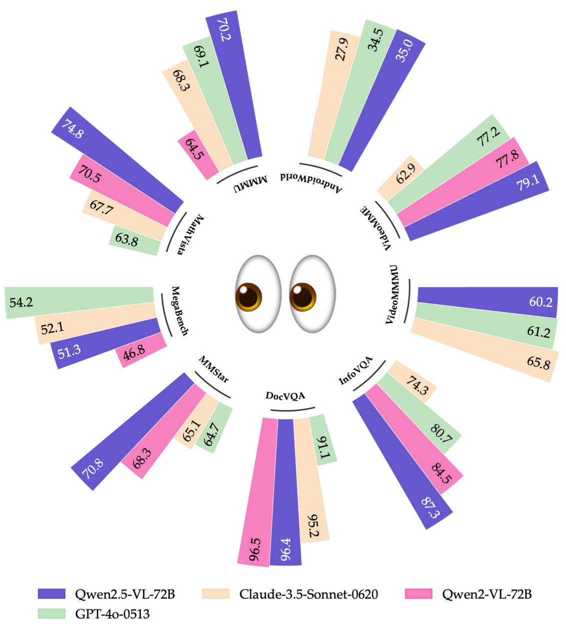
Qwen2.5-VL 技术报告

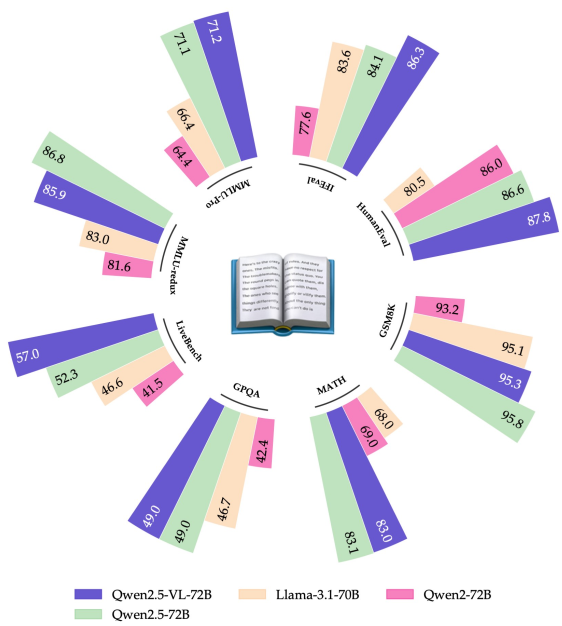
Qwen团队，阿里巴巴集团

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

摘要

我们介绍Qwen2.5-VL，这是Qwen视觉语言系列的最新旗舰模型，展示了基础能力和创新功能的显著进步。Qwen2.5-VL在理解和与世界互动方面取得了重大飞跃，通过增强的视觉识别、精确的物体定位、强大的文档解析和长视频理解。Qwen2.5-VL的一个突出特点是其能够准确地使用边界框或点来定位物体。它提供了从发票、表单和表格中提取强大的结构化数据，以及对图表、图示和布局的详细分析。为了处理复杂输入，Qwen2.5-VL引入了动态分辨率处理和绝对时间编码，使其能够处理不同大小的图像和延长时长（最长可达数小时）的视频，并实现秒级事件定位。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。通过从头开始训练一个原生动态分辨率视觉变换器（ViT）并结合窗口注意力，我们显著减少了计算开销，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，还作为一个互动视觉代理，能够在现实场景中进行推理、工具使用和任务执行，例如操作计算机和移动设备。该模型在各个领域实现了强大的泛化能力，无需特定任务的微调。Qwen2.5-VL提供三种尺寸，满足从边缘AI到高性能计算的多样化用例。旗舰Qwen2.5-VL-72B模型与最先进的模型如GPT-4o和Claude 3.5 Sonnet相匹配，特别是在文档和图示理解方面表现优异。较小的Qwen2.5-VL-7B和Qwen2.5-VL-3B模型在可比竞争对手中表现更佳，即使在资源受限的环境中也提供强大的能力。此外，Qwen2.5-VL保持了强大的语言性能，保留了Qwen2.5 LLM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 引言

大型视觉语言模型（LVLMs）（OpenAI，2024；Anthropic，2024a；Team等，2023；Wang等，2024f）代表了人工智能的一个关键突破，标志着对多模态理解和交互的变革性方法。通过无缝整合视觉感知与自然语言处理，这些先进模型从根本上重塑了机器如何解释和分析跨多个领域的复杂信息。尽管多模态大型语言模型取得了显著进展，但这些模型的当前能力可以比作三明治饼干的中间层——在各种任务中表现出色，但在卓越性能上仍显不足。细粒度视觉任务构成了这一类比的基础层。在这一版本的Qwen2.5-VL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造一个代理放大器。该框架的顶层是多模态推理，通过利用最新的Qwen2.5 LLM并采用多模态QA数据构建来增强。

一系列研究推动了多模态大型模型的发展，其特点是架构设计、视觉输入处理和数据策划。LVLMs进展的主要驱动力之一是架构的持续创新。文献中提出的研究（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）逐步塑造了当前的范式，该范式通常由视觉编码器、跨模态投影器和LLM组成。细粒度感知模型已成为另一个关键领域。像（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）这样的模型推动了详细视觉理解的可能性边界。Omni（Li et al., 2024g; 2025b; Ye et al., 2024）和MoE（Riquelme et al., 2021; Lee et al., 2024; Li et al., 2024h;c; Wu et al., 2024b）的架构也激励了LVLMs的未来演变。视觉编码器（Chen et al., 2023; Liu et al., 2024b; Liang et al., 2025）和分辨率缩放（Li et al., 2023c; Ye et al., 2023; Li et al., 2023a）的增强在提高实际视觉理解质量方面发挥了关键作用。策划更具多样化场景和更高质量的数据是训练先进LVLMs的重要步骤。文献中提出的努力（Guo et al., 2024; Chen et al., 2024d; Liu et al., 2024a; Chen et al., 2024a; Tong et al., 2024; Li et al., 2024a）对这一努力做出了极有价值的贡献。

然而，尽管它们取得了显著进展，视觉-语言模型目前仍面临发展瓶颈，包括计算复杂性、有限的上下文理解、较差的细粒度视觉感知以及在不同序列长度下表现不一致。

在本报告中，我们介绍了最新的工作 Qwen2.5-VL，它延续了 Qwen 系列的开源理念，在各种基准测试中实现并甚至超越了顶级闭源模型。从技术上讲，我们的贡献有四个方面：（1）我们在视觉编码器中实现了窗口注意力，以优化推理效率；（2）我们引入了动态 FPS 采样，将动态分辨率扩展到时间维度，使得在不同采样率下实现全面的视频理解；（3）我们通过对齐绝对时间在时间域中升级了 MRoPE，从而促进了更复杂的时间序列学习；（4）我们在策划高质量数据方面做出了重大努力，为预训练和监督微调提供支持，进一步将预训练语料库从 1.2 万亿个标记扩展到 4.1 万亿个标记。

Qwen2.5-VL的闪亮特性如下：

- 强大的文档解析能力：Qwen2.5-VL 将文本识别升级为全方位文档解析，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。
- 精确的对象定位跨格式：Qwen2.5-VL 解锁了在检测、指向和计数对象方面的更高准确性，支持绝对坐标和 JSON 格式，以便进行高级空间推理。
- 超长视频理解与细粒度视频定位：我们的模型将原生动态分辨率扩展到时间维度，增强了理解持续数小时的视频的能力，同时在几秒钟内提取事件片段。
- 增强的代理功能适用于计算机和移动设备：利用先进的基础知识、推理和决策能力，提升模型在智能手机和计算机上的卓越代理功能。

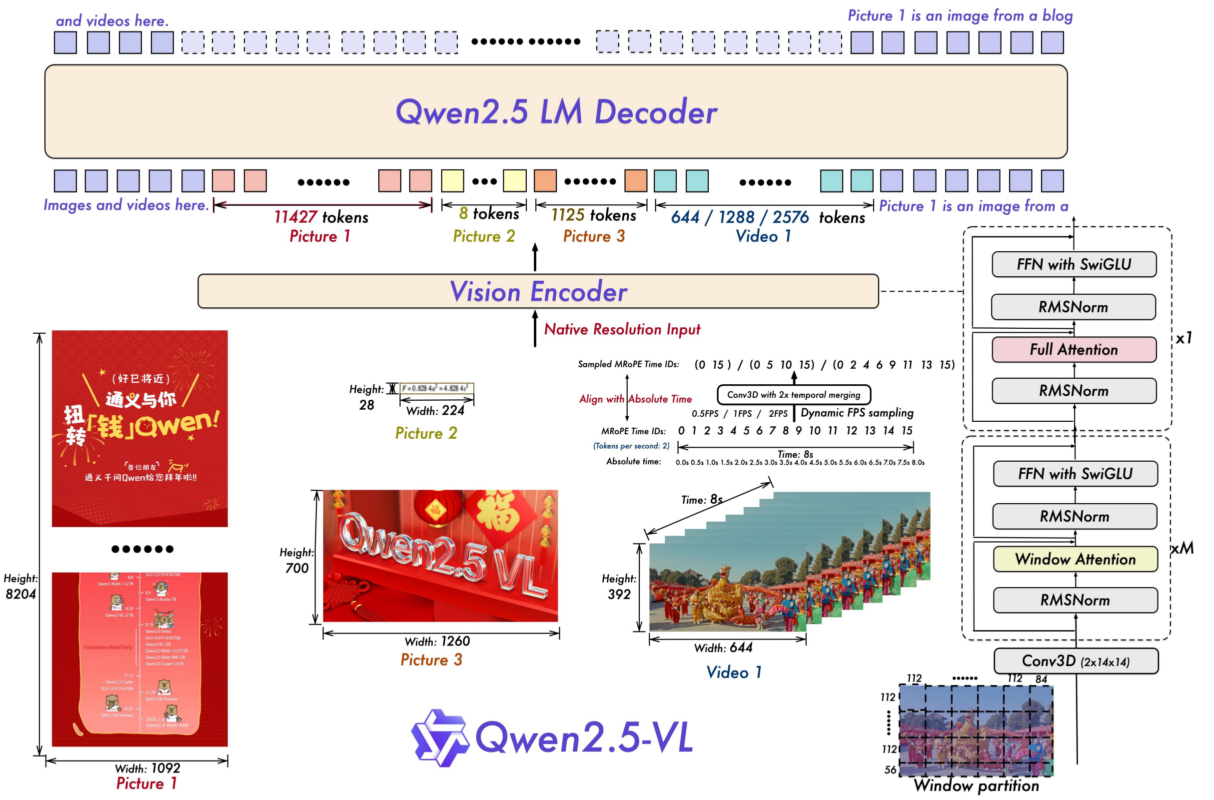


Figure 1: The Qwen2.5-VL framework demonstrates the integration of a vision encoder and a language model decoder to process multimodal inputs, including images and videos. The vision encoder is designed to handle inputs at their native resolution and supports dynamic FPS sampling. Images of varying sizes 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

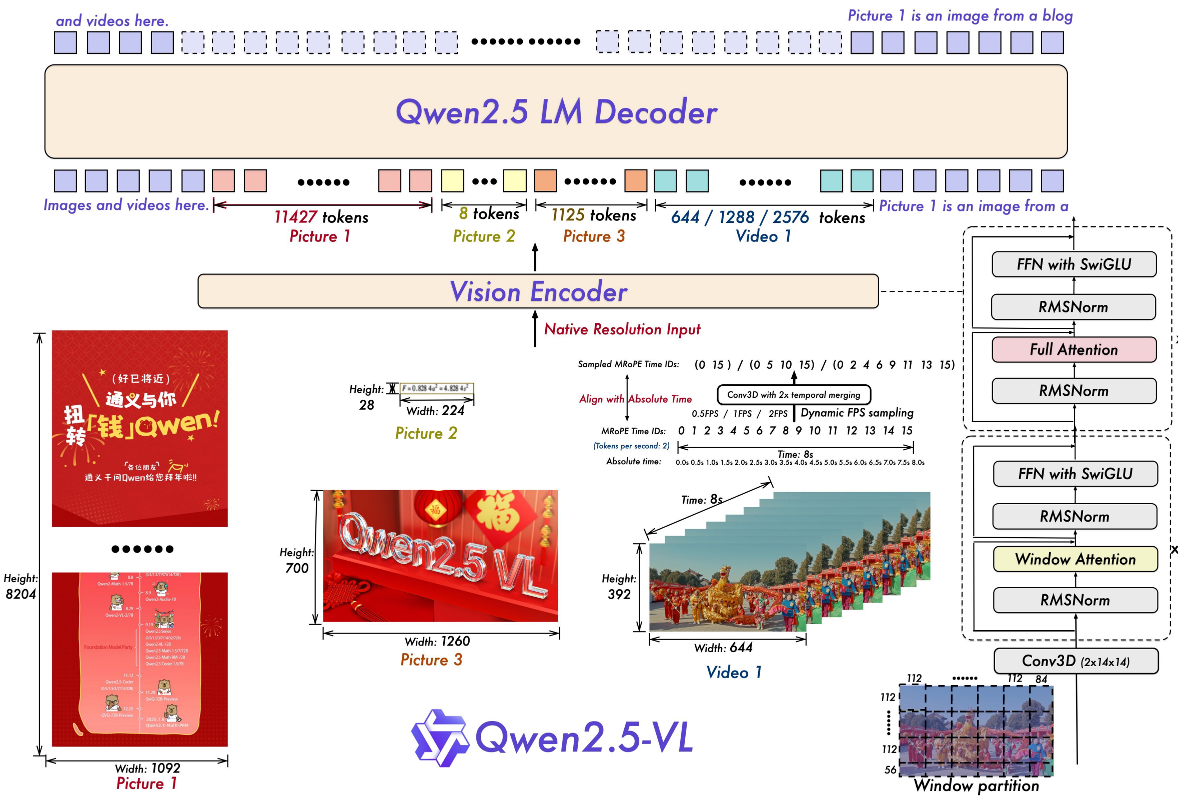


图1：Qwen2.5-VL框架展示了视觉编码器和语言模型解码器的集成，以处理多模态输入，包括图像和视频。视觉编码器旨在处理原生分辨率的输入，并支持动态FPS采样。不同大小的图像和具有不同FPS速率的视频帧被动态映射到不同长度的令牌序列。值得注意的是，MRoPE将时间ID与时间维度上的绝对时间对齐，使模型能够更好地理解时间动态，例如事件的节奏和精确的时刻定位。处理后的视觉数据随后被输入到Qwen2.5 LM解码器中。我们重新设计了视觉变换器（ViT）架构，结合了先进的组件，如带有SwiGLU激活的FFN、用于归一化的RMSNorm和基于窗口的注意力机制，以提高性能和效率。

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述。

2.1 模型架构

Qwen2.5-VL的整体模型架构由三个组件组成：

大型语言模型：Qwen2.5-VL系列采用大型语言模型作为其基础组件。该模型使用来自Qwen2.5 LLM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，我们对1D RoPE（旋转位置嵌入）进行了修改，采用了与绝对时间对齐的多模态旋转位置嵌入。

视觉编码器：Qwen2.5-VL 的视觉编码器采用了重新设计的视觉变换器（ViT）架构。在结构上，我们结合了 2D-RoPE 和窗口注意力，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在输入到 ViT 之前被调整为 28 的倍数。视觉编码器通过以 14 的步幅将图像拆分为补丁来处理图像，从而生成一组图像特征。我们在第 2.1.1 节中对视觉编码器进行了更详细的介绍。

基于MLP的视觉-语言合并：为了应对长序列图像特征带来的效率挑战，我们采用了一种简单而有效的方法，在将特征序列输入大型语言模型（LLM）之前对其进行压缩。具体而言，我们并不是直接使用原始补丁

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

通过视觉变换器（ViT）提取的特征，我们首先将空间上相邻的四个补丁特征进行分组。这些分组特征随后被连接并通过一个两层的多层感知器（MLP）传递，以将它们投影到与LLM中使用的文本嵌入对齐的维度。该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。

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

表1：Qwen2.5-VL的配置。

2.1.1 快速高效的视觉编码器

视觉编码器在多模态大型语言模型（MLLMs）中发挥着关键作用。为了应对由于原生分辨率输入而导致的训练和推理过程中计算负载不平衡所带来的挑战，我们重新设计了视觉变换器（ViT）架构。一个关键问题是处理不同大小图像时所涉及的二次计算复杂度。为了解决这个问题，我们在大多数层中引入了窗口注意力，这确保了计算成本与补丁数量呈线性关系，而不是二次关系。在我们的架构中，只有四层采用全自注意力，而其余层则利用最大窗口大小为112×112（对应于8×8补丁）的窗口注意力。小于112×112的区域在处理时不进行填充，保持其原始分辨率。这个设计使得模型能够在输入分辨率下原生运行，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，我们将方法扩展到三维补丁分区。具体而言，我们使用14×14图像补丁作为基本单元，这与传统的静态图像ViTs一致。对于视频数据，两个连续帧被组合在一起，显著减少了输入语言模型的标记数量。这一设计不仅保持了与现有架构的兼容性，还提高了处理序列视频数据的效率。

为了简化整体网络结构，我们将ViT架构与大型语言模型（LLMs）的设计原则更紧密地对齐。具体而言，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。这些选择增强了模型视觉和语言组件之间的计算效率和兼容性。

在训练方面，我们从头开始训练重新设计的ViT。训练过程包括几个阶段，涵盖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

训练。图像根据其原始纵横比随机采样，使模型能够有效地对不同分辨率的输入进行泛化。这种方法不仅提高了模型的适应性，还确保了在不同大小的视觉数据上进行稳定和高效的训练。

2.1.2 原生动态分辨率和帧率

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

在空间域中，Qwen2.5-VL 动态地将不同大小的图像转换为具有相应长度的令牌序列。与传统方法通过归一化坐标不同，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。这使得模型能够固有地学习尺度信息，提高其处理不同分辨率图像的能力。

对于视频输入，Qwen2.5-VL 结合了动态帧率 (FPS) 训练和绝对时间编码。通过适应可变帧率，该模型能够更好地捕捉视频内容的时间动态。与其他采用文本时间戳或利用额外头部来实现时间定位的方法不同，我们引入了一种新颖且高效的策略，直接将 MRoPE ID 与时间戳对齐。这种方法使模型能够通过时间维度 ID 之间的间隔理解时间的节奏，而无需任何额外的计算开销。

2.1.3 与绝对时间对齐的多模态旋转位置嵌入

位置嵌入对于在视觉和语言模态中建模序列数据至关重要。在Qwen2-VL中引入的多模态旋转位置嵌入（MRoPE）的基础上，我们扩展了其功能，以更好地处理视频中的时间信息。

Qwen2-VL中的MRoPE将位置嵌入分解为三个不同的组件：时间、身高和宽度，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置ID，使得MRoPE在功能上等同于传统的1D RoPE（Su等，2024）。对于图像，时间ID在视觉标记之间保持不变，而高度和宽度组件则根据每个标记在图像中的空间位置分配唯一的ID。在处理视频时，视频被视为帧的序列，时间ID在每帧之间递增，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中的时间位置ID与输入帧的数量相关，这并未考虑内容变化的速度或视频中事件的绝对时间。为了解决这一局限性，Qwen2.5-VL引入了一个关键改进：将MRoPE的时间组件与绝对时间对齐。如图1所示，通过利用时间ID之间的间隔，模型能够学习在不同FPS采样率的视频之间保持一致的时间对齐。

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，我们显著扩大了预训练数据的规模，将其从1.2万亿个标记增加到大约4万亿个标记。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。该数据集涵盖了各种多模态数据，例如图像标题、交错的图像-文本数据、光学字符识别（OCR）数据、视觉知识（例如，名人、地标、植物和动物识别）、多模态学术问题、定位数据、文档解析数据、视频描述、视频定位和基于代理的交互数据。在整个训练过程中，我们仔细调整了不同阶段这些数据类型的组成和比例，以优化学习效果。

交错的图像-文本数据 交错的图像-文本数据对于多模态学习至关重要，提供了三个关键好处：(1) 通过同时提供视觉和文本提示来实现上下文学习（Alayrac et al., 2022），(2) 在缺少图像时保持强大的仅文本能力（Lin et al., 2024），以及 (3) 包含广泛的通用信息。然而，现有的交错数据大部分

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，且通常噪声较大，限制了其在复杂推理和创造性生成中的实用性。

为了解决这些挑战，我们开发了一条评分和清理数据的流程，确保仅使用高质量、相关的交错数据。我们的过程包括两个步骤：标准数据清理（Li et al., 2024e），然后是使用内部评估模型的四阶段评分系统。评分标准包括：（1）仅文本质量，（2）图像-文本相关性，（3）图像-文本互补性，以及（4）信息密度平衡。这种细致的方法提高了模型进行复杂推理和生成连贯多模态内容的能力。

以下是这些图文评分标准的描述：

图像-文本相关性：更高的分数表示图像与文本之间的联系更强，其中图像有意义地补充、解释或扩展文本，而不仅仅是装饰它。

信息互补性：更高的分数反映了图像和文本之间更大的互补信息。每个部分应提供独特的细节，共同构建一个完整的叙述。

信息密度平衡：更高的分数意味着图像和文本之间信息的分布更加平衡，避免过多的文本或图像信息，并确保两者之间的适当平衡。

使用绝对位置坐标进行数据基础化 我们采用原生分辨率训练，旨在实现对世界的更准确感知。相比之下，相对坐标无法有效表示图像中物体的原始大小和位置。为了解决这一局限性，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真实世界尺度和空间关系，从而在物体检测和定位等任务中提高性能。

为了提高基础能力的泛化能力，我们开发了一个综合数据集，涵盖了带有指代表达的边界框和点，利用了公开可用的数据集和专有数据。我们的方法涉及将数据合成到各种格式中，包括 XML、JSON 和自定义格式，采用诸如复制粘贴增强（Ghiasi 等，2021）和与现成模型（如 Grounding DINO（Liu 等，2023c）和 SAM（Kirillov 等，2023））的合成等技术。这种方法促进了基础能力的更强评估和提升。

为了提高模型在开放词汇检测上的性能，我们扩展了训练数据集，包含超过10,000个物体类别。此外，为了提高模型在极端物体检测场景中的有效性，我们在查询中合成了不存在的物体类别，并构建了包含每个物体多个实例的图像数据。

为了确保卓越的基于点的物体定位能力，我们构建了一个全面的指向数据集，包括公开可用的数据和合成数据。具体而言，数据源包括来自PixMo（Deitke等，2024）的公共指向和计数数据、公开可获取的物体定位数据（来自物体检测和实例分割任务），以及通过自动化管道合成的针对特定图像细节的精确指向数据。

文档全解析数据 为了训练Qwen2.5-VL，我们合成了大量的文档数据语料库。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能力。具体而言，我们在文档中融入了多种元素，如表格、图表、方程式、自然或合成图像、乐谱和化学公式。这些元素统一采用HTML格式，整合了布局框信息和插图描述到HTML标签结构中。我们还根据典型的阅读顺序丰富了文档布局，并在基于HTML的真实数据中包含了每个模块（如段落和图表）对应的坐标。这种创新的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，都能够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，从而促进了更高效和准确的文档理解与转换。

Below is the QwenVL HTML format:

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

```
<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}"> 表格内容 </table> # 图表 <div class="图表" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 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> # 图片标题 <div class="图片标题" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> 图片标题 </p></div> # 图片OCR <div class="图片OCR" data-bbox="x1 y1 x2 y2"> <img data-bbox="x1 y1 x2 y2" /><p> 图片OCR </p></div> # 乐谱 <div class="乐谱" format="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" format="微笑" 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，从而使Qwen2.5-VL能够高效处理和理解。

OCR 数据 来自不同来源的数据被收集和整理，以增强 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，以在自然环境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，我们纳入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，利用高质量的合成图像和真实世界的自然场景图像。这种组合确保了在各种语言环境中的强大性能，并提高了模型对不同文本外观和环境条件的适应能力。对于图表类型的数据，我们使用包括 matplotlib、seaborn 和 plotly 在内的可视化库合成了 100 万个样本，涵盖了条形图、关系图和热图等图表类别。关于表格数据，我们通过离线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 为了确保在理解具有不同帧率（FPS）的视频数据时增强鲁棒性，我们在训练过程中动态采样FPS，以实现训练数据集中FPS的更均匀分布。此外，对于超过半小时的视频，我们通过针对性的合成管道合成多帧字幕，专门构建了一组长视频字幕。关于视频定位数据，我们以基于秒的格式和小时-分钟-秒-帧（hmsf）格式制定了时间戳，确保模型能够准确理解和输出各种格式的时间。

代理数据 我们增强感知和决策能力，以构建 Qwen2.5-VL 的代理能力。对于感知，我们在移动、网络和桌面平台上收集截图。使用合成数据引擎生成截图标题和 UI 元素定位注释。标题任务帮助 Qwen2.5-VL 理解图形界面，而定位任务则帮助它对齐元素的外观和功能。对于决策，我们首先将移动、网络和桌面平台的操作统一为具有共享动作空间的函数调用格式。从开源数据中收集的一组带注释的多步骤轨迹，通过代理框架（Wang et al., 2025; 2024b;c）在虚拟环境中合成，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.

每个步骤的推理过程通过人类和模型注释者进行（Xu et al., 2024）。具体而言，给定一个真实操作，我们在截图上突出显示它。然后，我们向注释者提供全局查询，以及该操作前后的截图，并要求他们撰写推理内容以解释该操作背后的意图。使用基于模型的过滤器来筛选低质量的推理内容。这种推理内容防止Qwen2.5-VL对真实操作的过拟合，使其在现实场景中更加稳健。

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

表2：不同阶段的训练数据量和组成。

2.2.2 训练方案

我们从头开始训练了一个视觉变换器（ViT），使用了DataComp（Gadre等，2023）和一些内部数据集作为视觉编码器的初始化，同时利用预训练的Qwen2.5大型语言模型（LLM）（Yang等，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的阶段，每个阶段采用不同的数据配置和训练策略，以逐步增强模型的能力。

在第一阶段，仅训练视觉变换器（ViT），以提高其与语言模型的对齐，为多模态理解奠定坚实基础。在此阶段的主要数据来源包括图像标题、视觉知识和OCR数据。这些数据集经过精心挑选，以促进ViT提取有意义的视觉表示的能力，这些表示可以有效地与文本信息集成。

在第二阶段，所有模型参数被解冻，模型在多样化的多模态图像数据集上进行训练，以增强其处理复杂视觉信息的能力。此阶段引入了更复杂和需要推理的数据库，例如交错数据、多任务学习数据集、视觉问答（VQA）、多模态数学、基于代理的任务、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次联系的能力，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更长序列上的推理能力，视频和基于代理的数据被纳入，同时增加了序列长度。这使得模型能够以更高的精度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特别有利。

为了应对不同图像大小和文本长度带来的挑战，这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于LLM和视觉编码器。考虑到视觉编码器的参数相对较少，并且我们引入了窗口注意力以进一步降低其计算需求，我们专注于在不同GPU之间平衡LLM的计算负载。具体而言，我们根据输入序列长度动态打包数据样本，以确保计算负载的一致性。在第一和第二阶段，数据被均匀打包到序列长度为8,192，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 训练后

Qwen2.5-VL的后训练对齐框架采用了双阶段优化范式，包括监督微调（SFT）和直接偏好优化（DPO）（Rafailov等，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了表征基础和行为精炼的问题。

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）旨在通过针对性的指令优化，弥合预训练表示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）来构建指令跟随数据，故意偏离预训练数据模式，同时保持与Qwen2-VL（Wang et al.，2024e）的架构一致性。这种格式转换使得三项关键适应成为可能：1）针对多模态轮流对话的明确对话角色标记，2）在文本指令旁边结构化注入视觉嵌入，以及3）通过格式感知打包保持跨模态位置关系。通过在这种增强模式下向模型暴露精心策划的多模态指令-响应对，SFT实现了高效的知识转移，同时保持了预训练特征的完整性。

2.3.1 指令数据

监督微调（SFT）阶段采用了一个精心策划的数据集，旨在增强模型在多种模态下的指令跟随能力。该数据集包含大约200万个条目，纯文本数据（50%）和多模态数据（50%）均匀分布，其中包括图像-文本和视频-文本组合。多模态数据的纳入使模型能够有效处理复杂输入。值得注意的是，尽管纯文本和多模态条目数量相等，但由于嵌入的视觉和时间信息，多模态条目在训练过程中消耗的令牌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，并附有多语言条目以支持更广泛的语言多样性。

数据集的结构反映了不同层次的对话复杂性，包括单轮和多轮互动。这些互动通过从单图像输入到多图像序列的场景进一步进行上下文化，从而模拟现实的对话动态。查询来源主要来自开源库，此外还包括经过策划的购买数据集和在线查询数据的贡献。这种组合确保了广泛的覆盖面，并增强了数据集的代表性。

为了应对广泛的应用场景，数据集包括针对一般视觉问答（VQA）、图像描述、数学问题解决、编码任务和安全相关查询的专业子集。此外，还构建了专门用于文档和光学字符识别（Doc 和 OCR）、基础定位、视频分析和代理交互的数据集，以增强特定领域的专业能力。有关数据的详细信息可以在论文的相关部分找到。这种结构化和多样化的组成确保了 SFT 阶段有效地将预训练表示与下游多模态任务的细微需求对齐，从而促进模型性能的稳健性和上下文意识。

2.3.2 数据过滤管道

训练数据的质量是影响视觉语言模型性能的关键因素。开源和合成数据集通常表现出显著的变异性，往往包含噪声、冗余或低质量的样本。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。低质量数据可能导致预训练表示与下游任务需求之间的次优对齐，从而降低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和可靠的模型性能至关重要。

为了应对这些挑战，我们实施了一个两阶段的数据过滤流程，旨在系统地提高监督微调（SFT）数据集的质量。该流程包括以下阶段:

阶段 1：特定领域分类 在初始阶段，我们采用 *Qwen2-VL-Instag*，一个源自 Qwen2-VL-72B 的专业分类模型，对问答 (QA) 对进行层次分类。该模型将 QA 对组织为八个主要领域，例如 *Coding* 和 *Planning*，这些领域进一步细分为 30 个细粒度子类别。例如，主要领域 *Coding* 被细分为包括 *Code_Debugging*、*Code_Generation*、*Code_Translation* 和 *Code_Understanding* 的子类别。这种层次结构促进了领域感知和子领域感知的过滤策略，使得管道能够优化针对每个类别特定特征的数据清理过程。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，它结合了基于规则和基于模型的方法，以全面提高数据质量。鉴于 {v*} 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

文档处理、光学字符识别（OCR）和视觉定位等领域的多样性特征可能需要独特的过滤策略。下面，我们提供了这些领域中应用的一般过滤策略的概述。

基于规则的过滤使用预定义的启发式方法来消除低质量或有问题的条目。具体而言，对于与文档处理、OCR 和视觉定位任务相关的数据集，识别并删除重复模式，以防止扭曲模型的学习过程并确保最佳性能。此外，包含不完整、截断或格式不正确的响应的条目——在合成数据集和多模态上下文中常见——也会被排除。为了保持相关性并维护伦理标准，与任务无关或可能导致有害输出的查询和答案也会被丢弃。这种结构化的方法确保数据集遵循伦理准则并满足特定任务的要求。

基于模型的过滤进一步通过利用在Qwen2.5-VL系列上训练的奖励模型来精炼数据集。这些模型评估跨多个维度的多模态问答对。查询根据复杂性和相关性进行评估，仅保留那些适当具有挑战性和上下文相关的示例。答案根据正确性、完整性、清晰度、与查询的相关性和帮助程度进行评估。在视觉基础任务中，特别关注验证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进入SFT阶段。

2.3.3 增强推理的拒绝采样

为了补充我们的结构化数据过滤管道，我们采用拒绝采样作为一种策略，以精炼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，例如数学问题解决、代码生成和特定领域的视觉问答（VQA）。先前的研究表明，结合链式思维（CoT）Wei et al.（2022）推理显著提高了模型的推理性能。（DeepSeek-AI et al.，2024）我们的后训练实验证实了这一点，强调了结构化推理过程在实现高质量结果中的重要性。

拒绝采样过程始于包含真实标签的丰富数据集。这些数据集经过精心策划，以包括需要多步骤推理的任务，例如数学问题解决、代码生成和特定领域的视觉问答（VQA）。使用Qwen2.5-VL模型的中间版本，我们将生成的响应与真实标签进行评估。只有模型输出与预期答案匹配的样本被保留，从而确保数据集仅由高质量、准确的示例组成。

为了进一步提高数据质量，我们应用额外的约束来过滤不良输出。具体来说，我们排除表现出代码切换、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，这对下游应用至关重要。

在将链式推理应用于视觉语言模型时，一个关键挑战是它们对文本和视觉模态的依赖。中间推理步骤可能未能充分整合视觉信息，要么忽略相关的视觉线索，要么误解它们。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，以验证中间推理步骤的准确性。这些机制确保了链式推理过程中的每一步有效整合视觉和文本模态。尽管做出了这些努力，实现最佳模态对齐仍然是一个持续的挑战，需要进一步的进展。

通过拒绝采样生成的数据显著增强了模型的推理能力。通过迭代地优化数据集并去除低质量或错误的样本，我们使模型能够从强调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，还为未来在视觉-语言建模方面的改进奠定了基础。

2.3.4 训练食谱

Qwen2.5-VL的后训练过程分为两个阶段：监督微调（SFT）和直接偏好优化（DPO），在这两个阶段中，视觉变换器（ViT）的参数保持不变。在SFT阶段，模型在多样的多模态数据上进行微调，包括图像-文本对、视频和纯文本，这些数据来源于一般的视觉问答（VQA）、拒绝采样以及专门的数据集，如文档和光学字符识别（OCR）、定位、视频和与代理相关的任务。DPO阶段专注于图像-文本和纯文本数据，利用偏好数据使模型与人类偏好对齐，每个样本仅处理一次，以确保优化的高效性。这个简化的过程增强了模型的

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
$\text{MMMU-Pro}_{\text{overall}}$ (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								
$\text{MathVista}_{\text{mini}}$ (Lu et al., 2024)	72.3 Chen et al. (2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$\text{MATH-Vision}_{\text{full}}$ (Wang et al., 2024d)	32.2 Chen et al. (2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$\text{MathVerse}_{\text{mini}}$ (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
$\text{MMBench-EN}_{\text{test}}$ (Liu et al., 2023d)	88.3 Chen et al. (2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$\text{MMBench-CN}_{\text{test}}$ (Liu et al., 2023d)	88.5 Chen et al. (2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$\text{MMBench-V1.1-EN}_{\text{test}}$ (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
$\text{BLINK}_{\text{val}}$ (Fu et al., 2024c)	63.8 Chen et al. (2024d)	-	**68.0**	63.8	-	64.4	56.4	47.6
$\text{CRPE}_{\text{relation}}$ (Wang et al., 2024h)	78.8 Chen et al. (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$\text{HallBench}_{\text{avg}}$ (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
$\text{RealWorldQA}_{\text{avg}}$ (X.AI, 2024)	78.7 Chen et al. (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$\text{MME-RealWorld}_{\text{en}}$ (Zhang et al., 2024f)	62.9 Chen et al. (2024d)	51.6	45.2	62.9	-	**63.2**	57.4	53.1
$\text{MMVet}_{\text{turbo}}$ (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.

跨模态推理和任务特定性能，同时保持与用户意图的一致性。

3 个实验

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

3.1 与最先进模型的比较

表3：Qwen2.5-VL和最先进技术的性能。

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
$\mathrm{MMMU\text{-}Pro}_{overall}$ (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								
$\mathrm{MathVista}_{mini}$ (Lu et al., 2024)	72.3 Chen et al. (2024d)	67.7	63.8	72.3	70.5	**74.8**	68.2	62.3
$\mathrm{MATH\text{-}Vision}_{full}$ (Wang et al., 2024d)	32.2 Chen et al. (2024d)	-	30.4	32.2	25.9	**38.1**	25.1	21.2
$\mathrm{MathVerse}_{mini}$ (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
$\mathrm{MMBench\text{-}EN}_{test}$ (Liu et al., 2023d)	88.3 Chen et al. (2024d)	82.6	83.4	88.3	86.9	**88.6**	83.5	79.1
$\mathrm{MMBench\text{-}CN}_{test}$ (Liu et al., 2023d)	88.5 Chen et al. (2024d)	83.5	82.1	**88.5**	86.7	87.9	83.4	78.1
$\mathrm{MMBench\text{-}V1.1\text{-}EN}_{test}$ (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
$\mathrm{CRPE}_{relation}$ (Wang et al., 2024h)	78.8 Chen et al. (2024d)	-	76.6	78.8	-	**79.2**	76.4	73.6
$\mathrm{HallBench}_{avg}$ (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
$\mathrm{RealWorldQA}_{avg}$ (X.AI, 2024)	78.7 Chen et al. (2024d)	60.1	75.4	**78.7**	77.8	75.7	68.5	65.4
$\mathrm{MME\text{-}RealWorld}_{en}$ (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

实验部分评估了Qwen2.5-VL在各种数据集上的表现，并将其与最先进的模型进行比较，如Claude-3.5-Sonnet-0620（Anthropic，2024a）、GPT-4o-0513（OpenAI，2024）、InternVL2.5（Chen等，2024d）以及不同规模的Qwen2-VL（Wang等，2024e）。在大学水平的问题中，Qwen2.5-VL-72B在MMM U（Yue等，2023）上获得了70.2的分数。对于MMM U-Pro（Yue等，2024），Qwen2.5-VL-72B得分51.1，超越了之前的开源最先进模型，并实现了与GPT-4o相当的性能。

在与数学相关的任务中，Qwen2.5-VL-72B 展现了强大的能力。在 MathVista (Lu et al., 2024) 上，它的得分为 74.8，超过了之前开源的最先进得分 72.3。对于 MATH-Vision (Wang et al., 2024d)，Qwen2.5-VL-72B 的得分为 38.1，而 MathVerse (Zhang et al., 2024c) 的得分为 57.6，两者与其他领先模型相比均显示出竞争力的结果。

对于一般的视觉问答，Qwen2.5-VL-72B 在多个基准测试中表现出色。在 MMbench-EN (Liu et al., 2023d) 中，它的得分为 88.6，略微超过了之前的最佳得分 88.3。该模型在 MuirBench (Wang et al., 2024a) 中也表现良好，得分为 70.7，在 BLINK (Fu et al., 2024c) 中得分为 64.4。在 MTVQA (Tang et al., 2024) 的多语言能力评估中，Qwen2.5-VL-72B 的得分为 31.7，展示了其强大的多语言文本识别能力。在 MMVet (Yu et al., 2024) 和 MM-MT-Bench (Agrawal et al., 2024) 等主观评估中，Qwen2.5-VL-72B 分别得分 76.2 和 7.6，展现了出色的自然对话体验和用户满意度。

3.2 纯文本任务的表现

为了批判性地评估指令调优模型在纯文本任务上的表现，如表4所示，我们选择了几个代表性的基准来评估模型在各种领域的能力，包括一般任务（Wang et al., 2024j; Gema et al., 2024; White et al., 2024）、数学和科学任务（Rein et al., 2023; Hendrycks et al., 2021; Cobbe et al., 2021）、编码任务（Chen et al., 2021; Cassano et al., 2023）和对齐任务（Zhou et al., 2023）。我们将Qwen2.5-VL与几种相似规模的大型语言模型（LLMs）进行了比较。结果表明，Qwen2.5-VL不仅在多模态任务上达到了最先进的（SoTA）性能，而且在纯文本任务上也表现出领先的性能，展示了其在多样化评估标准下的多功能性和稳健性。

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+指令模型和Qwen2.5-VL在纯文本任务上的表现。

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 定量结果

3.3.1 一般视觉问答

为了全面评估模型在一般视觉问答（VQA）和对话中的能力，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表现出色。具体而言，它在基准数据集如MMBench系列（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）和MM-MT-Bench（Agrawal et al., 2024）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数据集上达到了88.4%的准确率，超越了之前的最先进模型，如InternVL2.5（78B）和Claude-3.5 Sonnet-0620。同样，在MMStar数据集上，Qwen2.5-VL获得了70.8%的得分，表现优于该基准中的其他领先模型。这些结果强调了该模型在多样语言环境中的稳健性和适应性。

此外，在高分辨率的现实场景中，特别是在MME-RealWorld基准测试上，Qwen2.5-VL展示了领先的性能，得分为63.2，展现了其对现实环境的广泛适应能力。此外，在MuirBench数据集上评估的多图像理解任务中，Qwen2.5-VL取得了领先的得分70.7，进一步突显了其卓越的泛化能力。总体而言，这些结果展示了Qwen2.5-VL在各种场景中处理通用视觉问答（VQA）任务的强大多功能性和有效性。

值得注意的是，即使是小规模版本的 Qwen2.5-VL，特别是 Qwen2.5-VL-7B 和 Qwen2.5-VL-3B，表现也非常具有竞争力。例如，在 MMStar 数据集上，Qwen2.5-VL-7B 达到了 63.9%，而 Qwen2.5-VL-3B 的得分为 55.9%。这表明 Qwen2.5-VL 的架构不仅强大，而且具有可扩展性，即使参数较少也能保持强劲的性能。

3.3.2 文档理解与光学字符识别 (OCR)

我们在多种OCR、图表和文档理解基准上评估了我们的模型。表5展示了Qwen2.5-VL模型与顶级模型在以下与OCR相关的基准上的性能比较：AI2D（Kembhavi等，2016），TextVQA（Singh等，2019），DocVQA（Mathew等，2021b），InfoVQA（Mathew等，2021a），ChartQA（Masry等，2022），CharXiv（Wang等，2024k），SEED-Bench-2-Plus（Li等，2024b），OCRBench（Liu等，2023e），OCRBench_v2（Fu等，2024b），CC-OCR（Yang等，2024b），OmniDocBench（Ouyang等，2024），VCR（Zhang等，2024e）。

对于多场景、多语言和各种内置（手写、表格、图表、化学公式和数学表达式）文档的元素解析的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

由于经过精心策划的训练数据和LLM模型的出色能力，Qwen2.5-VL-72B模型在CC-OCR和OmniDocBench中设定了新的最先进水平。

对于场景文本、图表、图示和文档的OCR相关理解基准，Qwen2.5-VL模型在理解能力方面表现出色。值得注意的是，在复合OCR相关理解基准如OCRBench、专注于信息图的InfoVQA，以及涵盖包括图表、地图和网页在内的文本丰富场景的SEED-Bench-2-Plus上，Qwen2.5-VL-72B取得了显著的成绩，显著超越了强劲的竞争对手如InternVL2.5-78B。

此外，对于与OCR相关的综合基准测试，如OCRBench_v2，包括广泛的OCR相关解析和理解任务，Qwen2.5-VL模型也实现了最佳性能，分别比最佳模型Gemini 1.5-Pro在英语和中文赛道上超出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表现。

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 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人工智能模型至关重要。在大型视觉语言模型中，视觉定位允许根据自然语言查询或描述精确定位和识别图像中的特定对象、区域或元素。这种能力超越了传统的对象检测，通过在视觉内容和语言上下文之间建立语义关系，促进了更细致和具有上下文意识的视觉推理。我们在指称表达理解基准（Kazemzadeh et al., 2014; Mao et al., 2016）、野外对象检测（Li et al., 2022b）、自我策划的点定位基准和CountBench（Paiss et al., 2023）上评估了Qwen2.5-VL的定位能力。

我们将Qwen2.5-VL的视觉定位能力与其他领先的LVLM进行比较，包括Gemini、Grounding-DINO (Liu et al., 2023c)、Molmo (Deitke et al., 2024)和InternVL2.5。

Qwen2.5-VL 在从框定位、点定位到计数的不同基准测试中实现了领先的性能。通过为 Qwen2.5-VL 配备框和点定位能力，它能够理解、定位并推理图像某些部分的细节。对于开放词汇的物体检测，Qwen2.5-VL 在 ODinW-13 上取得了 43.1 mAP 的良好表现，超越了大多数 LVLM，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 解锁了基于点的定位能力，使其能够精确定位某个物体的细节，而这些细节在过去很难通过边界框表示。Qwen2.5-VL 的计数能力也取得了重大进展，在 CountBench 上以 Qwen2.5-VL-72B 使用“先检测再计数”的提示实现了 93.6 的领先准确率。

3.3.4 视频理解与定位

我们评估了我们的模型在多种视频理解和定位任务上的表现，利用的基准包括时长从几秒到几小时的视频。表8展示了Qwen2.5-VL模型与顶级专有模型在以下视频基准上的性能比较：Video-MME (Fu et al., 2024a)、Video-MMMU (Hu et al., 2025)、MMVU (Zhao

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及其他模型在基础上的表现。

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

表 7　Qwen2.5-VL及其他模型在co上的表现　　unting.

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）和Charades-STA（Gao et al., 2017）。值得注意的是，在LVBench和MLVU上，这些评估通过问答任务进行长视频理解能力的测试，Qwen2.5-VL-72B取得了显著的成绩，远远超过了强有力的竞争对手如GPT-4o。

通过利用提议的同步MRoPE，Qwen2.5-VL增强了其在时间敏感视频理解方面的能力，具有改进的时间戳引用、时间定位、密集字幕和其他功能。在评估准确定位事件或活动及其精确时间戳能力的Charades-STA数据集上，Qwen2.5-VL-72B取得了令人印象深刻的mIoU得分50.9，从而超越了GPT-4o的表现。对于所有评估的基准，我们将每个视频分析的最大帧数限制为768，总视频令牌数不超过24,576。

表8：Qwen2.5-VL及其他模型在视频基准测试上的表现。

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 代理人

在多模态模型中，代理能力对于使这些模型能够有效地与现实世界设备互动至关重要。我们通过多个方面评估Qwen2.5-VL的代理能力。用户界面

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$_{\text{EM}}$	20.8	28.5	12.5	66.4	59.1	**67.36**
Android Control Low$_{\text{EM}}$	19.4	60.2	19.4	84.4	59.2	**93.7**
AndroidWorld$_{\text{SR}}$	34.5% (SoM)	26% (SoM)	27.9%	26.1%	6% (SoM)	**35%**
MobileMiniWob++$_{\text{SR}}$	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.

元素的基础通过 ScreenSpot (Cheng et al., 2024) 和 ScreenSpot Pro (Li et al., 2025a) 进行评估。离线评估在 Android Control (Li et al., 2024f) 上进行，而在线评估则在包括 AndroidWorld (Rawles et al., 2024)、Mobile MiniWob++ (Rawles et al., 2024) 和 OSWorld (Xie et al., 2025) 等平台上进行。我们将 Qwen2.5-VL-72B 的性能与其他知名模型进行比较，如 GPT-4o (OpenAI, 2024)、Gemini 2.0 (Deepmind, 2024)、Claude (Anthropic, 2024b)、Aguvis-72B (Xu et al., 2024) 和 Qwen2-VL-72B (Wang et al., 2024e)。结果在表 9 中展示。

表9：穿孔 Qwen2.5-VL和其他模型在GUI Agen上的表现 t 基准测试。

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

Qwen2.5-VL-72B的性能在GUI基础基准测试中展示了卓越的进步。它在ScreenSpot上达到了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%的准确率显著设定了新的标准，远远超过了Aguvis-72B（23.6%）和其基础版本Qwen2-VL-72B（1.6%）。利用这些卓越的基础能力，Qwen2.5-VL-72B在所有离线评估基准测试中显著超越了基线，差距很大。在在线评估中，由于基础能力有限，一些基线在完成任务时遇到了困难。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++上能够超越基线，并在在线评估中在OSWorld上实现了可比的性能，而无需辅助标记。这一观察表明，Qwen2.5-VL-72B能够在真实和动态环境中作为一个代理进行操作。

4 结论

我们推出了Qwen2.5-VL，这是一系列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展。Qwen2.5-VL在视觉识别、物体定位、文档解析和长视频理解方面具备增强的能力，在静态和动态任务中表现出色。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，而窗口注意力则在不牺牲分辨率保真度的情况下减少了计算开销。Qwen2.5-VL适用于从边缘AI到高性能计算的广泛应用。旗舰产品Qwen2.5-VL-72B与领先模型如GPT-4o和Claude 3.5 Sonnet相匹配或超越，特别是在文档和图表理解方面，同时在纯文本任务上保持强劲表现。较小的Qwen2.5-VL-7B和Qwen2.5-VL-3B变体在同类竞争者中表现优越，提供了高效性和多功能性。Qwen2.5-VL为视觉语言模型设定了新的基准，展示了在各个领域的卓越泛化和任务执行能力。其创新为更智能和互动的系统铺平了道路，架起了感知与现实世界应用之间的桥梁。

5 位作者

核心贡献者：白帅，陈克勤，刘雪晶，王佳林，葛文彬，宋思博，邓凯，王鹏，王世杰，唐俊，钟虎门，朱元志，杨明坤，李兆海，万建强，王鹏飞，丁伟，傅哲仁，徐义恒，叶家博，张希，谢天宝，程泽森，张航，杨志博，徐海洋，林俊阳

贡献者[1]：安阳，惠宾源，余博文，程晨，刘大义恒，范洪，黄飞，刘家伟，徐金，涂建宏，曾建元，张杰，王金凯，张建伟，周景仁，杨克欣，李梅，严明，倪娜，门瑞，姜松涛，邓晓东，黄晓明，周西明，任兴章，范扬，张怡畅，朱宜凯，刘玉琼，郭志芳

[1] Alphabetical order.

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 等人。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 等人。Flamingo：一种用于少量学习的视觉语言模型。在 *NeurIPS*，2022。

人类学。Claude 3.5 颂，2024a。网址 https://www.anthropic.com/news/claude-3-5-sonnet。

人类本质公司。介绍计算机使用、新的claude 3.5十四行诗和claude 3.5俳句，2024b。网址 https://www.anthropic.com/news/3-5-models-and-computer-use。

费德里科·卡萨诺，约翰·古瓦尔，丹尼尔·阮，悉尼·阮，露娜·菲普斯-科斯廷，唐纳德·平克尼，明浩·易，杨天子，卡罗琳·简·安德森，莫莉·Q·费尔德曼，阿尔君·古哈，迈克尔·格林伯格，阿比纳夫·江达。MultiPL-E：一种可扩展的多语言神经代码生成基准测试方法。*IEEE Trans. Software Eng.*，49(7)：3675–3691，2023。

归名·哈迪·陈，顺年·陈，瑞飞·张，俊英·陈，向博·吴，志毅·张，志鸿·陈，建全·李，向万，和本友·王。Allava：利用gpt4v合成数据构建轻量级视觉-语言模型。*arXiv preprint arXiv:2402.11684*，2024a。

贾成陈，天浩梁，谢尔曼·肖，郑青王，凯·王，宇博·王，元生·倪，王竹，子妍·姜，博涵·吕，等。Mega-bench：将多模态评估扩展到超过500个真实世界任务。*arXiv preprint arXiv:2410.10563*，2024b。

林晨，李金松，董晓怡，张潘，臧宇航，陈泽辉，段浩东，王佳琪，乔宇，林大华，等。我们在评估大型视觉-语言模型的正确道路上吗？*arXiv:2403.20330*，2024c。

M陈克，杰瑞·特沃雷克，金辉宇，袁启明，亨里克·庞德·德·奥利维拉·平托，贾里德·卡普兰，哈里森·爱德华兹，尤里·布尔达，尼古拉斯·约瑟夫，格雷格·布罗克曼，亚历克斯·雷，劳尔·普里，格雷琴·克鲁格，迈克尔·彼得罗夫，海迪·克拉夫，吉里什·萨斯特里，帕梅拉·米什金，布鲁克·陈，斯科特·格雷，尼克·赖德，米哈伊尔·帕夫洛夫，阿莱西娅·鲍尔，卢卡斯·凯泽，穆罕默德·巴瓦里安，克莱门斯·温特，菲利普·蒂莱，费利佩·佩特罗斯基·苏奇，戴夫·卡明斯，马蒂亚斯·普拉普特，福提奥斯·钱齐斯，伊丽莎白·巴恩斯，阿里尔·赫伯特-沃斯，威廉·赫布根·古斯，亚历克斯·尼科尔，亚历克斯·佩诺，尼古拉斯·特扎克，唐杰，伊戈尔·巴布什金，苏奇尔·巴拉吉，尚坦努·贾因，威廉·桑德斯，克里斯托弗·赫斯，安德鲁·N·卡尔，扬·莱克，约书亚·阿基亚姆，维丹特·米斯拉，埃文·森川，亚历克·拉德福德，马修·奈特，迈尔斯·布伦达奇，米拉·穆拉蒂，凯蒂·梅耶，彼得·维林德，鲍勃·麦克格鲁，达里奥·阿莫代，萨姆·麦肯德利什，伊利亚·苏茨克弗，以及沃伊切赫·扎伦巴。评估在代码上训练的大型语言模型。*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 和 Jifeng Dai. Internvl: 扩大视觉基础模型并对齐通用视觉-语言任务。 *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 等人。通过模型、数据和测试时间扩展开源多模态模型的性能边界。*arXiv preprint arXiv:2412.05271*，2024d。

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, 和 Zhiyong Wu. Seeclick: 利用图形用户界面基础为高级视觉图形用户界面代理服务。 *arXiv preprint arXiv:2401.10935*, 2024.

卡尔·科布，维尼特·科萨拉朱，穆罕默德·巴瓦里安，马克·陈，许宇俊，卢卡斯·凯泽，马蒂亚斯·普拉普特，杰瑞·特沃雷克，雅各布·希尔顿，中野礼一郎，克里斯托弗·赫塞，约翰·舒尔曼。训练验证器以解决数学文字问题。 *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* 中, *Proceedings of Machine Learning Research* 第 70 卷, 第 933–941 页. PMLR, 2017.

谷歌Deepmind。介绍Gemini 2.0：我们为代理时代推出的新AI模型，2024年。网址 https://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，刘爱心，冯贝，薛冰，王冰轩，吴博超，陆承达，赵承刚，邓承奇，张晨宇，阮冲，戴达麦，郭大雅，杨德建，陈德利，季东杰，李尔航，林芳云，戴富聪，罗富丽，郝光博，陈冠廷，李国伟，张华，包汉，徐汉伟，王浩诚，张浩伟，丁洪辉，辛华建，高华佐，李辉，曲辉，蔡杰伦，梁健，郭建忠，倪佳琦，李佳世，王家伟，陈金，陈景昌，袁景阳，邱俊杰，李俊龙，宋俊霄，董凯，胡凯，高凯歌，关康，黄可欣，余快，王连，张乐聪，徐雷，夏乐怡，赵亮，王立通，张丽月，李萌，王妙君，张明川，张明华，唐明辉，李明明，田宁，黄盼盼，王佩怡，张鹏，王千城，朱启豪，陈沁宇，杜秋实，陈瑞杰，金瑞龙，葛瑞琦，张瑞松，潘瑞哲，王润基，徐润新，张若宇，陈如意，李思思，陆尚豪，周尚妍，陈山煌，吴少青，叶胜丰，叶胜丰，马世荣，王诗宇，周双，余水平，周顺丰，潘书婷，王T，云涛，裴天，孙天宇，肖文亮，曾望鼎。Deepseek-v3技术报告。*CoRR*，abs/2412.19437，2024。doi: 10.48550/ARXIV.2412.19437。网址 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 等人。Molmo 和 pixmo：用于最先进的多模态模型的开放权重和开放数据。*arXiv preprint arXiv:2409.17146*，2024。

方新宇，毛康瑞，段浩东，赵向宇，李怡宁，林大华，陈凯。Mmbench-video：一个用于整体视频理解的长格式多镜头基准。*arXiv preprint arXiv:2406.14515*，2024。

超优傅，裴贤陈，云航沈，玉雷秦，梦丹张，徐林，振宇邱，伟林，金瑞杨，夏武郑，等。Mme：多模态大语言模型的综合评估基准。*arXiv:2306.13394*，2023。

超优傅，余汉戴，永东罗，雷李，书怀任，仁瑞张，子涵王，晨宇周，云航沈，梦丹张，等。视频-mme：首个多模态大型语言模型在视频分析中的综合评估基准。*arXiv:2405.21075*，2024a。

凌傅，杨彪，邝哲彬，宋佳俊，李宇哲，朱凌浩，罗启迪，王新宇，卢浩，黄铭鑫，李章，唐国志，单斌，林春辉，刘琦，吴冰洪，冯浩，刘浩，黄灿，唐景群，陈伟，金连文，刘玉良，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方面的改进基准，2024b。网址 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, 和 Ranjay Krishna. Blink: 多模态大型语言模型可以看见但无法感知。在 *European Conference on Computer Vision*，第 148–166 页。施普林格，2024c。

萨米尔·伊扎克·加德雷，加布里埃尔·伊尔哈科，亚历克斯·方，乔纳森·哈亚塞，乔治奥斯·斯米尔尼斯，陶·阮，瑞安·马滕，米切尔·沃茨曼，德鲁巴·戈什，张杰宇，等。Datacomp：寻找下一代多模态数据集。*arXiv:2304.14108*，2023。

高季扬，孙晨，杨振恒，和拉姆·内瓦蒂亚。Tall: 通过语言查询进行时间活动定位。在 *Proceedings of the IEEE international conference on computer vision*，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 等人。我们完成 mmlu 了吗? *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数据增强方法，用于实例分割。在 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第 2918–2928 页，2021 年。

天瑞关，福晓刘，西阳吴，瑞琪先，宗霞李，晓宇刘，西军王，立昌陈，芙蓉黄，雅瑟·雅库布，迪内什·马诺查，和天逸周。Hallusionbench：一个用于大型视觉语言模型中纠缠语言幻觉和视觉幻觉的高级诊断套件。*arXiv:2310.14566*，2023。

Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, 和 Xiang Yue. Mammoth-vl: 通过大规模指令调优引发多模态推理。 *arXiv preprint 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.

丹·亨德里克斯、科林·伯恩斯、索拉夫·卡达瓦斯、阿库尔·阿罗拉、史蒂文·巴萨特、埃里克·唐、道恩·宋和雅各布·斯坦哈特。使用MATH数据集测量数学问题解决能力。在*NeurIPS Datasets and Benchmarks*，2021年。

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, 和 Ziwei Liu. 视频-mmmu：评估来自多学科专业视频的知识获取。 *arXiv preprint arXiv:2501.13826*, 2025.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten 和 Tamara Berg. Referitgame: 在自然场景的照片中指代物体。在 *EMNLP*，2014。

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi 和 Ali Farhadi. 一张图胜过十幅图像。在 *ECCV*，2016。

亚历山大·基里洛夫，埃里克·敏顿，尼基拉·拉维，毛汉子，克洛伊·罗兰，劳拉·古斯塔夫森，肖特特，斯宾塞·怀特海德，亚历山大·C·伯格，罗万燕，等人。分割任何东西。在 *ICCV*，2023。

李炳宽，朴范灿，金彩媛，罗永满。Moai：大型语言和视觉模型的所有智能的混合。在 *European Conference on Computer Vision*，第 273–302 页。施普林格，2024。

博李，裴源张，景康杨，元汉张，范怡普，和子维刘。Otterhd：一种高分辨率多模态模型。*arXiv:2311.04219*，2023a。博李，元汉张，董国，任瑞张，冯李，浩张，凯晨张，裴源张，彦伟李，子维刘等。Llava-onevision：轻松的视觉任务转移。*arXiv preprint arXiv:2408.03326*，2024a。博豪李，玉莹葛，易晨，逸霄葛，瑞毛张，和英山。Seed-bench-2-plus：基于文本丰富的视觉理解对多模态大型语言模型进行基准测试。*arXiv preprint arXiv:2404.16790*，2024b。董旭李，宇东刘，浩宁吴，岳王，志奇沈，博文曲，欣尧牛，国银王，贝陈，和俊楠李。Aria：一种开放的多模态本地专家混合模型。*arXiv preprint arXiv:2410.05993*，2024c。俊楠李，董旭李，蔡明熊，和史蒂文C.H.霍伊。Blip：为统一的视觉-语言理解和生成引导语言-图像预训练。在*ICML*，2022a。俊楠李，董旭李，西尔维奥·萨瓦雷斯，和史蒂文霍伊。Blip-2：使用冻结的图像编码器和大型语言模型引导语言-图像预训练。*arXiv:2301.12597*，2023b。凯欣李，子扬孟，洪展林，子扬罗，宇晨田，景马，志勇黄，和查达生。Screenspot-pro：专业高分辨率计算机使用的GUI定位，2025a。网址 https://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf。预印本。昆昌李，雅丽王，怡南何，逸卓李，怡王，怡刘，尊王，吉兰徐，国晨，平罗等。Mvbench：一个全面的多模态视频理解基准。在*CVPR*，2024d。

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang 等人。基于图像的语言预训练。在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 10965–10975 页，2022b。

李青云，陈哲，王维云，王文海，叶胜龙，金振江，陈冠洲，何怡南，高张伟，崔尔飞，等。Omnicorpus：一个统一的多模态语料库，包含10亿级别的图像与文本交错。*arXiv preprint arXiv:2406.08418*，2024年。

魏丽，威廉·毕晓普，爱丽丝·李，克里斯·罗尔斯，福拉维约·坎贝尔-阿贾拉，迪维亚·泰亚甘杜鲁，和奥里安娜·里瓦。关于数据规模对计算机控制代理的影响。*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, 等人。Baichuan-omni 技术报告。*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, 等人。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

李云鑫，姜申源，胡宝天，王龙跃，钟万琦，罗文汉，马林，张敏。Uni-moe：通过专家混合扩展统一的多模态大语言模型。*arXiv preprint arXiv:2405.11273*，2024h。

张丽，杨彪，刘强，马志尹，张硕，杨景旭，孙亚博，刘玉良，白翔。猴子：图像分辨率和文本标签是大型多模态模型的重要因素。*arXiv:2311.06607*，2023c。

Yuxuan Liang, Xu Li, Xiaolei Chen, Haotian Chen, Yi Zheng, Chenghang Lai, Bin Li, 和 Xiangyang Xue. 在高分辨率大型视觉语言模型中，全球语义引导的子图像特征权重分配。 *arXiv preprint arXiv:2501.14276*, 2025.

吉林，尹洪旭，魏平，帕夫洛·莫尔查诺夫，穆罕默德·肖伊比，和宋汉。Vila：关于视觉语言模型的预训练。在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*，第 26689–26699 页，2024 年。

刘浩天，李春源，李宇恒，和李永在。通过视觉指令调优改进基线。*arXiv:2310.03744*，2023a。

刘浩天，李春元，吴青阳，和李永宰。视觉指令调优。*arXiv:2304.08485*，2023b。

刘世龙, 曾兆阳, 任天赫, 李峰, 张浩, 杨杰, 李春月, 杨建伟, 苏航, 朱俊娟, 和 张磊. Grounding dino: 将 dino 与基础预训练结合用于开放集目标检测. *arXiv:2303.05499*, 2023c.

扬州刘，岳曹，张伟高，韦云王，哲陈，文海王，浩天，乐伟陆，西舟朱，彤陆，等。Mminstruct：一个具有广泛多样性的高质量多模态指令调优数据集。*Science China Information Sciences*，67(12)：1–16，2024a。

刘元, 段浩东, 李博, 张元汉, 张松阳, 赵王博, 袁逸可, 王佳琪, 何聪辉, 刘子维, 陈凯, 和林大华. 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. 文章: 用可负担的策略提升你的视觉-语言模型. *arXiv preprint arXiv:2409.04828*, 2024b.

刘元欣，李士成，刘怡，王宇翔，任书怀，李雷，陈思硕，孙旭，侯璐。Tempcompass：视频大型语言模型真的理解视频吗？*arXiv preprint arXiv:2403.00476*，2024c。

刘宇亮, 张丽, 黄明鑫, 杨彪, 于文文, 李春元, 尹旭成, 刘成林, 金连文, 和白翔. Ocrbench: 大型多模态模型中OCR的隐藏奥秘. *arXiv:2305.07895*, 2023e.

潘璐，Hritik Bansal，Tony Xia，刘嘉诚，李春元，Hannaneh Hajishirzi，Hao Cheng，Kai-Wei Chang，Michel Galley，和高剑锋。Mathvista：评估基础模型在视觉上下文中的数学推理。在 *ICLR*，2024。

Karttikeya Mangalam, Raiymbek Akshulakov 和 Jitendra Malik. Egoschema：一种用于超长视频语言理解的诊断基准。在 *NeurIPS*，2023。

毛俊华，黄Jonathan，托谢夫亚历山大，坎布鲁奥安娜，尤伊尔阿兰·L，和墨菲凯文。生成和理解明确的物体描述。在 *CVPR*，2016年。

艾哈迈德·马斯里，杜宣龙，贾庆坦，沙菲克·乔提，和恩阿穆尔·霍克。Chartqa：一个关于图表的问答基准，涉及视觉和逻辑推理。*arXiv:2203.10244*，2022。

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

Minesh Mathew, Dimosthenis Karatzas 和 CV Jawahar. Docvqa：一个用于文档图像的 VQA 数据集。在 *WACV*，2021b。

MiniMax，李奥尼安，龐偉公，博揚，博基山，常流，鄭成，張春浩，郭聰超，陳達，李東，焦恩偉，李更新，張國軍，孫浩海，董厚澤，朱佳岱，莊佳琦，宋家源，朱金，韓景濤，李景揚，謝俊彬，徐俊豪，顏俊傑，張凱順，蕭克誠，康克西，韓樂，王樂揚，余連飛，馮立恒，鄭林，柴林博，邢龍，居美智，池明源，張莫之，黃培凱，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, 彭宇 赵, 齐 杨, 奇迪 许, 切香 王, 秦 王, 秋辉 李, 瑞涛 冷, 盛敏 施, 书琦 余, 思辰 李, 松泉 朱, 涛 黄, 天润 梁, 伟高 孙, 伟轩 孙, 伟宇 程, 文凯 李, 向军 宋, 小 苏, 小东 韩, 新杰 张, 新竹 侯, 许 敏, 旬 邹, 旭阳 沈, 燕 龚, 英杰 朱, 逸鹏 周, 依然 钟, 永义 胡, 元翔 范, 月 余, 玉峰 杨, 余浩 李, 云南 黄, 云姬 李, 云鹏 黄, 云志 许, 玉鑫 毛, 泽涵 李, 泽康 李, 泽伟 陶, 泽文 应, 赵阳 丛, 振 秦, 振华 范, 志航 余, 卓 江, 和 子佳 吴. Minimax-01: 用闪电注意力扩展基础模型, 2025. URL https://arxiv.org/abs/2501.08313.

Openai. Chatml 文档，2024。网址 https://github.com/openai/openai-python/blob/main/chatml.md。

OpenAI。你好 gpt-4o，2024。网址 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, 和 Conghui He. Omnidocbench: 用全面注释对多样化 PDF 文档解析进行基准测试, 2024. URL https://arxiv.org/abs/2412.07626.

罗尼·派斯、阿里尔·埃夫拉特、奥梅尔·托夫、希兰·扎达、因巴尔·莫塞里、米哈尔·伊拉尼和塔利·德凯尔。教学视频教孩子数到十。在 *Proceedings of the IEEE/CVF International Conference on Computer Vision*，第 3170–3180 页，2023 年。

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch 等人。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: 将多模态大型语言模型与世界相结合。*arXiv:2306.14824*, 2023.

拉斐尔·拉法伊洛夫、阿基特·夏尔马、埃里克·米切尔、克里斯托弗·D·曼宁、斯特凡诺·埃尔蒙和切尔西·芬。直接偏好优化：您的语言模型实际上是一个奖励模型。在爱丽丝·哦、特里斯坦·瑙曼、阿米尔·格洛伯森、凯特·塞恩科、莫里茨·哈特和谢尔盖·莱文（编辑），*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年。网址 http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html。

克里斯托弗·罗尔斯，萨拉·克林克梅利，易凡·张，乔纳森·沃尔茨，嘉布丽尔·刘，玛丽贝丝·费尔，爱丽丝·李，威廉·比肖普，韦·李，福拉维约·坎贝尔-阿贾拉等。Androidworld：一个用于自主智能体的动态基准测试环境。*arXiv:2405.14573*，2024。

大卫·雷因，贝蒂·李·侯，阿萨·库珀·斯蒂克兰，杰克逊·佩蒂，理查德·元哲·庞，朱利安·迪拉尼，朱利安·迈克尔，和塞缪尔·R·博曼。GPQA：一个研究生级别的谷歌防护问答基准。*CoRR*，abs/2311.12022，2023。

天河任、青江、石龙刘、赵阳曾、文龙刘、韩高、洪杰黄、郑宇马、晓可江、逸豪陈等。基础DINO 1.5：推进开放集物体检测的"边缘"。*arXiv preprint arXiv:2405.10300*，2024。

卡洛斯·里克尔梅，乔安·普伊克塞尔维，巴西尔·穆斯塔法，马克西姆·诺伊曼，罗多夫·热纳通，安德烈·苏萨诺·平托，丹尼尔·凯瑟斯，和尼尔·霍尔斯比。通过稀疏专家混合扩展视觉。*Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

阿曼普里特·辛格，维韦克·纳塔拉詹，米特·沙阿，余江，辛雷·陈，德鲁夫·巴特拉，德维·帕里克，和马库斯·罗尔巴赫。朝着能够阅读的vqa模型迈进。在*CVPR*，2019年。

苏建林，穆尔塔达·H·M·艾哈迈德，卢宇，潘胜峰，温博，和刘云峰。Roformer：带有旋转位置嵌入的增强型变换器。*Neurocomputing*，568:127063，2024。

唐景群, 刘琦, 叶永杰, 陆晶辉, 魏舒, 林春辉, 李万青, 穆罕默德·菲特里·法伊兹·宾·马哈茂德, 冯浩, 赵震, 王彦杰, 刘玉良, 刘浩, 白翔, 和黄灿. Mtvqa: 多语言文本中心视觉问答的基准测试. *arXiv:2405.11985*, 2024.

双子团队，罗汉·阿尼尔，塞巴斯蒂安·博尔戈，吴永辉，尚-巴蒂斯特·阿拉亚克，余佳辉，拉杜·索里库特，约翰·沙尔克维克，安德鲁·M戴，安雅·豪斯等。双子：一系列高能力的多模态模型。*arXiv preprint 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.

盛邦通，埃利斯·布朗，彭浩·吴，尚贤·吴，马诺杰·米德波古，赛·查里塔·阿库拉，吉汉·杨，舒生·杨，阿迪提亚·艾耶尔，西辰·潘，等。Cambrian-1：一个完全开放、以视觉为中心的多模态大语言模型探索。*arXiv preprint arXiv:2406.16860*，2024年。费·王，邢宇·傅，詹姆斯·黄，泽坤·李，秦·刘，晓耕·刘，明宇·德里克·马，南·徐，文轩·周，凯·张，等。Muirbench：一个全面的强健多图像理解基准。*arXiv preprint arXiv:2406.09411*，2024a。俊阳·王，海洋·徐，海涛·贾，西·张，明·颜，伟洲·沈，吉·张，费·黄，和吉涛·桑。Mobile-agent-v2：通过多代理协作进行有效导航的移动设备操作助手。*arXiv preprint arXiv:2406.01014*，2024b。俊阳·王，海洋·徐，家博·叶，明·颜，伟洲·沈，吉·张，费·黄，和吉涛·桑。Mobile-agent：具有视觉感知的自主多模态移动设备代理。*arXiv preprint arXiv:2401.16158*，2024c。可·王，俊廷·潘，伟康·施，子穆·卢，明杰·詹，和洪生·李。使用数学视觉数据集测量多模态数学推理。*arXiv:2402.14804*，2024d。彭·王，帅·白，思南·谭，世杰·王，志豪·范，金泽·白，克勤·陈，雪晶·刘，家林·王，文彬·葛，杨·范，凯·党，梦飞·杜，轩成·任，瑞·门，大义恒·刘，常·周，景仁·周，和俊阳·林。Qwen2-vl：增强视觉-语言模型对世界的感知，适用于任何分辨率。*arXiv:2409.12191*，2024e。彭·王，帅·白，思南·谭，世杰·王，志豪·范，金泽·白，克勤·陈，雪晶·刘，家林·王，文彬·葛，等。Qwen2-vl：增强视觉-语言模型对世界的感知，适用于任何分辨率。*arXiv preprint arXiv:2409.12191*，2024f。伟汉·王，泽海·何，文怡·洪，燕·程，晓寒·张，吉·齐，晓涛·顾，世宇·黄，彬·徐，宇晓·董，等。Lvbench：一个极长视频理解基准。*arXiv preprint arXiv:2406.08035*，2024g。伟云·王，义铭·任，浩文·罗，天通·李，晨翔·严，哲·陈，文海·王，青云·李，乐伟·卢，西洲·朱，等。全视项目v2：朝着开放世界的一般关系理解迈进。*arXiv preprint arXiv:2402.19474*，2024h。文海·王，季峰·戴，哲·陈，振航·黄，志奇·李，西洲·朱，晓伟·胡，彤·卢，乐伟·卢，洪生·李，等。Internimage：通过可变形卷积探索大规模视觉基础模型。在*Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*，第14408–14419页，2023年。新龙·王，晓松·张，正雄·罗，权·孙，宇峰·崔，金生·王，范·张，跃泽·王，振·李，启颖·余，等。Emu3：下一个标记预测就是你所需要的一切。*arXiv preprint arXiv:2409.18869*，2024i。宇博·王，雪光·马，格·张，元生·倪，阿布拉尼尔·钱德拉，时光·郭，伟明·任，阿兰·阿鲁尔拉杰，轩·何，子妍·姜，天乐·李，马克斯·库，凯·王，亚历克斯·庄，荣琦·范，向·岳，和文虎·陈。MMLU-Pro：一个更强健和具有挑战性的多任务语言理解基准。*CoRR*，abs/2406.01574，2024j。振海龙·王，海洋·徐，俊阳·王，西·张，明·颜，吉·张，费·黄，和恒·吉。Mobile-agent-e：用于复杂任务的自我进化移动助手。*arXiv preprint arXiv:2501.11733*，2025年。子瑞·王，梦洲·夏，璐希·何，霍华德·陈，逸涛·刘，理查德·朱，凯曲·梁，欣迪·吴，浩天·刘，萨迪卡·马拉迪，亚历克西斯·谢瓦利耶，桑杰夫·阿罗拉，和丹琦·陈。Charxiv：绘制多模态大语言模型中现实图表理解的差距。*arXiv preprint arXiv:2406.18521*，2024k。杰森·魏，雪志·王，戴尔·舒尔曼斯，马尔滕·博斯马，埃德·H池，阮国·乐，和丹尼·周。思维链提示引发大型语言模型中的推理。*CoRR*，abs/2201.11903，2022年。网址 https://arxiv.org/abs/2201.11903。科林·怀特，塞缪尔·杜利，曼利·罗伯茨，阿卡·帕尔，本杰明·费尔，西达尔塔·贾因，拉维德·施瓦茨-齐夫，尼尔·贾因，哈立德·赛夫拉，西达尔塔·奈杜，钦梅·赫格德，扬·勒昆，汤姆·戈德斯坦，威利·内斯旺格，和米卡·戈德布鲁姆。LiveBench：一个具有挑战性、无污染的大语言模型基准。*CoRR*，abs/2406.19314，2024年。

Haoning Wu, Dongxu Li, Bei Chen, 和 Junnan Li. Longvideobench: 一种用于长上下文交错视频语言理解的基准，2024a. URL https://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.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang 等. Deepseek-vl2: 混合专家视觉语言模型用于高级多模态理解. *arXiv preprint 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 和 Lu Yuan. Florence-2: 推进多种视觉任务的统一表示 (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 等. Osworld: 在真实计算环境中对开放式任务的多模态代理进行基准测试. *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 和 Caiming Xiong. Aguvis: 统一的纯视觉代理用于自主 GUI 交互. *arXiv preprint arXiv:2412.04454*, 2024. An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang 等. Qwen2.5 技术报告. *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 和 Junyang Lin. Cc-ocr: 一个全面且具有挑战性的 OCR 基准，用于评估大规模多模态模型的识字能力, 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 等. X-vila: 大型语言模型的跨模态对齐. *arXiv preprint arXiv:2405.19335*, 2024. Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang 和 Jingren Zhou. mplug-owl2: 革新多模态大型语言模型与模态协作. *arXiv:2311.04257*, 2023. Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang 和 Lijuan Wang. Mm-vet: 评估大型多模态模型的综合能力. 在 *ICML*, 2024. Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun 等. Mmmu: 一个针对专家 AGI 的大规模多学科多模态理解与推理基准. *arXiv:2311.16502*, 2023. Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang 等. Mmmu-pro: 一个更强大的多学科多模态理解基准. *arXiv preprint arXiv:2409.02813*, 2024. Biao Zhang 和 Rico Sennrich. 均方根层归一化. 在 *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 和 Yinfei Yang. Ferret-v2: 一个改进的基线，用于大型语言模型的引用和定位. *arXiv:2404.07973*, 2024a. Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding 等. Internlm-xcomposer2.5-omnilive: 一个综合的多模态系统，用于长期流媒体视频和音频交互. *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 等. Mathverse: 你的多模态 LLM 是否真正看到了视觉数学问题中的图表? 在 *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 和 Shuicheng Yan. omg-llava: 连接图像级、对象级、像素级推理与理解. *arXiv preprint arXiv:2406.19389*, 2024d. Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu 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.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, 等. Mme-realworld: 你的多模态 llm 能否挑战对人类来说困难的高分辨率现实场景？ *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, 和 Arman Cohan. Mmvu: 测量专家级多学科视频理解, 2025. URL https://arxiv.org/abs/2501.12380.

杰弗里·周，田建·卢，斯瓦罗普·米什拉，西达尔塔·布拉马，苏乔伊·巴苏，易峦，丹尼·周，和乐·侯。大型语言模型的指令遵循评估。*CoRR*，abs/2311.07911，2023。

周俊杰，舒妍，赵博，吴博雅，肖士涛，杨熙，熊永平，张博，黄铁军，刘铮。Mlvu：一个用于多任务长视频理解的综合基准。*arXiv preprint arXiv:2406.04264*，2024。