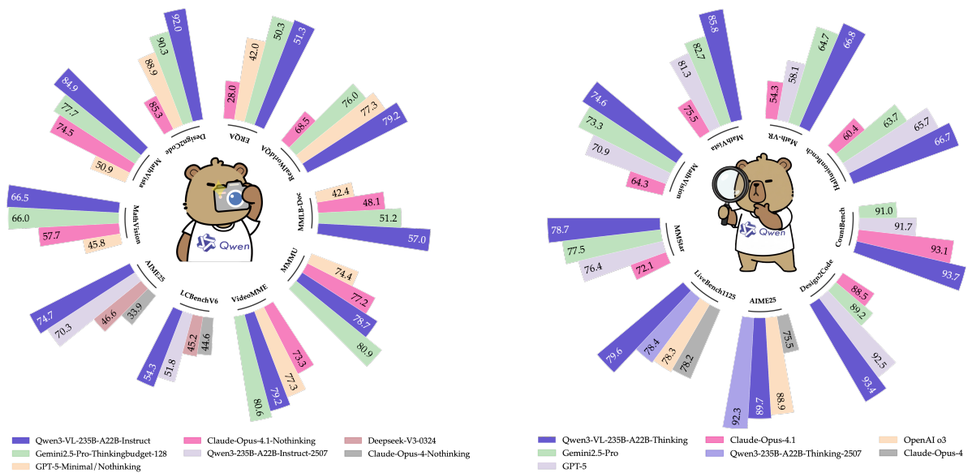
Qwen3-VL Technical Report

Qwen Team

https://chat.qwen.ai
https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3-VL

Abstract

We introduce Qwen3-VL, the most capable vision–language model in the Qwen series to date, achieving superior performance across a broad range of multimodal benchmarks. It natively supports interleaved contexts of up to 256K tokens, seamlessly integrating text, images, and video. The model family includes both dense (2B/4B/8B/32B) and mixture-of-experts (30B-A3B/235B-A22B) variants to accommodate diverse latency–quality trade-offs. Qwen3-VL delivers three core pillars: (i) markedly stronger pure-text understanding, surpassing comparable text-only backbones in several cases; (ii) robust long-context comprehension with a native 256K-token window for both text and interleaved multimodal inputs, enabling faithful retention, retrieval, and cross-referencing across long documents and videos; and (iii) advanced multimodal reasoning across single-image, multi-image, and video tasks, demonstrating leading performance on comprehensive evaluations such as MMMU and visual-math benchmarks (e.g., MathVista and MathVision). Architecturally, we introduce three key upgrades: (i) an enhanced interleaved-MRoPE for stronger spatial–temporal modeling across images and video; (ii) DeepStack integration, which effectively leverages multi-level ViT features to tighten vision–language alignment; and (iii) text-based time alignment for video, evolving from T-RoPE to explicit textual timestamp alignment for more precise temporal grounding. To balance text-only and multimodal learning objectives, we apply square-root reweighting, which boosts multimodal performance without compromising text capabilities. We extend pretraining to a context length of 256K tokens and bifurcate post-training into non-thinking and thinking variants to address distinct application requirements. Furthermore, we allocate additional compute resources to the post-training phase to further enhance model performance. Under comparable token budgets and latency constraints, Qwen3-VL achieves superior performance in both dense and Mixture-of-Experts (MoE) architectures. We envision Qwen3-VL serving as a foundational engine for image-grounded reasoning, agentic decision-making, and multimodal code intelligence in real-world workflows.



arXiv:2511.21631v2 [cs.CV] 27 Nov 2025

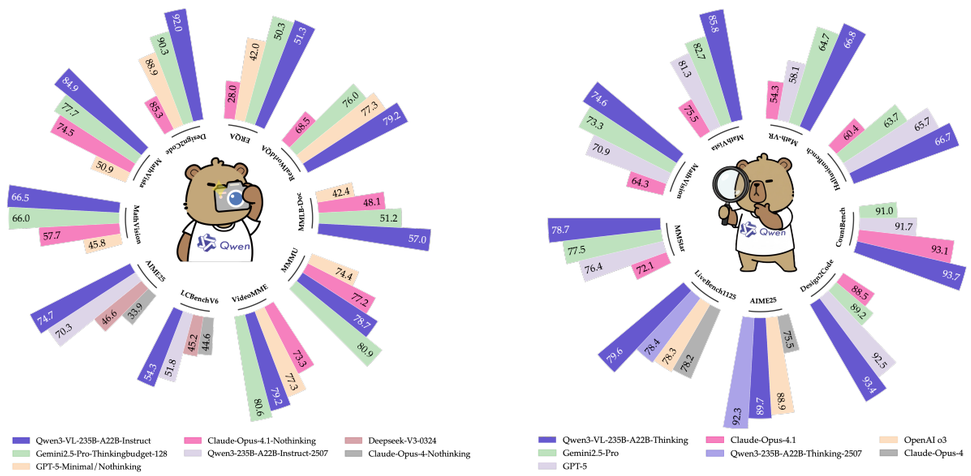
Qwen3-VL 技术报告

Qwen 团队

https://chat.qwen.ai https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen3-VL

摘要

我们介绍了Qwen3-VL，这是迄今为止Qwen系列中最强大的视觉-语言模型，在广泛的跨模态基准测试中实现了卓越性能。该模型原生支持高达256K个token的交错上下文，无缝集成文本、图像和视频。该模型家族包括密集型（2B/4B/8B/32B）和专家混合型（30B-A3B/235B-A22B）变体，以适应不同的时延-质量权衡。Qwen3-VL提供三大核心支柱：(i) 显著更强的纯文本理解能力，在某些情况下超越了类似的纯文本主干；(ii) 强大的长上下文理解能力，具有原生256K个token的窗口，适用于纯文本和交错多模态输入，能够忠实保留、检索和跨参考长文档和视频；(iii) 跨单图像、多图像和视频任务的先进多模态推理能力，在MMMU和视觉数学基准测试（例如Math-Vista和MathVision）等综合评估中表现出领先性能。在架构上，我们引入了三项关键升级：(i) 增强的交错-MRoPE，用于加强图像和视频的空间-时间建模；(ii) DeepStack集成，有效利用多级ViT特征来加强视觉-语言对齐；(iii) 基于文本的视频时间对齐，从T-RoPE演变为明确的文本时间戳对齐，以实现更精确的时间锚定。为了平衡纯文本和多模态学习目标，我们应用了平方根重加权，在提升多模态性能的同时不损害文本能力。我们将预训练扩展到256K个token的上下文长度，并将后训练分为非思考型和思考型变体，以应对不同的应用需求。此外，我们为后训练阶段分配了额外的计算资源，以进一步提升模型性能。在可比的token预算和时延限制下，Qwen3-VL在密集型和专家混合型（MoE）架构中均实现了卓越性能。我们期望Qwen3-VL作为图像锚定推理、代理决策和现实工作流程中的多模态代码智能的基础引擎。



1 Introduction

Vision–language models (VLMs) have achieved substantive progress in recent years, evolving from foundational visual perception to advanced multimodal reasoning across images and video. The rapid advancement of VLMs has given rise to a rapidly expanding landscape of downstream applications—such as long-context understanding, STEM reasoning, GUI comprehension and interaction, and agentic workflows. Crucially, these advances must not erode the underlying large language model's (LLM's) linguistic proficiency; multimodal models are expected to match or surpass their text-only counterparts on language benchmarks.

In this report, we present Qwen3-VL and its advances in both general-purpose and advanced applications. Built on the Qwen3 series (Yang et al., 2025a), we instantiate four dense models (2B/4B/8B/32B) and two mixture-of-experts (MoE) models (30B-A3B / 235B-A22B), each trained with a context window of up to 256K tokens to enable long-context understanding. By optimizing the training corpus and training strategy, we preserve the underlying LLM's language proficiency during vision–language (VL) training, thereby substantially improving overall capability. We release both non-thinking and thinking variants; the latter demonstrates significantly stronger multimodal reasoning capabilities, achieving superior performance on complex reasoning tasks.

We first introduce the architectural improvements, which span three components: 1) Enhanced positional encoding. In Qwen2.5-VL, we used MRoPE as a unified positional encoding scheme for text and vision. We observed that chunking the embedding dimensions into temporal (t), horizontal (h), and vertical (w) groups induces an imbalanced frequency spectrum and hampers long-video understanding. We therefore adopt an interleaved MRoPE that distributes t, h, and w uniformly across low- and high-frequency bands, yielding more faithful positional representations. 2) DeepStack for cross-layer fusion. To strengthen vision–language alignment, we incorporate the pioneering DeepStack (Meng et al., 2024) mechanism. Visual tokens from different layers of the vision encoder are routed to corresponding LLM layers via lightweight residual connections, enhancing multi-level fusion without introducing extra context length. 3) Explicit video timestamps. We replace the absolute-time alignment via positional encoding used in Qwen2.5-VL with explicit timestamp tokens to mark frame groups, providing a simpler and more direct temporal representation. In addition, on the optimization side, we move from a per-sample loss to a square-root-normalized per-token loss, which better balances the contributions of text and multimodal data during training.

To build a more capable and robust vision–language foundation model, we overhauled our training data in terms of quality, diversity, and structure. Key upgrades include enhanced caption supervision, expanded omni-recognition and OCR coverage, normalized grounding with 3D/spatial reasoning, and new corpora for code, long documents, and temporally grounded video. We further infused chain-of-thought reasoning and high-quality, diverse GUI-agent interaction data to bridge perception, reasoning, and action. Together, these innovations enable stronger multimodal understanding, precise grounding, and tool-augmented intelligence.

Our training pipeline consists of two stages: pretraining and post-training. Pretraining proceeds in four phases: a warm-up alignment phase that updates only the merger (vision–language projection) layers while keeping the rest of the model frozen, followed by full-parameter training with progressively larger context windows at 8K, 32K, and 256K sequence lengths. Post-training comprises three phases: (i) supervised fine-tuning on long chain-of-thought data, (ii) knowledge distillation from stronger teacher models, and (iii) reinforcement learning.

The above innovations equip Qwen3-VL with strong capabilities not only as a robust vision–language foundation model but also as a flexible platform for real-world multimodal intelligence—seamlessly integrating perception, reasoning, and action across diverse application domains. In the following sections, we present the model architecture, training framework, and extensive evaluations that demonstrate its consistent and competitive performance on text, vision, and multimodal reasoning benchmarks.

2 Model Architecture

Following Qwen2.5-VL (Bai et al., 2025), Qwen3-VL adopts a three-module architecture comprising a vision encoder, an MLP-based vision–language merger, and a large language model (LLM). Figure 1 depicts the detailed model structure.

Large Language Model: Qwen3-VL is instantiated in three dense variants (Qwen3-VL-2B/4B/8B/32B) and two MoE variants (Qwen3-VL-30B-A3B, Qwen3-VL-235B-A22B), all built upon Qwen3 backbones. The flagship model, Qwen3-VL-235B-A22B, has 235B total parameters with 22B activated per token. It

1 简介

视觉-语言模型（VLMs）近年来取得了实质性进展，从基础视觉感知发展到跨图像和视频的高级多模态推理。VLMs的快速发展催生了一个快速扩大的下游应用领域——如长上下文理解、STEM推理、GUI理解和交互以及代理工作流程。关键在于，这些进步不能削弱底层大型语言模型（LLM）的语言能力；多模态模型被期望在语言基准测试上与或超越其纯文本对应物。

在本报告中，我们介绍了 Qwen3-VL 及其在通用和高级应用方面的进展。基于 Qwen3 系列（杨等人，2025a），我们实例化了四个密集模型（2B/4B/8B/32B）和两个专家混合（MoE）模型（30B-A3B / 235B-A22B），每个模型均使用高达 256K 个 token 的上下文窗口进行训练，以实现长上下文理解。通过优化训练语料库和训练策略，我们在视觉-语言（VL）训练中保留了底层大语言模型（LLM）的语言能力，从而显著提升了整体能力。我们发布了非思考型和思考型两种变体；后者展示了显著更强的多模态推理能力，在复杂推理任务上取得了优异性能。

我们首先介绍了架构改进，涵盖三个组件：1) 增强型位置编码。在 Qwen2.5-VL 中，我们使用 MRoPE 作为文本和视觉的统一位置编码方案。我们观察到将嵌入维度分块为时间（t）、水平（h）和垂直（w）组会导致频率谱不平衡，并阻碍长视频理解。因此，我们采用交错 MRoPE，将 t、h 和 w 均匀分布在低频和高频带，从而生成更忠实的位置表示。2) DeepStack 用于跨层融合。为加强视觉-语言对齐，我们集成了开创性的 DeepStack（孟等人， 2024）机制。视觉编码器不同层的视觉 token 通过轻量级残差连接路由到相应的 LLM 层，增强了多级融合，而不会引入额外的上下文长度。3) 显式视频时间戳。我们用显式时间戳 token 替换了 Qwen2.5-VL 中通过位置编码进行的绝对时间对齐，以标记帧组，提供更简单、更直接的时间表示。此外，在优化方面，我们从按样本损失转向平方根归一化的按 token 损失，这更好地平衡了训练过程中文本和多模态数据的贡献。

为了构建一个更强大和鲁棒的视觉-语言基础模型，我们针对训练数据的质量、多样性和结构进行了全面升级。关键升级包括增强的标题监督、扩展的全方位感知和OCR覆盖范围、结合3D/空间推理的标准化Grounding，以及用于代码、长文档和时序关联视频的新语料库。我们进一步融入了思维链推理和高质量、多样化的GUI代理交互数据，以连接感知、推理和行动。这些创新共同促成了更强的多模态理解、精确Grounding和工具增强型智能。

我们的训练流程包含两个阶段：预训练和后训练。预训练分为四个阶段：一个预热对齐阶段，该阶段仅更新合并层（视觉-语言投影），其余模型保持冻结，随后进行全参数训练，使用逐渐增大的上下文窗口，序列长度为8K、32K和256K。后训练包含三个阶段：(i) 在长思维链数据上进行监督微调，(ii) 从更强的教师模型进行知识蒸馏，以及(iii) 强化学习。

上述创新使Qwen3-VL不仅具备强大的视觉-语言基础模型能力，还作为一个灵活的平台，支持现实世界中的多模态智能——无缝整合感知、推理和行动，应用于不同的应用领域。在接下来的章节中，我们将介绍模型架构、训练框架以及广泛的评估，展示其在文本、视觉和多模态推理基准测试上始终如一且具有竞争力的性能。

2 模型架构

遵循Qwen2.5-VL（Bai等人，2025），Qwen3-VL采用一个由视觉编码器、基于MLP的视觉-语言合并器和大语言模型（LLM）组成的三模块架构。图1描绘了详细的模型结构。

大型语言模型：Qwen3-VL 以三种密集变体（Qwen3-VL-2B/4B/8B/32B）和两种MoE变体（Qwen3-VL-30B-A3B、Qwen3-VL-235B-A22B）进行实例化，均基于Qwen3主干网络构建。旗舰模型Qwen3-VL-235B-A22B拥有235B总参数，每个token激活22B参数。它

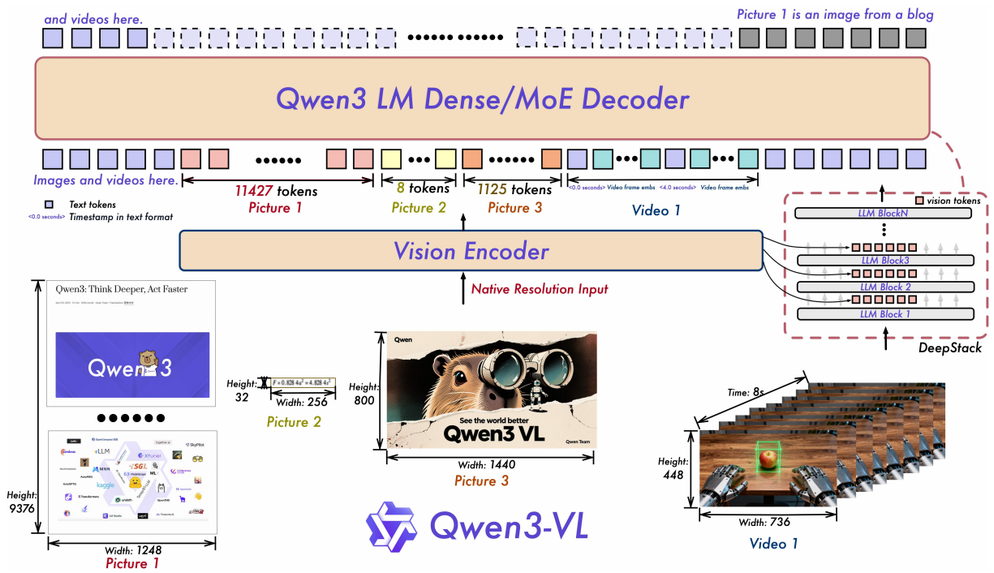


Figure 1: The Qwen3-VL framework integrates a vision encoder and a language model decoder to process multimodal inputs, including text, images, and video. The vision encoder is specifically designed to handle dynamic, native-resolution visual inputs, mapping them to visual tokens of variable length. To enhance perceptual capability and preserve rich visual information, we incorporate the pioneering DeepStack mechanism, which injects visual tokens from multiple layers of the vision encoder into corresponding layers of the LLM. Furthermore, we adopt Interleaved MRoPE to encode positional information for multimodal inputs with a balanced frequency spectrum, and introduce text-based timestamp tokens to more effectively capture the temporal structure of video sequences.

outperforms most VLMs across a broad set of multimodal tasks and surpasses its text-only counterpart on the majority of language benchmarks.

Vision Encoder: We utilize the SigLIP-2 architecture (Tschannen et al., 2025) as our vision encoder and continue training it with dynamic input resolutions, initialized from official pretrained checkpoints. To accommodate dynamic resolutions effectively, we employ 2D-RoPE and interpolate absolute position embeddings based on input size, following the methodology of CoMP (Chen et al., 2025). Specifically, we default to the SigLIP2-SO-400M variant and use SigLIP2-Large (300M) for small-scale LLMs (2B and 4B).

MLP-based Vision-Language Merger: As in Qwen2.5-VL, we use a two-layer MLP to compress 2×2 visual features from the vision encoder into a single visual token, aligned with the LLM's hidden dimension. Additionally, we deploy specialized mergers to support the DeepStack mechanism (Meng et al., 2024), the details of which are fully described in Section 2.2.

2.1 Interleaved MRoPE

Qwen2-VL (Wang et al., 2024c) introduced MRoPE to model positional information for multimodal inputs. In its original formulation, the embedding dimensions are partitioned into temporal (t), horizontal (h), and vertical (w) subspaces, each assigned distinct rotary frequencies. This results in an imbalanced frequency spectrum, which subsequent studies have shown to degrade performance on long-video understanding benchmarks. To address this, we redesign the frequency allocation by interleaving the t, h, and w components across the embedding dimensions (Huang et al., 2025). This ensures that each spatial–temporal axis is uniformly represented across both low- and high-frequency bands. The resulting balanced spectrum mitigates the original spectral bias and significantly improves long-range positional modeling for video.

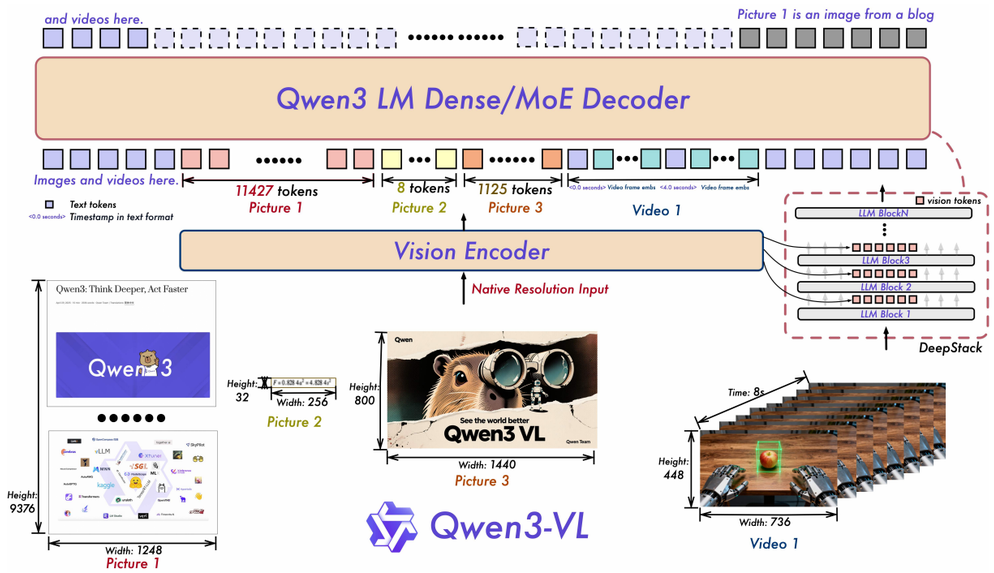


图1：Qwen3-VL 框架集成了视觉编码器和语言模型解码器，用于处理多模态输入，包括文本、图像和视频。视觉编码器专门设计用于处理动态、原始分辨率的视觉输入，将它们映射为长度可变的眼动标记。为了增强感知能力并保留丰富的视觉信息，我们采用了开创性的 DeepStack 机制，该机制将视觉编码器多个层级的视觉标记注入到 LLM 对应层级中。此外，我们采用 Interleaved MRoPE 对具有平衡频率谱的多模态输入进行位置信息编码，并引入基于文本的时间戳标记，以更有效地捕捉视频序列的时间结构。

在广泛的跨模态任务中表现优于大多数VLM，并在多数语言基准测试上超越其纯文本版本。

视觉编码器: 我们采用了 SigLIP-2 架构（Tschannen 等人。， 2025）作为我们的视觉编码器，并继续使用动态输入分辨率对其进行训练，初始状态来自官方预训练检查点。为了有效适应动态分辨率，我们采用了 2D-RoPE，并根据输入大小对绝对位置嵌入进行插值，遵循 CoMP（Chen 等人。， 2025）的方法。具体来说，我们默认采用 SigLIP2-SO-400M 变体，并使用 SigLIP2-Large (300M) 用于小规模的大语言模型（2B 和 4B）。

基于MLP的视觉-语言合并器：与Qwen2.5-VL类似，我们使用一个两层MLP来压缩 2×2来自视觉编码器的视觉特征，将其压缩成一个视觉token，并与大语言模型的隐藏维度对齐。此外，我们部署了专门的合并器来支持DeepStack机制（孟等人，2024），其详细信息在2.2节中完整描述。2.2。

2.1 交错MRoPE

Qwen2-VL (王等人， 2024c) 引入了 MRoPE 来对多模态输入中的位置信息进行建模。在其原始公式中，嵌入维度被划分为时间（t）、水平（h）和垂直（w）子空间，每个子空间被分配不同的旋转频率。这导致频率谱不平衡，后续研究表明这会降低在长视频理解基准测试上的性能。为了解决这个问题，我们通过在嵌入维度中交错 t、h 和 w 组件（黄等人， 2025）来重新设计频率分配。这确保了每个时空轴在低频和高频带中都能得到均匀表示。由此产生的平衡频谱减轻了原始的频谱偏差，并显著提高了视频的长距离位置建模能力。

2.2 DeepStack

We draw inspiration from DeepStack (Meng et al., 2024) and inject visual tokens into multiple layers of the LLM. Unlike the original DeepStack approach, which stacks tokens from multi-scale visual inputs, we extend DeepStack to extract visual tokens from intermediate layers of the Vision Transformer (ViT). This design preserves rich visual information, ranging from low- to high-level representations.

Specifically, as illustrated in Figure 1, we select features from three distinct levels of the vision encoder. Subsequently, dedicated vision–language merger modules project these multi-level features into visual tokens, which are then added directly to the corresponding hidden states of the first three LLM layers.

2.3 Video Timestamp

In Qwen2.5-VL, a time-synchronized variant of MRoPE is employed to endow the model with temporal awareness. However, we identify two key limitations of this approach: (1) By tying temporal position IDs directly to absolute time, the method produces excessively large and sparse temporal position ids for long videos, degrading the model's ability to understand long temporal contexts. (2) Effective learning under this scheme requires extensive and uniformly distributed sampling across various frame rates (fps), significantly increasing the cost of training data construction.

To address these issues, we adopt a textual token–based time encoding strategy (Chen et al., 2024b), wherein each video temporal patch is prefixed with a timestamp expressed as a formatted text string—e.g., `<3.0 seconds>`. Furthermore, during training, we generate timestamps in both seconds and HMS (hours:minutes:seconds) formats to ensure the model learns to interpret diverse timecode representations. Although this approach incurs a modest increase in context length, it enables the model to perceive temporal information more effectively and precisely, thereby facilitating time-aware video tasks such as video grounding and dense captioning.

3 Pre-Training

3.1 Training Recipe

We first enhance the vision encoder by conducting continuous training with dynamic resolutions based on the pre-trained SigLIP-2 model. The overall Qwen3-VL model adopts a three-module architecture, comprising this vision encoder, an MLP-based vision–language merger, and a Qwen3 large language model (LLM) backbone. Building on this architecture, our pre-training methodology is systematically structured into four distinct stages, designed to progressively build capabilities from basic alignment to long-context understanding. An overview of these stages is presented in Table 1.

Table 1: Training setup and hyperparameters across different stages for Qwen3-VL.

Stage	Objective	Training	Token Budget	Sequence Length
S0	Vision-Language Alignment	Merger	67B	8,192
S1	Multimodal Pre-Training	All	~1T	8,192
S2	Long-Context Pre-Training	All	~1T	32,768
S3	Ultra-Long-Context Adaptation	All	100B	262,144

Stage 0: Vision-Language Alignment. The initial stage (S0) focuses on efficiently bridging the modality gap between the vision encoder and the LLM. Crucially, only the parameters of the MLP merger are trained during this phase, while both the vision encoder and the LLM backbone remain frozen. We utilize a curated dataset of approximately 67B tokens, consisting of high-quality image-caption pairs, visual knowledge collections, and optical character recognition (OCR) data. All training is conducted with a sequence length of 8,192. This alignment-first approach establishes a solid foundation for cross-modal understanding before proceeding to full-parameter training.

Stage 1: Multimodal Pre-Training. Following the initial alignment, Stage 1 (S1) transitions to full-parameter Multimodal Pre-Training. In this phase, we unfreeze all model components—the vision encoder, the merger, and the LLM—for joint end-to-end training. The model is trained on a massive and diverse dataset of approximately 1 trillion (1T) tokens. To maintain the LLM's strong language abilities, the data mixture is composed of vision-language (VL) data and text-only data. The VL portion is rich and varied, adding interleaved image-text documents, visual grounding tasks, visual question

2.2 DeepStack

我们从 DeepStack (孟等人， 2024) 中汲取灵感，并将视觉标记注入到 LLM 的多个层中。与原始 DeepStack 方法不同，该方法堆叠来自多尺度视觉输入的标记，我们将 DeepStack 扩展到从视觉 Transformer (ViT) 的中间层中提取视觉标记。这种设计保留了丰富的视觉信息，范围从低级到高级表示。

具体来说，如图 1所示，我们从视觉编码器的三个不同层级选择特征。随后，专门的视觉-语言合并模块将这些多级特征投影到视觉标记中，这些标记随后被直接添加到前三个大语言模型层的对应隐藏状态中。

2.3 视频时间戳

在Qwen2.5-VL中，采用MRoPE的时间同步变体来赋予模型时间感知能力。然而，我们识别出这种方法的两个关键局限性：(1) 通过将时间位置ID直接与绝对时间绑定，该方法为长视频生成了过大且稀疏的时间位置ID，降低了模型理解长时序上下文的能力。(2) 在此方案下，有效学习需要跨各种帧率（fps）进行广泛且均匀分布的采样，显著增加了训练数据构建的成本。

为解决这些问题，我们采用基于文本标记的时间编码策略（Chen等人， 2024b），其中每个视频时间块前缀有一个表示为格式化文本字符串的时间戳——例如，<30秒>。此外，在训练过程中，我们生成秒和HMS（小时:分钟:秒）格式的时间戳，以确保模型学习解释多样化的时间码表示。尽管这种方法导致上下文长度略有增加，但它使模型能够更有效地感知时间信息，从而促进视频时间感知任务，如视频Grounding和密集描述。

3 预训练

3.1 训练配方

我们首先通过基于预训练的SigLIP-2模型进行动态分辨率连续训练来增强视觉编码器。整体Qwen3-VL模型采用三模块架构，包括该视觉编码器、基于MLP的视觉-语言合并模块以及Qwen3大语言模型（LLM）主干。在此架构基础上，我们的预训练方法系统性地分为四个不同阶段，旨在从基础对齐逐步构建到长上下文理解能力。这些阶段的概述见表1。

表1：Qwen3-VL在不同阶段的训练设置和超参数。

阶段	目标	训练	令牌预算	序列长度
S0	视觉语言对齐	合并	67B	8192
S1	多模态预训练	All	~1T	8192
S2	长上下文预训练	All	~1T	32768
S3	超长上下文适配	All	100B	262144

阶段 0：视觉语言对齐。 初始阶段（S0）专注于高效地弥合视觉编码器与大语言模型之间的模态差距。关键在于，在此阶段仅训练 MLP 合并器的参数，而视觉编码器和大语言模型主干则保持冻结。我们使用一个包含约 67B 个 token 的精选数据集，其中包含高质量的图像-文本对、视觉知识集合和光学字符识别（OCR）数据。所有训练均以 8,192 的序列长度进行。这种先对齐的方法在进入全参数训练之前，为跨模态理解奠定了坚实的基础。

第一阶段：多模态预训练。 在初始对齐之后，第一阶段（S1）过渡到全参数多模态预训练。在此阶段，我们解冻所有模型组件——视觉编码器、合并器和LLM——进行联合端到端训练。模型在约1万亿（1T）token的庞大且多样化的数据集上进行训练。为了保持LLM强大的语言能力，数据混合由视觉语言（VL）数据和纯文本数据组成。VL部分丰富多样，包括交错的图像-文本文档、视觉Grounding任务、视觉问答（VQA）、STEM领域的来自数据以及少量视频数据以引入时间理解。序列长度保持为8,192。

answering (VQA), data from STEM domains, and a small amount of video data to introduce temporal understanding. The sequence length remains at 8,192.

Stage 2: Long-Context Pre-Training. Stage 2 (S2) aims to significantly extend the model's contextual processing abilities. A key change in this stage is the quadrupling of the sequence length to 32,768, while all model parameters continue to be trainable. Training is conducted on a dataset of approximately 1T tokens, with an adjusted data mixture to support long-context tasks. The proportion of text-only data is increased to bolster long-form text comprehension, while the remaining VL data incorporates a significantly larger volume of video and agent-oriented instruction-following data. This stage is critical for enabling the model to process and reason over longer videos and complex, multi-step tasks.

Stage 3: Ultra-Long-Context Adaptation. The final stage (S3) is a specialized phase designed to push the model's context window to its operational limits. Here, we dramatically increase the sequence length to 262,144. The model is trained on a more focused 100B token dataset specifically curated for this purpose. The data is also composed of text-only data and VL data, with a strong emphasis on long-video and long-document understanding tasks. This final adaptation solidifies Qwen3-VL's proficiency in processing and analyzing extremely long sequential inputs, a key capability for applications like comprehensive document analysis and lengthy video summarization.

3.2 Pre-Training Data

3.2.1 Image Caption and Interleaved Text-Image Data

To build a robust foundation model for general-purpose vision–language understanding, we significantly expand and refine two core data modalities: image–caption pairs and interleaved text–image sequences. Our strategy emphasizes high-quality, diverse, and semantically rich multimodal grounding, supported by purpose-built models and rigorous filtering pipelines.

Image Caption Data: We curate a large-scale corpus of contemporary, predominantly Chinese–English multilingual image–text pairs from web sources and apply a multi-stage refinement pipeline centered on a specialized Qwen2.5-VL-32B model fine-tuned for recaptioning. This model leverages the original raw text associated with each image to generate more comprehensive, fluent, and fine-grained captions—enriching descriptions of visual elements (e.g., object attributes, spatial layouts, and contextual semantics) while simultaneously improving the linguistic quality and informativeness of the textual component.

Deduplication is performed exclusively on the recaptioned text using semantic similarity metrics, ensuring removal of redundant samples without sacrificing visual diversity. To further enhance coverage of underrepresented concepts, we apply clustering (Johnson et al., 2019; Douze et al., 2024; Diao et al., 2025) over visual embeddings to identify sparse regions in the data distribution and perform targeted augmentation. The result is a high-fidelity caption dataset that balances scale, diversity, and descriptive granularity.

Interleaved Text-Image Data: We collect diverse real-world multimodal documents sourced from recent Chinese and English websites (Laurençon et al., 2023; Zhu et al., 2023; Li et al., 2024c). All documents undergo domain classification (Wettig et al., 2025) using a lightweight Qwen-based scorer fine-tuned for fine-grained domain identification. Based on validation experiments across domains, we systematically exclude harmful or low-value categories—such as advertisements, promotional content, and clickbait—using the same efficient scorer to filter out undesirable samples.

For book-scale interleaved data, we employ a fine-tuned Qwen2.5-VL-7B model to perform high-accuracy multimodal parsing, precisely extracting and aligning text with embedded figures, diagrams, and photographs. To enable ultra-long context modeling, we construct a specialized subset by merging consecutive pages into sequences of up to 256K tokens, preserving natural page order and multimodal coherence. During preprocessing, we enforce strict quality controls: (i) pure-text or low-alignment segments are removed; (ii) for ultra-long book sequences, we require a minimum page count and a minimum image-to-text ratio to ensure meaningful visual–textual interaction throughout the context. This yields a clean, diverse, and layout-aware interleaved corpus optimized for both grounded understanding and long-range multimodal reasoning.

3.2.2 Knowledge

World knowledge is essential for multimodal large language models (MLLMs) to achieve robust visual understanding, grounded reasoning, and entity-aware generation across diverse downstream tasks. To equip Qwen3-VL with a comprehensive grasp of both real-world and fictional concepts, we construct a

answering（VQA）、STEM领域的来自数据，以及少量视频数据以引入时间理解。序列长度保持为8,192。

阶段 2：长上下文预训练。 阶段 2 (S2) 旨在显著扩展模型上下文处理能力。本阶段的一个关键变化是将序列长度增加到 32,768，同时所有模型参数继续可训练。训练在一个包含约 1T token 的数据集上进行，数据混合进行了调整以支持长上下文任务。文本数据比例增加以增强长文本理解能力，而其余的视觉语言 (VL) 数据则包含显著更多的视频和面向代理的指令遵循数据。此阶段对于使模型能够处理和推理更长的视频以及复杂的多步骤任务至关重要。

阶段 3：超长上下文适应。 最后阶段（S3）是一个专门设计的阶段，旨在将模型的上下文窗口推向其操作极限。在这里，我们将序列长度显著增加到 262,144。模型在一个更专注的 100B token 数据集上进行训练，该数据集专门为此目的而策划。数据也由纯文本数据和 VL 数据组成，特别强调长视频和长文档理解任务。这种最后的适应巩固了 Qwen3-VL 在处理和分析极长序列输入方面的能力，这对于综合文档分析和长视频摘要等应用来说是关键能力。

3.2 预训练数据

3.2.1 图像描述与交错文本-图像数据

为了构建一个用于通用视觉-语言理解的稳健基础模型，我们显著扩展并优化了两种核心数据模态：图像-描述对和交错文本-图像序列。我们的策略强调高质量、多样化且语义丰富的多模态Grounding，由专门构建的模型和严格的过滤流程支持。

图像描述数据：我们从网络来源精选了一个大规模的当代、以中文-英语为主的多元语言图像-文本对语料库，并应用了一个以专门为重述描述微调的Qwen2.5-VL-32B模型为中心的多阶段优化流程。该模型利用与每张图像关联的原始文本来生成更全面、流畅和细粒度的描述——丰富视觉元素（例如对象属性、空间布局和上下文语义）的描述，同时同时提高文本组件的语言质量和信息量。

去重操作仅对重新生成的文本使用语义相似度指标进行，确保移除冗余样本而不牺牲视觉多样性。为进一步增强代表性不足的概念覆盖率，我们应用聚类（Johnson 等人， 2019； Douze 等人， 2024； Diao 等人,2025）对视觉嵌入进行，以识别数据分布中的稀疏区域并执行针对性增强。结果是平衡规模、多样性和描述粒度的、高保真度标题数据集。

交错文本-图像数据：我们从近期的中文和英文网站收集多样化的真实世界多模态文档（Laurençon 等人， 2023； Zhu 等人， 2023； Li 等人， 2024c）。所有文档使用轻量级的基于Qwen的评分器进行领域分类（ Wettig 等人， 2025），该评分器针对细粒度领域识别进行了微调。基于跨领域的验证实验，我们系统地排除有害或低价值类别——如广告、推广内容和标题党——使用相同的效率评分器过滤掉不受欢迎的样本。

对于书规模交错数据，我们采用微调后的Qwen2.5-VL-7B模型进行高精度多模态解析，精确提取并对接文本与嵌入的图像、图表和照片。为支持超长上下文建模，我们将连续页面合并为最多256K个token的序列，保留自然的页面顺序和多模态连贯性。在预处理阶段，我们实施严格的质量控制：(i) 移除纯文本或低对齐片段；(ii) 对于超长的书籍序列，我们要求最低页数和最低图文比，以确保在整个上下文中实现有意义的视觉-文本交互。这产生了一个干净、多样且感知布局的交错语料库，优化了基于理解和长距离多模态推理的性能。

3.2.2 知识

世界知识对于多模态大语言模型（MLLMs）实现稳健的视觉理解、基于知识的推理以及在多样化下游任务中实现实体感知生成至关重要。为了使Qwen3-VL全面掌握现实世界和虚构概念，我们构建了一个以明确定义的实体为中心的大规模预训练数据集，涵盖超过十几个语义类别——包括动物、植物、地标、食物以及日常对象，如车辆、电子产品和服装。

large-scale pretraining dataset centered on well-defined entities spanning more than a dozen semantic categories—including animals, plants, landmarks, food, and everyday objects such as vehicles, electronics, and clothing.

Real-world entities follow a long-tailed distribution: prominent concepts appear frequently with high-quality annotations, while the majority are rare. To address this imbalance, we adopt an importance-based sampling strategy. High-prominence entities are sampled more heavily to ensure a sufficient learning signal, while low-prominence entities are included in smaller proportions to maintain broad coverage without overwhelming the training process. This approach effectively balances data quality, utility, and diversity.

All retained samples undergo a multi-stage refinement pipeline. In addition to standard filtering for noise and misalignment, we replace original or sparse captions—such as generic alt-text—with richer, LLM-generated descriptions. These enhanced captions not only identify the main entity but also describe its visual attributes, surrounding context, spatial layout, and interactions with other objects or people, thereby providing a more complete and grounded textual representation.

Together, these efforts yield a knowledge-rich, context-aware, and discrimination-focused training signal that significantly enhances Qwen3-VL's ability to recognize, reason about, and accurately describe visual concepts in real-world scenarios.

3.2.3 OCR, Document Parsing and Long Document Understanding

OCR: To enhance OCR performance on real-world images, we curate a dataset of 30 million in-house collected samples using a coarse-to-fine pipeline. This pipeline refines OCR annotations by integrating pseudo-labels from OCR-specialized models with refinements from Qwen2.5-VL—without any human annotation. Expanding beyond the 10 languages supported by Qwen2.5-VL (excluding Chinese and English), we incorporate an additional 29 languages, synthesizing approximately 30 million high-quality multilingual OCR samples and curating over 1 million internal real-world multilingual images.

Document Parsing: For document parsing, we collect 3 million PDFs from Common Crawl, evenly distributed across 10 document types (300K samples each), along with 4 million internal documents. An in-house layout model first predicts the reading order and bounding boxes for textual and non-textual regions; Qwen2.5-VL-72B then performs region-specific recognition. The outputs are reassembled into position-aware, layout-aligned parsing data.

To ensure robust parsing across heterogeneous formats, we design a unified annotation framework supporting two representations:

- QwenVL-HTML, which includes fine-grained, element-level bounding boxes;
- QwenVL-Markdown, where only images and tables are localized, with tables encoded in LaTeX.

We construct a large-scale synthetic HTML corpus with precise annotations and systematically convert it to Markdown format. To further improve model generalization, we generate pseudo-labels on extensive collections of real documents and filter them for quality. The final training set combines synthetic and high-quality pseudo-labeled data to enhance both scalability and robustness.

Long Document Understanding: To enhance the model's ability to understand multi-page PDFs—often spanning dozens of pages—we leverage a large-scale corpus of long-document data. First, we synthesize long-document parsing sequences by merging single-page document samples. In each sequence, multiple page images are placed at the beginning, followed by their corresponding text derived from OCR or HTML parsing. Second, we construct long-document visual question answering (VQA) data. Specifically, we sample high-quality multi-page PDFs and generate a diverse set of VQA examples that require the model to reason across multiple pages and heterogeneous document elements—such as charts, tables, figures, and body text. We carefully balance the distribution of question types and ensure that supporting evidence draws from a wide range of modalities and layout components, thereby promoting robust, grounded, and multi-hop reasoning over extended contexts.

3.2.4 Grounding and Counting

Visual grounding is a fundamental capability for multimodal models, enabling them to accurately identify, interpret, and localize a wide spectrum of visual targets from specific objects to arbitrary image regions. In Qwen3-VL, we systematically enhance grounding proficiency and support two grounding modalities: bounding boxes and points. These representations allow for precise and flexible interpretation of image content across diverse scenarios and downstream tasks. In addition, we extend the grounding capacity of

以明确定义的实体为中心的大规模预训练数据集，涵盖超过十几个语义类别——包括动物、植物、地标、食物以及日常对象，如车辆、电子产品和服装。

现实世界中的实体遵循长尾分布：重要的概念频繁出现且具有高质量的标注，而大多数实体则较为罕见。为了解决这种不平衡，我们采用基于重要性的采样策略。高重要性的实体会被更多地采样，以确保足够的学习信号，而低重要性的实体则以较小的比例被包含，以保持广泛的覆盖范围，同时避免过度冲击训练过程。这种方法有效地平衡了数据质量、效用和多样性。

全部保留的样本都经过多阶段精炼流程。除了标准的噪声和错位过滤外，我们还替换原始或稀疏的标题—例如通用替代文本—with 更丰富的大语言模型生成的描述。这些增强的标题不仅识别主要实体，还描述其视觉属性、周围环境、空间布局以及与其他对象或人的交互，从而提供更完整和基于文本的表示。

一起，这些努力产生了一个知识丰富、上下文感知和侧重于区分的训练信号，显著增强了Qwen3-VL在现实场景中识别、推理和准确描述视觉概念的能力。

3.2.3 OCR，文档解析和长文档理解

OCR: 为提升真实图像上的OCR性能，我们使用粗粒度到细粒度的流程，精选了一个包含3000万内部收集样本的数据集。该流程通过整合OCR专业模型的伪标签与Qwen2.5-VL的优化，对OCR标注进行精细化处理——无需人工标注。在Qwen2.5-VL支持的10种语言（不包括中文和英文）基础上，我们进一步整合了29种额外语言，合成约3000万高质量多语言OCR样本，并精选了超过100万张内部真实多语言图像。

文档解析： 对于文档解析，我们从 Common Crawl 收集了 300 万份 PDF 文件，均匀分布在 10 种文档类型中（每种 30 万份样本），此外还有 400 万份内部文档。内部布局模型首先预测文本和非文本区域的阅读顺序和边界框；Qwen2.5-VL-72B 然后执行区域特定的识别。输出被重新组装为位置感知、布局对齐的解析数据。

为确保跨异构格式实现稳健解析，我们设计了一个统一的标注框架，支持两种表示方式：

- QwenVL-HTML，其中包括细粒度的、元素级别的边界框；
- QwenVL-Markdown，其中仅对图像和表格进行定位，表格以LaTeX格式编码。

我们构建了一个大规模的合成HTML语料库，具有精确的标注，并将其系统地转换为Markdown格式。为进一步提升模型的泛化能力，我们在大量真实文档集合上生成伪标签，并筛选以保证质量。最终的训练集结合了合成数据和高质量伪标注数据，以增强可扩展性和稳健性。

长文档理解：为增强模型理解多页PDF（通常跨越数十页）的能力，我们利用了一个大规模的长文档数据集。首先，我们通过合并单页文档样本来合成长文档解析序列。在每个序列中，多个页面图像被放置在开头，随后是其对应的从OCR或HTML解析中得到的文本。其次，我们构建了长文档视觉问答（VQA）数据。具体而言，我们采样高质量的、多页PDF，并生成多样化的VQA示例，要求模型跨多页和异构文档元素（如图表、表格、图形和正文）进行推理——例如，图表、表格、图形和正文。我们仔细平衡问题类型的分布，并确保支持性证据来自广泛的模态和布局组件，从而促进在扩展上下文中的稳健、有据和多跳推理。

3.2.4 Grounding和计数

视觉 grounding 是多模态模型的一项基本能力，使其能够准确识别、解释和定位从特定对象到任意图像区域的广泛视觉目标。在 Qwen3-VL 中，我们系统地提升了 grounding 能力，并支持两种 grounding 模式：边界框和点。这些表示方法允许在各种场景和下游任务中精确且灵活地解释图像内容。此外，我们扩展了模型的 grounding 能力，以支持计数，从而实现对视觉实体的定量推理。在下文中，我们将简要描述 grounding 和计数的构建数据管道。

the model to support counting, enabling quantitative reasoning about visual entities. In the following, we briefly describe the data construction pipelines for grounding and counting.

Box-based Grounding: We begin by aggregating widely used open-source datasets, including COCO (Lin et al., 2014), Objects365 (Shao et al., 2019), OpenImages (Kuznetsova et al., 2020), and RefCOCO/+/g (Kazemzadeh et al., 2014; Mao et al., 2016). To further enrich data diversity, we developed an automated synthesis pipeline that generates high-quality object annotations across a broad range of scenarios. This pipeline operates in three stages: (i) object candidates are extracted from unlabeled images using Qwen2.5-VL; (ii) these candidates are localized and annotated using both open-vocabulary detectors (specifically, Grounding DINO (Liu et al., 2023a)) and Qwen2.5-VL; and (iii) the resulting annotations undergo quality assessment, with low-confidence or inaccurate ones systematically filtered out. Through this approach, we constructed a large-scale, highly diverse box-based grounding dataset spanning a wide variety of visual contexts and object categories.

Point-based Grounding: To ensure robust point-based grounding, we curated a comprehensive dataset combining publicly available and synthetically generated pointing annotations. It integrates three sources: (i) public pointing and counting annotations from PixMo (Deitke et al., 2024); (ii) object grounding data derived from public object detection and instance segmentation benchmarks; and (iii) high-precision pointing annotations generated by a dedicated synthesis pipeline designed to target fine-grained image details.

Counting: Building upon the grounding data, we curated a high-quality subset to form the basis of our counting dataset, which includes three distinct task formulations: direct counting, box-based counting, and point-based counting. Collectively, these three task types constitute a comprehensive counting dataset.

Different from Qwen2.5-VL, we adopt a normalized coordinate system scaled to the range $[0, 1000]$ in this version. This design improves robustness to variations in image resolution and aspect ratio across diverse inputs, while also simplifying post-processing and enhancing the usability of predicted coordinates in downstream applications.

3.2.5 Spatial Understanding and 3D Recognition

To facilitate sophisticated interaction with the physical world, Qwen3-VL is designed with a deep understanding of spatial context. This enables the model to interpret spatial relationships, infer object affordances, and perform action planning and embodied reasoning. It can also estimate the 3D spatial positions of objects from a single monocular image. To support these capabilities, we created two comprehensive datasets focused on Spatial Understanding and 3D Grounding.

Spatial Understanding. Beyond localizing objects, Qwen3-VL is trained to reason about spatial relationships, object affordances, and feasible actions in 2D scenes—capabilities essential for embodied AI and interactive applications. To this end, we construct a specialized dataset that goes beyond standard grounding by incorporating: (i) relational annotations (e.g., "the cup to the left of the laptop"), (ii) affordance labels (e.g., "graspable", "pressable", "sittable"), and (iii) action-conditioned queries that require planning (e.g., "What should I move first to reach the book behind the monitor?"). These samples are derived from both curated real-world scenes and synthetically generated layouts, with natural language queries automatically generated via templated and LLM-based methods to ensure diversity and complexity. Critically, all spatial references are expressed relative to other objects or scene frames, rather than absolute coordinates, encouraging robust relational reasoning. This training enables Qwen3-VL to not only answer "where" questions but also "how" and "what can be done" — forming a foundation for agentic interaction with visual environments.

3D Grounding. To further enhance the model's ability to understand the physical world from images, we constructed a specialized pretraining dataset for 3D visual grounding. We sourced data from public collections of diverse indoor and outdoor scenes and reformulated it into a visual question-answering format. Each sample consists of: 1) a single-view camera image, 2) a natural language referring expression, and 3) the corresponding 9-DoF 3D bounding box annotations in a structured JSON format, specifying the object's spatial position and semantic label. As the 3D bounding boxes are derived from multiple sensors and data sources, they exhibit varying camera intrinsic parameters and inherent noise. To this end, we filter out heavily occluded and inaccurate labels and follow Omni3D (Brazil et al., 2023) to unify all data into a virtual camera coordinate system. We also synthesized a large corpus of descriptive captions to create rich textual queries for 3D grounding. These descriptions go beyond naming the object's category to include detailed attributes, layout arrangements, spatial location, visual affordances, and interactions with surrounding objects—yielding more fine-grained and grounded referring expressions.

模型的 grounding 能力扩展到支持计数，从而实现对视觉实体的定量推理。在下文中，我们将简要描述 grounding 和计数的构建数据管道。

基于框的Grounding：我们首先汇集了广泛使用的开源数据集，包括COCO（Lin等人， 2014），Objects365（Shao等人， 2019），OpenImages（Kuznetsova等人， 2020），以及RefCOCO/+/g（Kazemzadeh等人， 2014； Mao等人，2016）。为了进一步丰富数据多样性，我们开发了一个自动合成管道，该管道可以在各种场景中生成高质量的物体标注。该管道分为三个阶段：(i) 使用Qwen2.5-VL从无标签图像中提取物体候选；(ii) 使用开放词汇检测器（特别是Grounding DINO（Liu等人， 2023a））和Qwen2.5-VL对候选进行定位和标注；(iii) 对生成的标注进行质量评估，系统地过滤掉低置信度或不准确的标注。通过这种方法，我们构建了一个大规模、高度多样化的基于框的Grounding数据集，涵盖了广泛的各种视觉场景和物体类别。

基于点的 Grounding：为确保基于点的 Grounding 的鲁棒性，我们精心策划了一个综合数据集，该数据集结合了公开可用的和合成生成的指向标注。它整合了三个来源：(i) 来自 PixMo 的公开指向和计数标注（Deitke 等人， 2024）；(ii) 来自公开的物体检测和实例分割基准测试的物体 Grounding 数据；(iii) 由专门设计的合成管道生成的用于针对细粒度图像细节的高精度指向标注。

计数：在 Grounding 数据的基础上，我们精心策划了一个高质量子集，以此作为我们计数数据集的基础，该数据集包括三种不同的任务公式：直接计数、基于框的计数和基于点的计数。这些三种任务类型共同构成一个全面的计数数据集。

与 Qwen2.5-VL 不同，在本版本中我们采用了一个归一化坐标系，其范围缩放为 $[0, 1000]$。这种设计提高了对图像分辨率和不同输入的纵横比变化的鲁棒性，同时简化了后处理，并增强了预测坐标在下游应用中的可用性。

3.2.5 空间理解与 3D 识别

为促进与物理世界的复杂交互，Qwen3-VL 经过设计，具备对空间上下文深度理解。这使得模型能够解释空间关系、推断对象可供性，并执行行动规划与具身推理。它还能从单目图像中估计对象的 3D 空间位置。为支持这些能力，我们创建了两个专注于空间理解与 3D Grounding 的综合数据集。

空间理解。除了解析对象位置外，Qwen3-VL 还被训练用于推理 2D 场景中的空间关系、对象可供性及可行行动——这些能力对于具身 AI 和交互式应用至关重要。为此，我们构建了一个超越标准 Grounding 的专用数据集，通过整合：(i) 关系性标注（例如，“笔记本电脑左侧的杯子”）、(ii) 可供性标签（例如，“可抓取”、“可按压”、“可坐下”）以及 (iii) 需要规划的行动条件查询（例如，“我应先移动什么才能拿到显示器后面的书？”）。这些样本源自精选的真实世界场景和合成生成的布局，通过模板化和基于大语言模型的方法自动生成自然语言查询，以确保多样性和复杂性。关键在于，所有空间参考都是相对于其他对象或场景框架表达的，而非绝对坐标，从而鼓励鲁棒的关系性推理。这种训练使 Qwen3-VL 不仅能够回答“在哪里”的问题，还能回答“如何”以及“能做什么”——为与视觉环境的代理式交互奠定基础。

3D边界框。为了进一步提升模型从图像中理解物理世界的能力，我们构建了一个专门用于3D视觉边界框的预训练数据集。我们从公开的多样化室内和室外场景集合中获取数据，并将其重新格式化为视觉问答格式。每个样本包含：1) 单视图相机图像，2) 自然语言指代表达式，以及3) 对应的9-DoF 3D边界框标注，以结构化的JSON格式指定对象的空间位置和语义标签。由于3D边界框来自多个传感器和数据源，它们表现出不同的相机内参和固有噪声。为此，我们过滤掉了严重遮挡和不准确的标签，并遵循 Omni3D（巴西等人。， 2023）将所有数据统一到虚拟相机坐标系中。我们还合成了一大批描述性字幕，以创建丰富的文本查询用于3D边界框。这些描述不仅包括命名对象的类别，还包括详细属性、布局排列、空间位置、视觉可供性和与周围对象的交互——从而产生更细粒度和更精确的指代表达式。

3.2.6 Code

We enhance the Qwen3-VL series with dedicated coding capabilities by incorporating two categories of code-related data into the training corpus, enabling the model to read, write, and reason about programs in both text-only and visually grounded contexts.

Text-Only Coding. We reuse the extensive code corpus from the Qwen3 and Qwen3-Coder series. This large-scale dataset spans a wide range of programming languages and domains—including software development, algorithmic problem solving, mathematical reasoning, and agent-oriented tasks—and establishes the model's foundational understanding of code syntax, algorithmic logic, and general-purpose program generation.

Multimodal Coding. To address tasks requiring both visual understanding and code generation, we curate data for a diverse suite of multimodal coding tasks. This dataset, sourced from both open-source datasets and internal synthesis pipelines, teaches the model to jointly understand visual inputs and generate functional code. The data covers several key tasks, including: converting UI screenshots into responsive HTML/CSS; generating editable SVG codes from images (Li et al., 2025c); solving visual programming challenges (Li et al., 2024a); answering multimodal coding questions (e.g., StackOverflow posts with images); and transcribing visual representations (such as flowcharts, diagrams, and LaTeX equations) into their respective code or markup. This novel data mixture enables Qwen3-VL to act as a bridge between visual perception and executable logic.

3.2.7 Video

The video comprehension capabilities of Qwen3-VL have been substantially advanced, enabling robust modeling of temporal dynamics across frames, fine-grained perception of spatial relationships, and coherent summarization of ultra-long video sequences. This enhancement is underpinned by a data processing pipeline featuring two principal innovations:

Temporal-Aware Video Understanding. (i) Dense Caption Synthesis: For long video sequences, we employ a short-to-long caption synthesis strategy to generate holistic, timestamp-interleaved, and temporally coherent story-level descriptions. Leveraging in-house captioning models, we further produce fine-grained annotations that jointly capture event-level temporal summaries and segment-specific visual details. (ii) Spatio-Temporal Video Grounding: We curate and synthesize large-scale video data annotated at the levels of objects, actions, and persons to strengthen the model's spatio-temporal grounding capabilities, thereby improving its capacity for fine-grained video understanding.

Video Data Balancing and Sampling. (i) Source Balancing: To ensure data balance and diversity, we assemble a large-scale dataset encompassing various video sources, including instructional content, cinematic films, egocentric recordings, etc. Dataset balance is achieved through systematic curation guided by metadata such as video titles, duration, and categorical labels. (ii) Length-Adaptive Sampling: During pre-training stages, we dynamically adjust sampling parameters, such as frames per second (fps) and the maximum number of frames, according to different sequence length constraints. This adaptive strategy mitigates information loss associated with suboptimal sampling practices (e.g., overly sparse frame selection or excessively low spatial resolution), thus preserving visual details and optimizing training efficacy.

3.2.8 Science, Technology, Engineering, and Mathematics (STEM)

Multimodal reasoning lies at the heart of Qwen3-VL, with STEM reasoning constituting its most essential part. Our philosophy follows a divide-and-conquer strategy: we first develop fine-grained visual perception and robust linguistic reasoning capabilities independently, and then integrate them in a synergistic manner to achieve effective multimodal reasoning.

Visual Perception Data. We develop a dedicated synthetic data generation pipeline that constructs geometric diagrams through programmatic (code-based) rendering. Using this pipeline, we generate: (i) 1 million point-grounding samples, such as intersection points, corners, and centers of gravity; and (ii) 2 million perception-oriented visual question answering pairs targeting fine-grained visual understanding of diagrams. To obtain high-fidelity textual descriptions, we further implement a two-stage captioning framework: an initial generation phase followed by rigorous model-based verification. Both stages employ ensembles of specialized models to ensure accuracy and descriptive granularity. This process yields a comprehensive dataset of 6 million richly annotated diagram captions spanning diverse STEM disciplines.

Multi-modal Reasoning Data. The majority of our multi-modal reasoning data consists of over 60

3.2.6 代码

我们通过将两类与代码相关的数据纳入训练语料库，增强了Qwen3-VL系列的专业编程能力，使模型能够在纯文本和视觉边界框两种上下文中读取、编写和推理程序。

文本仅编程。 我们重用了来自 Qwen3 和 Qwen3-Coder 系列的大量代码语料库。这个大规模数据集涵盖了多种编程语言和领域——包括软件开发、算法问题解决、数学推理和面向代理的任务——并建立了模型对代码语法、算法逻辑和通用程序生成的基础理解。

多模态编程。 为了应对需要视觉理解和代码生成相结合的任务，我们为一系列多模态编程任务收集了数据。该数据集来源于开源数据集和内部合成管道，教导模型联合理解视觉输入并生成功能性代码。数据涵盖了几个关键任务，包括：将UI截图转换为响应式HTML/CSS；从图像生成可编辑的SVG代码（Li等人。， 2025c）；解决视觉编程挑战（Li等人。,2024a）；回答多模态编程问题（例如，带有图像的StackOverflow帖子）；以及将视觉表示（如图表、流程图和LaTEX方程）转录为其相应的代码或标记。这种新颖的数据混合使Qwen3-VL能够充当视觉感知和可执行逻辑之间的桥梁。

3.2.7 视频

Qwen3-VL 的视频理解能力已大幅提升，能够对跨帧的时序动态进行稳健建模，对空间关系进行细粒度感知，并能对超长视频序列进行连贯的摘要。这一增强得益于一个包含两个主要创新的数据处理流程：

时序感知视频理解。 (i) 密集式字幕合成：对于长视频序列，我们采用短到长的字幕合成策略，生成整体性、时间戳交错且时序连贯的故事级描述。利用内部字幕模型，我们进一步生成细粒度标注，联合捕捉事件级时序摘要和片段级视觉细节。(ii) 时空视频 Grounding：我们策展和合成大规模在对象、行动和人物层级标注的视频数据，以增强模型的时空 Grounding 能力，从而提升其细粒度视频理解能力。

视频数据平衡和采样。 (i) 源平衡：为确保数据平衡和多样性，我们汇编了一个包含多种视频来源的大规模数据集，包括教学内容、电影、第一人称记录等。通过基于视频标题、时长和分类标签等元数据的系统策展，实现数据集平衡。(ii) 长度自适应采样：在预训练阶段，我们根据不同的序列长度约束动态调整采样参数，如每秒帧数 (fps) 和最大帧数。这种自适应策略减轻了与次优采样实践相关的信息损失（例如，过于稀疏的帧选择或过低的空间分辨率），从而保留视觉细节并优化训练效能。

3.2.8 科学、技术、工程和数学（STEM）

多模态推理是Qwen3-VL的核心，其中STEM推理是其最关键的部分。我们的理念遵循分而治之的策略：首先独立开发细粒度视觉感知和鲁棒的语言推理能力，然后以协同的方式将它们整合起来，以实现有效的多模态推理。

视觉感知数据。我们开发了一个专用的合成数据生成管道，通过程序化（基于代码）的渲染来构建几何图形。使用该管道，我们生成了：(i) 100万点 grounding 样本，例如交点、角和质心；(ii) 200万面向感知的视觉问答对，旨在对图表进行细粒度视觉理解。为了获得高保真度的文本描述，我们进一步实现了一个两阶段字幕框架：初始生成阶段和严格的基于模型的验证。这两个阶段都采用专用模型的集成来确保准确性和描述粒度。这个过程产生了一个包含600万个丰富标注的图表字幕的综合数据集，涵盖了不同的STEM学科。

多模 **dal 推理 数据**. 我们的多数多模态推理数据包含超过 6 0

million K–12 and undergraduate-level exercises, meticulously curated through a rigorous cleaning and reformulation pipeline. During quality filtering, we discard low-quality items, including those with corrupted images, irrelevant content, or incomplete or incorrect answers. During the reformulation stage, we translate exercises between Chinese and English and standardize the format of answers—such as step-by-step solution lists, mathematical expressions, and symbolic notations—to ensure consistency and uniform presentation. Regarding long CoT problem-solving data, we synthesize over 12 million multimodal reasoning samples paired with images. To ensure the continuity and richness of the reasoning process, we utilize the original rollouts generated by a strong reasoning model. To guarantee data reliability and applicability, each sample's reasoning trajectory undergoes rigorous validation—combining rule-based checks and model-based verification—and any instances containing ambiguous answers or code-switching are explicitly filtered out. Furthermore, to enhance reasoning quality, we retain only challenging problems via rejection sampling.

Linguistic Reasoning Data. In addition to multimodal reasoning data, we also incorporate reasoning data from Qwen3, as multimodal reasoning capabilities are largely derived from linguistic reasoning competence.

3.2.9 Agent

GUI: To endow Qwen3-VL with agentic capability for autonomous interaction with graphical user interfaces (GUIs), we curate and synthesize large-scale, cross-platform data spanning desktop, mobile, and web environments (Ye et al., 2025; Wang et al., 2025a; Lu et al., 2025). For GUI interface perception, we leverage metadata, parsing tools, and human annotations to construct tasks such as element description, dense captioning, and dense grounding, enabling robust understanding of diverse user interfaces. For agentic capability, we assemble multi-step task trajectories via a self-evolving trajectory-production framework, complemented by targeted human audits; we also carefully design and augment Chain-of-Thought rationales to strengthen planning, decision-making, and reflective self-correction during real-world execution.

Function Calling: For general function calling capabilities with multimodal contexts, we build a multimodal function calling trajectory synthesis pipeline. We first instruct capable models with images to generate user queries and their corresponding function definitions. We then sample model function calls with rationales and synthesize the function responses. This process is repeated until the user's query is judged to be solved. Between each step, trajectories can be filtered out due to formatting errors. Such a pipeline enables us to construct large-scale multimodal function-calling trajectories from vast images, without the need to implement executable functions.

Search: Among the general function calling capabilities, we regard the ability to perform searches as key to facilitating knowledge integration for long-tail entities in real-world scenarios. In this case, we collect multimodal factual lookup trajectories with online image search and text search tools, encouraging the model to perform searches for unfamiliar entities. By doing so, the model learns to gather information from the web to generate more accurate responses.

4 Post-Training

4.1 Training Recipe

Our post-training pipeline is a three-stage process designed to refine the model's instruction-following capabilities, bolster its reasoning abilities, and align it with human preferences. The specific data and methods for each stage are detailed in the subsequent sections.

Supervised Fine-Tuning (SFT). The first stage imparts instruction-following abilities and activates latent reasoning skills. This is conducted in two phases: an initial phase at a 32k context length, followed by an extension to a 256k context window that focuses on long-document and long-video data. To cater to different needs, we bifurcate the training data into standard formats for non-thinking models and Chain-of-Thought (CoT) formats for thinking models, the latter of which explicitly models the reasoning process.

Strong-to-Weak Distillation. The second stage employs knowledge distillation, where a powerful teacher model transfers its capabilities to our student models. Crucially, we perform this distillation using *text-only* data to fine-tune the LLM backbone. This method proves highly effective, yielding significant improvements in reasoning abilities across both text-centric and multimodal tasks.

Reinforcement Learning (RL). The final stage utilizes RL to further enhance model performance and alignment. This phase is divided into Reasoning RL and General RL. We apply large-scale reinforcement

关于长 CoT 问题解决数据，我们合成了超过 12 百万份的多模态推理样本并配以图像。为确保推理过程的连续性和丰富性，我们利用强大推理模型生成的原始输出。为保证数据可靠性和适用性，每个样本的推理轨迹都经过严格验证——结合基于规则的检查和基于模型的验证——并明确过滤掉包含模糊答案或代码转换的实例。此外，为了提升推理质量，我们通过拒绝采样仅保留有挑战性的问题。

语言推理数据。除了多模态推理数据外，我们还整合了来自Qwen3的推理数据，因为多模态推理能力很大程度上源于语言推理能力。

3.2.9 代理

GUI： 为了赋予Qwen3-VL在图形用户界面（GUI）上自主交互的代理能力，我们收集和合成了大量跨平台数据，涵盖桌面、移动和网页环境（Ye等人， 2025； Wang等人， 2025a； Lu等人， 2025）。对于GUI界面感知，我们利用元数据、解析工具和人工标注来构建元素描述、密集式字幕和密集式Grounding等任务，从而实现对多样化用户界面的稳健理解。对于代理能力，我们通过自进化轨迹生成框架组装多步任务轨迹，并辅以定向人工审核；我们还精心设计和增强Chain-of-Thought推理链，以加强在真实世界执行过程中的规划、决策和反思式自我纠正能力。

函数调用： 对于具有多模态上下文的一般函数调用能力，我们构建了一个多模态函数调用轨迹合成管道。我们首先使用图像指令让能够胜任的模型生成用户查询及其相应的函数定义。然后我们采样带有推理的模型函数调用，并合成函数响应。这个过程会一直重复，直到用户的查询被判断为已解决。在每一步之间，由于格式错误，轨迹可能会被过滤掉。这样的管道使我们能够从大量图像中构建大规模的多模态函数调用轨迹，而无需实现可执行的函数。

搜索： 在通用函数调用能力中，我们将执行搜索的能力视为关键，以促进现实场景中长期尾实体的知识整合。在这种情况下，我们收集了使用在线图像搜索和文本搜索工具的多模态事实查找轨迹，鼓励模型对不熟悉的实体进行搜索。通过这样做，模型学习从网络上收集信息以生成更准确的响应。

4 后训练

4.1 训练配方

我们的后训练流程是一个三阶段过程，旨在提升模型的指令遵循能力、增强其推理能力，并使其与人类偏好保持一致。每个阶段的具体数据和方法的细节将在后续章节中详细说明。

监督微调 (SFT)。 第一阶段赋予模型指令遵循能力，并激活潜在的推理技能。这分为两个阶段进行：首先在 32k 的上下文长度下进行初始阶段，然后扩展到 256k 的上下文窗口，专注于长文档和长视频数据。为了满足不同需求，我们将训练数据分为非思考模型的标准化格式和思考模型的思维链 (CoT) 格式，后者明确地对推理过程进行建模。

强到弱蒸馏。 第二阶段采用知识蒸馏，其中强大的教师模型将其能力传递给我们的学生模型。关键在于，我们使用纯文本 数据进行蒸馏，以微调大语言模型 (LLM) 主干。这种方法非常有效，在文本中心和多模态任务上都显著提升了推理能力。

强化对齐 **强化学习 (RL)**。 最终阶段利用RL进一步提升模型性能 d
这一阶段分为推理强化学习和通用强化学习。我们应用大规模强化学习，在涵盖数学、OCR、Grounding、指令遵循等文本和多模态领域进 t
行跨领域学习，以提升更细粒度的能力。

learning across a comprehensive set of text and multimodal domains, including but not limited to math, OCR, grounding, and instruction-following, to improve finer-grained capabilities.

4.2 Cold Start Data

4.2.1 SFT Data

Our principal objective is to endow the model with the capacity to address a wide spectrum of real-world scenarios. Building upon the foundational capabilities of Qwen2.5-VL, which is proficient in approximately eight core domains and 30 fine-grained subcategories, we have strategically expanded its functional scope. This expansion was achieved by integrating insights from community feedback, academic literature, and practical applications, facilitating the introduction of novel capabilities. These include, but are not limited to, spatial reasoning for embodied intelligence, image-grounded reasoning for fine-grained visual understanding, spatio-temporal grounding in videos for robust object tracking, and the comprehension of long-context technical documents spanning hundreds of pages. Guided by these target tasks and grounded in authentic use cases, we systematically curated the SFT dataset through the meticulous selection and synthesis of samples from open-source datasets and web resources. This targeted data engineering effort has been instrumental in establishing Qwen3-VL as a more comprehensive and robust multimodal foundation model.

This dataset comprises approximately 1,200,000 samples, strategically composed to foster robust multimodal capabilities. This collection is partitioned into unimodal and multimodal data, with one-third consisting of text-only entries and the remaining two-thirds comprising image-text and video-text pairs. The integration of multimodal content is specifically designed to enable the model to interpret complex, real-world scenarios. To ensure global relevance, the dataset extends beyond its primary Chinese and English corpora to include a diverse set of multilingual samples, thereby broadening its linguistic coverage. Furthermore, it simulates realistic conversational dynamics by incorporating both single-turn and multi-turn dialogues contextualized within various visual settings, from single-image to multi-image sequences. Crucially, the dataset also features interleaved image-text examples engineered to support advanced agentic behaviors, such as tool-augmented image search and visually-grounded reasoning. This heterogeneous data composition ensures comprehensive coverage and enhances the dataset's representativeness for training generalizable and sophisticated multimodal agents.

Given Qwen3-VL's native support for a 256K token context length, we employ a staged training strategy to optimize for computational efficiency. This strategy comprises two phases: an initial one-epoch training phase with a sequence length of 32K tokens, followed by a second epoch at the full 256K token length. During this latter stage, the model is trained on a curriculum that interleaves long-context inputs with data sampled at the 32K token length. The long-context inputs include materials such as hundreds of pages of technical documents, entire textbooks, and videos up to two hours in duration.

The quality of training data is a critical determinant of the performance of vision-language models. Datasets derived from open-source and synthetic origins are often plagued by substantial variability and noise, including redundant, irrelevant, or low-quality samples. To mitigate these deficiencies, the implementation of a rigorous data filtering protocol is indispensable. Accordingly, our data curation process incorporates a two-phase filtering pipeline: Query Filtering and Response Filtering.

Query Filtering. In this initial phase, we leverage Qwen2.5-VL to identify and discard queries that are not readily verifiable. Queries with ambiguous instructions are minimally revised to enhance clarity while preserving the original semantic intent. Furthermore, web-sourced queries lacking substantive content are systematically eliminated. Crucially, all remaining queries undergo a final assessment of their complexity and contextual relevance, ensuring only appropriately challenging and pertinent samples are retained for the next stage.

Response Filtering. This phase integrates two complementary strategies:

- **Rule-Based Filtering:** A set of predefined heuristics is applied to eliminate responses exhibiting qualitative deficiencies, such as repetition, incompleteness, or improper formatting. To maintain semantic relevance and uphold ethical principles, we also discard any query-response pairs that are off-topic or possess the potential to generate harmful content.
- **Model-Based Filtering:** The dataset is further refined by employing reward models derived from the Qwen2.5-VL series. These models conduct a multi-dimensional evaluation of multimodal question-answering pairs. Specifically: (a) answers are scored against a range of criteria, including correctness, completeness, clarity, and helpfulness; (b) for vision-grounded tasks, the evaluation places special emphasis on verifying the accurate interpretation and utilization of visual information; and (c) this model-based approach enables the detection of subtle issues that typically elude rule-based methods,

学习范围涵盖但不限于数学、OCR、Grounding和指令遵循等文本和多模态领域，以提升更细粒度的能力。

4.2 冷启动数据

4.2.1 SFT数据

我们的主要目标是赋予模型处理广泛真实场景的能力。在Qwen2.5-VL的基础能力之上，该模型擅长大约八个核心领域和30个细粒度子类别，我们通过整合来自社区反馈、学术文献和实际应用的见解，战略性地扩展了其功能范围。这一扩展通过引入新功能实现，包括但不限于：为具身智能的空间推理、基于图像的细粒度视觉理解推理、视频中时空Grounding以实现鲁棒的对象跟踪，以及对跨越数百页的长期技术文档的理解。在这些目标任务指导下，并基于真实用例，我们通过从开源数据集和网络资源中精心选择和合成样本，系统地构建了SFT数据集。这项有针对性的数据工程工作对于将Qwen3-VL建立为一个更全面、更强大的多模态基础模型起到了关键作用。

该数据集包含约1,200,000个样本，策略性地组合以培养强大的多模态能力。该集合分为单模态和多模态数据，其中三分之一由纯文本条目组成，其余三分之二由图像-文本和视频-文本对组成。多模态内容的集成特别设计用于使模型能够解释复杂的现实世界场景。为确保全球相关性，该数据集扩展到其主要的中文和英文语料库之外，包含多样化的多语言样本，从而扩大其语言覆盖范围。此外，它通过结合单轮和多轮对话，在单图像到多图像序列的各种视觉环境中进行模拟，以模拟真实的对话动态。至关重要的是，该数据集还包含交错图像-文本示例，旨在支持高级代理行为，例如工具增强图像搜索和视觉推理。这种异构数据组合确保了全面覆盖，并增强了数据集对训练可推广和复杂多模态代理的代表性和全面性。

鉴于 Qwen3-VL 对 256K 词元上下文长度的原生支持，我们采用分阶段训练策略以优化计算效率。该策略包含两个阶段：首先进行一个序列长度为 32K 词元的单 epoch 训练阶段，随后在完整的 256K 词元长度上进行第二个 epoch。在后者阶段，模型在课程学习上训练，该课程学习交替包含长上下文输入与 32K 词元长度的数据采样。长上下文输入包括数百页技术文档、整本教科书以及最长两小时的视频等材料。

训练数据质量是视觉语言模型性能的关键决定因素。源自开源和合成来源的数据集常存在显著变异性与噪声，包括冗余、无关或低质量样本。为缓解这些缺陷，实施严格的数据过滤协议至关重要。因此，我们的数据管理流程包含两阶段过滤管道：查询过滤和响应过滤。

查询过滤。 在这个初始阶段，我们利用 Qwen2.5-VL 来识别并丢弃难以验证的查询。具有模糊指令的查询会被最小化修改以增强清晰度，同时保留原始的语义意图。此外，缺乏实质性内容的网络来源查询被系统性地消除。关键的是，所有剩余的查询都经过对其复杂性和上下文相关性的最终评估，确保只有适当具有挑战性和相关的样本被保留用于下一阶段。

响应过滤。 此阶段集成了两种互补策略：

- **基于规则的过滤：** 应用一组预定义的启发式规则来消除表现出质量缺陷的响应，例如重复、不完整或不适当的格式。为了保持语义相关性和坚持道德原则，我们还丢弃任何离题或可能产生有害内容的查询-响应对。
- **基于模型的过滤：** 该数据集通过采用源自Qwen2.5-VL系列的奖励模型进一步优化。这些模型对多模态问答对进行多维度评估。具体而言：(a) 评估答案的正确性、完整性、清晰度和有帮助性等标准；(b) 对于基于视觉的任务，评估特别强调验证对视觉信息的准确解读和利用；(c) 这种基于模型的方法能够检测到规则方法通常难以发现的问题，

such as inappropriate language mixing or abrupt stylistic shifts.

This multi-dimensional filtering framework ensures that only data meeting stringent criteria for quality, reliability, and ethical integrity is advanced to the SFT phase.

4.2.2 Long-CoT Cold Start Data

The foundation of our thinking models is a meticulously curated Long Chain-of-Thought (CoT) cold start dataset, engineered to elicit and refine complex reasoning capabilities. This dataset is built upon a diverse collection of queries spanning both pure-text and multimodal data, maintaining an approximate 1:1 ratio between vision-language and text-only samples to ensure balanced skill development.

The multimodal component, while covering established domains such as visual question answering (VQA), optical character recognition (OCR), 2D/3D grounding, and video analysis, places a special emphasis on enriching tasks related to STEM and agentic workflows. This strategic focus is designed to push the model's performance on problems requiring sophisticated, multi-step inference. The pure-text portion closely mirrors the data used for Qwen3, featuring challenging problems in mathematics, code generation, logical reasoning, and general STEM.

To guarantee high quality and an appropriate level of difficulty, we implement a rigorous multi-stage filtering protocol.

- **Difficulty Curation:** We selectively retain instances where baseline models exhibited low pass rates or generated longer, more detailed responses. This enriches the dataset with problems that are genuinely challenging for current models.
- **Multimodal Necessity Filtering:** For vision-language mathematics problems, we introduce a critical filtering step: we discard any samples that our Qwen3-30B-*nothink* model could solve correctly without access to the visual input. This ensures that the remaining instances genuinely necessitate multimodal understanding and are not solvable via textual cues alone.
- **Response Quality Control:** Aligning with the methodology of Qwen3, we sanitize the generated responses. For queries with multiple candidate answers, we first remove those containing incorrect final results. Subsequently, we filter out responses exhibiting undesirable patterns, such as excessive repetition, improper language mixing, or answers that showed clear signs of guessing without sufficient reasoning steps.

This stringent curation process yields a high-quality, challenging dataset tailored for bootstrapping advanced multimodal reasoning.

4.3 Strong-to-Weak Distillation

We adopt the Strong-to-Weak Distillation pipeline as described in Qwen3 to further improve the performance of lightweight models. This distillation process consists of two main phases:

- **Off-policy Distillation:** In the first phase, outputs generated by teacher models are combined to provide response distillation. This helps lightweight student models acquire fundamental reasoning abilities, establishing a strong foundation for subsequent on-policy training.
- **On-policy Distillation:** In the second phase, the student model generates the responses based on the provided prompts. These on-policy sequences are then used for fine-tuning the student model. We align the logits predicted by the student and teacher by minimizing the KL divergence.

4.4 Reinforcement Learning

4.4.1 Reasoning Reinforcement Learning

We train models across a diverse set of text and multimodal tasks, including mathematics, coding, logical reasoning, visual grounding, and visual puzzles. Each task is designed so that solutions can be verified deterministically via rules or code executors.

Data Preparation We curate training data from both open-source and proprietary sources and apply rigorous preprocessing and manual annotation to ensure high-quality RL queries. For multimodal queries, we use a preliminary checkpoint of our most advanced vision–language model (Qwen3-VL-235B-A22B) to sample 16 responses per query; any query for which all responses are incorrect is discarded.

例如不恰当的语言混合或突然的风格转变。

This multi-dimensional 性、可靠性和伦理完整性的数据才会被推进到SFT阶段。
ty性,em>可靠性及伦理完整性/em>被推进到SFT阶段。

4.2.2 长链思维冷启动数据

我们思考模型的基础是一个经过精心策划的长链思维（CoT）冷启动数据集，旨在激发和提升复杂的推理能力。该数据集基于多样化的查询集合，涵盖纯文本和多模态数据，保持视觉-语言样本与纯文本样本的大致1:1比例，以确保均衡的技能发展。

多模态组件虽然涵盖视觉问答（VQA）、光学字符识别（OCR）、2D/3D 定位和视频分析等既定领域，但特别强调丰富与 STEM 和代理式工作流程相关的任务。这种战略重点旨在提升模型在需要复杂、多步推理的问题上的性能。纯文本部分与 Qwen3 使用的数据高度相似，包含数学、代码生成、逻辑推理和通用 STEM 领域的挑战性问题。

为确保高质量和适当的难度水平，我们实施了一套严格的多阶段过滤协议。

- **难度筛选：** 我们选择性地保留那些基线模型通过率低或生成更长、更详细响应的实例。这使数据集充实了当前模型真正难以解决的问题。
- **多模态必要性过滤：** 对于视觉-语言数学问题，我们引入了一个关键的过滤步骤：我们丢弃任何 Qwen3-30B-*nothink* 模型在没有视觉输入的情况下能够正确解决的样本。这确保了剩余的实例确实需要多模态理解，并且不能仅通过文本线索解决。
- **响应质量控制：** 遵循Qwen3的方法论，我们对生成的响应进行清洗。对于具有多个候选答案的查询，我们首先移除包含错误最终结果的那些。随后，我们过滤掉表现出不良模式的响应，例如过度重复、不适当的语言混合或明显缺乏推理步骤而猜出的答案。

这种严格的筛选过程产生了一个高质量、具有挑战性的数据集，专为启动高级多模态推理而定制。

4.3 强-弱蒸馏

我们采用 Qwen3 中描述的强到弱蒸馏流程，以进一步提升轻量级模型的性能。该蒸馏过程包含两个主要阶段：

- **离线策略蒸馏：** 在第一阶段，教师模型生成的输出被组合起来提供响应蒸馏。这有助于轻量级学生模型获取基本的推理能力，为后续的在线策略训练奠定坚实的基础。
- **策略内蒸馏：** 在第二阶段，学生模型根据提供的提示生成响应。然后使用这些策略内序列来微调学生模型。我们通过最小化KL散度来对齐学生和教师预测的logits。

4.4 强化学习

4.4.1 推理强化学习

我们在多种文本和多模态任务上训练模型，包括数学、编程、逻辑推理、视觉Grounding和视觉谜题。每个任务都设计得足够好，使得解决方案可以通过规则或代码执行器进行确定性验证。

数据准备我们从开源和专有来源收集训练数据，并应用严格的预处理和人工标注，以确保高质量的强化学习查询。对于多模态查询，我们使用我们最先进的视觉-语言模型（Qwen3-VL-235B-A22B）的初步检查点来为每个查询采样16个响应；任何所有响应都错误的查询都将被丢弃。

We then run preliminary RL experiments per task to identify and remove data sources with limited potential for improvement. This process yields approximately 30K RL queries covering a variety of text and multimodal tasks. For training each model, we sample 16 responses for all queries and filter out easy queries whose pass rate exceeds 90%. We shuffle and combine task-specific datasets to construct mixed-task batches, ensuring a consistent, predefined ratio of samples per task. The ratio is determined through extensive preliminary experiments.

Reward System We implement a unified reward framework that delivers precise feedback across all tasks. The system provides shared infrastructure—data preprocessing, utility functions, and a reward manager to integrate multiple reward types—while the core reward logic is implemented per task. We use task-specific format prompts to guide model outputs to the required formats and therefore do not rely on explicit format rewards. To mitigate code-switching, we apply a penalty when the response language differs from the prompt language.

RL Algorithm We employ SAPO (Gao et al., 2025), a smooth and adaptive policy-gradient method, for RL training. SAPO delivers consistent improvements across diverse text and multimodal tasks and across different model sizes and architectures.

4.4.2 General Reinforcement Learning

The General Reinforcement Learning (RL) stage is designed to enhance the model's generalization capabilities and operational robustness. To this end, we employ a multi-task RL paradigm where the reward function is formulated based on a comprehensive set of tasks from the SFT phase, including VQA, image captioning, OCR, document parsing, grounding, and clock recognition. The reward mechanism is structured to optimize two principal dimensions of model performance:

- **Instruction Following:** This dimension evaluates the model's adherence to explicit user directives. It assesses the ability to handle complex constraints on content, format, length, and structured outputs (e.g., JSON), ensuring the generated response precisely matches user requirements.
- **Preference Alignment:** For open-ended or subjective queries, this dimension aligns the model's outputs with human preferences by optimizing for helpfulness, factual accuracy, and stylistic appropriateness. This fosters a more natural and engaging user interaction.

Furthermore, this stage acts as a corrective mechanism to unlearn strong but flawed knowledge priors ingrained during SFT. We address this by introducing specialized, verifiable tasks designed to trigger these specific errors, such as counter-intuitive object counting and complex clock time recognition. This targeted intervention is designed to supplant erroneous priors with factual knowledge.

Another critical objective is to mitigate inferior behaviors like inappropriate language mixing, excessive repetition, and formatting errors. However, the low prevalence of these issues makes general RL a sample-inefficient correction strategy. To overcome this, we curate a dedicated dataset at this stage. This dataset isolates prompts known to elicit such undesirable behaviors. This focused training enables the application of targeted, high-frequency penalties, effectively suppressing these residual errors.

Feedback for the RL process is delivered via a hybrid reward system that combines two complementary approaches:

- **Rule-Based Rewards:** This approach provides unambiguous, high-precision feedback for tasks with verifiable ground truths, such as format adherence and instruction following. By using well-defined heuristics, this method offers a robust mechanism for assessing correctness and effectively mitigates reward hacking, where a model might exploit ambiguities in a learned reward function.
- **Model-Based Rewards:** This method employs Qwen2.5-VL-72B-Instruct or Qwen3 as sophisticated judgers. The judge models evaluate each generated response against a ground-truth reference, scoring its quality across multiple axes. This approach offers superior flexibility for assessing nuanced or open-ended tasks where strict, rule-based matching is inadequate. It is particularly effective at minimizing false negatives that would otherwise penalize valid responses with unconventional formatting or phrasing.

4.5 Thinking with Images

Inspired by the great prior works on "thinking with images" (Wu et al., 2025a; Jin et al., 2025; Zheng et al., 2025; Lai et al., 2025), we endow Qwen3-VL with similar agentic capabilities through a two-stage training paradigm.

我们随后针对每个任务运行初步的强化学习实验，以识别并移除改进潜力有限的数据源。此过程产生了约30K个强化学习查询，涵盖了各种文本和多模态任务。为训练每个模型，我们对所有查询采样16个响应，并过滤掉通过率超过90%的简单查询。我们将特定任务的数据集进行洗牌和组合，以构建混合任务批次，确保每个任务样本的恒定、预定义比例。该比例是通过广泛的初步实验确定的。

奖励系统 我们实现了一个统一的奖励框架，可在所有任务中提供精确的反馈。该系统提供共享基础设施——数据预处理、实用函数和奖励管理器以集成多种奖励类型——而核心奖励逻辑则按任务实现。我们使用特定任务的格式提示来指导模型输出所需格式，因此无需依赖显式的格式奖励。为缓解代码转换问题，当响应语言与提示语言不同时，我们应用惩罚。

强化学习算法 我们采用SAPO（高等人,2025），一种平滑且自适应的策略梯度方法，用于强化学习训练。SAPO在不同文本和多模态任务、不同模型规模和架构上均能提供持续改进。

4.4.2 一般强化学习

一般强化学习（RL）阶段旨在提升模型的泛化能力和运行鲁棒性。为此，我们采用多任务强化学习范式，其中奖励函数基于SFT阶段的一整套任务综合制定，包括VQA、图像描述、OCR、文档解析、Grounding和时钟识别。奖励机制结构化以优化模型性能的两个主要维度：

- **指令遵循：** 此维度评估模型对明确用户指令的遵守程度。它评估处理内容、格式、长度和结构化输出（例如，JSON）等复杂约束的能力，确保生成的响应精确匹配用户要求。
- **偏好对齐：** 对于开放式或主观查询，该维度通过优化有用性、事实准确性和风格适当性来使模型输出与人类偏好对齐。这促进了更自然和投入的用户交互。

此外，这一阶段充当一种纠正机制，以消除在SFT过程中根深蒂固的强但存在缺陷的知识先验。我们通过引入专门的可验证任务来解决这个问题，这些任务旨在触发这些特定错误，例如反直觉的对象计数和复杂的时钟时间识别。这种有针对性的干预旨在用事实知识取代错误的先验。

另一个关键目标是减轻不适当语言混合、过度重复和格式错误等低劣行为。然而，这些问题很少出现，使得通用强化学习成为一种样本效率低下的纠正策略。为了克服这一点，我们在这一阶段策划了一个专门的数据集。该数据集隔离了已知会引发此类不良行为的提示。这种专注的训练使得可以应用有针对性的、高频率的惩罚，有效地抑制这些残余错误。

RL过程的反馈通过混合奖励系统传递，该系统结合了两种互补的方法：

- **基于规则的奖励：** 这种方法为具有可验证真实标准（如格式遵循和指令遵循）的任务提供明确、高精度的反馈。通过使用定义良好的启发式方法，此方法为评估正确性提供了一种稳健的机制，并有效缓解了奖励攻击，即模型可能利用学习到的奖励函数中的模糊性。
- **基于模型的奖励：** 此方法采用Qwen2.5-VL-72B-Instruct或Qwen3作为复杂的评判者。评判模型将每个生成的响应与真实标准参考进行比较，在多个维度上对其质量进行评分。这种方法为评估细微差别或开放式任务提供了更高的灵活性，在这些任务中严格的基于规则的匹配是不充分的。它在最小化错误否定（即原本会因非传统格式或措辞而受到惩罚的有效响应）方面特别有效。

4.5 基于图像的思考

受启发于关于“用图像思考”的伟大先验工作（吴等人。，2025a；金等人。，2025；郑等人。，2025；赖等人。，2025），我们通过两阶段训练范式赋予Qwen3-VL类似的代理能力。

In the first stage, we synthesize a cold-start agentic dataset comprising approximately 10k grounding examples—primarily simple two-turn visual question answering tasks such as attribute detection. We then perform supervised fine-tuning (SFT) on Qwen2.5-VL-32B to emulate the behavior of a visual agent: *think* → *act* → *analyze feedback* → *answer*. To further enhance its reasoning abilities, we apply multi-turn, tool-integrated reinforcement learning (RL).

In the second stage, we distill the trained Qwen2.5-VL-32B visual agents from the first stage to generate a larger, more diverse dataset of approximately 120k multi-turn agentic interactions spanning a broader range of visual tasks. We then apply a similar cold-start SFT and tool-integrated RL pipeline (now using both distilled and synthesized data) for the post-training of Qwen3-VL.

The multi-turn, tool-integrated RL procedure is nearly identical across both stages, differing only in the underlying data. During RL, we employ three complementary reward signals to encourage robust, tool-mediated reasoning:

- **Answer Accuracy Reward** leverages Qwen3-32B to measure whether the final answer is correct.
- **Multi-Turn Reasoning Reward** leverages Qwen2.5-VL-72B to evaluate whether the assistant correctly interprets tool or environment feedback and arrives at the answer through coherent, step-by-step reasoning.
- **Tool-Calling Reward** encourages appropriate tool usage by comparing the actual number of tool calls to an expert-estimated target. This target is determined offline by Qwen2.5-VL-72B based on task complexity.

Early experiments reveal a tendency for models to degenerate into making only a single tool call to hack the first two rewards, regardless of task demands. To mitigate this, we explicitly incorporate the tool-calling reward to promote adaptive tool exploration aligned with task complexity.

4.6 Infrastructure

We train the Qwen3-VL series models on Alibaba Cloud's PAI-Lingjun AI Computing Service, which provides the high-performance computing power required for compute-intensive scenarios such as AI and high-performance computing.

During the pretraining phase, the system employs a hybrid parallelism strategy built upon the Megatron-LM framework, integrating Tensor Parallelism (TP), Pipeline Parallelism (PP), Context Parallelism (CP), Expert Parallelism (EP), and ZeRO-1 Data Parallelism (DP). This configuration achieves a fine-grained balance among model scale, computational load, and communication overhead, enabling high hardware utilization and sustaining both high throughput and low communication latency—even at scales of up to 10,000 GPUs.

For local deployment and performance evaluation, we adopt deployment strategies based on either vLLM or SGLang. vLLM utilizes PagedAttention to enable memory-efficient management and high-throughput inference, while SGLang excels at structured generation and handling complex prompts. Together, these backends provide efficient inference and evaluation with stable, efficient, and flexible model inference capabilities.

5 Evaluation

5.1 General Visual Question Answering

To comprehensively assess the general visual question answering (VQA) capabilities of the Qwen3-VL series, we conduct extensive evaluations on a diverse set of benchmarks, including MMBench-V1.1 (Liu et al., 2023b), RealWorldQA (xAI, 2024), MMStar (Chen et al., 2024a), and SimpleVQA (Cheng et al., 2025). As detailed in Table 2, Table 3 and Table 4, the Qwen3-VL family demonstrates robust and highly competitive performance across a wide spectrum of model sizes, from 2B to 235B parameters.

In the comparison of thinking mode, Qwen3-VL-235B-A22B-Thinking achieves the highest score of 78.7 on MMStar. Gemini-2.5-Pro's (Comanici et al., 2025) Thinking mode delivers the best overall performance, but Qwen3-VL-235B-A22B-Thinking is not far behind. In the non-reasoning mode comparison, Qwen3-VL-235B-A22B-Instruct obtains the highest scores on MMBench and RealWorldQA, with 89.3/88.9 and 79.2, respectively.

In the experiments with medium-sized models, Qwen3-VL-32B-Thinking achieves the highest scores on MMBench and RealWorldQA, with 89.5/89.5 and 79.4, respectively. Notably, Qwen3-VL-32B-Instruct

在第一阶段，我们合成一个冷启动代理数据集，其中包含约10k个Grounding示例——主要是简单的两轮视觉问答任务，如属性检测。然后我们对Qwen2.5-VL-32B进行监督微调（SFT），以模拟视觉代理的行为：思考 → 行动 → 分析反馈 → 回答。为了进一步增强其推理能力，我们应用了多轮、工具集成的强化学习（RL）。

在第二阶段，我们从第一阶段蒸馏出训练好的Qwen2.5-VL-32B视觉代理，以生成一个更大、更多样化的数据集，其中包含约120k多轮代理交互，涵盖更广泛的视觉任务。然后，我们应用类似的冷启动SFT和工具集成RL流程（现在使用蒸馏和合成数据）对Qwen3-VL进行后训练。

多轮、工具集成的RL流程在两个阶段中几乎完全相同，仅在底层数据上有所不同。在RL过程中，我们采用三种互补的奖励信号来鼓励稳健的工具介导推理：

- **答案准确性奖励**利用Qwen3-32B来衡量最终答案是否正确。
- **多轮推理奖励**利用Qwen2.5-VL-72B来评估助手是否正确解释工具或环境反馈，并通过连贯的逐步推理得出答案。

- **工具调用奖励**通过将实际工具调用次数与专家估计的目标进行比较，来鼓励适当的工具使用。此目标由Qwen2.5-VL-72B根据任务复杂度离线确定。

早期实验表明，模型存在一种倾向，即仅调用单个工具来破解前两个奖励，而不管任务需求如何。为了缓解这一问题，我们明确地加入了工具调用奖励，以促进与任务复杂度相匹配的工具自适应探索。

4.6 基础设施

我们在阿里云的 PAI-Lingjun AI 计算服务上训练 Qwen3-VL 系列模型，该服务提供了 AI 和高性能计算等计算密集型场景所需的计算能力。

在预训练阶段，系统采用基于 Megatron-LM 框架构建的混合并行策略，集成了张量并行（TP）、流水线并行（PP）、上下文并行（CP）、专家并行（EP）和 ZeRO-1 数据并行（DP）。这种配置在模型规模、计算负载和通信开销之间实现了细粒度的平衡，从而提高了硬件利用率，并保持了高吞吐量和低通信延迟——即使在高达 10,000 个 GPU 的规模下也是如此。

为了本地部署和性能评估，我们采用基于vLLM或SGLang的部署策略。vLLM利用PagedAttention实现内存高效管理和高吞吐量推理，而SGLang擅长结构化生成和处理复杂提示。这些后端协同提供高效推理和评估，具备稳定、高效且灵活的模型推理能力。

5 评估

5.1 通用视觉问答

为了全面评估Qwen3-VL系列的通用视觉问答（VQA）能力，我们在MMBench-V1.1（刘等， 2023b）、RealWorldQA（xAI， 2024）、MMStar（陈等， 2024a）和SimpleVQA（程等,2025）等多种基准测试上进行了广泛评估。如表2、表3 和表4所示，Qwen3-VL系列在不同参数规模（从2B到235B）的模型上均表现出强大且极具竞争力的性能。

在思考模式的比较中，Qwen3-VL-235B-A22B-Thinking 在 MMStar 上取得了最高分 78.7。Gemini-2.5-Pro 的 Comanici 等人， 2025）Thinking 模式整体表现最佳，但 Qwen3-VL-235B-A22B-Thinking 并不落后。在非推理模式的比较中，Qwen3-VL-235B-A22B-Instruct 在 MMBench 和 RealWorldQA 上获得最高分，分别为 89.3/88.9 和 79.2。

In the expe riments with medium-sized models, Qwen3-VL-32B-Thinking 实现了最高分 o n MMBench 以及 RealWorldQA，分别达到 89.5/89.5 和 79.4。值得注意的是，Qwen3-VL-32B-Instru ct

even outperforms the Thinking variant on RealWorldQA, scoring 79.0.

The scalability of the Qwen3-VL series is evident in the strong performance of our smaller models. Specifically, the largest model, Qwen3-VL-8B, achieves the highest performance across all five benchmarks. For example, on MMBench-EN, the score in "thinking" mode increases from 79.9 for the 2B model to 85.3 for the 8B model. A similar upward trend is observed on other benchmarks, such as MMStar, where the score rises from 68.1 (2B, thinking) to 75.3 (8B, thinking).

5.2 Multimodal Reasoning

We evaluate the Qwen3-VL series on a wide range of multimodal reasoning benchmarks, primarily focusing on STEM-related tasks and visual puzzles, including MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MathVision (Wang et al., 2024b), MathVision-Wild$_{\text{photo}}$ (hereafter MathVision$_{\text{WP}}$) , MathVista (Lu et al., 2023), We-Math (Qiao et al., 2024), MathVerse (Zhang et al., 2024), DynaMath (Zou et al., 2024), Math-VR (Duan et al., 2025), LogicVista (Xiao et al., 2024), VisualPuzzles (Song et al., 2025b), VLM are Blind (Rahmanzadehgervi et al., 2025), ZeroBench (Main/Subtasks) (Roberts et al., 2025), and VisuLogic (Xu et al., 2025). As shown in Table 2, the flagship Qwen3-VL model demonstrates outstanding performance across both “non-thinking” and “thinking” models. Notably, Qwen3-VL-235B-A22B-Instruct achieves the best reported results among non-thinking or low-thinking-budget models on multiple benchmarks, including MathVista$_{\text{mini}}$, MathVision, MathVerse$_{\text{mini}}$, DynaMath, ZeroBench, VLMsAreBlind, VisuLogic, and VisualPuzzles$_{\text{Direct}}$. While, Qwen3-VL-235B-A22B-Thinking achieves state-of-the-art results on MathVista$_{\text{mini}}$, MathVision, MathVerse$_{\text{mini}}$, ZeroBench, LogicVista, and VisuLogic.

Among medium-sized models, as shown in Table 3, Qwen3-VL-32B demonstrates significant advantages, consistently outperforming Gemini-2.5-Flash and GPT-5-mini. Compared to the previous-generation Qwen2.5-VL-72B model, the medium-sized Qwen3-VL model has already surpassed it on reasoning tasks. This highlights significant progress in VLMs. Additionally, our newly introduced Qwen3-VL-30B-A3B MoE model also delivers competitive results.

Among small-sized models, we compare Qwen3-VL-2B/4B/8B against GPT-5-Nano, with results presented in Table 4. The 8B variant maintains a clear advantage overall, while the 4B model achieves the highest scores on DynaMath and VisuLogic. Notably, even the smallest 2B model exhibits strong reasoning capabilities.

5.3 Alignment and Subjective Tasks

The ability to follow complex user instructions and reduce potential image-level hallucinations is indispensable for current large vision language models (VLMs). We assess our models on three representative benchmarks: MM-MT-Bench (Agrawal et al., 2024), HallusionBench (Guan et al., 2023) and MIA-Bench (Qian et al., 2024). MM-MT-Bench is a multi-turn LLM-as-a-judge evaluation benchmark for testing multimodal instruction-tuned models. HallusionBench aims at diagnosing image-context reasoning and poses great challenges for current VLMs. MIA-Bench is a more comprehensive benchmark to evaluate models’ reactions to users’ complex instructions (e.g., creative writing with character limit and compositional instructions).

As shown in Table 2, our flagship Qwen3-VL-235B-A22B model consistently outperforms other closed-source models. On HallusionBench, our thinking version surpasses Gemini-2.5-pro (Comanici et al., 2025), GPT-5 (OpenAI., 2025) and Claude opus 4.1 (Anthropic., 2025) by 3.0, 1.0, and 6.3 points, respectively. On MIA-Bench, Qwen3-VL-235B-A22B-Thinking achieves the overall best score across all the other models, showing our superior multimodal instruction following ability. We also investigate detailed subtask results of MIA-Bench: our model overtakes GPT-5-high-thinking version by 10.0 and 5.0 points in *math* and *textual* subtasks of MIA-Bench, respectively. The same trend can be observed on our smaller-sized models like Qwen3-VL-30B-A3B, and Qwen3-VL-32B, where they overtake other models with comparable sizes. Our 2B/4B/8B series also performs well and shows a negligible drop, especially on MIA-Bench.

5.4 Text Recognition and Document Understanding

We compare the Qwen3-VL series with other models of comparable size on document-related benchmarks, including OCR, document parsing, document question answering (QA), and document reasoning.

We evaluate our flagship model, Qwen3-VL-235B-A22B, against state-of-the-art VLMs on the benchmarks listed in Table 2. On OCR-focused parsing benchmarks — including CC-OCR (Yang et al., 2024b) and OmniDocBench (Ouyang et al., 2024) — as well as comprehensive OCR benchmarks such as OCR-Bench (Liu et al., 2024) and OCRBench_v2 (Fu et al., 2024b), the Qwen3-VL-235B-A22B-Instruct model

甚至在 RealWorldQA 上也优于思考变体，得分为 79.0。

Qwen3-VL 系列的可扩展性在我们的较小模型上的优异表现中得以体现。具体来说，最大的模型 Qwen3-VL-8B 在所有五个基准测试中均取得了最高性能。例如，在 MMBench-EN 上，"思考"模式下的得分从 2B 模型的 79.9 提升至 8B 模型的 85.3。在其他基准测试中也观察到类似的上升趋势，例如在 MMStar 上，得分从 68.1（2B，非思考）上升到 75.3（8B，思考）。

5.2 多模态推理

我们在一系列多模态推理基准测试上评估了 Qwen3-VL 系列，主要关注与 STEM 相关的任务和视觉谜题，包括 MMMU（Yue 等人，2024a）、MMMU-Pro（Yue 等人，2024b）、MathVision（Wang 等人，2024b）、MathVision-Wild$^{\text{photo}}$（以下简称 MathVision$_{\text{WP}}$）、MathVista（Lu 等人，2023）、We-Math（Qiao 等人，2024）、MathVerse（Zhang 等人，2024）、DynaMath（Zou 等人，2024）、Math-VR（Duan 等人，2025）、LogicVista（Xiao 等人，2024）、VisualPuzzles（Song 等人，2025b）、VLM are Blind（Rahmanzadehgervi 等人，2025）、ZeroBench（主/子任务）（Roberts 等人，2025）以及 VisuLogic（Xu 等人，2025）。如表 2 所示，旗舰 Qwen3-VL 模型在 "非思考" 和 "思考" 模型上均表现出色。值得注意的是，Qwen3-VL-235B-A22B-Instruct 在多个基准测试中，包括 MathVista$_{\text{mini}}$、MathVision、MathVerse$_{\text{mini}}$、DynaMath、ZeroBench、VLMsAreBlind、VisuLogic 和 VisualPuzzles$_{\text{Direct}}$，在非思考或低思考预算模型中取得了最佳报告结果。而 Qwen3-VL-235B-A22B-Thinking 在 MathVista$_{\text{mini}}$、MathVision、MathVerse$_{\text{mini}}$、ZeroBench、LogicVista 和 VisuLogic 上实现了最先进的结果。

在中等规模模型中，如表 3所示，Qwen3-VL-32B展现出显著优势，始终优于Gemini-2.5-Flash和GPT-5-mini。与上一代Qwen2.5-VL-72B模型相比，中等规模的Qwen3-VL模型在推理任务上已经超越。这突显了VLMs的显著进步。此外，我们新引入的Qwen3-VL-30B-A3B MoE模型也取得了具有竞争力的结果。

在小规模模型中，我们比较Qwen3-VL-2B/4B/8B与GPT-5-Nano，结果如表 4所示。8B版本整体保持明显优势，而4B模型在DynaMath和VisuLogic上获得最高分。值得注意的是，即使是2B模型也展现出强大的推理能力。

5.3 对齐和主观任务

遵循复杂用户指令和减少潜在图像级幻觉的能力，对当前大视觉语言模型（VLMs）而言不可或缺。我们在三个代表性基准上评估我们的模型：MM-MT-Bench（Agrawal等人， 2024），HallusionBench（Guan等人， 2023）和MIA-Bench（Qian等人， 2024）。MM-MT-Bench是一个用于测试多模态指令微调模型的多轮LLM作为评估者的基准。HallusionBench旨在诊断图像上下文推理，对当前VLMs提出了巨大挑战。MIA-Bench是一个更全面的基准，用于评估模型对用户复杂指令的反应（例如，带有字符限制和组合指令的创意写作）。

如表所示2，我们的旗舰模型Qwen3-VL-235B-A22B始终优于其他闭源模型。在HallusionBench上，我们的思考版本超越了Gemini-2.5-pro（Comanici等人，2025），GPT-5（OpenAI，2025）和Claude opus 4.1（Anthropic，2025），分别领先3.0、1.0和6.3分。在MIA-Bench上，Qwen3-VL-235B-A22B-Thinking在所有其他模型中获得了总分最佳成绩，展现了我们卓越的多模态指令跟随能力。我们还研究了MIA-Bench的详细子任务结果：我们的模型在MIA-Bench的数学和文本子任务中，分别以10.0和5.0分的优势超越了GPT-5-high-thinking版本。在我们的小尺寸模型如Qwen3-VL-30B-A3B和Qwen3-VL-32B上，也呈现相同趋势，它们在同等规模的模型中表现更优。我们的2B/4B/8B系列同样表现良好，性能下降可忽略不计，尤其是在MIA-Bench上。

5.4 文本识别与文档理解

我们将Qwen3-VL系列与其他规模相当的模型在文档相关基准测试上进行比较，包括OCR、文档解析、文档问答（QA）和文档推理。

我们评估了我们的旗舰模型Qwen3-VL-235B-A22B，在表2中列出的基准测试上与最先进的VLM进行对比2。在以OCR为重点的解析基准测试——包括CC-OCR（Yang等人， 2024b）和OmniDocBench（Ouyang等人， 2024）——以及OCR基准测试，如OCR- Bench（Liu等人， 2024）和OCRBench_v2（Fu等人， 2024b），Qwen3-VL-235B-A22B-Instruct模型

Table 2: **Performance of Qwen3-VL-235B-A22B and top-tier models on visual benchmarks.** The highest scores of the reasoning and non-reasoning models are shown in **bold** and underlined, respectively. Results marked with an $*$ are sourced from the technical report. $+$ denotes results with tool use.

	Benchmark	Qwen3-VL 235B-A22B		Gemini 2.5 Pro		OpenAI GPT-5		Claude Opus 4.1	
		thinking	instruct	thinking	budget-128	high	minimal	thinking	non-thinking
STEM Puzzle	MMMU	80.6	78.7	81.7*	80.9	**84.2***	74.4*	78.4	77.2
	MMMU-Pro	69.3	68.1	68.8*	71.2	**78.4***	62.7*	64.8	60.7
	$\text{MathVista}_{\text{mini}}$	**85.8**	84.9	82.7*	77.7	81.3	50.9	75.5	74.5
	MathVision	**74.6**	66.5	73.3*	66.0	70.9	45.8	64.3	57.7
	$\text{MathVision}_{\text{WP}}$	**63.8**	57.0	63.2	56.9	62.8	40.1	54.0	46.4
	We-Math	74.8	67.5	**80.6**	74.5	73.8	51.8	65.2	60.2
	$\text{MathVerse}_{\text{mini}}$	**85.0**	72.5	82.9	65.9	84.1	43.0	70.6	68.1
	DynaMath	82.8	79.4	80.0	78.5	**85.4**	74.0	75.1	72.0
	Math-VR	**66.8**	65.0	64.7*	54.3	58.1	21.7	54.3	38.0
	ZeroBench	**4**	2	3	1	2	2	3	1
	VlmsAreBlind	79.5	80.4	**86.1**	78.5	80.5	53.4	77.8	72.2
	LogicVista	**72.2**	65.8	72.0	68.7	71.8	46.3	67.3	63.5
	VisuLogic	**34.4**	29.9	31.6	26.9	28.5	27.2	27.9	27.2
	VisualPuzzles	57.2	54.7	**60.9**	56.9	57.3	47.9	48.8	47.6
General VQA	MMBench-EN	88.8	89.3	**90.1***	88.4	83.8	81.3	79.4	83.0
	MMBench-CN	88.6	88.9	**89.7***	86.4	83.5	79.9	84.9	74.3
	RealWorldQA	81.3	79.2	78.0*	76.0	**82.8**	77.3	69.9	68.5
	MMStar	**78.7**	78.4	77.5*	78.5	76.4	65.2	72.1	71.0
	SimpleVQA	61.3	63.0	**65.4**	66.9	61.8	56.7	56.7	55.7
Alignment	HallusionBench	**66.7**	63.2	63.7*	60.9	65.7	53.7	60.4	55.1
	MM-MT-Bench	**8.5**	8.5	8.4*	7.6	7.6	7.5	7.8	7.9
	MIA-Bench	**92.7**	91.3	92.3	91.3	92.4	92.6	91.2	90.0
Document Understanding	DocVQA_{test}	**96.5**	97.1	92.6	94.0	91.5	89.6	92.5	89.2
	InfoVQA_{test}	**89.5**	89.2	84.2	82.9	79.0	69.9	69.4	60.9
	$\text{AI2D}_{\text{w. M.}}$	89.2	89.7	**90.9**	90.0	89.7	84.1	86.4	84.4
	ChartQA_{test}	**90.3**	90.3	83.3	62.6	59.7	59.1	86.2	83.9
	OCRBench	**875**	920	866	872	810	787	764	750
	$\text{OCRBench\_v2}_{\text{en}}$	**66.8**	67.1	54.3	55.2	53.0	48.2	48.4	47.2
	$\text{OCRBench\_v2}_{\text{zh}}$	**63.5**	61.8	48.5	53.1	43.2	37.7	43.7	38.0
	CC-OCR	**81.5**	82.2	77.2	76.8	68.3	66.1	69.1	66.0
	$\text{OmniDocBench}_{\text{en}}$	**0.155**	0.143	0.347	0.206	0.356	0.174	0.194	-
	$\text{OmniDocBench}_{\text{zh}}$	**0.207**	0.207	0.238	0.249	0.472	0.389	0.293	-
	CharXiv(DQ)	90.5	89.4	**94.4**	87.8	89.2	79.5	88.5	87.8
	CharXiv(RQ)	66.1	62.1	67.9	62.9	**81.1***	57.8	63.6	60.2
	$\text{MMLongBench}_{\text{Doc}}$	**56.2**	57.0	55.6	51.2	51.5	42.4	54.5	48.1
2D/3D Grounding	RefCOCO-avg	**92.1**	91.9	74.6*	-	66.8	-	-	-
	CountBench	**93.7**	93.0	91.0*	91.0	91.7	87.8	93.1	91.9
	ODinW-13	**43.2**	48.6	33.7*	34.5	-	-	-	-
	ARKitScenes	**53.7**	56.9	-	-	-	-	-	-
	Hypersim	**11.0**	13.0	-	-	-	-	-	-
	SUNRGBD	**34.9**	39.4	29.7	-	-	-	-	-
Embodied/Spatial Understanding	ERQA	52.5	51.3	55.3	50.3	**65.7***	42.0*	34.8	28.0
	VSI-Bench	**60.0**	62.7	-	-	-	-	-	-
	EmbSpatialBench	**84.3**	83.1	79.1	73.3	82.9	75.1	69.2	66.0
	RefSpatialBench	**69.9**	65.5	36.5	35.6	23.8	23.1	-	-
	RoboSpatialHome	**73.9**	69.4	47.5	49.2	53.5	43.6	-	-
Multi-Image	BLINK	67.1	70.7	70.6*	70.0	**71.0**	62.8	64.1	62.9
	MUIRBENCH	**80.1**	73.0	77.2	74.0	77.5	66.5	-	-
Video Understanding	MVBench	75.2	76.5	69.9	65.8	**75.3**	64.6	61.4	59.0
	$\text{Video-MME}_{\text{w/o sub.}}$	79.0	79.2	**85.1**	80.6	84.7	77.3	75.6	73.3
	$\text{MLVU}_{\text{M-Avg}}$	83.8	84.3	85.6	81.2	**86.2**	78.3	73.5	71.2
	LVBench	63.6	67.7	**73.0**	69.0	-	-	-	-
	$\text{Charades-STA}_{\text{mIoU}}$	**63.5**	64.8	-	-	-	-	-	-
	VideoMMMU	80.0	74.7	83.6*	79.4	**84.6***	61.6*	76.2	70.1
	MMVU	71.1	68.1	**74.9**	72.2	73.0	68.1	66.4	61.4
Perception with Tool	V*	**85.9**	93.7+	83.8	72.7	72.8	56.7	-	-
	HRBench4K	84.3	85.4+	**87.3**	84.8	-	-	-	-
	HRBench8K	76.6	82.4+	**85.4**	80.1	-	-	-	-
Multi-Modal Coding	Design2Code	**93.4**	92.0	89.2	90.3	92.5	88.9	88.5	85.3
	ChartMimic	78.4	80.5	83.9	79.9	62.1	41.4	**85.2**	82.9
	UniSVG	65.8	69.8	70.0	67.9	71.7	74.5	**73.0**	72.5
Multi-Modal Agent	ScreenSpot Pro	**61.8**	62.0	-	-	-	-	-	-
	OSWorldG	**68.3**	66.7	45.2	-	-	-	-	-
	AndroidWorld	**62.0**	63.7	-	-	-	-	-	-
	OSWorld	**38.1**	31.6	-	-	-	-	-	44.4
	WindowsAA	**32.1**	28.9	-	-	-	-	-	-

表 2：**Qwen3-VL-235B-A22B 与顶级模型在视觉基准测试上的性能表现。**推理模型和非推理模型的最高分分别以**加粗**和下划线显示。标有 $*$ 的结果来源于技术报告。 $+$ 表示带有工具使用的结果。

	基准测试	Qwen3-VL 235B-A22B		Gemini 2.5 Pro		OpenAI GPT-5		Claude Opus 4.1	
		思考	指令	思考	预算-128	high	最小	思考	非思考
STEM 谜题	MMMU	80.6	78.7	81.7*	80.9	**84.2***	74.4*	78.4	77.2
	MMMU-Pro	69.3	68.1	68.8*	71.2	**78.4***	62.7*	64.8	60.7
	$\text{MathVista}_{\text{mini}}$	**85.8**	84.9	82.7*	77.7	81.3	50.9	75.5	74.5
	MathVision	**74.6**	66.5	73.3*	66.0	70.9	45.8	64.3	57.7
	$\text{MathVision}_{\text{WP}}$	**63.8**	57.0	63.2	56.9	62.8	40.1	54.0	46.4
	We-Math	74.8	67.5	**80.6**	74.5	73.8	51.8	65.2	60.2
	$\text{MathVerse}_{\text{mini}}$	**85.0**	72.5	82.9	65.9	84.1	43.0	70.6	68.1
	DynaMath	82.8	79.4	80.0	78.5	**85.4**	74.0	75.1	72.0
	Math-VR	**66.8**	65.0	64.7*	54.3	58.1	21.7	54.3	38.0
	ZeroBench	**4**	2	3	1	2	2	3	1
	VlmsAreBlind	79.5	80.4	**86.1**	78.5	80.5	53.4	77.8	72.2
	LogicVista	**72.2**	65.8	72.0	68.7	71.8	46.3	67.3	63.5
	VisuLogic	**34.4**	29.9	31.6	26.9	28.5	27.2	27.9	27.2
	VisualPuzzles	57.2	54.7	**60.9**	56.9	57.3	47.9	48.8	47.6
通用视觉问答	MMBench-EN	88.8	89.3	**90.1***	88.4	83.8	81.3	79.4	83.0
	MMBench-CN	88.6	88.9	**89.7***	86.4	83.5	79.9	84.9	74.3
	RealWorldQA	81.3	79.2	78.0*	76.0	**82.8**	77.3	69.9	68.5
	MMStar	**78.7**	78.4	77.5*	78.5	76.4	65.2	72.1	71.0
	SimpleVQA	61.3	63.0	**65.4**	66.9	61.8	56.7	56.7	55.7
对齐	HallusionBench	**66.7**	63.2	63.7*	60.9	65.7	53.7	60.4	55.1
	MM-MT-Bench	**8.5**	8.5	8.4*	7.6	7.6	7.5	7.8	7.9
	MIA-Bench	**92.7**	91.3	92.3	91.3	92.4	92.6	91.2	90.0
文档理解	DocVQA_{test}	**96.5**	97.1	92.6	94.0	91.5	89.6	92.5	89.2
	InfoVQA_{test}	**89.5**	89.2	84.2	82.9	79.0	69.9	69.4	60.9
	$\text{AI2D}_{\text{w. M.}}$	89.2	89.7	**90.9**	90.0	89.7	84.1	86.4	84.4
	ChartQA_{test}	**90.3**	90.3	83.3	62.6	59.7	59.1	86.2	83.9
	OCRBench	**875**	920	866	872	810	787	764	750
	$\text{OCRBench\_v2}_{\text{en}}$	**66.8**	67.1	54.3	55.2	53.0	48.2	48.4	47.2
	$\text{OCRBench\_v2}_{\text{zh}}$	**63.5**	61.8	48.5	53.1	43.2	37.7	43.7	38.0
	CC-OCR	**81.5**	82.2	77.2	76.8	68.3	66.1	69.1	66.0
	$\text{OmniDocBench}_{\text{en}}$	**0.155**	0.143	0.347	0.206	0.356	0.174	0.194	-
	$\text{OmniDocBench}_{\text{zh}}$	**0.207**	0.207	0.238	0.249	0.472	0.389	0.293	-
	CharXiv(DQ)	90.5	89.4	**94.4**	87.8	89.2	79.5	88.5	87.8
	CharXiv(RQ)	66.1	62.1	67.9	62.9	**81.1***	57.8	63.6	60.2
	$\text{MMLongBench}_{\text{Doc}}$	**56.2**	57.0	55.6	51.2	51.5	42.4	54.5	48.1
2D/3D Grounding	RefCOCO-avg	**92.1**	91.9	74.6*	-	66.8	-	-	-
	CountBench	**93.7**	93.0	91.0*	91.0	91.7	87.8	93.1	91.9
	ODinW-13	**43.2**	48.6	33.7*	34.5	-	-	-	-
	ARKitScenes	**53.7**	56.9	-	-	-	-	-	-
	Hypersim	**11.0**	13.0	-	-	-	-	-	-
	SUNRGBD	**34.9**	39.4	29.7	-	-	-	-	-
具身/空间理解	ERQA	52.5	51.3	55.3	50.3	**65.7***	42.0*	34.8	28.0
	VSI-Bench	**60.0**	62.7	-	-	-	-	-	-
	EmbSpatialBench	**84.3**	83.1	79.1	73.3	82.9	75.1	69.2	66.0
	RefSpatialBench	**69.9**	65.5	36.5	35.6	23.8	23.1	-	-
	RoboSpatialHome	**73.9**	69.4	47.5	49.2	53.5	43.6	-	-
多图像	BLINK	67.1	70.7	70.6*	70.0	**71.0**	62.8	64.1	62.9
	MUIRBENCH	**80.1**	73.0	77.2	74.0	77.5	66.5	-	-
视频理解	MVBench	75.2	76.5	69.9	65.8	**75.3**	64.6	61.4	59.0
	视频-MMEw/o sub.	79.0	79.2	**85.1**	80.6	84.7	77.3	75.6	73.3
	$\text{MLVU}_{\text{M-Avg}}$	83.8	84.3	85.6	81.2	**86.2**	78.3	73.5	71.2
	LVBench	63.6	67.7	**73.0**	69.0	-	-	-	-
	$\text{Charades-STA}_{\text{mIoU}}$	**63.5**	64.8	-	-	-	-	-	-
	VideoMMMU	80.0	74.7	83.6*	79.4	**84.6***	61.6*	76.2	70.1
	MMVU	71.1	68.1	**74.9**	72.2	73.0	68.1	66.4	61.4
感知带工具	V*	**85.9**	93.7+	83.8	72.7	72.8	56.7	-	-
	HRBench4K	84.3	85.4+	**87.3**	84.8	-	-	-	-
	HRBench8K	76.6	82.4+	**85.4**	80.1	-	-	-	-
多模态编程	设计到代码	**93.4**	92.0	89.2	90.3	92.5	88.9	88.5	85.3
	ChartMimic	78.4	80.5	83.9	79.9	62.1	41.4	**85.2**	82.9
	UniSVG	65.8	69.8	70.0	67.9	71.7	74.5	**73.0**	72.5
多模态代理	ScreenSpot Pro	**61.8**	62.0	-	-	-	-	-	-
	OSWorldG	**68.3**	66.7	45.2	-	-	-	-	-
	AndroidWorld	**62.0**	63.7	-	-	-	-	-	-
	OSWorld	**38.1**	31.6	-	-	-	-	-	44.4
	WindowsAA	**32.1**	28.9	-	-	-	-	-	-

Table 3: **Performance of medium-sized Qwen3-VL models and previous models on visual benchmarks.** The highest scores are shown in **bold**. Results marked with an $*$ are sourced from the technical report. $+$ denotes results with tool use.

	Benchmark	Qwen3-VL 30B-A3B		Qwen3-VL 32B		Gemini 2.5 Flash		GPT-5 mini	
		thinking	instruct	thinking	instruct	thinking	non-thinking	high	minimal
STEM Puzzle	MMMU	76.0	74.2	78.1	76.0	77.7	76.3	**79.0**	67.9
	MMMU-Pro	63.0	60.4	**68.1**	65.3	67.2	65.9	67.3	53.7
	$\text{MathVista}_{\text{mini}}$	81.9	80.1	**85.9**	83.8	79.4	75.3	79.1	59.6
	MathVision	65.7	60.2	70.2	63.4	64.3	60.7	**71.9**	46.6
	$\text{MathVision}_{\text{WP}}$	**58.9**	52.3	58.6	54.6	53.6	49.0	56.6	42.8
	We-Math	70.0	56.9	**71.6**	63.3	53.9	60.3	70.2	51.4
	$\text{MathVerse}_{\text{mini}}$	79.6	70.2	**82.6**	76.8	77.7	75.9	78.8	36.5
	DynaMath	80.1	73.4	**82.0**	76.7	75.9	69.7	81.4	71.3
	Math-VR	61.7	61.3	**62.3**	59.8	58.8	54.7	58.2	26.4
	ZeroBench	0	0	2	1	1	**3**	**3**	2
	VlmsAreBlind	72.5	67.5	85.1	**87.0**	77.5	75.9	75.8	62.0
	LogicVista	65.8	53.5	70.9	62.2	67.3	60.0	**71.4**	50.8
	VisuLogic	26.6	23.0	**32.4**	29.7	31.0	23.3	27.2	27.6
	VisualPuzzles	52.0	46.2	54.7	53.2	41.4	45.0	**59.3**	48.2
General VQA	MMBench-EN	87.0	86.1	**89.5**	87.6	87.1	86.6	86.6	78.5
	MMBench-CN	85.9	85.3	**89.4**	87.7	87.3	86.0	84.0	76.3
	RealWorldQA	77.4	73.7	78.4	**79.0**	76.0	75.7	**79.0**	73.3
	MMStar	75.5	72.1	**79.4**	77.7	76.5	75.8	74.1	61.3
	SimpleVQA	54.3	52.7	55.4	56.9	**63.2**	59.2	56.8	50.3
Alignment	HallusionBench	66.0	61.5	**67.4**	63.8	63.5	59.1	63.2	55.9
	MM-MT-Bench	7.9	8.0	8.3	**8.4**	8.1	8.0	7.7	7.4
	MIA-Bench	91.6	91.2	**92.3**	91.8	91.1	90.6	92.0	**92.3**
Document Understanding	DocVQA_{test}	95.5	95.0	96.1	**96.9**	92.8	93.0	90.5	90.6
	InfoVQA_{test}	85.6	81.8	**89.2**	87.0	82.5	81.7	77.6	72.8
	$\text{AI2D}_{\text{w. M.}}$	86.9	85.0	88.9	**89.5**	88.7	87.7	88.2	82.9
	ChartQA_{test}	**89.4**	86.8	89.0	88.5	60.6	69.0	57.5	57.8
	OCRBench	839	**903**	855	895	853	864	821	807
	$\text{OCRBench\_v2}_{\text{en}}$	62.6	63.2	**68.4**	67.4	52.2	50.6	52.6	45.7
	$\text{OCRBench\_v2}_{\text{zh}}$	60.4	57.8	**62.1**	59.2	43.8	43.9	45.1	41.0
	CC-OCR	77.8	**80.7**	79.6	80.3	75.4	74.8	70.8	61.6
	$\text{OmniDocBench}_{\text{en}}$	0.165	0.183	**0.148**	0.151	0.265	0.228	0.181	0.260
	$\text{OmniDocBench}_{\text{zh}}$	**0.233**	0.253	0.236	0.239	0.245	0.305	0.316	0.425
	CharXiv(DQ)	86.9	85.5	90.2	**90.5**	90.1	85.5	89.4	78.6
	CharXiv(RQ)	56.6	48.9	65.2	62.8	61.7	60.1	**68.6**	48.9
	$\text{MMLongBench}_{\text{Doc}}$	47.4	47.1	54.6	**55.4**	49.0	44.6	50.3	39.6
2D/3D Grounding	RefCOCO-avg	89.3	89.7	91.1	**91.9**	-	-	-	-
	CountBench	90.0	89.8	94.1	**94.9**	86.0	83.7	91.0	84.1
	ODinW-13	42.3	**47.5**	41.8	46.6	-	-	-	-
	ARKitScenes	55.6	**56.1**	46.1	55.6	-	-	-	-
	Hypersim	11.4	12.5	12.5	**14.0**	-	-	-	-
	SUNRGBD	34.6	**38.1**	33.9	37.0	-	-	-	-
Embodied/Spatial Understanding	ERQA	45.3	43.0	52.3	48.8	-	-	**54.0**	45.8
	VSI-Bench	56.1	**63.2**	61.2	61.5	-	-	31.5	30.5
	EmbSpatialBench	80.6	76.4	**82.7**	81.5	-	-	80.7	72.1
	RefSpatialBench	54.2	53.1	**67.2**	61.4	-	-	9.0	4.0
	RoboSpatialHome	65.5	62.9	**74.2**	64.6	-	-	54.3	44.6
Multi-Image	BLINK	65.4	67.7	**68.5**	67.3	68.1	66.8	-	56.7
	MUIRBENCH	77.6	62.9	**80.3**	72.8	72.7	67.5	-	57.5
Video Understanding	MVBench	72.0	72.3	**73.2**	72.8	-	-	-	-
	$\text{Video-MME}_{\text{w/o sub.}}$	73.3	74.5	77.3	76.6	**79.6**	75.6	78.9	71.0
	$\text{MLVU}_{\text{M-Avg}}$	78.9	81.3	82.3	82.1	82.1	77.8	**83.3**	71.7
	LVBench	59.2	62.5	62.6	63.8	**64.5**	62.2	-	-
	$\text{Charades-STA}_{\text{mIoU}}$	62.7	**63.5**	62.8	61.2	-	-	-	-
	VideoMMMU	75.0	68.7	79.0	71.9	73.9	65.2	**82.5***	56.7
	MMVU	66.1	59.8	67.9	66.8	**69.8**	68.2	**69.8**	64.8
Perception with Tool	V*	81.2	89.5^{+}	84.8	$\mathbf{91.1}^{+}$	-	-	78.6	63.9
	HRBench4K	77.8	82.5^{+}	82.1	$\mathbf{84.6}^{+}$	-	-	78.6	66.3
	HRBench8K	71.3	79.3^{+}	74.8	$\mathbf{81.6}^{+}$	-	-	74.4	60.9
Multi-Modal Agent	ScreenSpot Pro	57.3	**60.5**	57.1	57.9	-	-	-	-
	OSWorldG	59.6	61.0	64.0	**65.1**	-	-	-	-
	AndroidWorld	55.0	54.3	**63.7**	57.3	-	-	-	-
	OSWorld	30.6	30.3	**41.0**	32.6	-	-	-	-
	WindowsAA	24.2	24.9	**42.9**	30.9	-	-	-	-

表 3：**Qwen3-VL 中等规模模型与先前模型在视觉基准测试上的性能**。最高分以 **粗体** 显示。标有 $*$ 的结果源自技术报告。$+$ 表示带有工具使用的结果。

	基准测试	Qwen3-VL 30B-A3B		Qwen3-VL 32B		Gemini 2.5 Flash		GPT-5 mini	
		思考	指令	思考	指令	思考	非思考	high	最小
STEM 谜题	MMMU	76.0	74.2	78.1	76.0	77.7	76.3	**79.0**	67.9
	MMMU-Pro	63.0	60.4	**68.1**	65.3	67.2	65.9	67.3	53.7
	$\text{MathVista}_{\text{mini}}$	81.9	80.1	**85.9**	83.8	79.4	75.3	79.1	59.6
	MathVision	65.7	60.2	70.2	63.4	64.3	60.7	**71.9**	46.6
	$\text{MathVision}_{\text{WP}}$	**58.9**	52.3	58.6	54.6	53.6	49.0	56.6	42.8
	We-Math	70.0	56.9	**71.6**	63.3	53.9	60.3	70.2	51.4
	$\text{MathVerse}_{\text{mini}}$	79.6	70.2	**82.6**	76.8	77.7	75.9	78.8	36.5
	DynaMath	80.1	73.4	**82.0**	76.7	75.9	69.7	81.4	71.3
	Math-VR	61.7	61.3	**62.3**	59.8	58.8	54.7	58.2	26.4
	ZeroBench	0	0	2	1	1	**3**	**3**	2
	VlmsAreBlind	72.5	67.5	85.1	**87.0**	77.5	75.9	75.8	62.0
	LogicVista	65.8	53.5	70.9	62.2	67.3	60.0	**71.4**	50.8
	VisuLogic	26.6	23.0	**32.4**	29.7	31.0	23.3	27.2	27.6
	VisualPuzzles	52.0	46.2	54.7	53.2	41.4	45.0	**59.3**	48.2
通用视觉问答	MMBench-EN	87.0	86.1	**89.5**	87.6	87.1	86.6	86.6	78.5
	MMBench-CN	85.9	85.3	**89.4**	87.7	87.3	86.0	84.0	76.3
	RealWorldQA	77.4	73.7	78.4	**79.0**	76.0	75.7	**79.0**	73.3
	MMStar	75.5	72.1	**79.4**	77.7	76.5	75.8	74.1	61.3
	SimpleVQA	54.3	52.7	55.4	56.9	**63.2**	59.2	56.8	50.3
对齐	HallusionBench	66.0	61.5	**67.4**	63.8	63.5	59.1	63.2	55.9
	MM-MT-Bench	7.9	8.0	8.3	**8.4**	8.1	8.0	7.7	7.4
	MIA-Bench	91.6	91.2	**92.3**	91.8	91.1	90.6	92.0	**92.3**
文档理解	DocVQA_{test}	95.5	95.0	96.1	**96.9**	92.8	93.0	90.5	90.6
	InfoVQA_{test}	85.6	81.8	**89.2**	87.0	82.5	81.7	77.6	72.8
	$\text{AI2D}_{\text{w. M.}}$	86.9	85.0	88.9	**89.5**	88.7	87.7	88.2	82.9
	ChartQA_{test}	**89.4**	86.8	89.0	88.5	60.6	69.0	57.5	57.8
	OCRBench	839	**903**	855	895	853	864	821	807
	$\text{OCRBench\_v2}_{\text{en}}$	62.6	63.2	**68.4**	67.4	52.2	50.6	52.6	45.7
	$\text{OCRBench\_v2}_{\text{zh}}$	60.4	57.8	**62.1**	59.2	43.8	43.9	45.1	41.0
	CC-OCR	77.8	**80.7**	79.6	80.3	75.4	74.8	70.8	61.6
	$\text{OmniDocBench}_{\text{en}}$	0.165	0.183	**0.148**	0.151	0.265	0.228	0.181	0.260
	$\text{OmniDocBench}_{\text{zh}}$	**0.233**	0.253	0.236	0.239	0.245	0.305	0.316	0.425
	CharXiv(DQ)	86.9	85.5	90.2	**90.5**	90.1	85.5	89.4	78.6
	CharXiv(RQ)	56.6	48.9	65.2	62.8	61.7	60.1	**68.6**	48.9
	$\text{MMLongBench}_{\text{Doc}}$	47.4	47.1	54.6	**55.4**	49.0	44.6	50.3	39.6
2D/3D Grounding	RefCOCO-avg	89.3	89.7	91.1	**91.9**	-	-	-	-
	CountBench	90.0	89.8	94.1	**94.9**	86.0	83.7	91.0	84.1
	ODinW-13	42.3	**47.5**	41.8	46.6	-	-	-	-
	ARKitScenes	55.6	**56.1**	46.1	55.6	-	-	-	-
	Hypersim	11.4	12.5	12.5	**14.0**	-	-	-	-
	SUNRGBD	34.6	**38.1**	33.9	37.0	-	-	-	-
具身/空间理解	ERQA	45.3	43.0	52.3	48.8	-	-	**54.0**	45.8
	VSI-Bench	56.1	**63.2**	61.2	61.5	-	-	31.5	30.5
	EmbSpatialBench	80.6	76.4	**82.7**	81.5	-	-	80.7	72.1
	RefSpatialBench	54.2	53.1	**67.2**	61.4	-	-	9.0	4.0
	RoboSpatialHome	65.5	62.9	**74.2**	64.6	-	-	54.3	44.6
多图像	BLINK	65.4	67.7	**68.5**	67.3	68.1	66.8	-	56.7
	MUIRBENCH	77.6	62.9	**80.3**	72.8	72.7	67.5	-	57.5
视频理解	MVBench	72.0	72.3	**73.2**	72.8	-	-	-	-
	视频-$\text{MME}_{\text{w/o 子任务。}}$	73.3	74.5	77.3	76.6	**79.6**	75.6	78.9	71.0
	$\text{MLVU}_{\text{M-Avg}}$	78.9	81.3	82.3	82.1	82.1	77.8	**83.3**	71.7
	LVBench	59.2	62.5	62.6	63.8	**64.5**	62.2	-	-
	$\text{Charades-STA}_{\text{mIoU}}$	62.7	**63.5**	62.8	61.2	-	-	-	-
	VideoMMMU	75.0	68.7	79.0	71.9	73.9	65.2	82.5*	56.7
	MMVU	66.1	59.8	67.9	66.8	**69.8**	68.2	**69.8**	64.8
感知带工具	V*	81.2	89.5^{+}	84.8	$\mathbf{91.1}^{+}$	-	-	78.6	63.9
	HRBench4K	77.8	82.5^{+}	82.1	$\mathbf{84.6}^{+}$	-	-	78.6	66.3
	HRBench8K	71.3	79.3^{+}	74.8	$\mathbf{81.6}^{+}$	-	-	74.4	60.9
多模态代理	ScreenSpot Pro	57.3	**60.5**	57.1	57.9	-	-	-	-
	OSWorldG	59.6	61.0	64.0	**65.1**	-	-	-	-
	AndroidWorld	55.0	54.3	**63.7**	57.3	-	-	-	-
	OSWorld	30.6	30.3	**41.0**	32.6	-	-	-	-
	WindowsAA	24.2	24.9	**42.9**	30.9	-	-	-	-

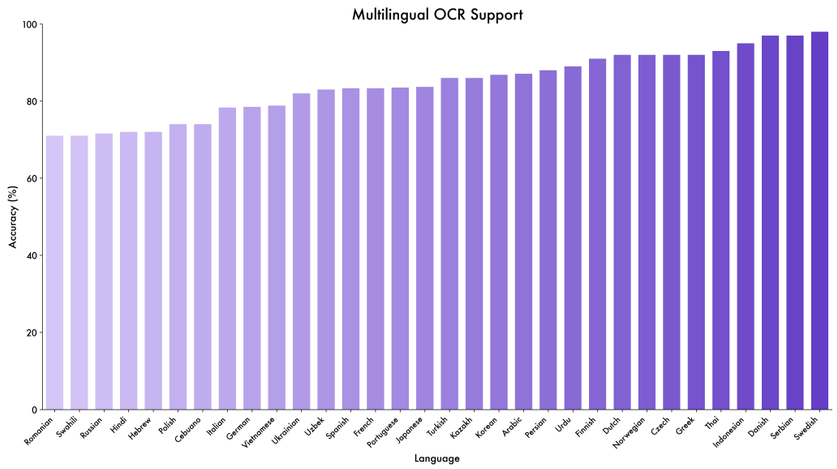


Figure 2: Multilingual OCR performance of our model on a self-built test set. The model achieves over 70% accuracy on 32 out of 39 supported languages, demonstrating strong and usable multilingual capabilities.

establishes a new state of the art, marginally outperforming its "thinking" counterpart, Qwen3-VL-235B-A22B-Thinking. On OCR-related visual question answering (VQA) benchmarks that require both OCR capability and keyword search — such as DocVQA (Mathew et al., 2021b), InfoVQA (Mathew et al., 2021a), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), and the CharXiv (Wang et al., 2024g) description subset — both the Instruct and Thinking variants achieve comparable performance, demonstrating consistently strong results across these tasks. Notably, on the reasoning subset of CharXiv — which demands deep chart comprehension and multi-step reasoning — the Thinking variant surpasses the Instruct version and ranks second only to GPT5-thinking and Gemini-2.5-Pro-Thinking.

Furthermore, among the smaller-sized variants in the Qwen3-VL series, both Qwen3-VL-30BA3B models and Qwen3-VL-32B models consistently outperform Gemini-2.5-Flash and GPT-5-mini across most evaluation metrics, as shown in Table 3. Even the compact dense models — Qwen3-VL-8B, Qwen3-VL-4B, and Qwen3-VL-2B — demonstrate remarkably competitive performance on OCR parsing, visual question answering (VQA), and comprehensive benchmark suites, as detailed in Table 4. This highlights the exceptional efficiency and strong scalability of the Qwen3-VL architecture across model sizes.

In this version of the Qwen3-VL, we have placed particular emphasis on enhancing its ability to understand long documents. As reported in Table 2, in the comparison within the flagship models on the MMLongBench-Doc benchmark (Ma et al., 2024), our Qwen3-VL-235B-A22B achieves overall accuracy of 57.0%/56.2% under the instruct/thinking settings, showcasing the SOTA performance on the long document understanding task.

Beyond its strong performance on established benchmarks, we have also made substantial strides in multilingual support. This represents a major expansion from the 10 non-English/Chinese languages supported by Qwen2.5-VL to 39 languages in Qwen3-VL. We assess this expanded capability on a newly constructed, in-house dataset. As illustrated in Figure 2, the model's accuracy surpasses 70%—a threshold we consider practical for real-world usability—on 32 out of the 39 languages tested. This demonstrates that the strong OCR capabilities of Qwen3-VL are not confined to a handful of languages but extend across a broad and diverse linguistic spectrum.

5.5 2D and 3D Grounding

In this section, we conduct a comprehensive evaluation of the Qwen3-VL series on both 2D and 3D grounding-related benchmarks and compare the models with state-of-the-art models that possess similar capabilities.

We evaluate Qwen3-VL's 2D grounding capabilities on the referring expression comprehension benchmarks RefCOCO/+/g (Kazemzadeh et al., 2014; Mao et al., 2016), the open-vocabulary object detection benchmark ODinW-13 (Li et al., 2022), and the counting benchmark CountBench (Paiss et al., 2023). For

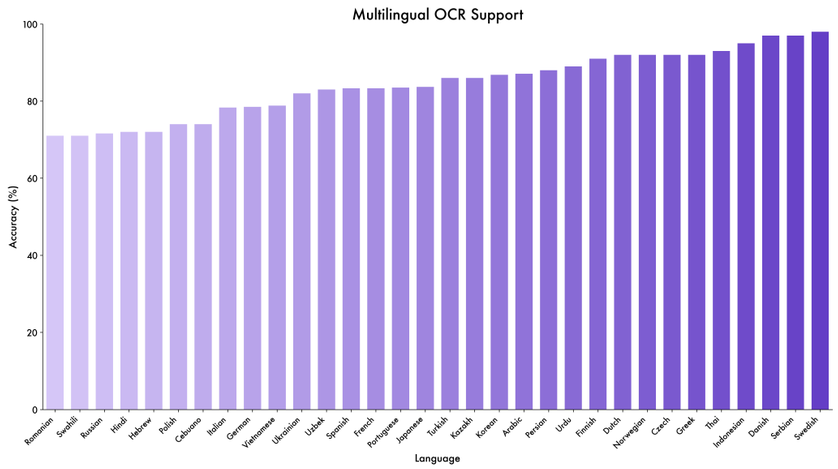


图2：我们的模型在自建测试集上的多语言OCR性能。该模型在39种支持语言的32种上达到了超过70%的准确性，展示了强大的可用多语言能力。

建立了一个新的SOTA，略微优于其"思考"对应的版本，Qwen3-VL-235B-A22B-Thinking。在需要同时具备OCR能力和关键词搜索的OCR相关视觉问答（VQA）基准测试中，例如DocVQA（Mathew等人，2021b），InfoVQA（Mathew等人，2021a），AI2D（Kembhavi等人,2016），ChartQA（Masry等人，2022），以及CharXiv（Wang等人,2024g）描述子集——指令和思考变体均取得了相当的性能，在这些任务中展现了持续强劲的结果。值得注意的是，在CharXiv的推理子集中——该子集要求深度图表理解和多步推理——思考变体超越了指令版本，仅次于GPT5-thinking和Gemini-2.5-Pro-Thinking。

此外，在Qwen3-VL系列中较小的模型变体中，Qwen3-VL-30BA3B模型和Qwen3-VL-32B模型在大多数评估指标上始终优于Gemini-2.5-Flash和GPT-5-mini，如表3所示。即使是紧凑的密集模型——Qwen3-VL-8B、Qwen3-VL-4B和Qwen3-VL-2B——在OCR解析、视觉问答（VQA）以及综合基准测试套件上表现出极具竞争力的性能，详情见表4。这突显了Qwen3-VL架构在各个模型尺寸上的卓越效率和强大可扩展性。

在本版本的Qwen3-VL中，我们特别着重于提升其理解长文档的能力。如表2所示，在MMLongBench-Doc基准测试（Ma等人，2024）中，我们的Qwen3-VL-235B-A22B在指令/思考设置下实现了57.0%/56.2%的整体准确率，展示了其在长文档理解任务上的SOTA性能。

在已建立的基准测试上表现优异之外，我们在多语言支持方面也取得了重大进展。这代表了从 Qwen2.5-VL 支持的 10 种非英语/中文语言扩展到 Qwen3-VL 支持的 39 种语言。我们在一个新构建的内部数据集上评估了这一扩展能力。如图2所示，模型在 39 种测试语言中的 32 种上的准确性超过了 70%——我们认为这是实际应用中的实用阈值——这表明 Qwen3-VL 强大的 OCR 能力不仅局限于少数语言，而是扩展到了广泛多样的语言谱系。

5.5 2D 和 3D Grounding

在本节中，我们对 Qwen3-VL 系列在 2D 和 3D 定位相关基准测试上的表现进行全面评估，并将模型与具有相似能力的最先进模型进行比较。

我们评估了 Qwen3-VL 的 2D 定位能力，基准测试包括 RefCOCO/+/g (Kazemzadeh 等人，2014； Mao 等人，2016)，开放词汇对象检测基准 ODinW-13 (Li 等人，2022)，以及计数基准 CountBench (Paiss 等人，2023)。对于

Table 4: **Performance of small-sized Qwen3-VL models and GPT-5-nano on visual benchmarks.**

	Benchmark	Qwen3-VL 2B		Qwen3-VL 4B		Qwen3-VL 8B		OpenAI GPT-5 nano	
		thinking	instruct	thinking	instruct	thinking	instruct	high	minimal
STEM Puzzle	MMMU	61.4	53.4	70.8	67.4	74.1	69.6	75.8	57.6
	MMMU-Pro	42.5	36.5	57.0	53.2	60.4	55.9	57.2	36.5
	$\text{MathVista}_{\text{mini}}$	73.6	61.3	79.5	73.7	81.4	77.2	71.5	40.9
	MathVision	45.9	31.6	60.0	51.6	62.7	53.9	62.2	33.2
	$\text{MathVision}_{\text{WP}}$	35.5	30.9	48.7	44.4	53.3	45.4	49.3	28.3
	$\text{MathVerse}_{\text{mini}}$	66.9	52.1	75.2	46.8	77.7	62.1	74.2	27.0
	DynaMath	66.7	54.2	74.4	65.3	73.2	67.7	78.0	62.0
	Math-VR	37.7	20.7	58.1	52.3	59.0	53.4	49.7	25.0
	ZeroBench	0	0	0	0	2	1	1	1
	VlmsAreBlind	50.0	56.0	68.6	71.9	69.1	74.0	66.7	40.2
	LogicVista	50.0	35.8	61.1	53.2	65.1	55.3	59.7	40.5
	VisuLogic	25.4	11.5	30.2	19.0	27.5	22.5	24.5	24.0
	VisualPuzzles	37.4	34.3	48.9	43.7	51.7	47.9	43.5	31.3
General VQA	MMBench-EN	79.9	78.4	84.6	83.9	85.3	84.5	78.4	50.8
	MMBench-CN	78.8	75.9	83.8	83.5	85.5	84.7	77.6	48.5
	RealWorldQA	69.5	63.9	73.2	70.9	73.5	71.5	71.8	60.7
	MMStar	68.1	58.3	73.2	69.8	75.3	70.9	68.6	41.3
	SimpleVQA	43.6	40.7	48.8	48.0	49.6	50.2	46.0	39.0
Alignment	HallusionBench	54.9	51.4	64.1	57.6	65.4	61.1	58.4	39.3
	MM-MT-Bench	6.9	5.9	7.7	7.5	8.0	7.7	6.6	6.2
	MIA-Bench	85.6	83.6	91.0	89.7	91.5	91.1	89.9	89.6
Document Understanding	DocVQA_{test}	92.9	93.3	94.2	95.3	95.3	96.1	88.2	78.3
	InfoVQA_{test}	77.1	72.4	83.0	80.3	86.0	83.1	68.6	49.2
	$\text{AI2D}_{\text{w. M.}}$	80.4	76.9	84.9	84.1	84.9	85.7	81.9	65.7
	ChartQA_{test}	86.6	79.1	88.8	84.6	88.6	89.6	52.1	48.6
	OCRBench	792	858	808	881	819	896	753	701
	$\text{OCRBench\_v2}_{\text{en}}$	56.4	56.3	61.8	63.7	63.9	65.4	48.1	37.9
	$\text{OCRBench\_v2}_{\text{zh}}$	51.9	53.0	55.8	57.6	59.2	61.2	33.6	27.3
	CC-OCR	68.3	72.8	73.8	76.2	76.3	79.9	58.9	52.9
	$\text{OmniDocBench}_{\text{en}}$	0.370	0.292	0.234	0.244	0.209	0.170	0.401	0.454
	$\text{OmniDocBench}_{\text{zh}}$	0.447	0.348	0.297	0.285	0.253	0.264	0.518	0.568
	CharXiv(DQ)	70.1	62.3	83.9	76.2	85.9	83.0	82.0	64.4
	CharXiv(RQ)	37.1	26.8	50.3	39.7	53.0	46.4	50.1	31.7
	$\text{MMLongBench}_{\text{Doc}}$	33.8	31.6	44.4	43.5	48.0	47.9	31.8	22.1
2D/3D Grounding	RefCOCO-avg	84.8	85.6	88.2	89.0	88.2	89.1	-	-
	CountBench	84.1	88.4	89.4	84.9	91.5	80.5	80.0	62.9
	ODinW-13	36.0	43.4	39.4	48.2	39.8	44.7	-	-
	ARKitScenes	47.7	56.2	46.3	56.6	46.6	56.8	-	-
	Hypersim	11.2	12.0	11.9	12.2	12.0	12.7	-	-
	SUNRGBD	28.6	33.8	28.0	34.7	30.4	36.2	-	-
Embodied/Spatial Understanding	ERQA	41.8	28.3	47.3	41.3	46.8	45.8	45.8	37.8
	VSI-Bench	48.0	53.9	55.2	59.3	56.6	59.4	15.4	27.0
	EmbSpatialBench	75.9	69.2	80.7	79.6	81.1	78.5	74.2	50.7
	RefSpatialBench	28.9	30.3	45.3	46.6	44.6	54.2	12.6	2.5
	RoboSpatialHome	45.3	49.1	63.2	61.7	62.0	66.9	46.1	44.8
Multi-Image	BLINK	57.2	53.8	63.4	65.8	64.7	69.1	58.3	42.2
	MUIRBENCH	68.1	47.4	75.0	63.8	76.8	64.4	65.7	45.7
Video Understanding	MVBench	64.5	61.7	69.3	68.9	69.0	68.7	-	-
	$\text{Video-MME}_{\text{w/o sub.}}$	62.1	61.9	68.9	69.3	71.8	71.4	66.2	49.4
	$\text{MLVU}_{\text{M-Avg}}$	69.2	68.3	75.7	75.3	75.1	78.1	69.2	52.6
	LVBench	47.6	47.4	53.5	56.2	55.8	58.0	-	-
	$\text{Charades-STA}_{\text{mIoU}}$	56.9	54.5	59.0	55.5	59.9	56.0	-	-
	VideoMMMU	54.1	41.9	69.4	56.2	72.8	65.3	63.0	40.2
	MMVU	48.9	41.7	58.6	50.5	62.0	58.7	63.1	51.0
Perception with Tool	V*	69.1	75.9^{+}	74.9	88.0^{+}	77.5	90.1^{+}	-	-
	HRBench4K	69.4	72.6^{+}	73.5	81.3^{+}	72.4	82.3^{+}	-	-
	HRBench8K	62.6	68.9^{+}	67.1	74.4^{+}	68.1	78.0^{+}	-	-
Multi-Modal Agent	ScreenSpot Pro	32.2	48.5	49.2	59.5	46.6	54.6	-	-
	OSWorldG	41.8	46.1	53.9	58.2	56.7	58.2	-	-
	AndroidWorld	46.1	36.4	52.0	45.3	50.0	47.6	-	-
	OSWorld	19.0	17.0	31.4	26.2	33.9	33.9	-	-
	WindowsAA	-	-	35.5	23.4	24.1	28.8	-	-

表 4: **Qwen3-VL小尺寸模型和GPT-5-nano在视觉基准测试上的性能** **s.**

	基准测试	Qwen3-VL 2B		Qwen3-VL 4B		Qwen3-VL 8B		OpenAI GPT-5 nano	
		思考	指令	思考	指令	思考	指令	high	最小
STEM 谜题	MMMU	61.4	53.4	70.8	67.4	74.1	69.6	75.8	57.6
	MMMU-Pro	42.5	36.5	57.0	53.2	60.4	55.9	57.2	36.5
	$\text{MathVista}_{\text{mini}}$	73.6	61.3	79.5	73.7	81.4	77.2	71.5	40.9
	MathVision	45.9	31.6	60.0	51.6	62.7	53.9	62.2	33.2
	$\text{MathVision}_{\text{WP}}$	35.5	30.9	48.7	44.4	53.3	45.4	49.3	28.3
	$\text{MathVerse}_{\text{mini}}$	66.9	52.1	75.2	46.8	77.7	62.1	74.2	27.0
	DynaMath	66.7	54.2	74.4	65.3	73.2	67.7	78.0	62.0
	Math-VR	37.7	20.7	58.1	52.3	59.0	53.4	49.7	25.0
	ZeroBench	0	0	0	0	2	1	1	1
	VlmsAreBlind	50.0	56.0	68.6	71.9	69.1	74.0	66.7	40.2
	LogicVista	50.0	35.8	61.1	53.2	65.1	55.3	59.7	40.5
	VisuLogic	25.4	11.5	30.2	19.0	27.5	22.5	24.5	24.0
	VisualPuzzles	37.4	34.3	48.9	43.7	51.7	47.9	43.5	31.3
通用视觉问答	MMBench-EN	79.9	78.4	84.6	83.9	85.3	84.5	78.4	50.8
	MMBench-CN	78.8	75.9	83.8	83.5	85.5	84.7	77.6	48.5
	RealWorldQA	69.5	63.9	73.2	70.9	73.5	71.5	71.8	60.7
	MMStar	68.1	58.3	73.2	69.8	75.3	70.9	68.6	41.3
	SimpleVQA	43.6	40.7	48.8	48.0	49.6	50.2	46.0	39.0
对齐	HallusionBench	54.9	51.4	64.1	57.6	65.4	61.1	58.4	39.3
	MM-MT-Bench	6.9	5.9	7.7	7.5	8.0	7.7	6.6	6.2
	MIA-Bench	85.6	83.6	91.0	89.7	91.5	91.1	89.9	89.6
文档 理解	DocVQA_{test}	92.9	93.3	94.2	95.3	95.3	96.1	88.2	78.3
	InfoVQA_{test}	77.1	72.4	83.0	80.3	86.0	83.1	68.6	49.2
	$\text{AI2D}_{\text{w. M。}}$	80.4	76.9	84.9	84.1	84.9	85.7	81.9	65.7
	ChartQA_{test}	86.6	79.1	88.8	84.6	88.6	89.6	52.1	48.6
	OCRBench	792	858	808	881	819	896	753	701
	$\text{OCRBench \_v2}_{\text{en}}$	56.4	56.3	61.8	63.7	63.9	65.4	48.1	37.9
	$\text{OCRBench\_v2}_{\text{zh}}$	51.9	53.0	55.8	57.6	59.2	61.2	33.6	27.3
	CC-OCR	68.3	72.8	73.8	76.2	76.3	79.9	58.9	52.9
	OmniDocBench en	0.370	0.292	0.234	0.244	0.209	0.170	0.401	0.454
	$\text{OmniDocBench}_{\text{zh}}$	0.447	0.348	0.297	0.285	0.253	0.264	0.518	0.568
	CharXiv(DQ)	70.1	62.3	83.9	76.2	85.9	83.0	82.0	64.4
	CharXiv(RQ)	37.1	26.8	50.3	39.7	53.0	46.4	50.1	31.7
	$\text{MMLongBench}_{\text{Doc}}$	33.8	31.6	44.4	43.5	48.0	47.9	31.8	22.1
2D/3D Grounding	RefCOCO-avg	84.8	85.6	88.2	89.0	88.2	89.1	-	-
	CountBench	84.1	88.4	89.4	84.9	91.5	80.5	80.0	62.9
	ODinW-13	36.0	43.4	39.4	48.2	39.8	44.7	-	-
	ARKitScenes	47.7	56.2	46.3	56.6	46.6	56.8	-	-
	Hypersim	11.2	12.0	11.9	12.2	12.0	12.7	-	-
	SUNRGBD	28.6	33.8	28.0	34.7	30.4	36.2	-	-
具身/空间 理解	ERQA	41.8	28.3	47.3	41.3	46.8	45.8	45.8	37.8
	VSI-Bench	48.0	53.9	55.2	59.3	56.6	59.4	15.4	27.0
	EmbSpatialBench	75.9	69.2	80.7	79.6	81.1	78.5	74.2	50.7
	RefSpatialBench	28.9	30.3	45.3	46.6	44.6	54.2	12.6	2.5
	RoboSpatialHome	45.3	49.1	63.2	61.7	62.0	66.9	46.1	44.8
多图像	BLINK	57.2	53.8	63.4	65.8	64.7	69.1	58.3	42.2
	MUIRBENCH	68.1	47.4	75.0	63.8	76.8	64.4	65.7	45.7
视频 理解	MVBench	64.5	61.7	69.3	68.9	69.0	68.7	-	-
	视频-$\text{MME}_{\text{w/o 子任务。}}$	62.1	61.9	68.9	69.3	71.8	71.4	66.2	49.4
	$\text{MLVU}^{\text{M-Avg}}$	69.2	68.3	75.7	75.3	75.1	78.1	69.2	52.6
	LVBench	47.6	47.4	53.5	56.2	55.8	58.0	-	-
	Charades-STA mIoU	56.9	54.5	59.0	55.5	59.9	56.0	-	-
	VideoMMMU	54.1	41.9	69.4	56.2	72.8	65.3	63.0	40.2
	MMVU	48.9	41.7	58.6	50.5	62.0	58.7	63.1	51.0
感知 带工具	V*	69.1	75.9^{+}	74.9	88.0^{+}	77.5	90.1^{+}	-	-
	HRBench4K	69.4	72.6^{+}	73.5	81.3^{+}	72.4	82.3^{+}	-	-
	HRBench8K	62.6	68.9^{+}	67.1	74.4^{+}	68.1	78.0^{+}	-	-
多模态 代理	ScreenSpot Pro	32.2	48.5	49.2	59.5	46.6	54.6	-	-
	OSWorldG	41.8	46.1	53.9	58.2	56.7	58.2	-	-
	AndroidWorld	46.1	36.4	52.0	45.3	50.0	47.6	-	-
	OSWorld	19.0	17.0	31.4	26.2	33.9	33.9	-	-
	WindowsAA	-	-	35.5	23.4	24.1	28.8	-	-

ODinW-13, we adopt mean Average Precision (mAP) as the evaluation metric by setting confidence scores to 1.0. To ensure comparability with conventional open-set object detection specialist models, we provide all dataset categories simultaneously within the prompt during evaluation. As shown in Table 2, our flagship model, Qwen3-VL-235B-A22B, demonstrates outstanding performance and achieves state-of-the-art (SOTA) results across 2D grounding and counting benchmarks. Notably, it achieves 48.6 mAP on ODinW-13, demonstrating strong performance in multi-target open-vocabulary object grounding. Detailed results for our smaller-scale variants, which also exhibit competitive performance in 2D visual grounding, are presented in Tables 3 and 4, respectively.

Moreover, in this version of Qwen3-VL, we enhance its spatial perception capabilities for 3D object localization. We evaluate the Qwen3-VL series against other models of comparable scale on Omni3D (Brazil et al., 2023), a comprehensive benchmark comprising datasets such as ARKitScenes (Baruch et al., 2021), Hypersim (Roberts et al., 2021), and SUN RGB-D (Song et al., 2015). We employ mean Average Precision (mAP) as our evaluation metric. Each input is an image-text pair consisting of the image and a textual prompt specifying the object category. To ensure a fair comparison with existing VLMs, we set the IoU threshold to 0.15 and report mAP@0.15 on the Omni3D test set, with detection confidence fixed at 1.0. As shown in Table 2, our flagship Qwen3-VL-235B-A22B model consistently outperforms other closed-source models across multiple datasets. Specifically, on the SUN RGB-D dataset (Song et al., 2015), the Qwen3-VL-235B-A22B-Thinking variants surpass the performance of Gemini-2.5-Pro by 5.2 points. Our smaller-scale variants (e.g., Qwen3-VL-30BA3B, -32B, -8B, -4B, -2B) also exhibit remarkably competitive performance in 3D object grounding, with detailed results provided in Tables 3 and 4, respectively.

5.6 Fine-grained Perception

We measure the models' fine-grained perception capabilities on three popular benchmarks. The Qwen3-VL series demonstrates a substantial leap in fine-grained visual understanding compared to its predecessor, Qwen2.5-VL-72B. Notably, Qwen3-VL-235B-A22B achieves the state-of-the-art performance across all three benchmarks when augmented with tools—reaching **93.7** on V* (Wu &Xie, 2024), 85.3 on HRBench-4k (Wang et al., 2024e), and 82.3 on HRBench-8k (Wang et al., 2024e). This consistent outperformance highlights the effectiveness of architectural refinements and training strategies introduced in Qwen3-VL, particularly in handling high-resolution inputs and subtle visual distinctions critical for fine-grained perception tasks. Second, and perhaps more surprisingly, the performance gains from integrating external tools consistently outweigh those from simply increasing model size. For example, within the Qwen3-VL family, the absolute improvement by adding tools is consistently ~ 5 points across V*. These findings reinforce our conviction that scaling tool-integrated agentic learning in multimodality is a highly promising path forward.

5.7 Multi-Image Understanding

Beyond single-image grounded dialogue evaluation, advancing VLMs to handle multi-image understanding is of significant value. This task requires higher-level contextual analysis across diverse visual patterns, enabling more advanced recognition and reasoning capabilities. To this end, we nourish Qwen3-VL with comprehensive cross-image pattern learning techniques, including multi-image referring grounding, visual correspondence, and multi-hop reasoning. We evaluated Qwen3-VL on two prominent multi-image benchmarks: BLINK (Fu et al., 2024c) and MuirBench (Wang et al., 2024a). As shown in Table 2, Qwen3-VL demonstrates overall superiority in multi-image understanding compared to other leading LVLMs. Specifically, Qwen3-VL-235B-A22B-Instruct achieves performance comparable to state-of-the-art models such as Gemini-2.5-pro, while Qwen3-VL-235B-A22B-Thinking attains a remarkable leading score of 80.1 on MuirBench, surpassing all other models.

5.8 Embodied and Spatial Understanding

For embodied and spatial understanding, Qwen3-VL's performance is rigorously benchmarked against leading SOTA models using a challenging suite of benchmarks: ERQA (Team et al., 2025), VSIBench (Yang et al., 2025b), EmbSpatial (Du et al., 2024), RefSpatial (Zhou et al., 2025), and RoboSpatialHome (Song et al., 2025a). Across these benchmarks, the model showcases exceptional capabilities, rivaling the performance of top-tier models like Gemini-2.5-Pro, GPT-5, and Claude-Opus-4.1. This success is largely driven by the model's profound spatial understanding, which stems from its training on high-resolution visual data with fine-grained pointing, relative-position annotations, and QA pairs. This capability is clearly validated by its strong results on EmbSpatial, RefSpatial, and RoboSpatialHome, where Qwen3-VL-235B-A22 achieves scores of 84.3, 69.9, and 73.9, respectively. Moreover, its embodied intelligence is significantly enhanced through the integration of pointing, grounding, and spatio-temporal perception data during training, leading to top-tier scores of 52.5 on ERQA (Team et al., 2025) and 60.0 on VSIBench (Yang et al.,

ODinW-13，我们采用平均精度均值（mAP）作为评估指标，并将置信度分数设置为1.0。为确保与传统的开放集目标检测专业模型具有可比性，我们在评估时将所有数据集类别同时提供在提示中。如表2所示，我们的旗舰模型Qwen3-VL-235B-A22B表现出色，在2D定位和计数基准测试中均取得了最先进（SOTA）的结果。值得注意的是，它在ODinW-13上达到了48.6 mAP，在多目标开放词汇目标定位方面表现出强大的性能。我们较小规模的变体在2D视觉定位方面也表现出具有竞争力的性能，详细结果分别呈现在表3 和4中。

此外，在本版本的Qwen3-VL中，我们增强了其用于3D对象定位的空间感知能力。我们使用Omni3D（Brazil等人，2023），一个包含ARKitScenes（Baruch等人，2021）、Hypersim（Roberts等人，2021）和SUN RGB-D（Song等人，2015）等多个数据集的综合基准，对Qwen3-VL系列与其他规模相当的模型进行评估。我们采用平均精度均值（mAP）作为评估指标。每个输入是一个图像-文本对，包含图像和指定对象类别的文本提示。为确保与现有视觉语言模型（VLMs）的公平比较，我们将IoU阈值设置为0.15，并在Omni3D测试集上报告mAP@0.15，检测置信度固定为1.0。如表2所示，我们的旗舰模型Qwen3-VL-235B-A22B在多个数据集上始终优于其他闭源模型。具体而言，在SUN RGB-D数据集（Song等人，2015）上，Qwen3-VL-235B-A22B-Thinking变体比Gemini-2.5-Pro高出5.2个百分点。我们较小规模的变体（例如Qwen3-VL-30BA3B、-32B、-8B、-4B、-2B）在3D对象定位方面也表现出极具竞争力的性能，详细结果分别呈现在表3 和4中。

5.6 细粒度感知

我们在三个流行的基准测试上测量了模型的细粒度感知能力。与前身Qwen2.5-VL-72B相比，Qwen3-VL系列在细粒度视觉理解方面实现了显著的飞跃。值得注意的是，当带工具增强时，Qwen3-VL-235B-A22B在所有三个基准测试中都达到了最先进性能——在V*（Wu &Xie，2024）上达到 **93.7**，在HRBench-4k（Wang et al.，2024e）上达到85.3，在HRBench-8k（Wang et al.，2024e）上达到82.3。这种持续的超常表现突出了Qwen3-VL中引入的架构改进和训练策略的有效性，特别是在处理高分辨率输入和细粒度感知任务中至关重要的微妙视觉区分方面。其次，也许更令人惊讶的是，集成外部工具带来的性能提升始终高于单纯增加模型规模带来的提升。例如，在Qwen3-VL系列中，添加工具带来的绝对提升在V*上始终是 ~ 5 点。这些发现加强了我们的信念，即在大多模态中扩展带工具的代理式学习是一条极具前景的道路。

5.7 多图像理解

超越单图像 grounding 对话评估，将 VLM 推向多图像理解是具有重要价值的。这项任务需要跨多种视觉模式进行更高级的上下文分析，从而实现更高级的识别和推理能力。为此，我们通过全面的跨图像模式学习技术滋养 Qwen3-VL，包括多图像指代 grounding、视觉对应以及多跳推理。我们在两个著名的多图像基准测试上评估了 Qwen3-VL：BLINK（Fu 等人， 2024c）和 MUIRBENCH（Wang 等人，2024a）。如表 2所示，与其它领先的 LVLM 相比，Qwen3-VL 在多图像理解方面表现出整体优势。具体而言，Qwen3-VL-235B-A22B-Instruct 的性能与 Gemini-2.5-pro 等最先进模型相当，而 Qwen3-VL-235B-A22B-Thinking 在 MUIRBENCH 上取得了 80.1 的领先分数，超越了所有其他模型。

5.8 体现与空间理解

对于具身和空间理解，Qwen3-VL 的性能通过一套具有挑战性的基准测试进行了严格的基准测试：ERQA (团队等，2025)，VSIBench (杨等， 2025b)，EmbSpatial (杜等， 2024)，RefSpatial (周等， 2025)，以及 RoboSpatialHome(宋等,2025a)。在这些基准测试中，该模型展示了卓越的能力，其性能可与顶级模型 Gemini-2.5-Pro、GPT-5 和 Claude-Opus-4.1 相媲美。这种成功很大程度上是由模型深刻的空间理解能力驱动的，这种能力源于其在高分辨率视觉数据上的训练，这些数据包含细粒度的指向、相对位置标注和问答对。这种能力通过其在 EmbSpatial、RefSpatial 和 RoboSpatialHome 上的强劲结果得到了明确验证，其中 Qwen3-VL-235B-A22 分别取得了 84.3、69.9 和 73.9 的分数。此外，通过在训练过程中整合指向、Grounding 和时空感知数据，其具身智能得到了显著增强，从而在 ERQA (团队等， 2025) 上取得了 52.5 的顶级分数，在 VSIBench (杨等，

2025b) for Qwen3-VL-235B-A22B.

5.9 Video Understanding

Benefiting from the scaling of training data and key architectural enhancements, Qwen3-VL demonstrates substantially improved video understanding capabilities. In particular, the integration of interleaved MRoPE, the insertion of textual timestamps, and scaling temporally dense video captions collectively enable the Qwen3-VL 8B variant to achieve performance competitive with the significantly larger Qwen2.5-VL 72B model.

We conduct a comprehensive evaluation across a diverse set of video understanding tasks, encompassing general video understanding (VideoMME (Fu et al., 2024a), MVBench (Li et al., 2024b)), temporal video grounding (Charades-STA (Gao et al., 2017)), video reasoning (VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025)), and long-form video understanding (LVBench (Wang et al., 2024d), MLVU (Zhou et al., 2024)). In comparison with state-of-the-art proprietary models — including Gemini 2.5 Pro, GPT-5, and Claude Opus 4.1, Qwen3-VL demonstrates competitive and, in several cases, superior performance. In particular, our flagship model, Qwen3-VL-235B-A22B-Instruct, achieves performance on par with leading models such as Gemini 2.5 Pro (with a thinking budget of 128) and GPT-5 minimal on standard video understanding benchmarks. By extending the context window to 256K tokens, it further attains or even surpasses Gemini-2.5-Pro on long-video evaluation tasks, most notably on MLVU.

Regarding evaluation details, we imposed a cap of 2,048 frames per video for all benchmarks, ensuring that the total number of video tokens did not exceed 224K. The maximum number of tokens per frame was set to 768 for VideoMMMU and MMVU, and to 640 for all other benchmarks. Additionally, videos from Charades-STA were sampled at 4 frames per second (fps), while a rate of 2 fps was used for all other benchmarks. For VideoMMMU, we employed a model-based judge for evaluation, as rule-based scoring proved insufficiently accurate. It is worth noting that our comparison cannot guarantee full fairness due to resource and API limitations, which constrained the number of input frames used during evaluation: 512 for Gemini 2.5 Pro, 256 for GPT-5, and 100 for Claude Opus 4.1.

5.10 Agent

We evaluate UI perception with GUI-grounding tasks (ScreenSpot (Cheng et al., 2024), ScreenSpot Pro (Li et al., 2025b), OSWorldG(Xie et al., 2025a)) and assess decision-making abilities through online environment evaluations (AndroidWorld (Rawles et al., 2024), OSWorld (Xie et al., 2025c;b)). For GUI grounding, Qwen3-VL-235B-A22B achieves state-of-the-art performance across multiple tasks, covering interactive interfaces on desktop, mobile, and PC, and demonstrating exceptionally strong UI perception capabilities. For online evaluations, Qwen3-VL 32B scores 41 on OSWorld and 63.7 on AndroidWorld, which surpasses the current foundation VLMs. Qwen3-VL demonstrates exceptionally strong planning, decision-making, and reflection abilities as a GUI agent. Furthermore, smaller Qwen3-VL models have demonstrated highly competitive performance on these benchmarks.

5.11 Text-Centric Tasks

To comprehensively evaluate the text-centric performance of Qwen3-VL, we adopt automatic benchmarks to assess model performance on both instruct and thinking models. These benchmarks can be categorized into the following key types: (1) **Knowledge:** MMLU-Pro (Wang et al., 2024f), MMLU-Redux (Gema et al., 2024), GPQA (Rein et al., 2023), SuperGPQA (Team, 2025), (2) **Reasoning:** AIME-25 (AIME, 2025), HMMT-25 (HMMT, 2025), LiveBench (2024-11-25) (White et al., 2024), (3) **Code:** LiveCodeBench v6 (Jain et al., 2024), CFEval, OJBench (Wang et al., 2025c), (4) **Alignment Tasks:** IFEval (Zhou et al., 2023), Arena-Hard v2 (Li et al., 2024d) [1], Creative Writing v3 (Paech, 2023) [2], WritingBench (Wu et al., 2025b), (5) **Agent:** BFCL-v3 (Patil et al., 2024), TAU2-Retail, TAU2-Airline, TAU2-Telecom, (6) **Multilingual:** MultiIF (He et al., 2024), MMLU-ProX, INCLUDE (Romanou et al., 2025), PolyMATH (Wang et al., 2025b).

Evaluation Settings For Qwen3-VL instruct models including 235B-A22B, 32B and 30B-A3B, we configure the sampling hyperparameters with temperature = 0.7, top-p = 0.8, top-k = 20, and presence penalty = 1.5. As for the small instruct models including 8B, 4B and 2B, we set the temperature = 1.0, top-p = 1.0, top-k = 40, and presence penalty = 2.0. We set the max output length to 32,768 tokens.

For Qwen3-VL thinking models with Mixture-of-Experts (MoE) architecture, we set the sampling temperature to 0.6, top-p to 0.95, and top-k to 20. For the dense thinking models, we set temperature = 1.0, top-p

[1]For reproducibility of Arena-Hard v2, we report the win rates evaluated by GPT-4.1.
[2]For reproducibility of Creative Writing v3, we report the scores evaluated by Claude 3.7 Sonnet.

2025b）用于 Qwen3-VL-235B-A22B。

5.9 视频理解

受益于训练数据的扩展和关键架构的增强，Qwen3-VL 在视频理解能力上表现出显著提升。特别是，交错 MRoPE 的集成、文本时间戳的插入以及时间密集型视频字幕的扩展，共同使 Qwen3-VL 8B 版本在标准视频理解基准测试中达到了与显著更大的 Qwen2.5-VL 72B 模型相当的性能。

我们针对多样化的视频理解任务进行了全面评估，涵盖通用视频理解（VideoMME (Fu 等人， 2024a），MVBench (Li 等人， 2024b))、时间视频 Grounding（Charades-STA (Gao 等人， 2017))、视频推理（VideoMMMU (Hu 等人,2025)，MMVU (Zhao 等人， 2025))以及长视频理解（LVBench (Wang 等人，2024d)，MLVU (Zhou 等人， 2024)））。与 Gemini 2.5 Pro、GPT-5 和 Claude Opus4.1 等最先进的专有模型相比，Qwen3-VL 表现出具有竞争力的性能，在某些情况下甚至更优。特别是，我们的旗舰模型 Qwen3-VL-235B-A22B-Instruct 在标准视频理解基准测试中达到了与 Gemini 2.5 Pro（思考预算为 128）和 GPT-5 mini 相当的性能。通过将上下文窗口扩展到 256K 个 token，它在长视频评估任务上进一步达到了或甚至超越了 Gemini-2.5-Pro，尤其是在 MLVU 上。

关于评估细节，我们对所有基准测试的视频均设置了2,048帧的上限，确保视频令牌总数不超过224K。VideoMMMU和MMVU的每帧最大令牌数设置为768，其他所有基准测试则设置为640。此外，Charades-STA的视频以每秒4帧（fps）的速率采样，而其他所有基准测试则使用2 fps的速率。对于VideoMMMU，我们采用基于模型的裁判进行评估，因为基于规则的评分被证明不够准确。值得注意的是，由于资源和API限制，我们的比较无法保证完全公平，这限制了评估过程中使用的输入帧数量：Gemini 2.5 Pro为512帧，GPT-5为256帧，Claude Opus 4.1为100帧。

5.10 代理

我们通过GUI-grounding任务（ScreenSpot (Cheng等人， 2024)，ScreenSpot Pro (Li等人，2025b)，OSWorldG(Xie等人， 2025a))评估UI感知能力，并通过在线环境评估（AndroidWorld (Rawles等人， 2024)，OSWorld (Xie等人， 2025c;b))评估决策能力。对于GUI grounding，Qwen3-VL-235B-A22B在多个任务中实现了最先进的性能，涵盖桌面、移动和PC上的交互式界面，并展现出极强的UI感知能力。对于在线评估，Qwen3-VL 32B在OSWorld上得分41，在Android World上得分63.7，这超过了当前的基础VLM。作为GUI代理，Qwen3-VL展现出极强的规划、决策和反思能力。此外，更小的Qwen3-VL模型在这些基准测试中也表现出极具竞争力的性能。

5.11 文本中心任务

为全面评估Qwen3-VL的文本中心性能，我们采用自动基准测试来评估模型在指令和思考模型上的表现。这些基准测试可以分为以下主要类型：(1) 知识：MMLU-Pro（Wang等人，2024f）、MMLU-Redux（Gema等人，2024）、GPQA（Rein等人，2023）、SuperGPQA（团队，2025）、(2) 推理：AIME-25（AIME，2025）、HMMT-25（HMMT，2025）、LiveBench（2024-11-25）（White等人，2024）、(3) 代码：LiveCodeBench v6（Jain等人，2024）、CFEval、OJBench（Wang等人，2025c）、(4) 对齐任务：IFEval（Zhou等人，2023）、Arena-Hard v2（Li等人，2024d）1、创意写作 v3（Paech，2023）2、WritingBench（Wu等人，2025b）、(5) 代理：BFCL-v3（Patil等人，2024）、TAU2-Retail、TAU2-Airline、TAU2-Telecom、(6) 多语言：MultiIF（He等人，2024）、MMLU-ProX、INCLUDE（Romanou等人，2025）、PolyMATH（Wang等人，2025b）。

评估设置 对于Qwen3-VL指令模型包括235B-A22B、32B和30B-A3B，我们使用温度 = 0.7、top-p = 0.8、top-k = 20和存在惩罚= 1.5配置采样超参数。对于小型指令模型包括8B、4B和2B，我们设置温度 = 1.0、top-p = 1.0、top-k = 40和存在惩罚 = 2.0。我们将最大输出长度设置为32,768个token。

对于采用专家混合架构的Qwen3-VL思考模型，我们将采样温度设置为0.6，top-p设置为0.95，top-k设置为20。对于密集型思考模型，我们设置温度为 = 1.0，top-p

[1]为了确保Arena-Hard v2的可重复性，我们报告了由GPT-4.1评估的胜率.[2]为了确保创意写作v3的可重复性，我们报告了由Claude3.7 Sonnet评估的分数。

Table 5: **Comparison among Qwen3-VL-235B-A22B (Instruct) and other baselines. The highest and second-best scores are shown in bold and underlined respectively.**

	Benchmark	Qwen3-VL 235B-A22B Instruct	Qwen3 235B-A22B Instruct-2507	Deepseek V3 0324	Claude-Opus-4 (Without thinking)
Knowledge	MMLU-Pro	81.8	83.0	81.2	**86.6**
	MMLU-Redux	92.2	93.1	90.4	**94.2**
	GPQA	74.3	**77.5**	68.4	74.9
	SuperGPQA	60.4	**62.6**	57.3	56.5
Reasoning	AIME-25	**74.7**	70.3	46.6	33.9
	HMMT-25	**57.4**	55.4	27.5	15.9
	LiveBench 2024-11-25	74.8	**75.4**	66.9	74.6
Alignment Tasks	IFEval	87.8	**88.7**	82.3	87.4
	Arena-Hard V2 (winrate)	77.4	**79.2**	45.6	51.5
	Creative Writing v3	86.5	**87.5**	81.6	83.8
	WritingBench	**85.5**	85.2	74.5	79.2
Coding & Agent	LiveCodeBench v6	**54.3**	51.8	45.2	44.6
	BFCL-v3	67.7	**70.9**	64.7	60.1
Multilingualism	MultiIF	76.3	**77.5**	66.5	-
	MMLU-ProX	77.8	**79.4**	75.8	-
	INCLUDE	80.0	79.5	**80.1**	-
	PolyMATH	45.1	**50.2**	32.2	30.0

= 0.95, top-k = 20, and additionally apply a presence penalty of 1.5 to encourage greater output diversity. We set the max output length to 32,768 tokens, except AIME-25, HMMT-25 and LiveCodeBench v6 where we extend the length to 81,920 tokens to provide sufficient thinking space.

The detailed results are as follows.

Qwen3-VL-235B-A22B We compare our flagship model Qwen3-VL-235B-A22B with the leading instruct and thinking models. For the Qwen3-VL-235B-A22B-Instruct, we take Qwen3-235B-A22B-Instruct-2507, DeepSeek V3 0324, and Claude-Opus-4 (without thinking) as the baselines. For the Qwen3-VL-235B-A22B-Thinking, we take Qwen3-235B-A22B-Thinking-2507, OpenAI o3 (medium), Claude-Opus-4 (with thinking) as baselines. We present the evaluation results in Table 5 and Table 6.

- From Table 5, Qwen3-VL-235B-A22B-Instruct achieves competitive results, comparable to or even surpassing the other leading models, including DeepSeek V3 0324, Claude-Opus-4 (without thinking), and our previous flagship model Qwen3-235B-A22B-Instruct-2507. Particularly, Qwen3-VL-235B-A22B-Instruct exceeds other models on reasoning-demand tasks (e.g., mathematics and coding). It is worth noting that DeepSeek V3 0324 and Qwen3-235B-A22B-Instruct-2507 are Large Language Models, while Qwen3-VL-235B-A22B-Instruct is a Vision Language model which can process visual and textual tasks. This means that Qwen3-VL-235B-Instruct has achieved the integration of visual and textual capabilities.
- From Table 6, Qwen3-VL-235B-A22B-Thinking also achieves competitive results compared with other leading thinking models. Qwen3-VL-235B-A22B-Thinking exceeds OpenAI o3 (medium) and Claude-Opus-4 (with thinking) on AIME-25 and LiveCodeBench v6, which means Qwen3-VL-235B-A22B-Thinking has better reasoning ability.

Qwen3-VL-32B / 30B-A3B We compare our Qwen3-VL-32B and Qwen3-VL-30B-A3B models with their corresponding text-only counterparts, namely Qwen3-32B, Qwen3-30B-A3B, and Qwen3-30B-A3B-2507. We present the evaluation results in Table 7 and Table 8.

- From Table 7, for instruct models, Qwen3-VL-32B and Qwen3-VL-30B-A3B show significant performance improvement compared with Qwen3-32B and Qwen3-30B-A3B on all the benchmarks. Qwen3-VL-30B-A3B achieves comparable or even better results compared with Qwen3-30B-A3B-2507, particularly AIME-25 and HMMT-25.
- From Table 8, for thinking models, Qwen3-VL-32B and Qwen3-VL-30B-A3B surpass the baselines in most of the benchmarks. Qwen3-VL-30B-A3B also shows comparable performance compared with Qwen3-30B-A3B-2507.

表5：**Co Qwen3-VL-235B-A22B（指令）与其他基线的比较。最高分和第二高分分别以粗体和下划线显示。nd第二高分数分别以粗体和下划线显示。**

	基准测试	Qwen3-VL 235B-A22B 指令	Qwen3 235B-A22B 指令-2507	Deepseek V3 0324	Claude-Opus-4 (不思考)
知识	MMLU-Pro	81.8	83.0	81.2	**86.6**
	MMLU-Redux	92.2	93.1	90.4	**94.2**
	GPQA	74.3	**77.5**	68.4	74.9
	SuperGPQA	60.4	**62.6**	57.3	56.5
推理	AIME-25	**74.7**	70.3	46.6	33.9
	HMMT-25	**57.4**	55.4	27.5	15.9
	LiveBench 2024-11-25	74.8	**75.4**	66.9	74.6
对齐任务	IFEval	87.8	**88.7**	82.3	87.4
	Arena-Hard V2 (winrate)	77.4	**79.2**	45.6	51.5
	创意写作 v3	86.5	**87.5**	81.6	83.8
	WritingBench	**85.5**	85.2	74.5	79.2
编程与代理	LiveCodeBench v6	**54.3**	51.8	45.2	44.6
	BFCL-v3	67.7	**70.9**	64.7	60.1
多语制	MultiIF	76.3	**77.5**	66.5	-
	MMLU-ProX	77.8	**79.4**	75.8	-
	INCLUDE	80.0	79.5	**80.1**	-
	PolyMATH	45.1	**50.2**	32.2	30.0

= 0.95，top-k = 20，并且额外应用存在惩罚1.5以鼓励更大的输出多样性。我们设置最大输出长度为32,768个token，但在AIME-25、HMMT-25和LiveCodeBench v6中，我们将长度扩展到81,920个token，以提供足够的思考空间。

详细结果如下。

Qwen3-VL-235B-A22B 我们对比旗舰模型Qwen3-VL-235B-A22B与领先指令和思考模型。对于Qwen3-VL-235B-A22B-Instruct，我们选取Qwen3-235B-A22B-Instruct-2507、DeepSeek V3 0324和Claude-Opus-4（不带思考）作为基线。对于Qwen3-VL-235B-A22B-Thinking，我们选取Qwen3-235B-A22B-Thinking-2507、OpenAI o3（中）、Claude-Opus-4（带思考）作为基线。我们在表5 和表6中展示评估结果。

- 从表5中可以看出，Qwen3-VL-235B-A22B-Instruct取得了具有竞争力的结果，表现与或甚至优于其他领先模型，包括DeepSeek V3 0324、Claude-Opus-4（不带思考）以及我们之前的旗舰模型Qwen3-235B-A22B-Instruct-2507。特别地，Qwen3-VL-235B-A22B-Instruct在推理需求任务（如数学和编程）上超越了其他模型。值得注意的是，DeepSeek V3 0324和Qwen3-235B-A22B-Instruct-2507是大语言模型，而Qwen3-VL-235B-A22B-Instruct是视觉语言模型，能够处理视觉和文本任务。这意味着Qwen3-VL-235B-Instruct已经实现了视觉和文本能力的融合。
- 从表 6来看，Qwen3-VL-235B-A22B-Thinking 在与其他领先思考模型相比中也取得了具有竞争力的结果。Qwen3-VL-235B-A22B-Thinking 在 AIME-25 和 LiveCodeBench v6 上超过了 OpenAI o3（中）和 Claude- Opus-4（带思考功能），这意味着 Qwen3-VL-235B-A22B-Thinking 具有更好的推理能力。

Qwen3-VL-32B / 30B-A3B 我们将我们的 Qwen3-VL-32B 和 Qwen3-VL-30B-A3B 模型与其对应的纯文本版本进行比较，即 Qwen3-32B、Qwen3-30B-A3B 和 Qwen3-30B-A3B-2507。我们在表7 和表 8中展示了评估结果。

- 从表 7来看，对于指令模型，Qwen3-VL-32B 和 Qwen3-VL-30B-A3B 在所有基准测试上与 Qwen3-32B 和 Qwen3-30B-A3B 相比都显示了显著的性能提升。Qwen3-VL-30B-A3B 在 AIME-25 和 HMMT-25 上取得了与 Qwen3-30B-A3B-2507 相当甚至更好的结果。
- 从表8，对于思考模型，Qwen3-VL-32B和Qwen3-VL-30B-A3B在大多数基准测试中超越了基线。Qwen3-VL-30B-A3B与Qwen3-30B-A3B-2507的性能也相当。

Table 6: **Comparison among Qwen3-VL-235B-A22B (Thinking) and other reasoning baselines. The highest and second-best scores are shown in bold and underlined respectively.**

	Benchmark	Qwen3-VL 235B-A22B Thinking	Qwen3 235B-A22B Thinking-2507	OpenAI o3 (medium)	Claude-Opus-4 (With thinking)
Knowledge	MMLU-Pro	83.8	84.4	**85.9**	-
	MMLU-Redux	93.7	93.8	**94.9**	94.6
	GPQA	77.1	81.1	**83.3(high)**	79.6
	SuperGPQA	64.3	**64.9**	-	-
Reasoning	AIME-25	89.7	**92.3**	88.9(high)	75.5
	HMMT-25	77.4	**83.9**	77.5	58.3
	LiveBench 2024-11-25	**79.6**	78.4	78.3	78.2
Coding	LiveCodeBench v6	70.1	**74.1**	58.6	48.9
	CFEval	1964	**2134**	2043	-
	OJBench	27.5	**32.5**	25.4	-
Alignment Tasks	IFEval	88.2	87.8	**92.1**	89.7
	Arena-Hard V2 (winrate)	74.8	79.7	**80.8**	59.1
	Creative Writing v3	85.7	86.1	**87.7**	83.8
	WritingBench	86.7	**88.3**	85.3	79.1
Agent	BFCL-v3	71.8	71.9	**72.4**	61.8
	TAU2-Retail	67.0	71.9	**76.3**	-
	TAU2-Airline	62.0	58.0	**70.0**	-
	TAU2-Telecom	44.7	45.6	**60.5**	-
Multilingualism	MultiIF	79.1	**80.6**	80.3	-
	MMLU-ProX	80.6	81.0	**83.3**	-
	INCLUDE	80.0	81.0	**86.6**	-
	PolyMATH	57.8	**60.1**	49.7	-

Table 7: **Comparison among Qwen3-VL-32B-Instruct, Qwen3-VL-30B-A3B-Instruct, and corresponding baselines.**

	Benchmark	Qwen3-VL 32B Instruct	Qwen3 32B Instruct	Qwen3-VL 30B-A3B Instruct	Qwen3 30B-A3B Instruct	Qwen3 30B-A3B Instruct-2507
Knowledge	MMLU-Pro	78.6	71.9	77.8	69.1	78.4
	MMLU-Redux	89.8	85.7	88.4	84.1	89.3
	GPQA	68.9	54.6	70.4	54.8	70.4
	SuperGPQA	54.6	43.2	53.1	42.2	53.4
Reasoning	AIME-25	66.2	20.2	69.3	21.6	61.3
	HMMT-25	46.1	10.9	50.6	12.0	43.0
	LiveBench 2024-11-25	72.2	31.3	65.4	59.4	69.0
Alignment Tasks	IFEval	84.7	83.2	85.8	83.7	84.7
	Arena-Hard V2 (winrate)	64.7	37.4	58.5	24.8	69.0
	Creative Writing v3	85.6	80.6	84.6	68.1	86.0
	WritingBench	82.9	81.3	82.6	72.2	85.5
Coding & Agent	LiveCodeBench v6	43.8	29.1	42.6	29.0	43.2
	BFCL-v3	70.2	63.0	66.3	58.6	65.1
Multilingualism	MultiIF	72.0	70.7	66.1	70.8	67.9
	MMLU-ProX	73.4	69.3	70.9	65.1	72.0
	INCLUDE	74.0	69.6	71.6	67.8	71.9
	PolyMATH	40.5	22.5	44.3	23.3	43.1

Qwen3-VL-8B / 4B / 2B We present the evaluation results of Qwen3-VL-2B, Qwen3-VL-4B, and Qwen3-VL-8B in Table 9 and Table 10. For Qwen3-VL-2B and Qwen3-VL-8B, we compare them with Qwen3-1.7B and Qwen3-8B. For Qwen3-VL-4B, we compare it with Qwen3-4B and Qwen3-4B-2507. Overall, these edge-side models exhibit impressive performance and outperform baselines. These results demonstrate

表6：**Co Qwen3-VL-235B-A22B（思考）与其他推理基线的比较。T 最高和 第二好的分数分别以粗体和下划线显示。**

	基准测试	Qwen3-VL 235B-A22B 思考	Qwen3 235B-A22B 思考-2507	OpenAI o3 (中等)	Claude-Opus-4 (带思考)
知识	MMLU-Pro	83.8	84.4	**85.9**	-
	MMLU-Redux	93.7	93.8	**94.9**	94.6
	GPQA	77.1	81.1	**83.3(高)**	79.6
	SuperGPQA	64.3	**64.9**	-	-
推理	AIME-25	89.7	**92.3**	88.9(高)	75.5
	HMMT-25	77.4	**83.9**	77.5	58.3
	LiveBench 2024-11-25	**79.6**	78.4	78.3	78.2
编程	LiveCodeBench v6	70.1	**74.1**	58.6	48.9
	CFEval	1964	**2134**	2043	-
	OJBench	27.5	**32.5**	25.4	-
对齐任务	IFEval	88.2	87.8	**92.1**	89.7
	Arena-Hard V2 (winrate)	74.8	79.7	**80.8**	59.1
	创意写作 v3	85.7	86.1	**87.7**	83.8
	WritingBench	86.7	**88.3**	85.3	79.1
代理	BFCL-v3	71.8	71.9	**72.4**	61.8
	TAU2-Retail	67.0	71.9	**76.3**	-
	TAU2-Airline	62.0	58.0	**70.0**	-
	TAU2-Telecom	44.7	45.6	**60.5**	-
多语制	MultiIF	79.1	**80.6**	80.3	-
	MMLU-ProX	80.6	81.0	**83.3**	-
	INCLUDE	80.0	81.0	**86.6**	-
	PolyMATH	57.8	**60.1**	49.7	-

表7： **Qwen3-VL-32B-Instruct、Qwen3-VL-30B-A3B-Instruct与对应基线的比较。**

	基准测试	Qwen3-VL 32B 指令	Qwen3 32B 指令	Qwen3-VL 30B-A3B 指令	Qwen3 30B-A3B 指令	Qwen3 30B-A3B 指令-2507
知识	MMLU-Pro	78.6	71.9	77.8	69.1	78.4
	MMLU-Redux	89.8	85.7	88.4	84.1	89.3
	GPQA	68.9	54.6	70.4	54.8	70.4
	SuperGPQA	54.6	43.2	53.1	42.2	53.4
推理	AIME-25	66.2	20.2	69.3	21.6	61.3
	HMMT-25	46.1	10.9	50.6	12.0	43.0
	LiveBench 2024-11-25	72.2	31.3	65.4	59.4	69.0
对齐任务	IFEval	84.7	83.2	85.8	83.7	84.7
	Arena-Hard V2 (winrate)	64.7	37.4	58.5	24.8	69.0
	创意写作 v3	85.6	80.6	84.6	68.1	86.0
	WritingBench	82.9	81.3	82.6	72.2	85.5
编程与代理	LiveCodeBench v6	43.8	29.1	42.6	29.0	43.2
	BFCL-v3	70.2	63.0	66.3	58.6	65.1
多语制	MultiIF	72.0	70.7	66.1	70.8	67.9
	MMLU-ProX	73.4	69.3	70.9	65.1	72.0
	INCLUDE	74.0	69.6	71.6	67.8	71.9
	PolyMATH	40.5	22.5	44.3	23.3	43.1

Qwen3-VL-8B / 4B / 2B 我们展示了 Qwen3-VL-2B、Qwen3-VL-4B 和 Qwen3-VL-8B 的评估结果，见表9 和表10。对于 Qwen3-VL-2B 和 Qwen3-VL-8B，我们将其与 Qwen3-1.7B 和 Qwen3-8B 进行比较。对于 Qwen3-VL-4B，我们将其与 Qwen3-4B 和 Qwen3-4B-2507 进行比较。总体而言，这些边缘端模型表现出色，并优于基线。这些结果证明

Table 8: **Comparison among Qwen3-VL-32B (Thinking), Qwen3-VL-30B-A3B (Thinking), and corresponding baselines.**

	Benchmark	Qwen3-VL 32B Thinking	Qwen3 32B Thinking	Qwen3-VL 30B-A3B Thinking	Qwen3 30B-A3B Thinking	Qwen3 30B-A3B Thinking-2507
Knowledge	MMLU-Pro	82.1	79.1	80.5	78.5	80.9
	MMLU-Redux	91.9	90.9	90.9	89.5	91.4
	GPQA	73.1	68.4	74.4	65.8	73.4
	SuperGPQA	59.0	54.1	56.4	51.8	56.8
Reasoning	AIME-25	83.7	72.9	83.1	70.9	85.0
	HMMT-25	64.6	51.8	67.6	49.8	71.4
	LiveBench 2024-11-25	74.7	65.7	72.1	74.3	76.8
Coding	LiveCodeBench v6	65.6	60.6	64.2	57.4	66.0
	CFEval	1842	1986	1894	1940	2044
	OJBench	20.0	24.1	23.4	20.7	25.1
Alignment Tasks	IFEval	87.8	85.0	81.7	86.5	88.9
	Arena-Hard V2 (winrate)	60.5	50.3	56.7	36.3	56.0
	Creative Writing v3	83.3	84.4	82.5	79.1	84.4
	WritingBench	86.2	78.4	85.2	77.0	85.0
Agent	BFCL-v3	71.7	70.3	68.6	69.1	72.4
	TAU2-Retail	59.4	59.6	64.0	34.2	58.8
	TAU2-Airline	52.5	38.0	48.0	36.0	58.0
	TAU2-Telecom	46.9	26.3	27.2	22.8	26.3
Multilingualism	MultiIF	78.0	73.0	73.0	72.2	76.4
	MMLU-ProX	77.2	74.6	76.1	73.1	76.4
	INCLUDE	76.3	73.7	74.5	71.9	74.4
	PolyMATH	52.0	47.4	51.7	46.1	52.6

Table 9: **Comparison among Qwen3-VL-2B (Instruct), Qwen3-VL-4B (Instruct), Qwen3-VL-8B (Instruct) and corresponding baselines.**

	Benchmark	Qwen3-VL 2B Instruct	Qwen3-VL 4B Instruct	Qwen3-VL 8B Instruct	Qwen3 1.7B Instruct	Qwen3 4B Instruct	Qwen3 8B Instruct	Qwen3 4B Instruct-2507
Knowledge	MMLU-Pro	49.0	67.1	71.6	42.3	58.0	63.4	69.6
	MMLU-Redux	66.5	81.5	84.9	63.6	77.3	79.5	84.2
	GPQA	42.0	55.9	61.9	34.7	41.7	39.3	62.0
	SuperGPQA	24.3	40.3	44.5	22.8	32.0	35.8	42.8
Reasoning	AIME-25	22.2	46.6	45.9	10.6	19.1	20.9	47.4
	HMMT-25	10.9	30.7	32.5	6.2	12.1	11.8	31.0
	LiveBench 2024-11-25	39.5	60.9	62.0	35.6	48.4	53.5	63.0
Alignment Tasks	IFEval	68.2	82.3	83.7	67.1	81.2	83.0	83.4
	Arena-Hard V2 (winrate)	6.4	30.4	46.3	4.1	9.5	15.5	43.4
	Creative Writing v3	48.6	72.3	77.0	49.1	53.6	69.0	83.5
	WritingBench	73.0	82.5	83.1	65.1	68.5	71.4	83.4
Coding & Agent	LiveCodeBench v6	20.3	37.9	39.3	16.1	26.4	25.5	35.1
	BFCL-v3	55.4	63.3	66.3	52.2	57.6	60.2	61.9
Multilingualism	MultiIF	43.2	61.5	66.8	43.2	61.3	69.2	69.0
	MMLU-ProX	38.8	59.4	65.4	33.5	49.6	58.0	61.6
	INCLUDE	45.8	61.4	67.0	42.6	53.8	62.5	60.1
	PolyMATH	14.9	28.8	30.4	10.3	16.6	18.8	31.1

the efficacy of our Strong-to-Weak Distillation approach, making it possible for us to build the lightweight models with remarkably reduced costs and efforts.

表8：**Qwen3-VL-32B（思考）、Qwen3-VL-30B-A3B（思考）及其对应基线的比较。**

	基准测试	Qwen3-VL 32B 思考	Qwen3 32B 思考	Qwen3-VL 30B-A3B 思考	Qwen3 30B-A3B 思考	Qwen3 30B-A3B Thinking-2507
知识	MMLU-Pro	82.1	79.1	80.5	78.5	80.9
	MMLU-Redux	91.9	90.9	90.9	89.5	91.4
	GPQA	73.1	68.4	74.4	65.8	73.4
	SuperGPQA	59.0	54.1	56.4	51.8	56.8
推理	AIME-25	83.7	72.9	83.1	70.9	85.0
	HMMT-25	64.6	51.8	67.6	49.8	71.4
	LiveBench 2024-11-25	74.7	65.7	72.1	74.3	76.8
编程	LiveCodeBench v6	65.6	60.6	64.2	57.4	66.0
	CFEval	1842	1986	1894	1940	2044
	OJBench	20.0	24.1	23.4	20.7	25.1
对齐任务	IFEval	87.8	85.0	81.7	86.5	88.9
	Arena-Hard V2 (winrate)	60.5	50.3	56.7	36.3	56.0
	创意写作 v3	83.3	84.4	82.5	79.1	84.4
	WritingBench	86.2	78.4	85.2	77.0	85.0
代理	BFCL-v3	71.7	70.3	68.6	69.1	72.4
	TAU2-Retail	59.4	59.6	64.0	34.2	58.8
	TAU2-Airline	52.5	38.0	48.0	36.0	58.0
	TAU2-Telecom	46.9	26.3	27.2	22.8	26.3
多语制	MultiIF	78.0	73.0	73.0	72.2	76.4
	MMLU-ProX	77.2	74.6	76.1	73.1	76.4
	INCLUDE	76.3	73.7	74.5	71.9	74.4
	PolyMATH	52.0	47.4	51.7	46.1	52.6

表9：**Qwen3-VL-2B（指令）、Qwen3-VL-4B（指令）、Qwen3-VL-8B（指令）与对应基线的比较。**

	基准测试	Qwen3-VL 2B 指令	Qwen3-VL 4B 指令	Qwen3-VL 8B 指令	Qwen3 1.7B 指令	Qwen3 4B 指令	Qwen3 8B 指令	Qwen3 4B 指令-2507
知识	MMLU-Pro	49.0	67.1	71.6	42.3	58.0	63.4	69.6
	MMLU-Redux	66.5	81.5	84.9	63.6	77.3	79.5	84.2
	GPQA	42.0	55.9	61.9	34.7	41.7	39.3	62.0
	SuperGPQA	24.3	40.3	44.5	22.8	32.0	35.8	42.8
推理	AIME-25	22.2	46.6	45.9	10.6	19.1	20.9	47.4
	HMMT-25	10.9	30.7	32.5	6.2	12.1	11.8	31.0
	LiveBench 2024-11-25	39.5	60.9	62.0	35.6	48.4	53.5	63.0
对齐任务	IFEval	68.2	82.3	83.7	67.1	81.2	83.0	83.4
	Arena-Hard V2 (winrate)	6.4	30.4	46.3	4.1	9.5	15.5	43.4
	创意写作 v3	48.6	72.3	77.0	49.1	53.6	69.0	83.5
	WritingBench	73.0	82.5	83.1	65.1	68.5	71.4	83.4
编程与代理	LiveCodeBench v6	20.3	37.9	39.3	16.1	26.4	25.5	35.1
	BFCL-v3	55.4	63.3	66.3	52.2	57.6	60.2	61.9
多语制	MultiIF	43.2	61.5	66.8	43.2	61.3	69.2	69.0
	MMLU-ProX	38.8	59.4	65.4	33.5	49.6	58.0	61.6
	INCLUDE	45.8	61.4	67.0	42.6	53.8	62.5	60.1
	PolyMATH	14.9	28.8	30.4	10.3	16.6	18.8	31.1

我们强到弱蒸馏方法的有效性，使我们能够以显著降低的成本和精力构建轻量级模型。

Table 10: **Comparison among Qwen3-VL-2B (Thinking), Qwen3-VL-4B (Thinking), Qwen3-VL-8B (Thinking) and corresponding baselines.**

	Benchmark	Qwen3-VL 2B Thinking	Qwen3-VL 4B Thinking	Qwen3-VL 8B Thinking	Qwen3 1.7B Thinking	Qwen3 4B Thinking	Qwen3 8B Thinking	Qwen3 4B Thinking-2507
Knowledge	MMLU-Pro	62.3	73.6	77.3	58.1	70.4	74.6	74.0
	MMLU-Redux	76.9	86.0	88.8	73.9	83.7	87.5	86.1
	GPQA	49.5	64.1	69.9	27.9	55.9	62.0	65.8
	SuperGPQA	34.6	46.8	51.2	31.2	42.7	47.6	47.8
Reasoning	AIME-25	39.0	74.5	80.3	36.8	65.6	67.3	81.3
	HMMT-25	22.8	53.1	60.6	24.3	42.1	43.2	55.5
	LiveBench 2024-11-25	50.1	68.4	69.8	51.1	63.6	67.1	71.8
Alignment Tasks	IFEval	75.1	82.6	83.2	72.5	81.9	85.0	87.4
	Arena-Hard V2 (winrate)	12.0	36.8	51.1	4.7	13.7	29.1	34.9
	Creative Writing v3	55.6	76.1	82.4	50.6	61.1	78.5	75.6
	WritingBench	77.9	84.0	85.5	68.9	73.5	75.0	83.3
Coding & Agent	LiveCodeBench v6	29.3	51.3	58.6	31.3	48.4	51.0	55.2
	BFCL-v3	57.2	67.3	63.0	56.6	65.9	68.1	71.2
Multilingualism	MultiIF	58.9	73.6	75.1	51.2	66.3	71.2	77.3
	MMLU-ProX	55.1	65.0	70.7	50.4	61.0	68.1	64.2
	INCLUDE	53.3	64.6	69.5	51.8	61.8	67.8	64.4
	PolyMATH	28.0	44.6	47.5	25.2	40.0	42.7	46.2

5.12 Ablation Study

5.12.1 Vision Encoder

We conduct comparative experiments against the original SigLIP-2. As shown in Table 11, in zero-shot evaluation at the CLIP pretraining stage, Qwen3-ViT maintains competitive performance on standard benchmarks while achieving substantial gains on OmniBench, our in-house holistic evaluation suite designed to assess world knowledge integration under diverse and challenging conditions. Furthermore, when integrated with the same 1.7B Qwen3 language model and trained for 1.5T tokens, Qwen3-ViT consistently outperforms the SigLIP-2-based baseline across multiple key tasks and remains significantly ahead on OmniBench, demonstrating its superiority and effectiveness as a stronger visual backbone.

Table 11: **Ablation on Qwen3-ViT.** We compare the performance metrics of Qwen3-ViT and SigLIP-2 during the CLIP pre-training stage, and further evaluate their downstream performance in the vision-language modeling (VLM) stage when paired with the same 1.7B Qwen3 language model.

ViT	Clip Bench							VLM Bench				
	ImageNet-1K	ImageNet-V2	ImageNet-A	ImageNet-R	ImageNet-S	ObjectNet	Omni	OCRB	AI2D	RLWDQA	InfoVQA	Omni
SigLIP-2	84.2	78.6	87.0	96.1	76.2	79.9	36.9	77.2	74.1	58.7	65.3	50.1
Qwen3-ViT	84.6	78.8	87.1	95.7	74.5	81.0	45.5	78.7	76.2	66.1	67.0	53.0

5.12.2 DeepStack

We conduct an ablation study to verify the effectiveness of the DeepStack mechanism. As demonstrated in Table 12, the model equipped with DeepStack achieved an overall performance gain across various benchmarks, strongly affirming its effectiveness. This gain is attributed to DeepStack's ability to integrate rich visual information, which effectively boosts the capability in fine-grained visual understanding, such as on the InfoVQA and DocVQA benchmarks.

Table 12: **Ablation on DeepStack.** We conduct the ablation study on the DeepStack using an internal 15B-A2B LLM, with all experiments pretrained on 200 billion tokens. We directly evaluate these pretrained models on the validation sets, without any post-training.

Method	AVG	AI2D	OCRB	TVQA	InfoVQA	ChartQA	DocVQA	MMMU	MMStar	RLWDQA	MMB_{EN}	MMB_{CN}
Baseline	74.7	81.8	81.0	80.6	71.9	81.5	89.5	52.9	55.5	67.7	81.0	78.1
DeepStack	76.0	83.2	83.6	80.5	74.2	83.3	91.1	54.1	57.7	68.1	81.2	78.5

表10： C Qwen3-VL-2B (思考), Qwen3-VL-4B (思考), Qwen3-VL 8B (思考)对应基线。

	基准测试	Qwen3-VL 2B 思考	Qwen3-VL 4B 思考	Qwen3-VL 8B 思考	Qwen3 1.7B 思考	Qwen3 4B 思考	Qwen3 8B 思考	Qwen3 4B Thinking-2507
知识	MMLU-Pro	62.3	73.6	77.3	58.1	70.4	74.6	74.0
	MMLU-Redux	76.9	86.0	88.8	73.9	83.7	87.5	86.1
	GPQA	49.5	64.1	69.9	27.9	55.9	62.0	65.8
	SuperGPQA	34.6	46.8	51.2	31.2	42.7	47.6	47.8
推理	AIME-25	39.0	74.5	80.3	36.8	65.6	67.3	81.3
	HMMT-25	22.8	53.1	60.6	24.3	42.1	43.2	55.5
	LiveBench 2024-11-25	50.1	68.4	69.8	51.1	63.6	67.1	71.8
对齐任务	IFEval	75.1	82.6	83.2	72.5	81.9	85.0	87.4
	Arena-Hard V2 (winrate)	12.0	36.8	51.1	4.7	13.7	29.1	34.9
	创意写作 v3	55.6	76.1	82.4	50.6	61.1	78.5	75.6
	写作Bench	77.9	84.0	85.5	68.9	73.5	75.0	83.3
编程与代理	LiveCodeBench v6	29.3	51.3	58.6	31.3	48.4	51.0	55.2
	BFCL-v3	57.2	67.3	63.0	56.6	65.9	68.1	71.2
多语制	MultiIF	58.9	73.6	75.1	51.2	66.3	71.2	77.3
	MMLU-ProX	55.1	65.0	70.7	50.4	61.0	68.1	64.2
	INCLUDE	53.3	64.6	69.5	51.8	61.8	67.8	64.4
	PolyMATH	28.0	44.6	47.5	25.2	40.0	42.7	46.2

5.12 消融研究

5.12.1 视觉编码器

我们与原始SigLIP-2进行了对比实验。如表11所示，在CLIP预训练阶段的零样本评估中，Qwen3-ViT在标准基准测试上保持了具有竞争力的性能，同时在我们的内部整体评估套件OmniBench上实现了显著提升。该套件旨在评估在不同和具有挑战性的条件下整合世界知识的能力。此外，当与相同的1.7B Qwen3语言模型集成并训练1.5T令牌时，Qwen3-ViT在多个关键任务上始终优于SigLIP-2基线，并在OmniBench上保持显著领先，这证明了它作为更强的视觉主干的优势和有效性。

表11：**Qwen3-ViT的消融实验**。我们比较了Qwen3-ViT和SigLIP-2在CLIP预训练阶段的性能指标，并进一步评估它们在与相同的1.7B Qwen3语言模型配对时的下游性能（视觉-语言建模阶段）。

ViT	CLIP Bench							视觉语言模型测试平台				
	ImageNet-1K	ImageNet-V2	ImageNet-A	ImageNet-R	ImageNet-S	对象Net	Omni	OCRB	AI2D	RLWDQA	InfoVQA	Omni
SigLIP-2	84.2	78.6	87.0	96.1	76.2	79.9	36.9	77.2	74.1	58.7	65.3	50.1
Qwen3-ViT	84.6	78.8	87.1	95.7	74.5	81.0	45.5	78.7	76.2	66.1	67.0	53.0

5.12.2 DeepStack

我们进行消融研究以验证 DeepStack 机制的有效性。如表12所示，配备 DeepStack 的模型在多个基准测试中实现了整体性能提升，有力地证实了其有效性。这种提升归因于 DeepStack 集成丰富视觉信息的能力，这有效地提升了在细粒度视觉理解方面的能力，例如在 InfoVQA 和 DocVQA 基准测试上。

表12：**DeepStack消融实验**。 我们在DeepStack上进行了消融研究，使用内部15B-A2B大语言模型，所有实验都在2000亿token上进行预训练。我们直接在验证集上评估这些预训练模型，没有任何后训练。

方法	AVG	AI2D	OCRB	TVQA	InfoVQA	ChartQA	DocVQA	MMMU	MMStar	RLWDQA	MMB_{EN}	MMB_{CN}
基线	74.7	81.8	81.0	80.6	71.9	81.5	89.5	52.9	55.5	67.7	81.0	78.1
DeepStack	76.0	83.2	83.6	80.5	74.2	83.3	91.1	54.1	57.7	68.1	81.2	78.5

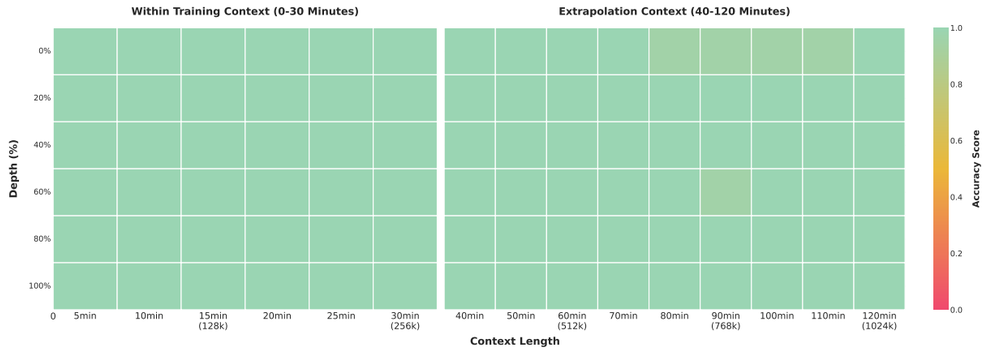


Figure 3: Needle-in-a-Haystack performance heatmap for Qwen3-VL-235B-A22B-Instruct across varying video durations and needle positions. Each cell shows accuracy (%) for locating and answering questions about the inserted "needle" frame.

5.12.3 Needle-in-a-Haystack

To evaluate the model's capability in processing long-context inputs, we construct a video "Needle-in-a-Haystack" evaluation on Qwen3-VL-235B-A22B-Instruct. In this task, a semantically salient "needle" frame—containing critical visual evidence—is inserted at varying temporal positions within a long video. The model is then tasked with accurately locating the target frame from the long video and answering the corresponding question. During evaluation, videos are uniformly sampled at 1 FPS, and frame resolution is dynamically adjusted to maintain a constant visual token budget.

As shown in Figure 3, the model achieves a perfect 100% accuracy on videos up to 30 minutes in duration—corresponding to a context length of 256K tokens. Remarkably, even when extrapolating to sequences of up to 1M tokens (approximately 2 hours of video) via YaRN-based positional extension, the model retains a high accuracy of 99.5%. These results strongly demonstrate the model's powerful long-sequence modeling capabilities.

6 Conclusion

In this work, we present Qwen3-VL, a state-of-the-art series of vision–language foundation models that advances the frontier of multimodal understanding and generation. By integrating high-quality multimodal data iteration and architectural innovations—such as enhanced interleaved-MRoPE, DeepStack vision-language alignment, and text-based temporal grounding—Qwen3-VL achieves unprecedented performance across a broad spectrum of multimodal benchmarks while maintaining strong pure-text capabilities. Its native support for 256K-token interleaved sequences enables robust reasoning over long, complex documents, image sequences, and videos, making it uniquely suited for real-world applications demanding high-fidelity cross-modal comprehension. The availability of both dense and Mixture-of-Experts variants ensures flexible deployment across diverse latency and quality requirements, and our post-training strategy—including non-thinking and thinking modes.

Looking forward, we envision Qwen3-VL as a foundational engine for embodied AI agents capable of seamlessly bridging the digital and physical worlds. Such agents will not only perceive and reason over rich multimodal inputs but also execute decisive, context-aware actions in dynamic environments—interacting with users, manipulating digital interfaces, and guiding robotic systems through grounded, multimodal decision-making. Future work will focus on extending Qwen3-VL's capabilities toward interactive perception, tool-augmented reasoning, and real-time multimodal control, with the ultimate goal of enabling AI systems that learn, adapt, and collaborate alongside humans in both virtual and physical domains. Additionally, we are actively exploring unified understanding-generation architectures, leveraging visual generation capabilities to elevate overall intelligence further. By openly releasing the entire model family under the Apache 2.0 license, we aim to catalyze community-driven innovation toward the vision of truly integrated, multimodal AI agents.

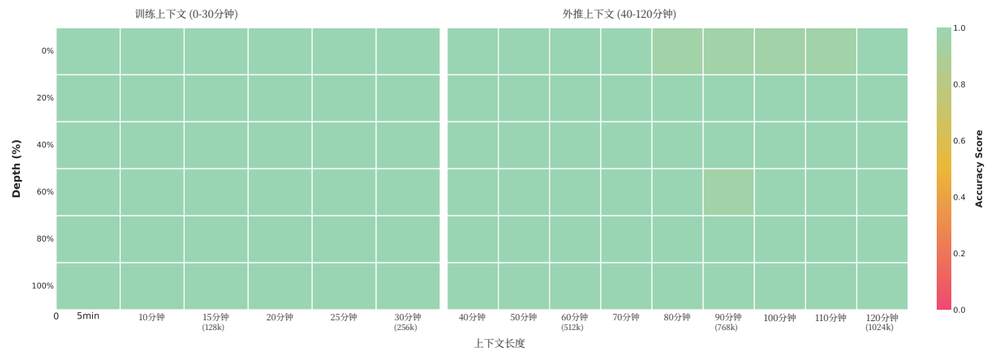


图3：Qwen3-VL-235B-A22B-Instruct 在不同视频时长和针位置下的"针 haystack"性能热力图。每个单元格显示定位并回答关于插入的"针"帧问题的准确性 (%)。

5.12.3 针 haystack

为了评估模型处理长上下文输入的能力，我们在 Qwen3-VL-235B-A22B-Instruct 上构建了一个视频"大海捞针"评估任务。在这个任务中，一个语义上显著的"针"帧——包含关键视觉证据——被插入到长视频的不同时间位置。然后，模型需要从长视频中准确定位目标帧并回答相应的问题。在评估过程中，视频以 1 FPS 的速率均匀采样，帧分辨率会动态调整以保持恒定的视觉令牌预算。

如图 3所示，该模型在时长最长可达30分钟的视频上实现了完美的100%准确率——对应于256K个token的上下文长度。值得注意的是，即使通过基于YaRN的位置扩展将序列扩展到高达1M个token（约2小时的视频），模型仍保持了高达99.5%的高准确率。这些结果有力地证明了该模型强大的长序列建模能力。

6 结论

在这项工作中，我们提出了Qwen3-VL，这是一系列最先进的视觉-语言基础模型，推动了多模态理解和生成的边界。通过整合高质量的多模态数据迭代和架构创新——例如增强的交错-MRoPE、DeepStack视觉-语言对齐以及基于文本的时间 grounding——Qwen3-VL在广泛的多模态基准测试中实现了前所未有的性能，同时保持了强大的纯文本能力。其对256K个token交错序列的原生支持能够对长、复杂文档、图像序列和视频进行稳健的推理，使其特别适合需要高保真跨模态理解的现实世界应用。密集型和专家混合变体的可用性确保了在不同延迟和质量要求下的灵活部署，我们的后训练策略——包括非思考模式和思考模式。

展望未来，我们期望Qwen3-VL成为能够无缝连接数字世界和物理世界的具身AI代理的基础引擎。这类代理不仅能够对丰富的多模态输入进行感知和推理，还能在动态环境中执行决策性、上下文感知的行动——与用户交互、操作数字界面，并通过基于多模态的决策引导机器人系统。未来的工作将聚焦于扩展Qwen3-VL在交互式感知、工具增强推理和实时多模态控制方面的能力，最终目标是实现能够在虚拟和物理领域与人类共同学习、适应和协作的AI系统。此外，我们正在积极探索统一的理解生成架构，利用视觉生成能力进一步提升整体智能。通过在Apache 2.0许可证下公开发布整个模型家族，我们旨在推动社区驱动的创新，朝着真正集成化、多模态AI代理的愿景前进。

7 Contributions and Acknowledgments

All contributors of Qwen3-VL are listed in alphabetical order by their last names.

Core Contributors: Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, Ke Zhu

Contributors: Yizhong Cao, Bei Chen, Chen Cheng, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Rongyao Fang, Tongkun Guan, Jinzheng He, Miao Hong, Songtao Jiang, Zheng Li, Xiaochuan Li, Junrong Lin, Yuqiong Liu, Yantao Liu, Na Ni, Xinyao Niu, Yatian Pang, Zihan Qiu, Tianhao Shen, Tianyi Tang, Yu Wan, Jinxi Wei, Chenfei Wu, Buxiao Wu, Xiao Xu, Mingfeng Xue, Ming Yan, Yuhuan Yang, Jiaxi Yang, Kexin Yang, Le Yu, Hao Yu, Jianke Zhang, Jianwei Zhang, Yichang Zhang, Zhenru Zhang, Siqi Zhang, Peiyang Zhang, Beichen Zhang, Hongbo Zhao, Xianwei Zhuang

Acknowledgments: We gratefully acknowledge the unwavering support provided by the teams led by Zulong Chen, Bing Deng, Feiyu Gao, Guanjun Jiang, Yue Liu, Hangdi Xing and Daijun Yu.

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.

AIME. Aime problems and solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIMEProblemsandSolutions.

Anthropic. Claude opus 4.1, 2025. URL https://www.anthropic.com/news/claude-opus-4-1.

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, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13154–13164, 2023.

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*, 2024a.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. Timemarker: A versatile video-llm for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024b.

Yitong Chen, Lingchen Meng, Wujian Peng, Zuxuan Wu, and Yu-Gang Jiang. Comp: Continual multimodal pre-training for vision foundation models. *arXiv preprint arXiv:2503.18931*, 2025.

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.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4637–4646, 2025.

7 贡献与致谢

Qwen3-VL的所有贡献者按其姓氏字母顺序列出。

核心贡献者： 白帅，蔡宇轩，陈瑞哲，陈科钦，陈雄辉，程泽森，邓亮浩，丁伟，高长，葛春江，葛文斌，郭志方，黄启东，黄杰，黄飞，惠宾元，蒋舒通，李兆海，李明生，李梅，李凯欣，林子成，林俊阳，刘雪晶，刘佳伟，刘成龙，刘杨，刘代恒，刘世玄，陆敦杰，罗瑞林，吕晨旭，门睿，孟令晨，任宣成，任兴章，宋思博，孙宇冲，唐军，涂建红，万建强，王鹏，王鹏飞，王秋月，王宇轩，谢天宝，徐一恒，徐海洋，徐金，杨志博，杨明坤，杨建新，杨安，余 Bowen，张飞，张航，张希，郑波，钟虎门，周景仁，周帆，周静，朱元志，朱科

贡献者： 曹一忠，北辰，程晨，褚云飞，崔泽宇，唐凯，邓晓东，范杨，方荣瑶，关同坤，何金政，洪妙，蒋松涛，李政，李晓川，林俊荣，刘宇琼，刘言涛，倪娜，牛新瑶，庞亚天，邱志涵，沈天浩，唐天毅，万宇，魏金溪，吴晨飞，吴不晓，徐晓，薛明峰，闫明，杨宇环，杨嘉熙，杨克欣，余乐，余浩，张建科，张建伟，张一畅，张振如，张思琪，张沛阳，张北辰，赵洪波，庄县伟

致谢： 我们衷心感谢由陈祖龙、邓冰、高飞宇、蒋冠军、刘越、刑杭迪和余戴军领导的研究团队提供的坚定支持。

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, et al. Pixtral 12b。*arXiv* 预印本 *arXiv:2410.07073*, 2024。

AIME。AIME 试题及解答，2025。URL https://artofproblemsolving.com/wiki/index.php/AIMEProblemsandSolutions。

Anthropic。Claude opus 4.1，2025。URLhttps://www.anthropic.com/news/claude-opus-4-1。

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, 和 Junyang Lin。Qwen2.5-VL 技术报告，2025。

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, 等. Arkitscenes：一个用于使用移动RGB-D数据对3D室内场景进行理解的多样化真实世界数据集。*arXiv*预印本 *arXiv:2111.08897*, 2021.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson 和 Georgia Gkioxari. Omni3d：一个用于野外3D对象检测的大型基准和模型。在 *IEEE/CVF*计算机视觉与模式识别会议论文集，第13154–13164页，2023年。

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, 等。我们是否正在正确地评估大视觉语言模型? *arXiv:2403.20330*, 2024a。

沈明, 兰晓寒, 袁奕天, 贾泽群, 和马琳. Timemarker: 一种通用的视频-大语言模型，用于长视频和短视频理解，具有卓越的时间定位能力. *arXiv*预印本*arXiv:2411.18211*, 2024b.

陈奕彤, 孟令晨, 彭武健, 吴祖玄, 和蒋宇刚. Comp: 用于视觉基础模型的持续多模态预训练. *arXiv*预印本 *arXiv:2503.18931*, 2025.

程侃之, 孙启舒, 储友刚, 许方智, 李言涛, 张建兵, 和吴志勇. Seeclick: 利用guigrounding为高级视觉gui代理赋能. *arXiv*预印本*arXiv:2401.10935*, 2024.

程先福、张伟、张世伟、杨建、关祥元、吴先杰、李祥、张格、刘家恒、麦宇莹等。SimpleVQA：用于多模态大语言模型的多模态事实性评估。在《IEEE/CVF国际计算机视觉会议论文集》中，第4637–4646页，2025年。

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, Zijia Chen, Peter Belcak, Yoshi Suhara, Hongxu Yin, et al. Climb: Clustering-based iterative data mixture bootstrapping for language model pre-training. *arXiv preprint arXiv:2504.13161*, 2025.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. 2024.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024.

Chengqi Duan, Kaiyue Sun, Rongyao Fang, Manyuan Zhang, Yan Feng, Ying Luo, Yufang Liu, Ke Wang, Peng Pei, Xunliang Cai, et al. Codeplot-cot: Mathematical visual reasoning by thinking with code-driven images. *arXiv preprint arXiv:2510.11718*, 2025.

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.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

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, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. Are we done with mmlu? *CoRR*, abs/2406.04127, 2024. doi: 10.48550/ARXIV.2406.04127. URL https://doi.org/10.48550/arXiv.2406.04127.

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, 2023.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *CoRR*, abs/2410.15553, 2024. doi: 10.48550/ARXIV.2410.15553. URL https://doi.org/10.48550/arXiv.2410.15553.

HMMT. Hmmt 2025. https://www.hmmt.org, 2025.

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.

Gheorghe Comanici、Eric Bieber、Mike Schaekermann、Ice Pasupat、Noveen Sachdeva、Inderjit Dhillon、Marcel Blistein、Ori Ram、张丹、Evan Rosen等。Gemini 2.5：通过高级推理、多模态、长上下文和下一代代理能力推动前沿。*arXiv*预印本*arXiv:2507.06261*，2025年。

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

Shizhe Diao、Yu Yang、Yonggan Fu、Xin Dong、Dan Su、Markus Kliegl、Zijia Chen、Peter Belcak、Yoshi Suhara、Hongxu Yin 等。Climb：基于聚类的迭代数据混合自举语言模型预训练。*arXiv* 预印本 *arXiv:2504.13161*，2025 年。

Matthijs Douze、Alexandr Guzhva、Chengqi Deng、Jeff Johnson、Gergely Szilvasy、Pierre-Emmanuel Mazaré、Maria Lomeli、Lucas Hosseini 和 Hervé Jégou。Faiss 库。2024 年。

杜梦飞, 吴彬豪, 李泽俊, 黄宣径, 魏中宇. Embspatial-bench: 基于大视觉语言模型对具身任务的空间理解进行基准测试. *arXiv* 预印本 *arXiv:2406.05756*, 2024.

段成琪, 孙凯悦, 方荣耀, 张曼远, 冯岩, 罗颖, 刘宇方, 王科, 裴鹏, 蔡训良, 等. Codeplot-cot: 通过代码驱动图像进行数学视觉推理. *arXiv* 预印本 *arXiv:2510.11718*, 2025.

付超宇, 戴宇涵, 罗永东, 李雷, 任书豪, 张仁睿, 王之涵, 周晨宇, 沈云航, 张梦丹, 等. Video-mme: 视频分析中多模态大语言模型的首个综合评估基准. *arXiv:2405.21075*, 2024a.

付兴宇，胡宇施，李邦正，冯宇，王浩宇，林旭东，罗丹，诺亚·A·史密斯，马伟秋，拉贾伊·克里希纳。BLINK：多模态大语言模型能看但不能感知。在欧洲计算机视觉会议，第148-166页。Springer，2024c。https://arxiv.org/abs/2501.00321。

李邦正，罗丹，诺亚·*A*·史密斯，马伟秋，拉贾伊·克里希纳。*BLINK*：多模态大语言模型能看但不能感知。在欧洲计算机视觉会议，第148-166页。Springer，2024c。

高长, 郑初杰, 陈雄辉, 唐凯, 刘世玄, 余 Bowen, 杨安, 白帅, 周景仁, 和 林俊阳. 软自适应策略优化. *arXiv* 预印本 *arXiv:2511.20347*, 2025.

高继阳, 孙晨, 杨振恒, 和 Nevatia 拉姆. Tall: 基于语言查询的时间活动定位. 在*IEEE* 国际计算机视觉会议论文集, pp. 5267–5275, 2017.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Hong Giwon, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, He Xuanli, 赵宇, 杜晓堂, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, 和 Pasquale Minervini. 我们完成 mmlu 吗？ *CoRR*, abs/2406.04127, 2024. doi: 10.48550/ARXIV.2406.04127. URL https://doi.org/10.48550/arXiv.2406.04127.

关天瑞，刘福晓，吴希阳，先瑞琪，李宗霞，刘晓宇，王西军，陈立昌，黄福荣，Yaser Yacoob，Dinesh Manocha，以及周天毅。HallusionBench：一个用于大视觉语言模型中纠缠语言幻觉 & 视觉错觉的高级诊断套件，2023。

何云，金迪，王超奇，Chloe Bi，Karishma Mandyam，张鹤佳，朱晨，李宁，徐腾宇，吕洪江，Shruti Bhosale，朱晨光，Karthik Abinav Sankararaman，Eryk Helenowski，Melanie Kambadur，Aditya Tayade，马浩，方汉，和王_sinong. Multi-if: 在多轮和多语言指令遵循上基准测试 llms. *CoRR*, abs/2410.15553, 2024. doi: 10.48550/ARXIV.2410.15553. URLhttps://doi.org/10.48550/arXiv.2410.15553.

HMMT. Hmmt 2025. https://www.hmmt.org, 2025.

胡凯瑞, 吴鹏浩, 浦繁艺, 王晓, 张元汉, 薛晔, 李波, 和 刘子为. Video- mmmu: 从多学科专业视频中评估知识获取. *arXiv* 预印本 *arXiv:2501.13826*, 2025.

Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision-language models, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *CoRR*, abs/2403.07974, 2024. doi: 10.48550/ARXIV.2403.07974. URL https://doi.org/10.48550/arXiv.2403.07974.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

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

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396, 2016.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, pp. 1956–1981, 2020.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36: 71683–71702, 2023.

Jinke Li, Jiarui Yu, Chenxing Wei, Hande Dong, Qiang Lin, Liangjing Yang, Zhicai Wang, and Yanbin Hao. Unisvg: A unified dataset for vector graphic understanding and generation with multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13156–13163, 2025a.

Kaixin Li, Yuchen Tian, Qisheng Hu, Ziyang Luo, Zhiyong Huang, and Jing Ma. Mmcode: Benchmarking multimodal large language models for code generation with visually rich programming problems. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 736–783, 2024a.

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, 2025b. URL https://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf. Preprint.

Kaixin Li et al. Iconstack, 2025c. URL https://huggingface.co/datasets/likaixin/IconStack-48M-Rendered-Train.

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*, 2024b.

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, 2022.

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*, 2024c.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *CoRR*, abs/2406.11939, 2024d. doi: 10.48550/ARXIV.2406.11939. URL https://doi.org/10.48550/arXiv.2406.11939.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

黄杰, 刘雪晶, 宋思博, 侯瑞冰, 长红, 林俊阳, 和 白帅. 重访视觉-语言模型中的多模态位置编码, 2025. 娜曼·贾因, 韩京, 郭磊, 李文丁, 颜范嘉, 张天军, 王思达, 阿曼多·索拉-莱扎马, 申库什克, 和 离昂·斯托伊卡. Livecodebench: 对代码的大语言模型进行整体且无污染的评估. *CoRR*,abs/2403.07974, 2024. doi: 10.48550/ARXIV.2403.07974. URLhttps://doi.org/10.48550/arXiv.2403.07974. 金 Bowen, 曾汉思, 越振瑞, 尹晋松, 阿里 Sercan, 王东, 汉姆 Zamani, 和 韩佳伟. Search-r1: 使用强化学习训练大语言模型进行推理和利用搜索引擎. ***arXiv*** 预印本 ***arXiv:2503.09516***, 2025. 杰夫·约翰逊, 马修斯·杜泽, 和 赫尔维·杰古. 使用 GPU 进行十亿规模的相似性搜索. ***IEEE*** 大数据汇刊, 7(3):535–547, 2019. 萨哈尔·卡泽马兹德, 奥尔登 · 维森特 , 马克 · 马滕 , 和 塔马拉 · 伯格 . Referitgame: 在自然场景照片中指代对象 . 收录于 *EMNLP*, 2014. 安尼鲁达·坎巴维, 迈克尔·萨尔瓦托, 艾瑞克·科夫, 崔珉俊, 哈南尼·哈吉希尔兹, 和 阿里·法哈迪. 图表胜千言. *ArXiv*, abs/1603.07396, 2016.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, 等. 开放式图像数据集 v4：大规模统一图像分类、目标检测和视觉关系检测. 国际计算机视觉杂志, 第 1956–1981 页, 2020.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, 和 Hengshuang Zhao. Mini-o3：为视觉搜索扩展推理模式和交互轮次. ***arXiv*** 预印本 ***arXiv:2509.07969***, 2025.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, 等. Obelics: 一个开放的网络规模过滤图像-文本文档数据集. 神经信息处理系统进展, 36: 71683–71702, 2023.

金杰李, 余嘉瑞, 魏晨星, 董汉德, 林强, 杨良静, 王志才, 和郝燕斌. Unisvg: 基于多模态大语言模型的矢量图形理解与生成统一数据集. 收录于《第33届ACM国际多媒体会议论文集》, 第13156–13163页, 2025a.

李凯欣, 田宇晨, 胡启胜, 罗子阳, 黄志勇, 和马静. Mmcode: 基于视觉丰富编程问题的多模态大语言模型代码生成基准测试. 收录于《计算语言学协会发现：EMNLP 2024》, 第736–783页, 2024a.

李凯欣, 孟子阳, 林红赞, 罗子阳, 田宇晨, 马静, 黄志勇, 和蔡天锡. Screenspot-pro: 专业高分辨率计算机使用的界面Grounding, 2025b. URLhttps://likaixin2000.github.io/papers/ScreenSpot_Pro.pdf. 预印本.

李凯欣等. Iconstack, 2025c. URLhttps://huggingface.co/datasets/likaixin/IconStack-48M-Rendered-Train. 李坤昌, 王亚丽, 何一南, 李一卓, 王一, 刘一, 王尊, 许继兰, 陈果, 罗平, 等. Mvbench: 一个综合性的多模态视频理解基准测试. 在 *CVPR*, 2024b.

李伦年·哈利·李, 张鹏川, 张浩天, 杨建伟, 李春元, 钟一武, 王丽娟, 袁路, 张雷, 黄仁能, 等. Grounded language-image 预训练. 在***IEEE/CVF*** 计算机视觉与模式识别会议论文集, pp. 10965–10975, 2022.

李庆云, 陈哲, 王伟云, 王文海, 叶胜龙, 金振江, 陈冠舟, 何一南, 高张伟, 崔尔飞, 等. Omnicorpus: 一个包含 100 亿级图像与文本交织的统一多模态语料库. ***arXiv*** 预印本 ***arXiv:2406.08418***, 2024c.

李天乐，蒋伟林，刘文轩，丽莎·邓拉普，吴天浩，朱邦华，约瑟夫·E·冈萨雷斯和伊昂·斯托伊卡。从众包数据到高质量基准测试：Arena-Hard和benchbuilder管道。 *CoRR*，abs/2406.11939, 2024d. doi: 10.48550/ARXIV.2406.11939. URL https://doi.org/10.48550/arXiv.2406.11939.

林宗毅、迈克尔·迈尔、塞尔日·贝尔尼、詹姆斯·海斯、皮埃特罗·佩罗纳、德瓦·拉曼南、皮奥特·多尔ár和C·劳伦斯·齐特尼克。微软coco：上下文中的常见对象。在*ECCV*，2014年。

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*, 2023a.

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*, 2023b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6.

Dunjie Lu, Yiheng Xu, Junli Wang, Haoyuan Wu, Xinyuan Wang, Zekun Wang, Junlin Yang, Hongjin Su, Jixuan Chen, Junda Chen, Yuchen Mao, Jingren Zhou, Junyang Lin, Binyuan Hui, and Tao Yu. Videoagenttrek: Computer use pretraining from unlabeled videos, 2025. URL https://arxiv.org/abs/2510.19488.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010, 2024.

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.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 23464–23487, 2024.

OpenAI. Gpt-5 system card, 2025. URL https://cdn.openai.com/gpt-5-system-card.pdf.

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.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models. *CoRR*, abs/2312.06281, 2023. doi: 10.48550/ARXIV.2312.06281. URL https://doi.org/10.48550/arXiv.2312.06281.

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.

Shishir G. Patil, Huanzhi Mao, Charlie Cheng-Jie Ji, Fanjia Yan, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (bfcl): From tool use to agentic evaluation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

刘世龙、曾昭阳、任天鹤、李峰、张浩、杨杰、李春月、杨建伟、苏杭、朱俊娟和张雷。Grounding dino：将dino与有基础预训练结合用于开放集目标检测。*arXiv:2303.05499*，2023a。刘元、段浩东、李博、张远汉、张宋阳、赵望博、袁一克、王佳琪、何聪辉、刘子伟、陈凯和林大华。Mmbench：你的多模态模型是全能选手吗？*arXiv:2307.06281*，2023b。刘玉良、张铮、黄明新、杨彪、余文文、李春元、尹旭程、刘成林、金连文和白翔。Ocrbench：大型多模态模型中ocr隐藏的奥秘。中国科学信息科学，67(12)，2024年12月。ISSN 1869-1919。doi: 10.1007/s11432-024-4235-6。URLhttp://dx.doi.org/10.1007/s11432-024-4235-6。

董洁、徐一恒、王俊丽、吴浩远、王新元、王泽坤、杨俊林、苏红金、陈继轩、陈俊达、毛宇辰、周景仁、林俊阳、惠宾源，以及余涛。Videoagenttrek：从无标签视频中进行的计算机使用预训练，2025。URLhttps://arxiv.org/abs/2510.19488。

潘路、Hritik Bansal、夏 Tony、刘家成、李春元、Hannaneh Hajishirzi、程浩、张凯伟、Michel Galley，以及高建锋。Mathvista：在视觉环境中评估基础模型数学推理能力。*arXiv* 预印本 *arXiv:2310.02255*，2023。

马宇博、臧宇航、陈亮宇、陈美奇、焦一珠、李新泽、陆新元、刘子余、马岩、董晓怡，等。Mmlongbench-doc：通过可视化进行长上下文文档理解基准测试。神经信息处理系统进展，37:95963–96010，2024。

毛军华，黄俊，亚历山大·托谢夫，奥纳·坎布鲁，艾伦·李·尤尔，以及凯文·墨菲。非歧义对象描述的生成与理解。在*CVPR*，2016年。

阿明德·马斯里，杜宣龙，谭嘉庆，沙菲克·乔蒂，以及恩努尔·霍克。Chartqa：一个关于图表的问答基准，包含视觉和逻辑推理。*arXiv:2203.10244*，2022年。

Minesh Mathew，Viraj Bagal，Rubèn Pérez Tito，迪莫斯忒尼斯·卡拉塔斯，埃尔南迪·瓦尔韦尼，以及C.V. 贾瓦哈尔。Infographicvqa。*2022 IEEE/CVF*冬季计算机视觉应用会议（*WACV*），第2582-2591页，2021a。

Minesh Mathew，迪莫斯忒尼斯·卡拉塔斯，以及CV 贾瓦哈尔。Docvqa：一个用于文档图像的VQA数据集。在*WACV*，2021b年。

孟令晨, 杨建伟, 田瑞, 戴希阳, 吴祖玄, 高建锋, 和蒋宇刚. Deepstack: 深度堆叠视觉标记对 lmms 来说出奇地简单且有效。在 神经信息处理系统进展 第 37 卷，第 23464–23487 页，2024 年.

OpenAI. Gpt-5 系统卡，2025 年。URLhttps://cdn.openai.com/gpt-5-system-card.pdf.

欧阳林克, 邱原, 周红斌, 朱佳伟, 张瑞, 林群舒, 王斌, 赵志远, 蒋曼, 赵晓萌, 石金, 吴帆, 崔培, 刘明浩, 李振祥, 许超, 张博, 石博天, 屠中颖, 和何聪辉. OmniDocBench: 基准测试具有全面注释的多样化 PDF 文档解析，2024 年。URL https://arxiv.org/abs/2412.07626.

Samuel J. Paech. Eq-bench：一个用于大语言模型的情绪智能基准。*CoRR*，abs/2312.06281，2023。doi: 10.48550/ARXIV.2312.06281。URLhttps://doi.org/10.48550/arXiv.2312.06281。

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

Shishir G. Patil、Huanzhi Mao、Charlie Cheng-Jie Ji、Fanjia Yan、Vishnu Suresh、Ion Stoica和Joseph E. Gonzalez。伯克利函数调用排行榜（BFCL）：从工具使用到大语言模型的代理评估。发表于《神经信息处理系统进展》，2024年。

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, 和 Zhe Gan. MIA-Bench：迈向更好的多模态大语言模型指令遵循评估。*arXiv* 预印本 *arXiv:2407.01509*, 2024。

乔润琪、谭启娜、董冠廷、吴明辉、孙崇、宋晓帅、龚曲卓玛、雷尚林、魏哲、张妙璇等。We-Math：你的大型多模态模型是否实现了类人的数学推理？*arXiv*预印本*arXiv:2407.01284*, 2024。

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. URL https://arxiv.org/abs/2407.06581.

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. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, et al. Zerobench: An impossible visual benchmark for contemporary large multimodal models, 2025. URL https://arxiv.org/abs/2502.09696.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10912–10922, 2021.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, et al. INCLUDE: evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8430–8439, 2019.

Chenglei Si, Yanzhe Zhang, Ryan Li, Zhengyuan Yang, Ruibo Liu, and Diyi Yang. Design2code: Benchmarking multimodal code generation for automated front-end engineering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3956–3974, 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15768–15780, 2025a.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025b. URL https://arxiv.org/abs/2504.10342.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

M-A-P Team. Supergpqa: Scaling LLM evaluation across 285 graduate disciplines. *CoRR*, abs/2502.14739, 2025. doi: 10.48550/ARXIV.2502.14739. URL https://doi.org/10.48550/arXiv.2502.14739.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024b.

Pooyan Rahmanzadehgervi、Logan Bolton、Mohammad Reza Taesiri和Anh Totti Nguyen。视觉语言模型是盲的：未能将详细的视觉特征翻译成文字，2025。URL https://arxiv.org/abs/2407.06581。

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, 和 Samuel R. Bowman. GPQA: 一个研究生级别的谷歌验证问答基准测试. *CoRR*,abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URLhttps://doi.org/10.48550/arXiv.2311.12022.

Jonathan Roberts, Mohammad Reza Taesiri, Ansh Sharma, Akash Gupta, Samuel Roberts, Ioana Croitoru, Simion-Vlad Bogolin, Jialu Tang, Florian Langer, 等. ZeroBench: 一个不可能的视觉基准测试，用于当代大型多模态模型, 2025. URL https://arxiv.org/abs/2502.09696.

Mike Roberts，Jason Ramapuram，Anurag Ranjan，Atulit Kumar，Miguel Angel Bautista，Nathan Paczan，Russ Webb，和 Joshua M Susskind。Hypersim：一个用于整体室内场景理解的逼真合成数据集。在 *IEEE/CVF*国际计算机视觉会议论文集，第 10912–10922 页，2021年。

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, et al. INCLUDE: 评估具有区域知识的跨语言语言理解。在 第*13*届国际学习表征会议，*ICLR 2025*，新加坡，*2025*年*4*月*24-28*日。OpenReview.net, 2025.

Shuai Shao，Zeming Li，Tianyuan Zhang，Chao Peng，Gang Yu，Xiangyu Zhang，Jing Li，和 Jian Sun。Objects365：一个大规模、高质量的对象检测数据集。在 *IEEE/CVF*国际计算机视觉会议论文集，第 8430–8439 页，2019年。

程雷思, 张岩哲, 李瑞安, 杨正源, 刘瑞波, 杨迪一. 设计到代码: 自动化前端工程的多模态代码生成基准测试. 在 美洲国家计算语言学协会第*2025*年会议论文集：人机语言技术（第一卷：长文）, 第3956-3974页, 2025.

宋灿希, 布卢基斯·瓦尔茨, 特雷布莱·乔纳森, 泰里·斯蒂芬, 苏雨, 比尔菲尔德·斯坦. 机器人空间感知: 为机器人教学二维和三维视觉语言模型的空间理解. 在 计算机视觉与模式识别会议论文集, 第15768-15780页, 2025a.

宋舒然, 李克腾·塞缪尔·P, 肖建雄. 苏恩RGB-D: RGB-D场景理解基准测试套件. 在 *IEEE*计算机视觉与模式识别会议论文集, 第567-576页, 2015.

宋月奇, 欧天越, 孔一博, 李泽成, Graham Neubig, 和 薛阳. VisualPuzzles：将多模态推理评估与领域知识解耦. *arXiv* 预印本 *arXiv:2504.10342*,2025b. URLhttps://arxiv.org/abs/2504.10342.

Gemini 机器人团队, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, 等等. Gemini 机器人: 将 AI 带入物理世界. *arXiv* 预印本 *arXiv:2503.20020*, 2025.

M-A-P 团队. SuperGPQA：跨 285 个研究生学科扩展大语言模型评估. *CoRR*, abs/2502.14739,2025. doi: 10.48550/ARXIV.2502.14739. URLhttps://doi.org/10.48550/arXiv.2502.14739.

Michael Tschannen, Alexey Gritsenko, 王晓, 穆罕默德·费贾德·纳伊姆, 伊布拉欣·阿尔阿卜杜尔穆斯林, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, 叶娅, Basil Mustafa 等. Siglip 2: 具有改进语义理解、定位和密集特征的多语言视觉语言编码器.*arXiv* 预印本 *arXiv:2502.14786*, 2025.

王飞, 付兴宇, 黄嘉毅, 李泽坤, 刘琴, 刘晓刚, 马明宇 Derek, 许楠, 周文轩, 张凯 等. Muirbench: 用于鲁棒多图像理解的综合性基准测试. *arXiv* 预印本 *arXiv:2406.09411*, 2024a.

王凯，潘俊婷，石伟康，ZimuLu，侯兴仁，周爱军，战明杰，和李红生。使用 math-vision 数据集测量多模态数学推理。神经信息处理系统进展, 37:95095–95169, 2024b.

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*, 2024c.

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*, 2024d.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *arXiv preprint*, 2024e. URL https://arxiv.org/abs/2408.15556.

Xinyuan Wang, Bowen Wang, Dunjie Lu, Junlin Yang, Tianbao Xie, Junli Wang, Jiaqi Deng, Xiaole Guo, Yiheng Xu, Chen Henry Wu, et al. Opencua: Open foundations for computer-use agents. *arXiv preprint arXiv:2508.09123*, 2025a.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, et al. Polymath: Evaluating mathematical reasoning in multilingual contexts. *CoRR*, abs/2504.18428, 2025b. doi: 10.48550/ARXIV.2504.18428. URL https://doi.org/10.48550/arXiv.2504.18428.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *CoRR*, abs/2406.01574, 2024f.

Zhexu Wang, Yiping Liu, Yejie Wang, Wenyang He, Bofei Gao, Muxi Diao, Yanxu Chen, Kelin Fu, Flood Sung, Zhilin Yang, Tianyu Liu, and Weiran Xu. Ojbench: A competition level code benchmark for large language models. *CoRR*, abs/2506.16395, 2025c. doi: 10.48550/ARXIV.2506.16395. URL https://doi.org/10.48550/arXiv.2506.16395.

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*, 2024g.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, et al. Livebench: A challenging, contamination-free LLM benchmark. *CoRR*, abs/2406.19314, 2024. doi: 10.48550/ARXIV.2406.19314. URL https://doi.org/10.48550/arXiv.2406.19314.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025a.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, June 2024.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing. *CoRR*, abs/2503.05244, 2025b. doi: 10.48550/ARXIV.2503.05244. URL https://doi.org/10.48550/arXiv.2503.05244.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2025-04-26.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Tianbao Xie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Zekun Wang, Yiheng Xu, Junli Wang, Doyen Sahoo, Tao Yu, and Caiming Xiong. Scaling computer-use grounding via user interface decomposition and synthesis, 2025a. URL https://arxiv.org/abs/2505.13227.

Peng Wang, 白帅, Tan Sinan, 王时杰, 范志浩, 白金泽, 陈克勤, 刘雪晶, 王佳林, 郭文斌, 范杨, 唐凯, 杜梦飞, 任宣程, 闻睿, 刘代恒, 周畅, 周景仁, 林俊阳. Qwen2-vl: 增强视觉语言模型在任何分辨率下对世界的感知. arXiv:2409.12191, 2024c. 王伟涵, 何泽海, 洪文怡, Cheng Yean, 张晓寒, Qi Ji, 古晓涛, 黄诗宇, 许斌 , 董宇晓 , 等 . Lvbench: 一个极端长视频理解基准 . arXiv preprint arXiv:2406.08035, 2024d. 王文斌, 丁亮, 曾民彦, 周西斌, 沈丽, 罗勇, 和戴晨. 分而治之， 合而为一：多模态大语言模型中高分辨率图像感知的无训练框架. arXiv preprint, 2024e. URL https://arxiv.org/abs/2408.15556. 王新元, 王Bowen, 陆敦洁, 杨俊林, 谢天宝, 王俊丽, 邓家奇, 郭晓乐, 许一恒, 吴晨 Henry, 等. Opencua: 计算机使用代理的开放基础. arXiv preprint arXiv:2508.09123, 2025a. 王一鸣, 张培, 唐嘉龙, 魏浩然, 杨宝松, 王瑞, 孙晨舒, 孙飞彤, 张继然, 吴俊玄, 藏启倩, 张一畅, 黄飞, 林俊阳, 等. PolyMATH: 评估多语言环境下的数学推理. CoRR, abs/2504.18428, 2025b. doi: 10.48550/ARXIV.2504.18428. URL https://doi.org/10.48550/arXiv.2504.18428. 王宇博, 马学广, 张格, 倪元生, Chandra Abhranil, 郭世光, 任为民, Arulraj Aaran, 何轩, 姜子岩, 李天乐, 等. MMLU-Pro: 一个更鲁棒和具有挑战性的多任务语言理解基准. CoRR, abs/2406.01574, 2024f. 王哲旭, 刘一波, 王叶洁, 何文阳, 高伯飞, 肖沐熙, 陈岩旭, 傅克麟 , Sung Flood, 杨志林 , 刘天宇 , 许为然 . Ojbench: 大语言模型的竞赛级别代码基准 . CoRR, abs/2506.16395, 2025c. doi: 10.48550/ARXIV.2506.16395. URL https://doi.org/10.48550/arXiv.2506.16395. 王子睿, 夏梦舟, 何Luxi, Chen Howard, 刘奕涛, 朱理查德, 梁开琪, 吴新迪, 刘浩天, Malladi Sadhika, Chevalier Alexis, Arora Sanjeev, 和陈戴奇. Charxiv: 多模态大语言模型中现实图表理解的差距图示. arXiv preprint arXiv:2406.18521, 2024g. Wettig Alexander, Lo Kyle, Min Sewon, Hajishirzi Hannaneh, 陈戴奇, 和 Soldaini Luca. 组织网络：构建领域增强预训练数据管理. arXiv preprint arXiv:2502.10341, 2025. White Colin, Dooley Samuel, Roberts Manley, Pal Arka, Feuer Benjamin, Jain Siddhartha, Shwartz-Ziv Ravid, J ain Neel, 等. Livebench: 一个具有挑战性、无污染的LLM基准. CoRR, abs/2406.19314, 2024. doi: 10.48550/ARXIV.2406.19314. URL https://doi.org/10.48550/arXiv.2406.19314. 吴金明, 邓志豪, 李伟, 刘艺定, 您博, 李博, Ma Zejun, 和 Liu Ziwei. Mmsearch-r1: 激励LLM进行搜索. arXiv preprint arXiv:2506.20670, 2025a. 吴彭浩和谢山宁. V*: 作为多模态大语言模型核心机制的有引导视觉搜索. 在IEEE/CVF计算机视觉与模式识别会议论文集 (CVPR) 中, pp. 13084– 13094, 2024年6月. 吴宇宁, 梅嘉豪, 严明, 李晨亮, 赖少鹏, 任宇然, 王子嘉, 张继, 吴梦越, 金秦, 和黄飞. Writingbench: 生成性写作的综合基准. CoRR, abs/2503.05244, 2025b. doi: 10.48550/ARXIV.2503.05244. URL https://doi.org/10. 48550/arXiv.2503.05244. xAI. Realworldqa: 真实世界空间理解基准. https://huggingface.co/ datasets/xai-org/RealworldQA, 2024. 访问时间: 2025-04-26. 肖一佳, Sun Edward, 刘天宇, 和王伟. Logicvista: 视觉环境中的多模态大语言模型逻辑推理基准. arXiv preprint arXiv:2407.04973, 2024. 谢天宝, 邓家奇, 李晓川, 杨俊林, 吴昊远, 陈继轩, 胡文静, 王新元, 许宇辉, 王泽坤, 许一恒, 王俊丽, Sahoo Doyen, 余涛, 和 Xiong Caiming. 通过用户界面分解和合成扩展计算机使用 Grounding, 2025a. URL https: //arxiv.org/abs/2505.13227.

Tianbao Xie, Mengqi Yuan, Danyang Zhang, Xinzhuang Xiong, Zhennan Shen, Zilong Zhou, Xinyuan Wang, Yanxu Chen, Jiaqi Deng, Junda Chen, Bowen Wang, Haoyuan Wu, Jixuan Chen, Junli Wang, Dunjie Lu, Hao Hu, and Tao Yu. Introducing osworld-verified. *xlang.ai*, July 2025b. URL https://xlang.ai/blog/osworld-verified.

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

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025. URL https://arxiv.org/abs/2504.15279.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report, 2025a.

Cheng Yang, Chufan Shi, Yaxin Liu, Bo Shui, Junjie Wang, Mohan Jing, Linran Xu, Xinyu Zhu, Siheng Li, Yuxiang Zhang, et al. Chartmimic: Evaluating lmm's cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

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.

Jiabo Ye, Xi Zhang, Haiyang Xu, Haowei Liu, Junyang Wang, Zhaoqing Zhu, Ziwei Zheng, et al. Mobile-agent-v3: Fundamental agents for gui automation. *arXiv preprint arXiv:2508.15144*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

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

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, Zhijian Xu, Chengye Wang, et al. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. URL https://arxiv.org/abs/2501.12380.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025.

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. doi: 10.48550/ARXIV.2311.07911. URL https://doi.org/10.48550/arXiv.2311.07911.

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.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

谢天宝，袁梦琪，张丹阳，熊新庄，沈振南，周子龙，王新元，陈岩旭，邓嘉奇，陈俊达，王 Bowen，吴浩远，陈继轩，王俊丽，陆敦杰，胡浩，余涛。介绍 osworld-verified。*xlang.ai*，2025年7月b版。URL https://xlang.ai/blog/osworld-verified。谢天宝，张丹阳，陈继轩，李晓川，赵思恒，曹瑞生，等。Osworld：在真实计算机环境中对开放式任务的多模态代理进行基准测试。神经信息处理系统进展，37:52040–52094，2025年c版。徐伟业，王嘉豪，王伟云，陈哲，周文刚，杨爱军，陆乐伟，李厚强，王晓华，朱锡舟，等。Visulogic：评估多模态大语言模型中视觉推理的基准，2025年。URL https://arxiv.org/abs/2504.15279。杨安，李安凤，杨保松，张贝辰，会炳元，郑博，余 Bowen，等。Qwen3 技术报告，2025年a版。杨成阳，石楚凡，刘亚欣，水波博，王俊杰，景茂，徐林然，朱新宇，李思恒，张宇翔，等。Chartmimic：通过图表到代码生成评估 LLM 的跨模态推理能力。*arXiv* 预印本 *arXiv:2406.09961*，2024年a版。杨继涵，杨舒生，W Gupta Anjali，韩瑞麟，李飞飞，谢山宁。空间中的思考：多模态大语言模型如何看、记忆和回忆空间。在 计算机视觉与模式识别会议论文集，第 10632–10643 页，2025年b版。杨志博，唐俊，李兆海，王鹏飞，万建强，钟胡敏，刘雪静，杨明坤，王鹏，白帅，金连文，林俊阳。Cc-ocr：为评估大型多模态模型在识字方面的综合性和挑战性而设计的 OCR 基准，2024年b版。URL https://arxiv.org/abs/2412.02210。叶嘉宝，张熙，徐海阳，刘浩伟，王俊阳，朱赵庆，郑子伟，等。Mobile-agent-v3：用于 GUI 自动化的基本代理。*arXiv* 预印本 *arXiv:2508.15144*，2025年。岳祥，倪源升，张凯，郑天宇，刘若琪，张格，等。Mmmu：为专家级 AGI 设计的大规模多学科多模态理解和推理基准。在 *IEEE/CVF* 计算机视觉与模式识别会议论文集，第 9556–9567 页，2024年a版。岳祥，郑天宇，倪源升，王宇博，张凯，童盛邦，孙宇轩，余波涛，张格，孙欢，等。Mmmu-pro：一个更鲁棒的多学科多模态理解基准。*arXiv* 预印本 *arXiv:2409.02813*，2024年b版。张仁睿，蒋东智，张奕驰，林浩坤，郭子余，邱鹏硕，周澳俊，陆磐，张凯伟，乔宇，等。Mathverse：您的多模态 LLM 真正能看到视觉数学问题中的图表吗？在 欧洲计算机视觉会议，第 169–186 页。Springer，2024年。赵一陇，谢路静，张浩伟，甘郭，龙奕涛，胡智远，胡通岩，陈伟源，李初寒，宋俊阳，徐志坚，王成业，等。Mmvu：测量专家级多学科视频理解，2025年。URL https://arxiv.org/abs/2501.12380。郑子伟，杨迈克尔，Hong Jack，赵晨晓，徐国海，杨乐，沈超，余行。Deepeyes：通过强化学习激励“用图像思考”。*arXiv* 预印本 *arXiv:2505.14362*，2025年。周恩森，安景坤，程驰，韩毅，荣山雨，张驰，王鹏伟，王中元，黄铁军，陆升，等。Roborefer：为机器人学中的视觉语言模型实现空间指代推理。*arXiv* 预印本 *arXiv:2506.04308*，2025年。周Jeffrey，陆天健，Swaroop Mishra，Brahma Siddhartha，Basu Sujoy，LUAN Yi，Zhou Denny，和 Le Hou。大型语言模型的指令遵循评估。*CoRR*，abs/2311.07911，2023年。doi: 10.48550/ARXIV.2311.07911。URLhttps://doi.org/10.48550/arXiv.2311.07911。周俊杰，舒岩，赵波，吴伯雅，肖时涛，杨希，熊永平，张波，黄铁军，刘正。Mlvu：一个用于多任务长视频理解的综合性基准。*arXiv* 预印本 *arXiv:2406.04264*，2024年。朱万荣，Hessel Jack，Awadalla Anas，Gadre Samir Yitzhak，Dodge Jesse，Fang Alex，Yu Youngjae，Schmidt Ludwig，Wang William Yang，Choi Yejin。Multimodal c4：一个开放的、十亿规模的、图像与文本交替的语料库。神经信息处理系统进展，36:8958–8974，2023 年。邹程科，郭兴刚，杨瑞，张俊宇，胡斌，和张欢。DynaMath：一个动态视觉基准，用于评估视觉语言模型的数学推理鲁棒性。*arXiv* 预印本 *arXiv:2411.00836*，2024年。

A Benchmarks

We evaluate Qwen3-VL on a wide range of public benchmarks across distinct capabilities: multimodal reasoning, general visual question answering, subjective experience & instruction following, document understanding (including OCR), 2D/3D visual grounding and counting, spatial reasoning, video understanding, GUI agent, and Text-Centric tasks. Below, we provide a detailed list of all the benchmarks used.

- **Multimodal Reasoning:** We evaluate the models on 12 benchmarks spanning a diverse range of domains—from mathematics and STEM to visual reasoning and puzzle-solving tasks: MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MathVision (Wang et al., 2024b), MathVision-Wild$_{\text{photo}}$, MathVista (Lu et al., 2023), We-Math (Qiao et al., 2024), MathVerse (Zhang et al., 2024), DynaMath (Zou et al., 2024), Math-VR (Duan et al., 2025), LogicVista (Xiao et al., 2024), VisualPuzzles (Song et al., 2025b), VLM are Blind (Rahmanzadehgervi et al., 2025), ZeroBench (Main/Subtasks) (Roberts et al., 2025), and VisuLogic (Xu et al., 2025).
- **General Visual Question Answering:** We evaluate the models on 4 General VQA benchmarks: MMBench-V1.1 (Liu et al., 2023b), RealWorldQA (xAI, 2024), MMStar (Chen et al., 2024a), and SimpleVQA Cheng et al. (2025).
- **Subjective Experience and Instruction Following:** We evaluate the model on 3 benchmarks, across subject experience and complex instruction following: HallusionBench (Guan et al., 2023), MM-MT-Bench (Agrawal et al., 2024), and MIA-Bench (Qian et al., 2024).
- **Document Understanding:** We perform comprehensive evaluation on OCR and document understanding ability of Qwen3-VL series across a diverse range OCR related benchmarks: DocVQA (Mathew et al., 2021b), InfoVQA (Mathew et al., 2021a), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), OCRBench (Liu et al., 2024), OCRBench_v2 (Fu et al., 2024b), CC-OCR (Yang et al., 2024b), OmniDocBench (Ouyang et al., 2024), CharXiv (Wang et al., 2024g), and MMLongBench-Doc (Ma et al., 2024).
- **2D/3D Grounding and Spatial Understanding:** We evaluate the models on 11 benchmarks include 2D grounding, 3D grounding and spatial understanding: RefCOCO/+/g (Kazemzadeh et al., 2014; Mao et al., 2016), ODinW-13 (Li et al., 2022), CountBench (Paiss et al., 2023), ARKitScenes (Baruch et al., 2021), Hypersim (Roberts et al., 2021), SUN RGB-D (Song et al., 2015), ERQA (Team et al., 2025), VSIBench (Yang et al., 2025b), EmbSpatial (Du et al., 2024),RefSpatial (Zhou et al., 2025), and RoboSpatialHome (Song et al., 2025a).
- **Video Understanding:** We use seven benchmarks to evaluate the model's video understanding capabilities: VideoMME (Fu et al., 2024a), MVBench (Li et al., 2024b), VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025), LVBench (Wang et al., 2024d), MLVU (Zhou et al., 2024), Charades-STA (Gao et al., 2017).
- **Coding:** We evaluate the model's multi-modal coding capabilities, particularly in front-end reconstruction and SVG generation, using the Design2Code (Si et al., 2025), ChartMimic (Yang et al., 2024a), and UniSVG (Li et al., 2025a) benchmarks.
- **GUI Agent:** We evaluate GUI agent capabilities using benchmarks that test both perception and decision-making. For perception, we use ScreenSpot (Cheng et al., 2024), ScreenSpot Pro (Li et al., 2025b), and OSWorldG (Xie et al., 2025a) to measure GUI grounding and understanding of interface layouts across devices. For decision-making, we use AndroidWorld (Rawles et al., 2024) and OSWorld (Xie et al., 2025c;b) to evaluate interactive control, planning, and execution within real or simulated operating environments.
- **Text-Centric Tasks:** We evaluate the models on a wide range of text-centric datasets. (1) **Knowledge:** MMLU-Pro (Wang et al., 2024f), MMLU-Redux (Gema et al., 2024), GPQA (Rein et al., 2023), SuperGPQA (Team, 2025), (2) **Reasoning:** AIME-25 (AIME, 2025), HMMT-25 (HMMT, 2025), LiveBench (2024-11-25) (White et al., 2024), (3) **Code:** LiveCodeBench v6 (Jain et al., 2024), CFEval, OJBench (Wang et al., 2025c), (4) **Alignment Tasks:** IFEval (Zhou et al., 2023), Arena-Hard v2 (Li et al., 2024d) , Creative Writing v3 (Paech, 2023), WritingBench (Wu et al., 2025b), (5) **Agent:** BFCL-v3 (Patil et al., 2024), TAU2-Retail, TAU2-Airline, TAU2-Telecom, (6) **Multilingual:** MultiIF (He et al., 2024), MMLU-ProX, INCLUDE (Romanou et al., 2025), PolyMATH (Wang et al., 2025b).

基准测试

我们在多个公共基准测试上评估了 Qwen3-VL，涵盖不同的能力：多模态推理、通用视觉问答、主观体验&指令遵循、文档理解（包括 OCR）、2D/3D 视觉 Grounding 和计数、空间推理、视频理解、GUI 代理以及 Text-Centrictasks。下面，我们提供了所有使用的基准测试的详细列表。

- 多模态推理：我们在12个基准测试上评估模型，这些测试涵盖了广泛的不同领域——从数学和STEM到视觉推理和谜题解决任务：MMMU（Yue等人，2024a），MMMU-Pro（Yue等人，2024b），MathVision（Wang等人，2024b），MathVision-Wildphoto，MathVista（Lu等人，2023），We-Math（Qiao等人，2024），MathVerse（Zhang等人，2024），DynaMath（Zou等人，2024），Math-VR（Duan等人，2025），LogicVista（Xiao等人，2024），VisualPuzzles（Song等人，2025b），VLM are Blind（Rahmanzadehgervi等人，2025），ZeroBench（Main/Subtasks）（Roberts等人，2025），以及VisuLogic（Xu等人，2025）。
- **通用视觉问答：**我们在4个通用视觉问答基准测试上评估了模型：MMBench-V1.1（Liu等人，2023b），RealWorldQA（xAI，2024），MMStar（Chen等人，2024a），以及SimpleVQACheng等人(2025）。
- **主观体验和指令遵循：**我们在3个基准测试上评估模型，涵盖主观体验和复杂指令遵循：HallusionBench（Guan等人，2023），MM-MT-Bench（Agrawal等人，2024），以及MIA-Bench（Qian等人，2024）。
- 文档理解：我们在OCR和文档理解能力方面对Qwen3-VL系列进行了全面评估，涵盖多种OCR相关基准测试：DocVQA（Mathew等人，2021b），InfoVQA（Mathew等人，2021a），AI2D（Kembhavi等人，2016），ChartQA（Masry等人，2022），OCRBench（Liu等人，2024），OCRBench_v2（Fu等人，2024b），CC-OCR（Yang等人，2024b），OmniDocBench（Ouyang等人，2024），CharXiv（Wang等人，2024g），以及MMLongBench-Doc（Ma等人，2024）。
- 2D/3D Grounding和空间理解：我们在11个基准测试上评估了模型，包括2D grounding、3D grounding和空间理解：RefCOCO/+/g（Kazemzadeh等人，2014年；Mao等人，2016年）、ODinW-13（Li等人，2022年）、CountBench（Paiss等人，2023年）、ARKitScenes（Baruch等人，2021年）、Hypersim（Roberts等人，2021年）、SUN RGB-D（Song等人，2015年）、ERQA（团队，2025年）、VSIBench（Yang等人，2025b）、EmbSpatial（Du等人，2024年）、RefSpatial（Zhou等人，2025年）以及RoboSpatialHome（Song等人，2025a）。
- **视频理解：**我们使用七个基准测试来评估模型的视频理解能力：VideoMME（Fu等人，2024a），MVBench（Li等人，2024b），VideoMMMU（Hu等人，2025），MMVU（Zhao等人，2025），LVBench（Wang等人,2024d），MLVU（Zhou等人，2024），Charades-STA（Gao等人，2017）。
- **编程：** 我们使用设计到代码 (Si 等人.,2025)、图表模仿 (Yang 等人.,2024a) 和 UniSVG (Li 等人., 2025a) 基准测试，评估模型的多模态编程能力，特别是在前端重建和 SVG 生成方面。
- **GUI 代理：** 我们使用测试感知和决策能力的基准测试来评估 GUI 代理功能。对于感知，我们使用屏幕捕捉 (Cheng 等人., 2024)、屏幕捕捉专业版 (Li 等人.,2025b) 和 OSWorldG (Xie 等人., 2025a) 来测量 GUI 定位和对跨设备界面布局的理解。对于决策，我们使用 AndroidWorld (Rawles 等人., 2024) 和 OSWorld (Xie 等人., 2025c;b) 来评估在真实或模拟操作环境中进行交互控制、规划和执行的能力。
- 以文本为中心的任务：我们在一系列以文本为中心的数据集上评估了模型。(1) 知识：MMLU-Pro (Wang 等人, 2024f)，MMLU-Redux (Gema 等人，2024)，GPQA (Rein 等人，2023)，SuperGPQA (团队，2025)，(2) 推理：AIME-25 (AIME，2025)，HMMT-25 (HMMT，2025)，LiveBench (2024-11-25) (White 等人，2024)，(3) 代码：LiveCodeBench v6 (Jain 等人，2024)，CFEval，OJBench (Wang 等人，2025c)，(4) 对齐任务：IFEval (Zhou 等人，2023)，Arena-Hard v2 (Li 等人，2024d)，创意写作 v3 (Paech，2023)，WritingBench (Wu 等人，2025b)，(5) 代理：BFCL-v3 (Patil 等人，2024)，TAU2-Retail，TAU2-Airline，TAU2-Telecom，(6) 多语言：MultiIF (He 等人，2024)，MMLU-ProX，INCLUDE (Romanou 等人，2025)，PolyMATH (Wang 等人，2025b)。

B Evaluation Prompts

To ensure reproducibility and facilitate future research, we provide here the complete set of prompts used to evaluate our model across all benchmarks. These prompts were consistently applied during inference to maintain fairness and comparability.

B.1 STEM & Puzzle

MMMU

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

MMMUPro_Standard

```
<image>
{question}
{options}
Please select the correct answer from the options.
```

MMMUPro_Vision

```
<image>
Identify the problem and solve it. Think step by step before answering.
```

MathVista | MathVision | MathVerse | LogicVista

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

We-Math

```
<image>
Now, we require you to solve a multiple-choice math question. Please briefly describe your
thought process and provide the final answer(option).
Question: {question}
Option: {options}
Regarding the format, please answer following the template below, and be sure to include
two <> symbols:
<Thought process>: «your thought process» <Answer>: «your option»
```

ZeroBench

```
<image>
{question}
Let's think step by step and give the final answer in curly braces, like this: {final
answer}
```


B 评估提示

为确保可复现性并促进未来研究，我们在此提供用于在所有基准测试中评估我们模型的完整提示集。这些提示在推理过程中始终如一地应用，以保持公平性和可比性。

B.1 STEM & 谜题

MMMU

```
<图像>
问题：{问题}
选项：
{选项}
请从上方选项中选择正确答案。
```

MMMUPro 标准_

```
<图像>{问题} {选项} 请从上方选项中选择正确答案。
```

MMMUPro 视觉_

```
<图像>识别问题并解决。回答前请逐步思考。
```

MathVista | MathVision | MathVerse | LogicVista

```
<图像>{问题}
```

We-Math

ZeroBench

```
<图像>{问题} 让我们一步步思考，并用大括号给出最终答案，例如这样：{最终答案}
```

DynaMath

```
<image>
## Question
{question}
## Answer Instruction: Please provide an answer to the question outlined above. Your response should adhere to the following JSON format, which includes two keys: 'solution' and 'short answer'. The 'solution' key can contain detailed steps needed to solve the question, and the 'short answer' key should provide a concise response.
Example of expected JSON response format:
{
"solution": "[Detailed step-by-step explanation]",
"short answer": "[Concise Answer]"
}
```

VLMBlind

```
<image>
Question: {question}
```

VisuLogic

```
<image>
{question}
Solve the complex visual logical reasoning problem through step-by-step reasoning. Think about the reasoning process first and answer the question following this format: Answer://boxed{$LETTER}
```

VisualPuzzles-Direct

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

VisualPuzzles-CoT

```
<image>
Question: {question}
Options:
{options}
Solve the multiple-choice question and then answer with the option letter from the given choices. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes), where LETTER is one of the options. Think step by step before answering.
```

B.2 GeneralVQA

MMBench | RealWorldQA | MMStar

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

SimpleVQA

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


DynaMath

VLMBlind

```
<图像>问题：{question}
```

VisuLogic

```
<图像>{question} 通过逐步推理解决复杂的视觉逻辑推理问题。先思考推理过程，然后按照以下格式回答问题：Answer://boxed{$字母}
```

VisualPuzzles-Direct

```
<图像>问题：{问题} 选项：{选项} 请直接用给定选项中的字母回答问题。
```

VisualPuzzles-CoT

B.2 GeneralVQA

MMBench | RealWorldQA | MMStar

```
<图像>问题：{问题} 选项：{选项} 请从上面的选项中选择正确答
案。
```

SimpleVQA

```
<图像>{问题}
```

B.3 Alignment

HallusionBench | MM_MT_Bench | MIA-Bench

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

B.4 Document-Understanding

MMLongBench-Doc

```
<image_1>
<image_2>
...
<image_n>
{question}
```

DocVQA | InfoVQA | ChartQA_TEST

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

AI2D

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

OCRBench | OCRBench_v2 | CC-OCR | CharXiv

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

OmniDocBench

```
<image>
You are an AI assistant specialized in converting PDF images to Markdown format. Please
follow these instructions for the conversion:
1. Text Processing: - Accurately recognize all text content in the PDF image without
guessing or inferring. - Convert the recognized text into Markdown format. - Maintain the
original document structure, including headings, paragraphs, lists, etc.

2. Mathematical Formula Processing:
- Convert all mathematical formulas to LaTeX format.
- Enclose inline formulas with \( \). For example: This is an inline formula \( E = mc^2 \)
- Enclose block formulas with \[ \]. For example: \[ \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \]

3. Table Processing: - Convert tables to HTML format. - Wrap the entire table with <table>
and </table>.
4. Figure Handling: - Ignore figures in the PDF image. Do not attempt to describe or
convert images.
5. Output Format: - Ensure the output Markdown document has a clear structure with
appropriate line breaks between elements. - For complex layouts, try to maintain the
original document's structure and format as closely as possible.
Please strictly follow these guidelines to ensure accuracy and consistency in the conversion.
Your task is to accurately convert the content of the PDF image into Markdown format without
adding any extra explanations or comments.
```


B.3 对齐

HallusionBench |MM MT 基准 |MIA-Bench_ _

```
<图像>{问题}
```

B.4 文档理解

MMLongBench-Doc

```
<图像_1>
<图像_2>
...
<图像_n>
{问题}
```

DocVQA | InfoVQA | ChartQA_TEST

```
<图像>{问题} 使用单个词语或短语回答问题。
```

AI2D

```
<图像>问题: {问题} 选项: {选项} 请从上方选项中选择正确答
案。
```

OCRBench | OCRBench v2 | CC-OCR | CharXiv_

```
<图像>{问题}
```

OmniDocBench

```
<图像>
你是 一位专门将PDF图像转换为Markdown格式的AI助手。请
遵循t 这些转换说明:
1. 文本处理: - 准确识别PDF图像中的所有文本内容，不得猜测或推断。- 将识别的文本转换为
Markdown格式。- 保持原始文档结构，包括标题、段落、列表等。

2. 数学公式处理: - 将所有数学公式转换为LaTeX格式。- 行内公式用\( \)包围。例如: 这是一个行
内公式 \( E = mc^2 \)- 块级公式用\[ \]包围。例如: \[ \frac{-b \pm \sqrt{b^2- 4ac}}{2a} \]


3. 表格处理: - 将表格转换为HTML格式。- 用 <table>包裹整个表格，并用 </table>结束。

4. 图形处理: - 忽略PDF图像中的图形。 不要试图描述或
转换图像。
5. 输出格式: - 确保输出的Markdown文档具有清晰的层次结构，并在元素之间使用适当的换行符。
- 对于复杂布局，尽量保持原始文档的结构和格式。
请严格遵循这些指南，以确保转换的准确性和一致性。您的任务是准确地将PDF图像的内容转换为
Markdown格式，不得添加任何额外的解释或评论。
```

B.5 2D/3D Grounding

RefCOCO

```
<image>
Locate every object that matches the description "{ref_sentence}" in the image. Report bbox coordinates in JSON format.
```

CountBench

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

ODinW-13

```
<image>
Locate every instance that belongs to the following categories: {obj_names}: Report bbox coordinates in JSON format.
```

ARKitScenes | Hypersim | SUNRGBD

```
<image>
Locate the {class_name } in the provided image and output their positions and dimensions using 3D bounding boxes. The results must be in the JSON format: ["bbox_3d":[x_center, y_center, z_center, x_size, y_size, z_size, roll, pitch, yaw],"label":"category"].
```

B.6 Embodied/Spatial Understanding

ERQA

```
<image_1>
<image_2>
...
<image_n>
{question}
```

VSI-Bench

multiple-choice:

```
<video>
These are frames of a video.
{question}
Options:
{options}
Answer with the option's letter from the given choices directly.
```

open-ended:

```
<video>
These are frames of a video.
{question}
Please answer the question using a single word or phrase.
```

EmbSpatialBench

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


B.5 2D/3D 定位

RefCOCO

<图像>
在图像中定位所有与描述 "{ref_sentence}" 匹配的对象。以JSON格式报告边界框坐标。

CountBench

<图像>问题：{question} 选项：{options} 请从上方选项中选择
正确答案。

ODinW-13

<图像>
定位e 报告属于以下类别： ´{obj_名字}·的所有实例的边界框
坐标 以JSON格式报告

ARKitScenes |Hypersim | SUNRGBD

<图像>
在提供的图像中定位 {class_名称 }，并使用 3D 边界框输出它们的位置和尺寸。结果必须以 JSON 格式呈现： ["bbox_3d":[x_中心, y_中心, z_中心, x_尺寸, y_尺寸, z_尺寸, roll, pitch, yaw],"label":"类别"].

B.6 具身/空间理解

ERQA

<图像_1>
<图像_2>
...
<图像_n>
{问题}

VSI-Bench

选择题：
<视频>
这些是视频的帧。
{问题}
选项:
{options}
直接用给定选项中的字母作答。

开放式：
<视频>
这些是视频的帧。
{问题}
请用单个词语或短语回答问题。

EmbSpatialBench

<图像>{问题}

RoboSpatialHome

```
<image>
Locate {object_name} in this image. Output the point coordinates in JSON format.
For example:
[
{"point_2d": [x, y], "label": "point_1"}
]
```

RefSpatialBench

```
<image>
{question} Output the point coordinates in JSON format.
For example:
[
{"point_2d": [x, y], "label": "point_1"}
]
```

B.7 Multi-Image

BLINK

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

MUIRBENCH

```
<image_1>
<text_1>
<image_2>
<text_2>
...
<image_n>
<text_n>
Answer with the option's letter from the given choices directly.
```

B.8 Video Understanding

MVBench | VideoMME | MLVU | LVBench - For instruct models

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

MVBench | VideoMME | MLVU | LVBench - For thinking models

```
<video>
Select the best answer to the following multiple-choice question based on the video.
Respond with only the letter (A, B, C, or D) of the correct option.
Question: {question}
{options}
Please reason step-by-step, identify relevant visual content, analyze key timestamps and
clues, and then provide the final answer.
```


RoboSpatialHome

```
<图像>在图像中定位{对象_名称}。以JSON格式输出点坐标。例如:[{"点_2坐标": [x, y], "
标签": "点_1"}]
```

RefSpatialBench

```
<图像>{问题} 以JSON格式输出点坐标。例如:[{"点_2d": [x,
y], "标签": "点_1"}]
```

B.7 多图像

BLINK

```
<图像>
问题：{question}
选项：
{options}
请从上方选项中选择正确答案。
```

MUIRBENCH

```
<图像_1><文本_1><图像_2><文本_2>...<图像_n><文本_n>直接用给定选项
的字母回答。
```

B.8 视频理解

MVBench | VideoMME | MLVU | LVBench - For 指令模型

```
<视频>选择最佳答案以回答以下基于视频的问答题。仅以正确选项的字母（A、B、C 或 D）形
式回答。问题：{question} 可能的答案选项：{options} 最佳答案是：
```

MVBench | VideoMME | MLVU | LVBench - For 思考模型

Charades-STA

```
<video>
Give you a textual query: {query_text}
When does the described content occur in the video?
Please return the timestamp in seconds.
```

VideoMMMU

Perception & Comprehension:

```
<video>
{question}
{options}
Please ignore the Quiz question in last frame of the video.
```

Adaptation-multiple-choice:

```
<video>
<image>
You should watch and learn the video content. Then apply what you learned to answer the
following multi-choice question. The image for this question is at the end of the video.
{question}
{options}
```

Adaptation-open-ended:

```
<video>
<image>
You should watch and learn the video content. Then apply what you learned to answer the
following open-ended question. The image for this question is at the end of the video.
{question}
```

MMVU

multiple-choice:

```
<video>
{question}
{options}
Visual Information: processed video
Answer the given multiple-choice question step by step. Begin by explaining your
reasoning process clearly. Conclude by stating the final answer using the following
format: "Therefore, the final answer is: $LETTER" (without quotes), where $LETTER is one
of the options. Think step by step before answering.
```

open-ended:

```
<video>
{question}
Visual Information: processed video
Answer the given question step by step. Begin by explaining your reasoning process
clearly.
Conclude by stating the final answer using the following format: "Therefore, the final
answer is: "Answer: $ANSWER" (without quotes), where $ANSWER is the final answer of the
question. Think step by step before answering.
```


Charades-STA

<视频>
给你一个文本查询：{query_文本} 描述的内容在视频中何
时发生？请以秒为单位返回时间戳。

VideoMMMU

感知 & 理解：

<视频>
{问题}
{选项}
请忽略视频最后一帧中的测验问题。

多选题适应：

<视频><图像>你应该观看并学习视频内容。然后将你学到的知识应用到回答以下多选题中。该问题
的图像位于视频末尾。{问题} {选项}

开放式适应：

<视频>
<图像>
你应该观看并学习视频内容。然后应用你所学到的东西来回答
以下开放式问题。这个问题的图像在视频的末尾。
{问题}

MMVU

多项选择：

<视频>
{问题}
{选项}
视觉信息：已处理的视频
针对给定的选择题，逐步作答。首先需清晰解释你的
推理过程。最后需使用以下格式给出最终答案：
："因此，最终答案是：$LETTER"（不带引号），其中 $LETTER 是选项之一
回答前请逐步思考。

开放式：

<视频>
{问题}
视觉信息：已处理的视频
逐步回答给定的问题。首先解释你的推理过程
清晰地。
最后使用以下格式陈述最终答案："因此，最终
答案是："答案：$ANSWER"（不带引号），其中$ANSWER是问题的最终答案
回答前请逐步思考。

B.9 Perception with Tool

V*

```
Your role is that of a research assistant specializing in visual information. Answer questions about images by looking at them closely and then using research tools. Please follow this structured thinking process and show your work.

Start an iterative loop for each question:

- **First, look closely:** Begin with a detailed description of the image, paying attention to the user's question. List what you can tell just by looking, and what you'll need to look up.
- **Next, find information:** Use a tool to research the things you need to find out.
- **Then, review the findings:** Carefully analyze what the tool tells you and decide on your next action.

Continue this loop until your research is complete.

To finish, bring everything together in a clear, synthesized answer that fully responds to the user's question.

#Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:
<tools>
{ "type":"function", "function": {"name": "image_zoom_in_tool", "description": "Zoom in on a specific region of an image by cropping it based on a bounding box (bbox) and an optional object label", "arguments": {"type": "object", "properties": {"bbox_2d": {"type": "array", "items": {"type": "number"}, "minItems": 4, "maxItems": 4, "description": "The bounding box of the region to zoom in, as [x1, y1, x2, y2], where (x1, y1) is the top-left corner and (x2, y2) is the bottom-right corner"}, "label": {"type": "string", "description": "The name or label of the object in the specified bounding box"}, "img_idx": {"type": "number", "description": "The index of the zoomed-in image (starting from 0)"}}, "required": ["bbox_2d", "label", "img_idx"]}}}
</tools>
For each function call, return a JSON object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{{"name": <function-name>, "arguments": <args-json-object>}}
</tool_call>
<image>
{question}
```


B.9 感知工具

V*

```
您的角色是一名专门从事 视觉信息的研究助理。 回答
关于图像的问题，方法是仔细观察图像，然后使用研究工具。
请遵循结构化思维过程，并展示你的思考过程。

针对每个问题，启动一个迭代循环：

- **首先， 仔细观察：** 从图像的详细描述开始，注意
用户的问题。 列出仅通过观察就能得知的内容， 以及你需要
查找的内容。
- **接下来，查找信息：** 使用工具研究你需要了解的事情。
- **然后， 审查结果：** 仔细分析工具告诉你的内容，并决定
你的下一步行动。

继续这个循环，直到你的研究完成。

要完成， 将所有内容整合到一个清晰、 综合的答案中，完全回答
用户的问题。

#工具

你可以调用一个或多个函数来协助处理用户查询。

你可以在 <tools></tools> XML 标签中获取函数签名：
<tools>
{ "type":"函数", "function": {"name": "图像_放大_到_工具", "描述": "放大到
图像的特定区域，通过裁剪边界框（bbox）和可选的对象标签
进行操作", "参数": {"类型": "对象", "属性": {"bbox_2d": {"类型":}}
"数组", "items": {"type": "数字"}, "minItems": 4，"最大项": 4，"描述": "The
边界 的 the 区域 以缩放 in, as [x1, y1, x2, y2], 哪里 (x1, y1) is the
左上角和 (x2, y2) 是右下角"}, "标签": {"类型": "字符串",
"描述": "指定边界框中对象的名称或标签"}, "img_idx":
{"type": "数字", "描述": "缩放图像的索引（从0开始）"}},
"必需": ["bbox_2d" "标签", "img_idx"]}}}
</工具>
对于每个函数调用， 返回一个包含函数名和参数的JSON对象
<工具_调用></工具_调用> XML标签：
<工具-调用>
{{"name": <函数名>, "参数": <参数JSON对象>}}
</工具-调用>
<图像>
{问题}
```

HRBench4K | HRBench8K

Your role is that of a research assistant specializing in visual information. Answer questions about images by looking at them closely and then using research tools. Please follow this structured thinking process and show your work.

Start an iterative loop for each question:

- **First, look closely:** Begin with a detailed description of the image, paying attention to the user's question. List what you can tell just by looking, and what you'll need to look up.
- **Next, find information:** Use a tool to research the things you need to find out.
- **Then, review the findings:** Carefully analyze what the tool tells you and decide on your next action.

Continue this loop until your research is complete.

To finish, bring everything together in a clear, synthesized answer that fully responds to the user's question.

#Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:
<tools>
{ "type":"function", "function": {"name": "image_zoom_in_tool", "description": "Zoom in on a specific region of an image by cropping it based on a bounding box (bbox) and an optional object label", "arguments": {"type": "object", "properties": {"bbox_2d": {"type": "array", "items": {"type": "number"}, "minItems": 4, "maxItems": 4, "description": "The bounding box of the region to zoom in, as [x1, y1, x2, y2], where (x1, y1) is the top-left corner and (x2, y2) is the bottom-right corner"}, "label": {"type": "string", "description": "The name or label of the object in the specified bounding box"}, "img_idx": {"type": "number", "description": "The index of the zoomed-in image (starting from 0)"}}, "required": ["bbox_2d", "label", "img_idx"]}}}
</tools>
For each function call, return a JSON object with function name and arguments within <tool_call></tool_call> XML tags:
<tool_call>
{{"name": <function-name>, "arguments": <args-json-object>}}
</tool_call>
<image>
{question}
{options}

B.10 Coding

Design2Code (Generation)

<image>
You are an expert web developer who specializes in HTML and CSS. A user will provide you with a screenshot of a webpage. You need to return a single HTML file that uses HTML and CSS to reproduce the given website. Include all CSS code in the HTML file itself. If it involves any images, use "rick.jpg" as the placeholder. Some images on the webpage are replaced with a blue rectangle as the placeholder, and use "rick.jpg" for those as well. Do not hallucinate any dependencies on external files. You do not need to include JavaScript scripts for dynamic interactions. Pay attention to things like size, text, position, and color of all the elements, as well as the overall layout. Respond with the content of the HTML+CSS file:

HRBench4K | HRBench8K

您的角色是专注于视觉信息的研究助理。通过仔细观察图像，并使用研究工具来回答有关图像的问题。请遵循这种结构化思维过程，并展示您的工作。

为每个问题启动一个迭代循环：

- **首先，仔细观察：** 从图像的详细描述开始，注意用户的问题。仅凭观察就能列出你能知道的内容，以及需要查找的内容。
- **接下来，查找信息：** 使用工具来研究你需要了解的事情。
- **然后，审查研究结果：**仔细分析工具告诉你的内容，并决定你的下一步行动。

继续这个循环，直到你的研究完成。

最后，将所有内容整合成一个清晰、综合性的答案，以全面回应针对用户的问题。

#工具

您可以调用一个或多个函数来协助处理用户查询。

您可以在 <tools></tools>XML 标签中找到函数签名：
<tools>
{ "type":"函数", "函数": {"name": "图像_放大_到_工具" 描述 <code>放大</code>在图像的特定区域上通过基于边界框（bbox）裁剪来裁剪可选对象标签" "参数": {"type": "对象", "属性": {"bbox_2d": {"type": "数组", "元素": {"type": "数字"}, "minItems": 4, "maxItems": 4, "描述": "The 边界 区域 the 区域 进行缩放 in, as [x1, y1, x2, y2], 在哪里 (x1, y1) is the 左上角和 (x2, y2) 是右下角"}, "标签": {"类型": "字符串", "描述": "指定边界框中对象的名称或标签"}, "img_idx": {"type": "数字", "描述": "缩放图像的索引（从0开始）"}}, "必需": ["bbox_2d", "标签", "img_idx"]}}}
</工具>
对于每个函数调用，返回一个包含函数名和参数的JSON对象<工具_调用></工具_调用> XML标签：
<工具-调用>
{{"name": <函数名>, "参数": <args-json-object>}}
</工具-调用>
<图像>

{问题}
{选项}

B.10 编程

设计到代码 (生成)

<图像>
你是一位精通 HTML 和 CSS 的专业网页开发者。用户将为你提供网页的截图。你需要返回一个单独的 HTML 文件，该文件使用 HTML 和 CSS 来重现给定的网站。将所有 CSS 代码都包含在 HTML 文件本身中。如果涉及图像，请使用 "rick.jpg" 作为占位符。网页上的一些图像被替换为蓝色矩形作为占位符，也使用 "rick.jpg"。不要虚构对外部文件的依赖。你不需要包含用于动态交互的 JavaScript 脚本。注意所有元素的大小、文本、位置和颜色，以及整体布局。以 HTML+CSS 文件的内容进行回复：

Design2Code (GPT-o4-mini Evaluation)

I will give you two images. The first is the reference, and the second is generated from the first via code rendering. Please rate their similarity from 0-100, where 0 means completely different and 100 means identical. Provide the score inside a LaTeX $\square$ and briefly explain your reasoning.
<reference_image>
<generated_image>

B.11 Agent

Screenspot | Screenspot-Pro | OSWorld-G

Tools
You may call one or more functions to assist with the user query.
You are provided with function signatures within <tools> ... </tools> XML tags:
<tools> { "name":"computer_use", "description": "Use a mouse to interact with a computer. The screen's resolution is <display_width_px>x <display_height_px>." "notes": "Click with the cursor tip centered on targets; avoid edges unless asked. Do not use other tools (type, key, scroll, left_click_drag). Only left_click and mouse_move are allowed. If you can't find the element, terminate and report failure.", "parameters":{ "type":"object", "required":["action"], "properties":{ "action":{ "type":"string", "enum":["mouse_move","left_click"], "description":"The action to perform." }, "coordinate":{ "type":"array", "description":"(x, y): pixels from left/top. Required for action=mouse_move and action=left_click." } } } }
</tools>
For each function call, return a JSON object with function name and arguments within <tool_call> ... </tool_call> XML tags:
<tool_call>
{{"name": <function-name>, "arguments": <args-json-object>}}
</tool_call>
Additionally, if you think the task is infeasible (e.g., the task is not related to the image), return:
<tool_call>
{"name": "computer_use", "arguments": {"action": "terminate", "status": "failure"}}
</tool_call>

设计到代码 (GPT-o4-mini 评估)

我将给你两张图像。第一张是参考图像，第二张是从第一张通过代码渲染生成的。通过代码渲染生成的。请评价它们的相似度从0到100，其中0表示完全不同且100表示相同。在LaTeX内提供分数。$\square$ and 简要解释你的推理。
<参考_图像>
<生成_图像>

B.11 代理

ScreenSpot |ScreenSpot-Pro |OSWorld-G

工具您可以通过调用一个或多个函数来协助处理用户查询。您可以在 <tools> 。 . . </tools> XML 标签:<tools> { "name":"computer_use", "description": "使用鼠标与计算机交互。屏幕的分辨率是 <display_width_px>x <display_height_px>。" "notes": "将光标尖端居中于目标处点击；除非被要求，否则避免边缘；不要使用其他工具（输入、按键、滚动、左_点击_拖动）。仅允许左_点击和鼠标_移动。如果您找不到元素，请终止并报告失败。", "parameters":{ "type":"对象", "必需":["action"], "属性":{ "action":{ "type":"字符串", "枚举":["mouse_移动","左_点击"], "描述":"要执行的操作。" }, "coordinate":{ "type":"数组", "描述":"(x, y): 从左/顶部的像素坐标。对于 action=mouse_移动 和 action=左_点击 是必需的。" } } }</tools>对于每个函数调用，在 <tool_call>。 . . </tool_call> XML 标签:<tool_call>{"name": <函数名>, "参数": <args-json对象>}}</tool_call>此外，如果您认为任务不可行（例如，任务与图像无关），请返回:
<tool_call>{"name": "computer_use", "参数": {"action": "终止", "状态": "失败"}}</tool_call>