本文由 AINLP 公众号整理翻译，更多 LLM 资源请扫码关注!

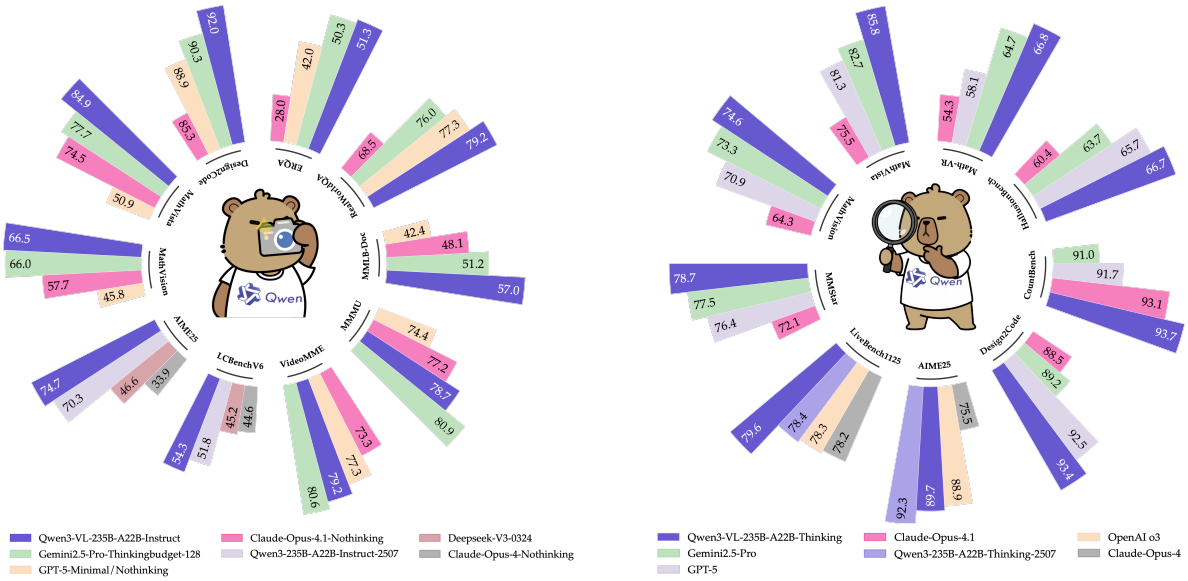
Qwen3-VL Technical Report

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.



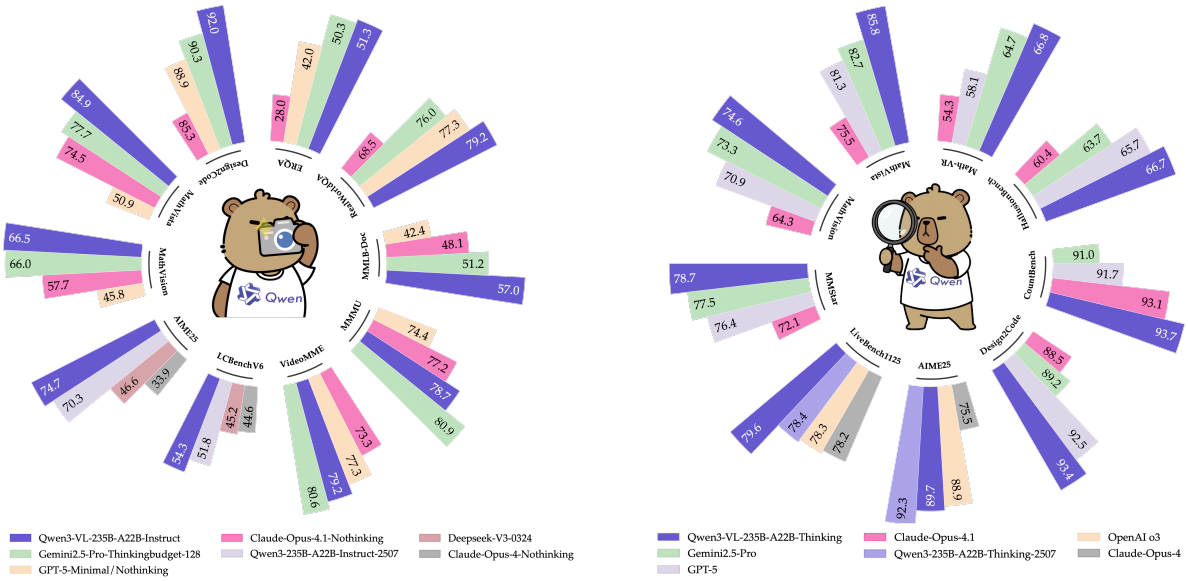
Qwen-VL 技术报告

通义千问团队

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

摘要

我们推出了 Qwen3-VL，这是迄今为止 Qwen 系列中最强大的视觉-语言模型，在各种多模态基准测试中均表现出色。它原生支持高达 256K tokens 的交错上下文，无缝集成文本、图像和视频。该模型系列包括密集型 (2B/4B/8B/32B) 和混合专家型 (30B-A3B/235B-A22B) 变体，以适应不同的延迟-质量权衡。Qwen3-VL 具有三个核心支柱：（i）显着增强的纯文本理解能力，在某些情况下超过了同类纯文本 backbone；（ii）强大的长上下文理解能力，具有原生的 256K-token 窗口，可用于文本和交错多模态输入，从而能够在长文档和视频中进行忠实保留、检索和交叉引用；以及（iii）跨单图像、多图像和视频任务的高级多模态推理，在 MMMU 和视觉数学基准（例如，Math-Vista 和 MathVision）等综合评估中表现出领先的性能。在架构上，我们引入了三个关键升级：（i）增强的 interleaved-MRoPE，用于增强图像和视频中的空间-时间建模；（ii）DeepStack 集成，它有效地利用多层 ViT 特征来加强视觉-语言对齐；以及（iii）基于文本的视频时间对齐，从 T-RoPE 演变为显式文本时间戳对齐，以实现更精确的时间定位。为了平衡纯文本和多模态学习目标，我们应用平方根重加权，这提高了多模态性能，而不会影响文本能力。我们将预训练扩展到 256K tokens 的上下文长度，并将后训练分为非思考型和思考型变体，以满足不同的应用需求。此外，我们为后训练阶段分配了额外的计算资源，以进一步提高模型性能。在相当的 token 预算和延迟约束下，Qwen3-VL 在密集型和混合专家 (MoE) 架构中均实现了卓越的性能。我们设想 Qwen3-VL 将成为现实世界工作流程中图像基础推理、代理决策和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 介绍

近年来，视觉-语言模型（VLMs）取得了实质性进展，从基础的视觉感知发展到跨图像和视频的高级多模态推理。VLM的快速发展催生了迅速扩展的下游应用领域——例如长上下文理解、STEM推理、GUI理解和交互以及代理工作流。至关重要的是，这些进步不能削弱底层大型语言模型（LLM）的语言能力；多模态模型预计在语言基准测试中达到或超过其纯文本对应模型（{v*}）。

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 (Yang et al., 2025a)，我们实例化了四个稠密模型（2B/4B/8B/32B）和两个混合专家 (MoE) 模型（30B-A3B / 235B-A22B），每个模型都使用高达 256K tokens 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) 训练期间保留了底层 LLM 的语言能力，从而大大提高了整体能力。我们发布了非思考和思考变体；后者表现出明显更强的多模态推理能力，在复杂的推理任务上实现了卓越的性能。

我们首先介绍架构上的改进，这些改进涵盖三个组件：1) 增强的位置编码。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) 组会导致不平衡的频谱，并阻碍长视频理解。因此，我们采用交错的 MRoPE，将 t、h 和 w 均匀地分布在低频和高频频段上，从而产生更真实的位置表示。2) 用于跨层融合的 DeepStack。为了加强视觉-语言对齐，我们采用了开创性的 DeepStack (Meng et al., 2024) 机制。来自视觉编码器不同层的视觉 tokens 通过轻量级的残差连接被路由到相应的 LLM 层，从而增强了多级融合，而无需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 tokens 替换了 Qwen2.5-VL 中使用的通过位置编码进行的绝对时间对齐，以标记帧组，从而提供更简单、更直接的时间表示。此外，在优化方面，我们从每个样本的损失转变为平方根归一化的每个 token 损失，这更好地平衡了训练期间文本和多模态数据的贡献。

为了构建更强大、更稳健的视觉-语言基础模型，我们从质量、多样性和结构方面全面改进了训练数据。主要升级包括增强的标题监督、扩展的 omni-recognition 和 OCR 覆盖范围、通过 3D/空间推理进行归一化的 grounding，以及用于代码、长文档和时间 grounding 视频的新语料库。我们进一步注入了思维链推理和高质量、多样化的 GUI 代理交互数据，以桥接感知、推理和行动。总之，这些创新能够实现更强大的多模态理解、精确的 grounding 和工具增强的智能。

我们的训练流程包括两个阶段：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-语言投影层），同时保持模型其余部分冻结；然后是在 8K、32K 和 256K 序列长度下，使用逐渐增大的上下文窗口进行全参数训练。后训练包括三个阶段：（i）在长链式思维数据上进行监督微调，（ii）从更强大的教师模型中进行知识蒸馏，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，而且成为一个灵活的现实世界多模态智能平台——在不同的应用领域无缝集成感知、推理和行动。在以下章节中，我们将介绍模型架构、训练框架和广泛的评估，这些评估证明了其在文本、视觉和多模态推理基准测试中始终如一且具有竞争力的性能。

2 模型架构

与 Qwen2.5-VL (Bai et al., 2025) 类似，Qwen3-VL 采用了三模块架构，包括视觉编码器、基于 MLP 的视觉-语言融合器和一个大型语言模型 (LLM)。图 1 描述了详细的模型结构。

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）实例化，所有变体均基于 Qwen3 主干构建。旗舰模型 Qwen3-VL-235B-A22B 拥有 235B 总参数，每个 token 激活 22B 参数。

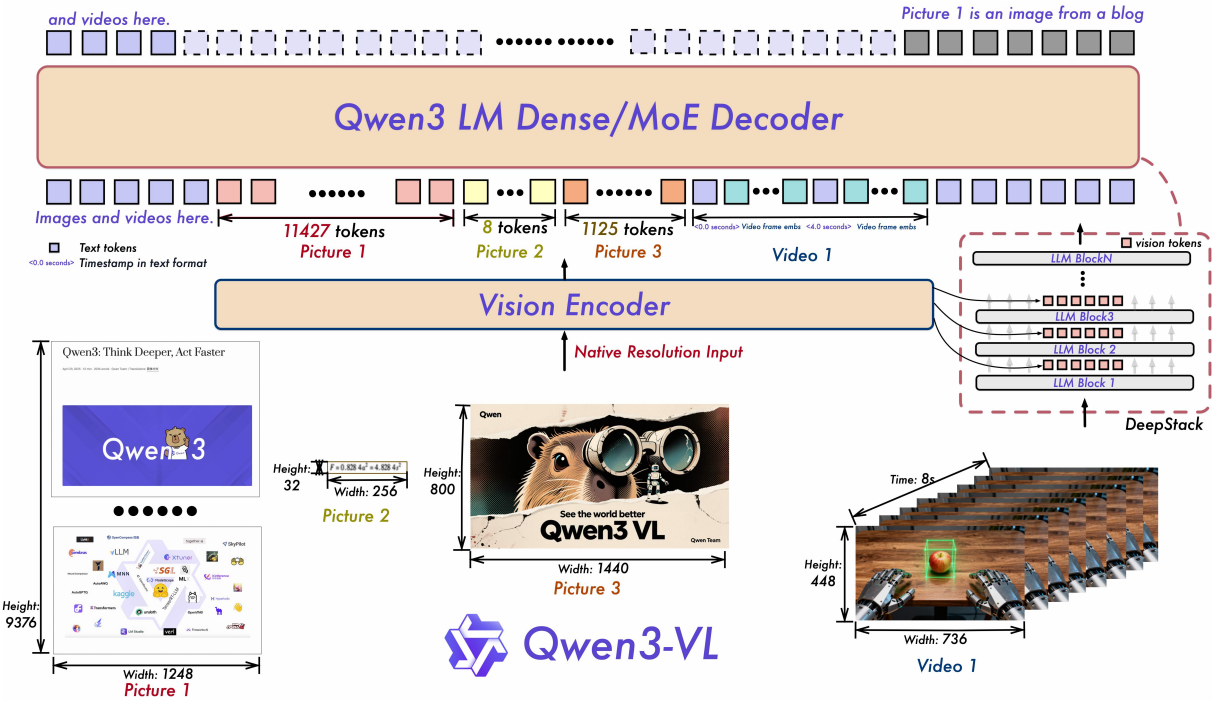


Figure 1: The Qwen3-VL framework integrates a vision encoder and a language model decoder to process multimodal inputs, including text, images, and video. The vision encoder is specifically designed to handle dynamic, native-resolution visual inputs, mapping them to visual tokens of variable length. 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.

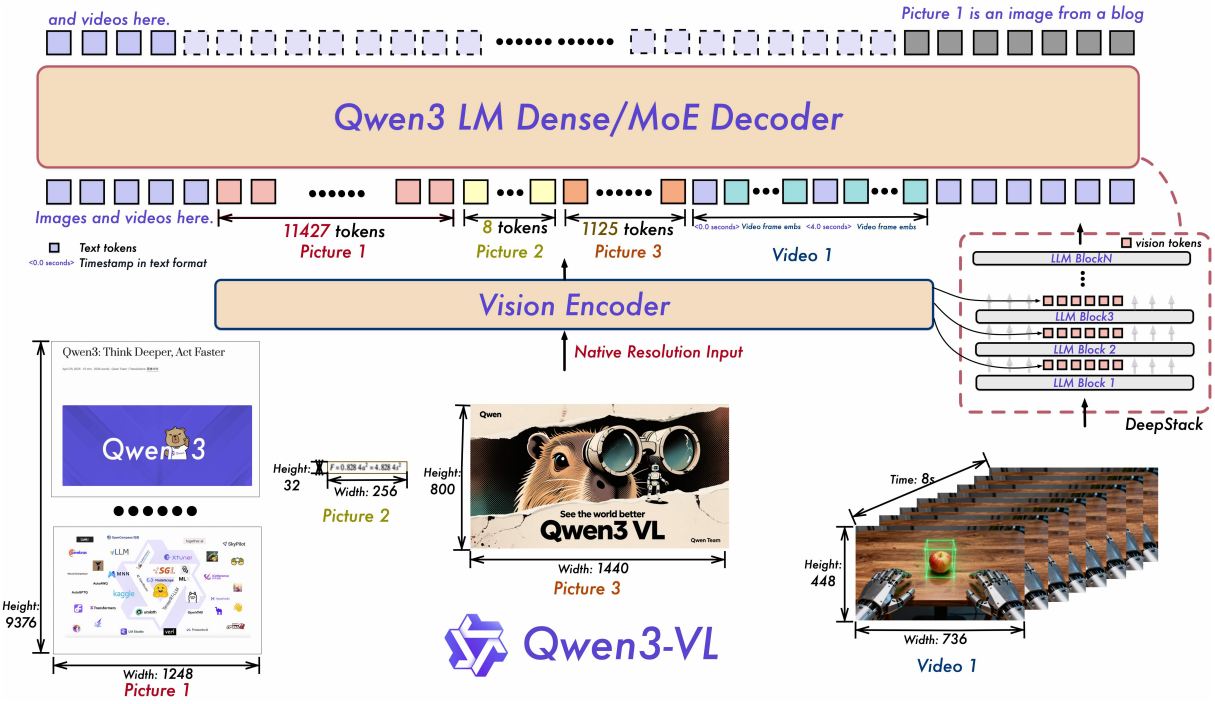


图 1：Qwen3-VL 框架集成了视觉编码器和语言模型解码器，以处理包括文本、图像和视频在内的多模态输入。视觉编码器专门设计用于处理动态的、原生分辨率的视觉输入，将其映射到可变长度的视觉 tokens。为了增强感知能力并保留丰富的视觉信息，我们采用了开创性的 DeepStack 机制，该机制将来自视觉编码器多个层的视觉 tokens 注入到 LLM 的相应层中。此外，我们采用 Interleaved MRoPE 来编码具有平衡频谱的多模态输入的位置信息，并引入基于文本的时间戳 tokens，以更有效地捕获视频序列的时间结构。

在广泛的多模态任务中，其性能优于大多数 VLM，并且在大多数语言基准测试中超过了其纯文本版本。

视觉编码器：我们采用 SigLIP-2 架构 (Tschannen et al., 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，并从官方预训练检查点进行初始化。为了有效地适应动态分辨率，我们采用 2D-RoPE 并根据输入大小插值绝对位置嵌入，遵循 CoMP (Chen et al., 2025) 的方法。具体来说，我们默认使用 SigLIP2-SO-400M 变体，并为小型 LLM（2B 和 4B）使用 SigLIP2-Large (300M)。

基于MLP的视觉-语言融合器：与Qwen2.5-VL中一样，我们使用一个两层MLP将视觉编码器中的2×2个视觉特征压缩成一个视觉token，与LLM的隐藏维度对齐。此外，我们部署了专门的融合器来支持DeepStack机制（Meng et al., 2024），其细节在第2.2节中完整描述。

2.1 交错式 MRoPE

Qwen2-VL (Wang et al., 2024c) 引入了 MRoPE 来建模多模态输入的位置信息。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，每个子空间分配不同的旋转频率。这导致了不平衡的频谱，后续研究表明，这会降低长视频理解基准测试的性能。为了解决这个问题，我们重新设计了频率分配，通过交错 t、h 和 w 分量在嵌入维度上进行分配 (Huang et al., 2025)。这确保了每个时空轴在低频和高频频段中都得到均匀表示。由此产生的平衡频谱减轻了原始频谱偏差，并显着提高了视频的远程位置建模能力。

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 (Meng et al., 2024) 中汲取灵感，并将视觉 tokens 注入到 LLM 的多个层中。与原始 DeepStack 方法堆叠来自多尺度视觉输入的 tokens 不同，我们扩展了 DeepStack 以从 Vision Transformer (ViT) 的中间层提取视觉 tokens。这种设计保留了丰富的视觉信息，范围从低级到高级表示 {v*}。

具体而言，如图 1 所示，我们从视觉编码器的三个不同层级中选择特征。随后，专用的视觉-语言融合模块将这些多层级特征投影到视觉 tokens 中，然后将这些 tokens 直接添加到前三个 LLM 层的相应隐藏状态 {v*} 中。

2.3 视频时间戳

在 Qwen2.5-VL 中，采用了一种时间同步的 MRoPE 变体，以赋予模型时间感知能力。然而，我们发现这种方法存在两个关键限制：（1）通过将时间位置 ID 直接与绝对时间绑定，该方法会为长视频生成过大且稀疏的时间位置 ID，从而降低模型理解长时序上下文的能力。（2）在这种方案下进行有效的学习需要在各种帧率 (fps) 上进行广泛且均匀的采样，从而显著增加训练数据构建的成本。

为了解决这些问题，我们采用了一种基于文本token的时间编码策略 (Chen et al., 2024b)，其中每个视频时间patch都以一个格式化的文本字符串作为前缀来表示时间戳——例如，<3.0 seconds>。此外，在训练期间，我们生成秒和HMS（小时:分钟:秒）格式的时间戳，以确保模型学会解释不同的时间码表示。虽然这种方法会适度增加上下文长度，但它使模型能够更有效和精确地感知时间信息，从而促进时间感知的视频任务，例如视频定位和密集字幕。

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的持续训练来增强视觉编码器。 整个 Qwen3-VL 模型采用三模块架构，包括该视觉编码器、一个基于 MLP 的视觉-语言融合器和一个 Qwen3 大型语言模型 (LLM) 主干。 在此架构的基础上，我们的预训练方法系统地构建为四个不同的阶段，旨在逐步构建从基本对齐到长上下文理解的能力。 这些阶段的概述如表 1 所示。

表 1: Qwen-VL 不同阶段的训练设置和超参数。

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

阶段 0：视觉-语言对齐。初始阶段 (S0) 侧重于高效地弥合视觉编码器和 LLM 之间的模态差距。 关键是，在此阶段仅训练 MLP 合并器的参数，而视觉编码器和 LLM 主干都保持冻结。 我们利用大约 67B 个 token 的精选数据集，该数据集由高质量的图像-标题对、视觉知识集合和光学字符识别 (OCR) 数据组成。所有训练均以 8,192 的序列长度进行。 这种对齐优先的方法为在进行全参数训练之前建立跨模态理解的坚实基础。

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——用于联合端到端训练。该模型在约1万亿（1T）token的海量多样化数据集上进行训练。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，增加了交错的图像-文本文档、视觉 grounding 任务、视觉问题 {v*}

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

回答 (VQA)、来自 STEM 领域的数据，以及少量视频数据，以引入时间理解。序列长度保持在 8,192。

第二阶段：长文本预训练。第二阶段 (S2) 旨在显著扩展模型的上下文处理能力。此阶段的一个关键变化是将序列长度增加到四倍，达到 32,768，同时所有模型参数继续保持可训练。训练在一个大约包含 1T tokens 的数据集上进行，并调整了数据混合比例以支持长文本任务。纯文本数据的比例增加，以增强长篇文本的理解能力，而剩余的 VL 数据则包含更大数量的视频和面向 Agent 的指令跟随数据。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。

第三阶段：超长上下文适应。最后阶段（S3）是一个专门的阶段，旨在将模型的上下文窗口推向其操作极限。在这里，我们将序列长度大幅增加到 262,144。该模型在一个更集中的 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，并强烈强调长视频和长文档理解任务。这种最终的适应巩固了 Qwen3-VL 在处理和分析极长顺序输入方面的能力，这是诸如综合文档分析和长视频摘要等应用的关键能力。

3.2 预训练数据

3.2.1 图像描述和交错文本-图像数据

为了构建一个用于通用视觉-语言理解的强大基础模型，我们显著扩展和改进了两种核心数据模式：图像-标题对和交错的文本-图像序列。我们的策略强调高质量、多样化和语义丰富的多模态基础，并由专门构建的模型和严格的过滤流程提供支持。

图像描述数据：我们从网络来源整理了一个大规模的当代中英文多语种图像-文本对语料库，并应用了一个多阶段的优化流程，该流程以专门为重新描述图像而微调的 Qwen2.5-VL-32B 模型为中心。该模型利用与每张图像相关的原始文本来生成更全面、流畅和细粒度的描述——丰富视觉元素的描述（例如，对象属性、空间布局和上下文语义），同时提高文本成分的语言质量和信息量。

去重仅在重新标注的文本上使用语义相似性指标执行，确保在不牺牲视觉多样性的前提下删除冗余样本。为了进一步提高对代表性不足概念的覆盖率，我们对视觉嵌入应用聚类（Johnson et al., 2019; Douze et al., 2024; Diao et al., 2025），以识别数据分布中的稀疏区域并执行有针对性的增强。最终得到的是一个高保真度的字幕数据集，它平衡了规模、多样性和描述粒度。

交错的文本-图像数据：我们收集了来自近期中文和英文网站的各种真实世界多模态文档（Laurençon et al., 2023; Zhu et al., 2023; Li et al., 2024c）。所有文档都经过领域分类（Wettig et al., 2025），使用基于Qwen的轻量级评分器进行微调，以进行细粒度的领域识别。基于跨领域的验证实验，我们系统地排除有害或低价值的类别——例如广告、促销内容和标题党——使用相同的有效评分器来过滤掉不良样本。

对于书籍规模的交错数据，我们采用微调后的 Qwen2.5-VL-7B 模型进行高精度多模态解析，精确提取文本并将其与嵌入的图形、图表和照片对齐。为了实现超长上下文建模，我们构建了一个专门的子集，通过将连续页面合并成最多 256K 个 token 的序列，从而保留自然的页面顺序和多模态连贯性。在预处理过程中，我们实施严格的质量控制：（i）删除纯文本或低对齐的片段；（ii）对于超长书籍序列，我们要求最小页面计数和最小图像文本比，以确保整个上下文中具有有意义的视觉-文本交互。这产生了一个干净、多样且布局感知的交错语料库，该语料库针对基础理解和长程多模态推理进行了优化。

3.2.2 知识

世界知识对于多模态大型语言模型 (MLLM) 在各种下游任务中实现稳健的视觉理解、有根据的推理和实体感知生成至关重要。为了使 Qwen3-VL 具备对现实世界和虚构概念的全面掌握，我们构建了一个

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

大规模预训练数据集，以明确定义的实体为中心，涵盖十几个语义类别——包括动物、植物、地标、食物以及车辆、电子产品和服装等日常物品。{v*}

现实世界的实体遵循长尾分布：突出的概念频繁出现，并带有高质量的标注，而大多数概念则很罕见。为了解决这种不平衡，我们采用了一种基于重要性的抽样策略。高突出度的实体被更频繁地抽样，以确保足够的学习信号，而低突出度的实体则以较小的比例包含在内，以保持广泛的覆盖范围，而不会使训练过程不堪重负。这种方法有效地平衡了数据质量、效用和多样性。

所有保留的样本都经过一个多阶段的优化流程。除了标准的噪声和未对齐过滤之外，我们还用更丰富的、由 LLM 生成的描述来替换原始或稀疏的标题——例如通用的 alt-text。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物体或人的互动，从而提供更完整和更可靠的文本表示 {v*}。

总而言之，这些努力共同产生了一个知识丰富、具有上下文感知能力且侧重于区分的训练信号，从而显著增强了 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：为了提高 OCR 在真实场景图像上的性能，我们使用由粗到精的流程，整理了一个包含 3000 万个内部收集样本的数据集。该流程通过整合 OCR 专用模型的伪标签和 Qwen2.5-VL 的优化结果来改进 OCR 注释，无需任何人工标注。除了 Qwen2.5-VL 支持的 10 种语言（不包括中文和英文）之外，我们还增加了 29 种语言，合成了大约 3000 万个高质量的多语言 OCR 样本，并整理了超过 100 万张内部真实场景多语言图像。

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（每种类型 30 万个样本），以及 400 万份内部文档。一个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；然后 Qwen2.5-VL-72B 执行特定区域的识别。输出被重新组装成具有位置感知、布局对齐的解析数据。

为了确保跨异构格式的稳健解析，我们设计了一个统一的标注框架，支持两种表示形式：{v*}

- QwenVL-HTML，它包含细粒度的、元素级别的边界框；
- QwenVL-Markdown，其中仅图像和表格被本地化，表格以 LaTeX 编码。

我们构建了一个带有精确注释的大规模合成 HTML 语料库，并系统地将其转换为 Markdown 格式。为了进一步提高模型的泛化能力，我们在大量的真实文档集合上生成伪标签，并对其进行质量过滤。最终的训练集结合了合成数据和高质量的伪标签数据，以增强可扩展性和鲁棒性。

长文档理解：为了增强模型理解多页 PDF 的能力（通常跨越数十页），我们利用了大规模的长文档数据语料库。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，多个页面图像放置在开头，然后是它们对应的文本，这些文本来源于 OCR 或 HTML 解析。其次，我们构建长文档视觉问答 (VQA) 数据。具体来说，我们采样高质量的多页 PDF，并生成各种各样的 VQA 示例，这些示例要求模型跨多个页面和异构文档元素（如图表、表格、图形和正文文本）进行推理。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进行稳健、有根据和多跳推理。

3.2.4 实例化和计数

视觉定位是多模态模型的一项基本能力，使它们能够准确地识别、解释和定位各种视觉目标，从特定对象到任意图像区域。在 Qwen3-VL 中，我们系统地增强了定位能力，并支持两种定位方式：边界框和点。这些表示形式允许在各种场景和下游任务中精确而灵活地解释图像内容。此外，我们扩展了 {v*} 的定位能力

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.

该模型支持计数，从而能够对视觉实体进行定量推理。下面，我们将简要介绍用于 grounding 和计数的数据构建流程。

基于框的 Grounding：我们首先聚合广泛使用的开源数据集，包括 COCO (Lin et al., 2014)、Objects365 (Shao et al., 2019)、OpenImages (Kuznetsova et al., 2020) 和 RefCOCO/+/g (Kazemzadeh et al., 2014; Mao et al., 2016)。为了进一步丰富数据的多样性，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对象注释。该管道分三个阶段运行：（i）使用 Qwen2.5-VL 从未标记的图像中提取对象候选；（ii）使用开放词汇检测器（特别是 Grounding DINO (Liu et al., 2023a)）和 Qwen2.5-VL 对这些候选对象进行定位和注释；（iii）对生成的注释进行质量评估，系统地过滤掉低置信度或不准确的注释。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，涵盖了各种视觉环境和对象类别。

基于点的 Grounding：为了确保基于点的 grounding 的稳健性，我们整理了一个综合数据集，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（Deitke et al., 2024）的公共指向和计数标注；（ii）从公共目标检测和实例分割基准测试中导出的目标 grounding 数据；以及（iii）由专门的合成管道生成的高精度指向标注，该管道旨在针对细粒度的图像细节。

在 grounding 数据的基石之上，我们精心挑选了一个高质量的子集，构成了我们的计数数据集的基础，该数据集包含三种不同的任务形式：直接计数、基于框的计数和基于点的计数。这三种任务类型共同构成了一个全面的计数数据集。

与 Qwen2.5-VL 不同，我们在此版本中采用了归一化的坐标系，将其缩放到 [0, 1000] 范围内。这种设计提高了对不同输入中图像分辨率和宽高比变化的鲁棒性，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

为了促进与物理世界的复杂交互，Qwen3-VL 被设计为对空间环境有深刻的理解。这使得模型能够解释空间关系，推断物体的可供性，并执行动作规划和具身推理。它还可以从单个单目图像估计物体的 3D 空间位置。为了支持这些能力，我们创建了两个全面的数据集，专注于空间理解和 3D 定位。

空间理解。除了定位物体之外，Qwen3-VL 还经过训练，能够推理 2D 场景中的空间关系、物体可供性和可行操作——这些能力对于具身人工智能和交互式应用至关重要。为此，我们构建了一个专门的数据集，该数据集通过结合以下内容超越了标准的基础：(i) 关系注释（例如，“笔记本电脑左边的杯子”），(ii) 可供性标签（例如，“可抓握”、“可按压”、“可坐”），以及 (iii) 需要规划的动作条件查询（例如，“我应该先移动什么才能拿到显示器后面的书？”）。这些样本来自精心策划的真实场景和合成生成的布局，自然语言查询通过基于模板和基于 LLM 的方法自动生成，以确保多样性和复杂性。至关重要的是，所有空间参考都相对于其他对象或场景框架表达，而不是绝对坐标，从而鼓励强大的关系推理。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”和“可以做什么”的问题——为与视觉环境的能动交互奠定基础。

3D 接地（Grounding）。为了进一步增强模型从图像中理解物理世界的能力，我们构建了一个专门的预训练数据集，用于 3D 视觉接地。我们从各种室内和室外场景的公共集合中获取数据，并将其重新构建为视觉问答格式。每个样本包括：1) 单视角相机图像，2) 自然语言指代表达式，以及 3) 相应的 9 自由度 3D 边界框注释，采用结构化的 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，因此它们表现出不同的相机内在参数和固有噪声。为此，我们过滤掉严重遮挡和不准确的标签，并遵循 Omni3D (Brazil et al., 2023) 将所有数据统一到虚拟相机坐标系中。我们还合成了大量的描述性标题，为 3D 接地创建丰富的文本查询。这些描述超越了命名对象的类别，还包括详细的属性、布局安排、空间位置、视觉可供性以及与周围对象的交互——从而产生更细粒度和接地的指代表达式。

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

我们通过将两类代码相关数据纳入训练语料库，增强了 Qwen3-VL 系列的专用编码能力，使模型能够在纯文本和视觉接地的上下文中读取、编写和推理程序，其中 {v*} 保持不变。

纯文本编码。我们复用了 Qwen3 和 Qwen3-Coder 系列中大量的代码语料库。这个大规模数据集涵盖了广泛的编程语言和领域，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。

多模态编码。为了解决既需要视觉理解又需要代码生成的任务，我们为各种多模态编码任务整理了数据。该数据集来源于开源数据集和内部合成管道，旨在教会模型联合理解视觉输入并生成功能代码。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 et al., 2025c）；解决视觉编程挑战（Li et al., 2024a）；回答多模态编码问题（例如，带有图像的StackOverflow帖子）；以及将视觉表示（例如流程图、图表和L{v*}TEX方程）转录为它们各自的代码或标记。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

Qwen-VL 的视频理解能力得到了显著提升，能够对跨帧的时间动态进行稳健建模、对空间关系进行细粒度感知，并对超长视频序列进行连贯的总结。 这种增强的基础是一个数据处理流程，其中包含两项主要创新：

时序感知视频理解。(i) 密集字幕合成：对于长视频序列，我们采用由短到长的字幕合成策略，以生成整体的、时间戳交错的、以及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的注释，共同捕获事件级的时间摘要和特定片段的视觉细节。(ii) 时空视频定位：我们整理和合成大规模视频数据，并在对象、动作和人物级别进行标注，以加强模型的时空定位能力，从而提高其细粒度视频理解能力。

视频数据平衡与采样。(i) 来源平衡：为了确保数据的平衡性和多样性，我们构建了一个大规模数据集，涵盖各种视频来源，包括教学内容、电影、以自我为中心的录像等。通过系统地管理视频标题、时长和类别标签等元数据来实现数据集的平衡。(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，例如每秒帧数 (fps) 和最大帧数。这种自适应策略减轻了与次优采样实践（例如，过于稀疏的帧选择或过低的 spatial resolution）相关的信息丢失，从而保留视觉细节并优化训练效果。

3. 2. 8 科学、技术、工程和数学 (STEM)

多模态推理是 Qwen3-VL 的核心，而 STEM 推理是其中最关键的部分。我们的理念遵循分而治之的策略：我们首先独立开发细粒度的视觉感知和强大的语言推理能力，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。我们开发了一个专门的合成数据生成流程，该流程通过程序化（基于代码）渲染来构建几何图。使用此流程，我们生成：（i）100万个点-grounding样本，例如交点、角点和重心；以及（ii）200万个面向感知的视觉问答对，目标是精细地视觉理解图表。为了获得高保真度的文本描述，我们进一步实施了一个两阶段的标题生成框架：一个初始生成阶段，然后是严格的基于模型的验证。这两个阶段都采用专门模型的集成，以确保准确性和描述粒度。此过程产生了一个包含600万个丰富注释的图表标题的综合数据集，涵盖了不同的STEM学科。

多模态推理数据。我们的大部分多模态推理数据包含超过 60 个 {v*}。

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

数百万道 K–12 和本科级别的练习题，通过严格的清洗和重构流程精心整理。在质量过滤期间，我们会丢弃低质量的项目，包括那些包含损坏图像、不相关内容或不完整或不正确答案的项目。在重构阶段，我们会在中文和英文之间翻译练习题，并标准化答案的格式——例如逐步解决方案列表、数学表达式和符号表示法——以确保一致性和统一呈现。关于长 CoT 问题解决数据，我们合成了超过 1200 万个与图像配对的多模态推理样本。为了确保推理过程的连续性和丰富性，我们利用了强大的推理模型生成的原始 rollouts。为了保证数据的可靠性和适用性，每个样本的推理轨迹都经过严格的验证——结合基于规则的检查和基于模型的验证——并且明确过滤掉任何包含模糊答案或代码转换的实例。此外，为了提高推理质量，我们仅通过拒绝抽样保留具有挑战性的问题。{v*}

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 与图形用户界面 (GUI) 自主交互的 Agent 能力，我们整理并合成了跨桌面、移动和 Web 环境的大规模跨平台数据（Ye et al., 2025; Wang et al., 2025a; Lu et al., 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诸如元素描述、密集字幕和密集接地的任务，从而能够稳健地理解各种用户界面。对于 Agent 能力，我们通过自我演化的轨迹生成框架组装多步任务轨迹，并辅以有针对性的人工审核；我们还精心设计并增强了思维链 (Chain-of-Thought) 推理，以加强实际执行过程中的规划、决策和反思性自我纠正。

函数调用：对于具有多模态上下文的通用函数调用能力，我们构建了一个多模态函数调用轨迹合成管道。我们首先指示具有图像的强大模型生成用户查询及其对应的函数定义。然后，我们采样带有理由的模型函数调用，并合成函数响应。重复此过程，直到判断用户的查询已解决。在每个步骤之间，由于格式错误，可以过滤掉轨迹。这种管道使我们能够从大量图像构建大规模多模态函数调用轨迹，而无需实现可执行函数。

搜索：在通用函数调用能力中，我们认为执行搜索的能力是促进现实场景中长尾实体知识整合的关键。在这种情况下，我们使用在线图像搜索和文本搜索工具收集多模态事实查找轨迹，鼓励模型对不熟悉的实体执行搜索。通过这样做，模型学会从网络收集信息，从而生成更准确的响应。

4 训练后

4. 1 训练配方

我们的后训练流程是一个三阶段过程，旨在完善模型的指令遵循能力，增强其推理能力，并使其与人类偏好对齐。每个阶段的具体数据和方法将在后续章节中详细介绍。

监督式微调（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：初始阶段在 32k 上下文长度下进行，然后扩展到 256k 上下文窗口，重点关注长文档和长视频数据。为了满足不同的需求，我们将训练数据分为非思考模型的标准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转移给我们的学生模型。至关重要的是，我们使用 *text-only* 数据执行此蒸馏，以微调 LLM 主干。 这种方法非常有效，在以文本为中心和多模态任务中，都能显著提高推理能力。

强化学习（RL）。最后阶段利用强化学习来进一步提升模型性能和对齐。此阶段分为推理 RL 和通用 RL。我们应用大规模强化

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,

学习涵盖全面的文本和多模态领域，包括但不限于数学、OCR、基础和指令跟随，以提高更细粒度的能力，其中能力包括 {v*}。

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的真实世界场景的能力。在Qwen2.5-VL的基本能力之上，它精通大约八个核心领域和30个细粒度子类别，我们战略性地扩展了其功能范围。这一扩展是通过整合来自社区反馈、学术文献和实际应用的见解来实现的，从而促进了新功能的引入。这些功能包括但不限于：用于具身智能的空间推理、用于细粒度视觉理解的图像-基础推理、用于鲁棒对象跟踪的视频中的时空基础，以及对跨越数百页的长上下文技术文档的理解。在这些目标任务的指导下，并以真实用例为基础，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，系统地策划了SFT数据集。这种有针对性的数据工程工作对于将Qwen3-VL建立为更全面和强大的多模态基础模型至关重要。

该数据集包含约 1,200,000 个样本，经过策略性地组合以培养强大的多模态能力。该集合被划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。多模态内容的集成专门用于使模型能够解释复杂的现实世界场景。为了确保全球相关性，该数据集超越了其主要的中文和英文语料库，包括各种多语言样本，从而扩大了其语言覆盖范围。此外，它通过结合在各种视觉环境中（从单张图像到多张图像序列）上下文化的单轮和多轮对话来模拟真实的对话动态。至关重要的是，该数据集还包含交错的图像-文本示例，这些示例旨在支持高级代理行为，例如工具增强的图像搜索和视觉基础推理。这种异构数据组成确保了全面的覆盖范围，并增强了数据集在训练可泛化和复杂的多模态代理方面的代表性。

鉴于 Qwen-VL 原生支持 256K token 的上下文长度，我们采用分阶段训练策略来优化计算效率。该策略包括两个阶段：初始阶段使用 32K token 的序列长度进行一个 epoch 的训练，然后第二个 epoch 使用完整的 256K token 长度。在后一个阶段，模型在一个课程上进行训练，该课程将长上下文输入与以 32K token 长度采样的数据交错。长上下文输入包括数百页的技术文档、完整的教科书以及长达两小时的视频等材料。

训练数据的质量是决定视觉-语言模型性能的关键因素。源自开源和合成来源的数据集通常受到显著的变异性和噪声的影响，包括冗余、不相关或低质量的样本。为了减轻这些缺陷，实施严格的数据过滤协议是必不可少的。因此，我们的数据管理过程包含一个两阶段的过滤流程：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 来识别和丢弃不易验证的查询。 对于具有模糊指令的查询，我们会进行最小程度的修改以提高清晰度，同时保留原始语义意图。 此外，我们会系统地删除缺乏实质内容的网络来源查询。 至关重要的是，所有剩余的查询都会经过对其复杂性和上下文相关性的最终评估，确保仅保留适当具有挑战性和相关性的 {v*} 样本以供下一阶段使用。

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、不完整或格式不正确。为了保持语义相关性并维护道德原则，我们还会丢弃任何离题或具有生成有害内容潜力的查询-响应对。
- 基于模型的过滤：通过使用从 Qwen2.5-VL 系列衍生的奖励模型，数据集得到进一步的提炼。这些模型对多模态问答对进行多维度评估。具体来说：（a）答案会根据一系列标准进行评分，包括正确性、完整性、清晰度和有用性；（b）对于视觉相关的任务，评估会特别强调对视觉信息的准确解释和利用的验证；（c）这种基于模型的方法能够检测到通常难以通过基于规则的方法发现的细微问题，例如幻觉（{v*}）。

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兀的风格转变。

这个多维过滤框架确保只有满足质量、可靠性和伦理完整性严格标准的数据才能进入 SFT 阶段。

4.2.2 长程 CoT 冷启动数据

我们思维模型的基础是一个精心策划的长链思维（CoT）冷启动数据集，旨在引出和完善复杂的推理能力。该数据集建立在跨越纯文本和多模态数据的各种查询集合之上，保持视觉-语言和纯文本样本之间大约 1:1 的比例，以确保平衡的技能发展。{v*}

多模态组件涵盖了视觉问答 (VQA)、光学字符识别 (OCR)、2D/3D 定位和视频分析等已建立的领域，同时特别强调丰富与 STEM 和代理工作流程相关的任务。 这种战略重点旨在推动模型在需要复杂、多步骤推理的问题上的表现。 纯文本部分与 Qwen3 使用的数据非常相似，包含数学、代码生成、逻辑推理和一般 STEM 方面的具有挑战性的问题。

为了保证高质量和适当的难度，我们实施了严格的多阶段过滤协议。

- 难度筛选：我们有选择地保留了基线模型通过率较低或生成更长、更详细回复的实例。这使得数据集富含对当前模型而言真正具有挑战性的问题。
- 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：我们丢弃任何我们的 Qwen3-30B-*nothink* 模型在没有视觉输入的情况下也能正确解决的样本。这确保了剩余的实例真正需要多模态理解，而不是仅通过文本线索就能解决。
- 响应质量控制：与 Qwen3 的方法一致，我们对生成的响应进行清理。对于具有多个候选答案的查询，我们首先删除包含不正确最终结果的答案。 随后，我们过滤掉表现出不良模式的响应，例如过度重复、不正确的语言混合或显示出明显猜测迹象而没有充分推理步骤的答案。

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，该数据集专为引导高级多模态推理而定制。

4.3 强到弱的知识蒸馏

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

- 离线策略蒸馏：在第一阶段，将教师模型生成的输出结合起来，以提供响应蒸馏。这有助于轻量级学生模型获得基本的推理能力，为后续的在线策略训练奠定坚实的基础。
- 在策略蒸馏：在第二阶段，学生模型根据提供的提示生成响应。然后，这些在策略序列用于微调学生模型。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

我们训练模型以执行各种文本和多模态任务，包括数学、编码、逻辑推理、视觉定位和视觉谜题。每个任务都经过精心设计，以便可以通过规则或代码执行器确定性地验证解决方案。

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手动标注，以确保高质量的RL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，为每个查询采样16个响应；对于所有响应都不正确的任何查询，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.

然后，我们针对每个任务运行初步的 RL 实验，以识别并移除改进潜力有限的数据源。此过程产生大约 3 万个 RL 查询，涵盖各种文本和多模态任务。对于每个模型的训练，我们为所有查询采样 16 个响应，并过滤掉通过率超过 90% 的简单查询。我们对特定于任务的数据集进行洗牌和组合，以构建混合任务批次，确保每个任务的样本具有一致的、预定义的比例。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，可以在所有任务中提供精确的反馈。该系统提供共享的基础设施——数据预处理、实用函数和一个奖励管理器来整合多种奖励类型——而核心奖励逻辑则根据每个任务进行实现。我们使用特定于任务的格式提示来引导模型输出到所需的格式，因此不依赖于显式的格式奖励。为了减轻代码切换，当响应语言与提示语言不同时，我们会施加惩罚。

RL 算法：我们采用 SAPO (Gao et al., 2025)，一种平滑且自适应的策略梯度方法，进行 RL 训练。SAPO 在不同的文本和多模态任务以及不同的模型大小和架构上都能提供持续的改进。

4.4.2 通用强化学习

通用强化学习（RL）阶段旨在增强模型的泛化能力和操作鲁棒性。为此，我们采用多任务 RL 范式，其中奖励函数是基于 SFT 阶段的一组综合任务制定的，包括 VQA、图像描述、OCR、文档解析、grounding 和时钟识别。奖励机制的结构旨在优化模型性能的两个主要维度：{v*}

- 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，确保生成的响应与用户要求精确匹配。
- 偏好对齐：对于开放式或主观查询，此维度通过优化有用性、事实准确性和风格适当性，使模型的输出与人类偏好对齐。这有助于建立更自然和引人入胜的用户互动。

此外，此阶段还充当一种纠正机制，以消除在 SFT 期间根深蒂固的强大但有缺陷的知识先验。我们通过引入专门的、可验证的任务来解决这个问题，这些任务旨在触发这些特定错误，例如违反直觉的物体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知识 {v*}。

另一个关键目标是减轻诸如不恰当的语言混合、过度重复和格式错误等不良行为。然而，这些问题发生率较低，使得通用强化学习成为一种样本效率低下的纠正策略。为了克服这个问题，我们在这个阶段整理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。这种有针对性的训练能够应用有针对性的、高频率的惩罚，从而有效地抑制这些残留错误。{v*}

强化学习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的真实结果的任务提供明确、高精度的反馈，例如格式遵守和指令遵循。通过使用明确定义的启发式方法，这种方法提供了一种强大的机制来评估正确性，并有效缓解奖励篡改，即模型可能利用学习到的奖励函数中的模糊性。{v*}
- 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。评判模型根据真实参考答案评估每个生成的响应，并从多个维度对其质量进行评分。这种方法为评估细微或开放式任务提供了卓越的灵活性，在这些任务中，严格的、基于规则的匹配是不够的。它尤其擅长最大限度地减少误报 {v*}，否则会惩罚具有非常规格式或措辞的有效响应。

4. 5 图像化思考

受先前关于“用图像思考”的伟大著作（Wu et al., 2025a; Jin et al., 2025; Zheng et al., 2025; Lai et al., 2025）的启发，我们通过一个两阶段的训练范式，赋予 Qwen3-VL 类似的智能体能力。

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，我们合成了包含大约 1 万个 grounding 示例的冷启动 Agentic 数据集——主要是一些简单的两轮视觉问答任务，例如属性检测。然后，我们对 Qwen2.5-VL-32B 进行监督微调 (SFT)，以模拟视觉 Agent 的行为：*think → act → analyze feedback → answer*。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，我们将第一阶段训练好的 Qwen2.5-VL-32B 视觉智能体进行提炼，以生成一个更大、更多样化的数据集，其中包含大约 12 万个多轮智能体交互，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（现在同时使用提炼和合成的数据）对 Qwen3-VL 进行后训练。

多轮、工具集成的强化学习过程在两个阶段几乎相同，仅在底层数据上有所不同。在强化学习过程中，我们采用三种互补的奖励信号来鼓励稳健的、工具介导的推理：

- 答案准确率奖励利用 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读工具或环境反馈，并通过连贯的、逐步的推理得出答案。公式表示法 {v*} 保持不变。
- 工具调用奖励通过比较实际的工具调用次数与专家估计的目标次数来鼓励适当的工具使用。该目标由 Qwen2.5-VL-72B 根据任务复杂性离线确定。{v*}

早期实验表明，模型倾向于退化为仅进行一次工具调用来破解前两个奖励，而不管任务需求如何。为了缓解这种情况，我们明确地纳入了工具调用奖励，以促进与任务复杂性相适应的自适应工具探索。{v*}

4.6 基础设施

我们使用阿里云PAI-灵骏智算服务训练 Qwen3-VL 系列模型，该服务为 AI 和高性能计算等计算密集型场景提供所需的高性能算力。

在预训练阶段，系统采用基于Megatron-LM框架的混合并行策略，整合了张量并行（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，从而实现了高硬件利用率，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，也能保持{v*}。

为了本地部署和性能评估，我们采用了基于 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。总之，这些后端提供了高效的推理和评估，并具有稳定、高效和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系列在通用视觉问答（VQA）方面的能力，我们在一系列不同的基准上进行了广泛的评估，包括MMBench-V1.1 (Liu et al., 2023b)、RealWorldQA (xAI, 2024)、MMStar (Chen et al., 2024a)和SimpleVQA (Cheng et al., 2025)。如表2、表3和表4所示，Qwen3-VL系列在从2B到235B参数的各种模型尺寸上都表现出强大且极具竞争力的性能。

在思维模式的比较中，Qwen3-VL-235B-A22B-Thinking 在 MMStar 上获得了最高的 78.7 分。Gemini-2.5-Pro (Comanici et al., 2025) 的思维模式提供了最佳的整体性能，但 Qwen3-VL-235B-A22B-Thinking 紧随其后。在非推理模式的比较中，Qwen3-VL-235B-A22B-Instruct 在 MMBench 和 RealWorldQA 上获得了最高的得分，分别为 89.3/88.9 和 79.2。

在中等规模模型的实验中，Qwen3-VL-32B-Thinking 在 MMBench 和 RealWorldQA 上取得了最高的成绩，分别为 89.5/89.5 和 79.4。值得注意的是，Qwen3-VL-32B-Instruct

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 Thinking 变体，得分为 79.0。

Qwen3-VL系列的可扩展性在我们较小模型的强大性能中可见一斑。具体来说，最大的模型Qwen3-VL-8B在所有五个基准测试中都取得了最高的性能。例如，在MMBench-EN上，“思考”模式下的分数从2B模型的79.9提高到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MMStar，其分数从68.1（2B，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上评估 Qwen3-VL 系列，主要侧重于 STEM 相关任务和视觉谜题，包括 MMMU (Yue et al., 2024a), MMMU- Pro (Yue et al., 2024b), MathVision (Wang et al., 2024b), MathVision-Wild $_{photo}$ (，以下简称 MathVision$_{WP}$) , 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), 和 VisuLogic (Xu et al., 2025)。如表 2 所示，旗舰版 Qwen3-VL 模型在“非思考”和“思考”模型中均表现出卓越的性能。值得注意的是，Qwen3-VL-235B-A22B-Instruct 在多个基准测试中，在非思考或低思考预算模型中取得了最佳报告结果，包括 MathVista$_{mini}$、MathVision、MathVerse$_{mini}$、DynaMath、ZeroBench、VLMsAreBlind、VisuLogic 和 VisualPuzzles$_{Direct}$。同时，Qwen3-VL-235B-A22B-Thinking 在 MathVista$_{mini}$、MathVision、MathVerse$_{mini}$、ZeroBench、LogicVista 和 VisuLogic 上取得了最先进的结果。

在中等规模的模型中，如表 3 所示，Qwen3-VL-32B 表现出显著的优势，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 模型已经在推理任务上超越了它。这突显了 VLM 的显著进步。此外，我们新推出的 Qwen3-VL-30B-A3B MoE 模型也提供了具有竞争力的结果。

在小型模型中，我们将 Qwen3-VL-2B/4B/8B 与 GPT-5-Nano 进行比较，结果如表 4 所示。8B 变体总体上保持明显的优势，而 4B 模型在 DynaMath 和 VisuLogic 上获得了最高分。值得注意的是，即使是最小的 2B 模型也表现出强大的推理能力。

5.3 对齐和主观任务

对于当前的大型视觉语言模型（VLMs）来说，遵循复杂的用户指令并减少潜在的图像级别幻觉的能力是不可或缺的。我们在三个具有代表性的基准上评估我们的模型：MM-MT-Bench (Agrawal et al., 2024)、HallusionBench (Guan et al., 2023) 和 MIA-Bench (Qian et al., 2024)。MM-MT-Bench 是一个多轮 LLM 作为评判者的评估基准，用于测试多模态指令调整模型。HallusionBench 旨在诊断图像上下文推理，并对当前的 VLM 提出了巨大的挑战。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，具有字符限制的创意写作和组合指令）。

如表2所示，我们的旗舰模型Qwen3-VL-235B-A22B始终优于其他闭源模型。在HallusionBench上，我们的思考版本分别超过Gemini-2.5-pro (Comanici et al., 2025)、GPT-5 (OpenAI., 2025) 和Claude opus 4.1 (Anthropic., 2025) 3.0、1.0和6.3个点。在MIA-Bench上，Qwen3-VL-235B-A22B-Thinking在所有其他模型中取得了总体最佳成绩，表明我们具有卓越的多模态指令遵循能力。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*math*和*textual*子任务中分别超过GPT-5-high-thinking版本10.0和5.0个点。在我们的较小尺寸模型（如Qwen3-VL-30B-A3B和Qwen3-VL-32B）上也可以观察到相同的趋势，它们超过了其他具有可比尺寸的模型。我们的2B/4B/8B系列也表现良好，并且下降幅度可忽略不计，尤其是在MIA-Bench上。

5.4 文本识别与文档理解

我们将 Qwen3-VL 系列与文档相关基准测试中其他大小相当的模型进行比较，这些基准测试包括 OCR、文档解析、文档问答 (QA) 和文档推理。

我们使用表 2 中列出的基准测试来评估我们的旗舰模型 Qwen3-VL-235B-A22B，并将其与最先进的 VLM 进行比较。在以 OCR 为中心的解析基准测试（包括 CC-OCR (Yang et al., 2024b) 和 OmniDocBench (Ouyang et al., 2024)）以及综合 OCR 基准测试（如 OCR-Bench (Liu et al., 2024) 和 OCRBench_v2 (Fu et al., 2024b)）中，Qwen3-VL-235B-A22B-Instruct 模型表现出色。

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：Qwen-VL-235B-A22B 和顶级模型在视觉基准测试上的性能。推理模型和非推理模型的最高分分别以粗体和下划线显示。标有 $*$ 的结果来自技术报告。$+$ 表示使用工具的结果。

	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	-	-	-	-	-	-

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：中等规模 Qwen3-VL 模型和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 $*$ 的结果来源于技术报告。$+$ 表示使用工具的结果。

	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	-	-	-	-

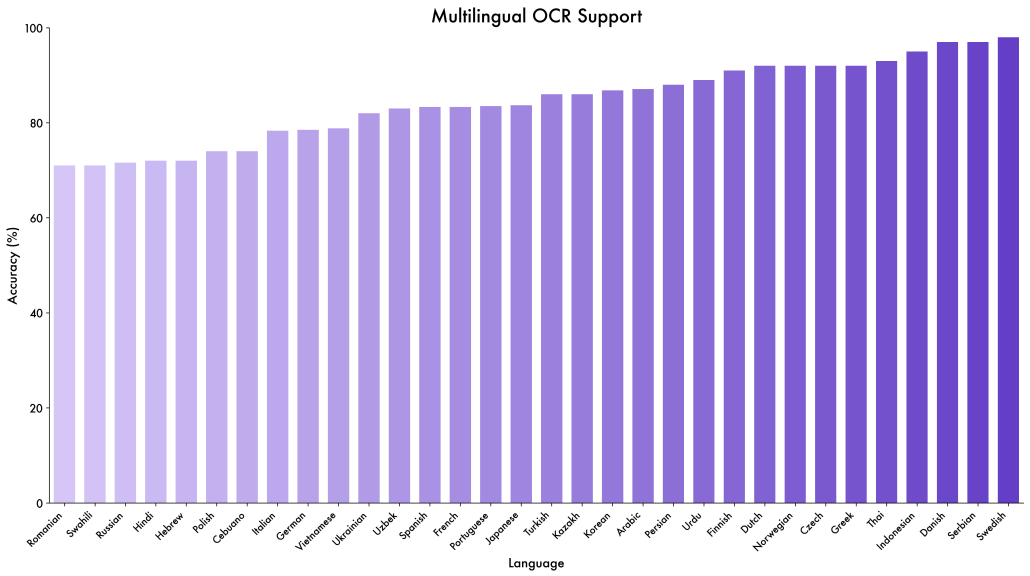


Figure 2: Multilingual OCR performance of our model on a self-built test set. The model achieves over 70% accuracy on 32 out of 39 supported languages, demonstrating strong and usable multilingual capabilities.

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

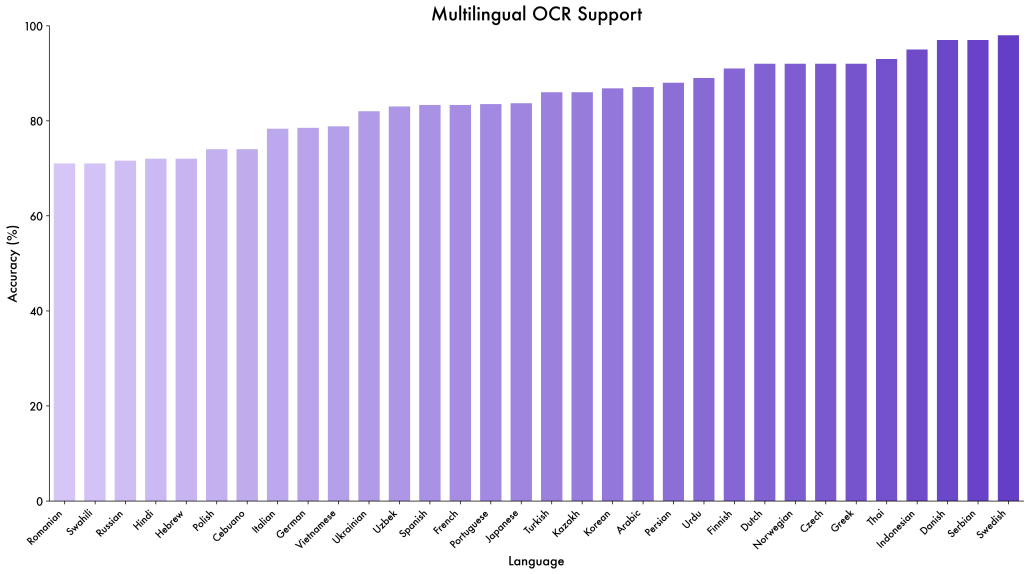


图 2：我们的模型在自建测试集上的多语言 OCR 性能。该模型在 39 种支持的语言中的 32 种语言上实现了超过 70% 的准确率，展示了强大且可用的多语言能力。

确立了新的技术水平，略微优于其“思考”版本 Qwen-VL-235B-A22B-Thinking。在需要 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 et al., 2021b)、InfoVQA (Mathew et al., 2021a)、AI2D (Kembhavi et al., 2016)、ChartQA (Masry et al., 2022) 和 CharXiv (Wang et al., 2024g) 描述子集，Instruct 和 Thinking 变体都取得了相当的性能，表明在这些任务中始终如一地取得了优异的成绩。值得注意的是，在 CharXiv 的推理子集上——该子集需要深入的图表理解和多步骤推理——Thinking 变体超越了 Instruct 版本，仅次于 GPT5-thinking 和 Gemini-2.5-Pro-Thinking。

此外，在Qwen3-VL系列中，无论是Qwen3-VL-30BA3B模型还是Qwen3-VL-32B模型，都始终在大多数评估指标上优于Gemini-2.5-Flash和GPT-5-mini，如表3所示。即使是紧凑的稠密模型——Qwen3-VL-8B、Qwen3-VL-4B和Qwen3-VL-2B——也在OCR解析、视觉问答（VQA）和综合基准测试套件上表现出极具竞争力的性能，如表4详细说明。这突显了Qwen3-VL架构在各种模型尺寸上的卓越效率和强大的可扩展性。

在此版本的 Qwen3-VL 中，我们特别强调增强其理解长文档的能力。如表 2 所示，在 MMLongBench-Doc 基准测试（Ma et al., 2024）上与旗舰模型的比较中，我们的 Qwen3-VL-235B-A22B 在 instruct/thinking 设置下实现了 57.0%/56.2% 的总体准确率，展示了其在长文档理解任务上的 SOTA 性能。

除了在已建立的基准测试中表现出色之外，我们在多语言支持方面也取得了长足的进步。这代表着从 Qwen2.5-VL 支持的 10 种非英语/中文语言到 Qwen3-VL 支持的 39 种语言的重大扩展。我们使用新构建的内部数据集评估了这种扩展的能力。如图 2 所示，该模型在 39 种测试语言中的 32 种语言上的准确率超过 70%——我们认为这是实际可用性的一个实用阈值。这表明 Qwen3-VL 强大的 OCR 功能不仅限于少数几种语言，而且扩展到广泛而多样的语言范围。

5.5 2D 和 3D 接地

在本节中，我们对 Qwen3-VL 系列在 2D 和 3D 接地相关基准上进行了全面评估，并将这些模型与具有类似功能的最新模型进行了比较。

我们评估了 Qwen3-VL 在指代表达式理解基准 RefCOCO/+/g (Kazemzadeh et al., 2014; Mao et al., 2016)、开放词汇目标检测基准 ODinW-13 (Li et al., 2022) 和计数基准 CountBench (Paiss et al., 2023) 上的 2D 定位能力。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: 小型 Qwen3-VL 模型和 GPT-5-nano 在视觉基准测试上的性能。

	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	-	-

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.,

在ODinW-13上，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。为了确保与传统的开放集目标检测专业模型具有可比性，我们在评估期间在提示中同时提供所有数据集类别。如表2所示，我们的旗舰模型Qwen3-VL-235B-A22B表现出色，并在2D grounding和计数基准测试中取得了最先进（SOTA）的结果。值得注意的是，它在ODinW-13上实现了48.6 mAP，展示了在多目标开放词汇目标grounding方面的强大性能。我们较小规模变体的详细结果分别在表3和表4中给出，它们在2D视觉grounding方面也表现出竞争优势。

此外，在这个版本的 Qwen3-VL 中，我们增强了其对 3D 对象定位的空间感知能力。我们使用平均精度均值 (mAP) 作为评估指标，在 Omni3D (Brazil et al., 2023) 上评估 Qwen3-VL 系列与其他同等规模的模型，Omni3D 是一个综合基准，包含 ARKitScenes (Baruch et al., 2021)、Hypersim (Roberts et al., 2021) 和 SUN RGB-D (Song et al., 2015) 等数据集。每个输入都是一个图像-文本对，由图像和指定对象类别的文本提示组成。为了确保与现有 VLM 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。具体来说，在 SUN RGB-D 数据集 (Song et al., 2015) 上，Qwen3-VL-235B-A22B-Thinking 变体的性能超过 Gemini-2.5-Pro 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 grounding 方面也表现出非常强的竞争力，详细结果分别在表 3 和表 4 中提供。

5.6 细粒度感知

我们通过三个流行的基准来衡量模型的细粒度感知能力。Qwen3-VL 系列在细粒度视觉理解方面相比其前身 Qwen2.5-VL-72B 实现了显著飞跃。值得注意的是，Qwen3-VL-235B-A22B 在所有三个基准上都实现了最先进的性能，并且在使用工具的情况下，在 V* (Wu &Xie, 2024) 上达到了 93.7，在 HRBench-4k (Wang et al., 2024e) 上达到了 85.3，在 HRBench-8k (Wang et al., 2024e) 上达到了 82.3。这种持续的优异表现突显了 Qwen3-VL 中引入的架构改进和训练策略的有效性，尤其是在处理高分辨率输入和对于细粒度感知任务至关重要的细微视觉差异方面。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超过了单纯增加模型尺寸带来的性能提升。例如，在 Qwen3-VL 系列中，添加工具带来的绝对改进在 V* 上始终 ∼ 5 个点。这些发现加强了我们的信念，即在多模态中扩展工具集成的代理学习是一条非常有希望的前进道路。

5.7 多图像理解

除了单图像的grounded dialogue评估之外，推进VLMs来处理多图像理解具有重要价值。这项任务需要在不同的视觉模式中进行更高级别的上下文分析，从而实现更高级的识别和推理能力。为此，我们使用全面的跨图像模式学习技术来增强Qwen3-VL，包括多图像指代grounding、视觉对应和多跳推理。我们在两个著名的多图像基准测试上评估了Qwen3-VL：BLINK (Fu et al., 2024c) 和 MuirBench (Wang et al., 2024a)。如表2所示，与其他领先的LVLM相比，Qwen3-VL在多图像理解方面表现出整体优势。具体来说，Qwen3-VL-235B-A22B-Instruct的性能与Gemini-2.5-pro等最先进的模型相当，而Qwen3-VL-235B-A22B-Thinking在MuirBench上获得了80.1的卓越领先分数，超过了所有其他模型。

5.8 具身与空间理解

为了具身和空间理解，Qwen3-VL 的性能通过一系列具有挑战性的基准，针对领先的 SOTA 模型进行了严格的基准测试：ERQA (Team et al., 2025)、VSIBench (Yang et al., 2025b)、EmbSpatial (Du et al., 2024)、RefSpatial (Zhou et al., 2025) 和 RoboSpatialHome (Song et al., 2025a)。在这些基准测试中，该模型展示了卓越的能力，可与 Gemini-2.5-Pro、GPT-5 和 Claude-Opus-4.1 等顶级模型的性能相媲美。这一成功主要归功于该模型深刻的空间理解能力，这源于其在高分辨率视觉数据上的训练，这些数据具有细粒度的指向、相对位置注释和 QA 对。其在 EmbSpatial、RefSpatial 和 RoboSpatialHome 上的出色结果清楚地验证了这一能力，其中 Qwen3-VL-235B-A22 分别获得了 84.3、69.9 和 73.9 的分数。此外，通过在训练期间集成指向、定位和时空感知数据，其具身智能得到了显著增强，从而在 ERQA (Team et al., 2025) 上获得了 52.5 的顶级分数，在 VSIBench (Yang et al., {v*}) 上获得了 60.0 的顶级分数。

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) 适用于 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 展示了大幅提升的视频理解能力。 特别是，交错式 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 变体能够达到与更大的 Qwen2.5-VL 72B 模型相媲美的性能。

我们对一系列不同的视频理解任务进行了全面评估，包括通用视频理解（VideoMME (Fu et al., 2024a), MVBench (Li et al., 2024b)），时间视频定位（Charades-STA (Gao et al., 2017)），视频推理（VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025)）和长视频理解（LVBench (Wang et al., 2024d), MLVU (Zhou et al., 2024)）。与最先进的专有模型（包括 Gemini 2.5 Pro、GPT-5 和 Claude Opus 4.1）相比，Qwen3-VL 表现出具有竞争力的性能，并且在某些情况下表现更优。 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（具有 128 的思考预算）和 GPT-5 minimal 等领先模型相当的性能。 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甚至超过了 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每个视频施加了 2,048 帧的上限，确保视频令牌的总数不超过 224K。VideoMMMU 和 MMVU 的每帧最大令牌数设置为 768，所有其他基准测试设置为 640。此外，Charades-STA 的视频以每秒 4 帧 (fps) 的速率采样，而所有其他基准测试使用 2 fps 的速率。对于 VideoMMMU，我们采用基于模型的评判进行评估，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，这些限制约束了评估期间使用的输入帧数：Gemini 2.5 Pro 为 512，GPT-5 为 256，Claude Opus 4.1 为 100。

5.10 智能体

我们使用 GUI 接地任务（ScreenSpot (Cheng et al., 2024), ScreenSpot Pro (Li et al., 2025b), OSWorldG(Xie et al., 2025a)）评估 UI 感知，并通过在线环境评估（AndroidWorld (Rawles et al., 2024), OSWorld (Xie et al., 2025c;b)）评估决策能力。对于 GUI 接地，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 上的交互界面，并展示了非常强大的 UI 感知能力。对于在线评估，Qwen3-VL 32B 在 OSWorld 上的得分为 41，在 AndroidWorld 上的得分为 63.7，超过了当前的基准 VLMs。Qwen3-VL 作为 GUI 代理展示了非常强大的规划、决策和反思能力。此外，较小的 Qwen3-VL 模型在这些基准测试中也表现出了极具竞争力的性能。

5.11 以文本为中心的任务

为了全面评估Qwen3-VL以文本为中心的能力，我们采用自动基准来评估模型在指令和思维模型上的性能。这些基准可以归纳为以下几个关键类型：（1）知识：MMLU-Pro (Wang et al., 2024f), MMLU-Redux (Gema et al., 2024), GPQA (Rein et al., 2023), SuperGPQA (Team, 2025), (2) 推理：AIME-25 (AIME, 2025), HMMT-25 (HMMT, 2025), LiveBench (2024-11-25) (White et al., 2024), (3) 代码：LiveCodeBench v6 (Jain et al., 2024), CFEval, OJBench (Wang et al., 2025c), (4) 对齐任务：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) 代理：BFCL-v3 (Patil et al., 2024), TAU2-Retail, TAU2-Airline, TAU2-Telecom, (6) 多语言：MultiIF (He et al., 2024), MMLU-ProX, INCLUDE (Romanou et al., 2025), PolyMATH (Wang et al., 2025b).

对于 Qwen3-VL instruct 模型，包括 235B-A22B、32B 和 30B-A3B，我们配置采样超参数为：temperature = 0.7，top-p = 0.8，top-k = 20，以及 presence penalty = 1.5。对于小型 instruct 模型，包括 8B、4B 和 2B，我们设置 temperature = 1.0，top-p = 1.0，top-k = 40，以及 presence penalty = 2.0。我们将最大输出长度设置为 32,768 个 tokens。

对于采用混合专家 (MoE) 架构的 Qwen-VL 思维模型，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。对于密集思维模型，我们将温度 = 设置为 1.0，top-p 设置为 {v*}。

[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.

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：Qwen-VL-235B-A22B (Instruct) 与其他基线模型的比较。最高分和第二高分分别以粗体和下划线显示。

	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 = 设为 20，并额外应用 1.5 的存在惩罚，以鼓励更大的输出多样性。我们将最大输出长度设置为 32,768 个 tokens，除了 AIME-25、HMMT-25 和 LiveCodeBench v6，在这些情况下，我们将长度扩展到 81,920 个 tokens，以提供足够的思考空间。

详细结果如下。

Qwen3-VL-235B-A22B 我们将我们的旗舰模型 Qwen3-VL-235B-A22B 与领先的指令和思考模型进行了比较。对于 Qwen3-VL-235B-A22B-Instruct，我们以 Qwen3-235B-A22B-Instruct-2507、DeepSeek V3 0324 和 Claude-Opus-4（无思考）作为基线。对于 Qwen3-VL-235B-A22B-Thinking，我们以 Qwen3-235B-A22B-Thinking-2507、OpenAI o3 (medium)、Claude-Opus-4（有思考）作为基线。我们在表 5 和表 6 中展示了评估结果。

- 从表5可以看出，Qwen3-VL-235B-A22B-Instruct取得了具有竞争力的结果，与其他领先模型相比，具有可比性甚至超越了它们，包括DeepSeek V3 0324、Claude-Opus-4（不进行思考）以及我们之前的旗舰模型Qwen3-235B-A22B-Instruct-2507。 特别是，Qwen3-VL-235B-A22B-Instruct在推理需求型任务（例如，数学和编码）上超过了其他模型。 值得注意的是，DeepSeek V3 0324和Qwen3-235B-A22B-Instruct-2507是大型语言模型，而Qwen3-VL-235B-A22B-Instruct是视觉语言模型，可以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集成。
- 从表 6 可以看出，Qwen3-VL-235B-A22B-Thinking 与其他领先的思考模型相比，也取得了具有竞争力的结果。Qwen3-VL-235B-A22B-Thinking 在 AIME-25 和 LiveCodeBench {v*} 上超过了 OpenAI o3 (medium) 和 Claude-Opus-4 (with thinking)，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 Qwen3-VL-32B 和 Qwen3-VL-30B-A3B 模型与其对应的纯文本模型进行比较，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 可以看出，对于 instruct 模型，Qwen3-VL-32B 和 Qwen3-VL-30B-A3B 在所有基准测试中都显示出比 Qwen3-32B 和 Qwen3-30B-A3B 显著的性能提升。Qwen3-VL-30B-A3B 取得了与 Qwen3-30B-A3B-2507 相当甚至更好的结果，尤其是在 AIME-25 和 HMMT-25 上。
- 从表 8 可以看出，对于思维模型，Qwen3-VL-32B 和 Qwen3-VL-30B-A3B 在大多数基准测试中都超过了基线模型。Qwen3-VL-30B-A3B 也表现出与 Qwen3-30B-A3B-2507 相当的性能。

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: Qwen-VL-235B-A22B (Thinking) 与其他推理基线的比较。最高分和第二高分分别以粗体和下划线显示。

	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	-

表 7：Qwen3-VL-32B-Instruct、Qwen3-VL-30B-A3B-Instruct 以及相应基线之间的比较。

	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 我们在表 9 和表 10 中展示了 Qwen3-VL-2B、Qwen3-VL-4B 和 Qwen3-VL-8B 的评估结果。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，这些边缘侧模型表现出令人印象深刻的性能，并且优于基线。这些结果表明

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）以及相应基线之间的比较。

	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

表 9：Qwen3-VL-2B（Instruct）、Qwen3-VL-4B（Instruct）、Qwen3-VL-8B（Instruct）以及相应基线模型之间的比较。

	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

我们强大的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。{v*}

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：Qwen3-VL-2B（思考）、Qwen3-VL-4B（思考）、Qwen3-VL-8B（思考）以及相应基线之间的比较。

	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 消融研究

5. 12.1 视觉编码器

我们进行了与原始 SigLIP-2 的对比实验。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竞争性的性能，同时在 OmniBench 上取得了显著的提升。OmniBench 是我们内部的整体评估套件，旨在评估在各种具有挑战性的条件下对世界知识的整合能力。此外，当与相同的 1.7B Qwen3 语言模型集成并训练 1.5T tokens 后，Qwen3-ViT 在多个关键任务上始终优于基于 SigLIP-2 的基线，并且在 OmniBench 上仍然显著领先，证明了其作为更强大的视觉骨干的优越性和有效性。

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，并进一步评估了它们在视觉-语言建模 (VLM) 阶段与相同的 1.7B Qwen3 语言模型配对时的下游性能。

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

我们进行了一项消融研究，以验证 DeepStack 机制的有效性。如表 12 所示，配备 DeepStack 的模型在各种基准测试中都取得了整体性能提升，有力地证实了其有效性。这种提升归功于 DeepStack 整合丰富视觉信息的能力，从而有效地提升了细粒度视觉理解能力，例如在 InfoVQA 和 DocVQA 基准测试中。

表 12：DeepStack 上的消融实验。我们使用内部的 15B-A2B LLM 在 DeepStack 上进行消融研究，所有实验均在 2000 亿个 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，无需任何后训练。

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

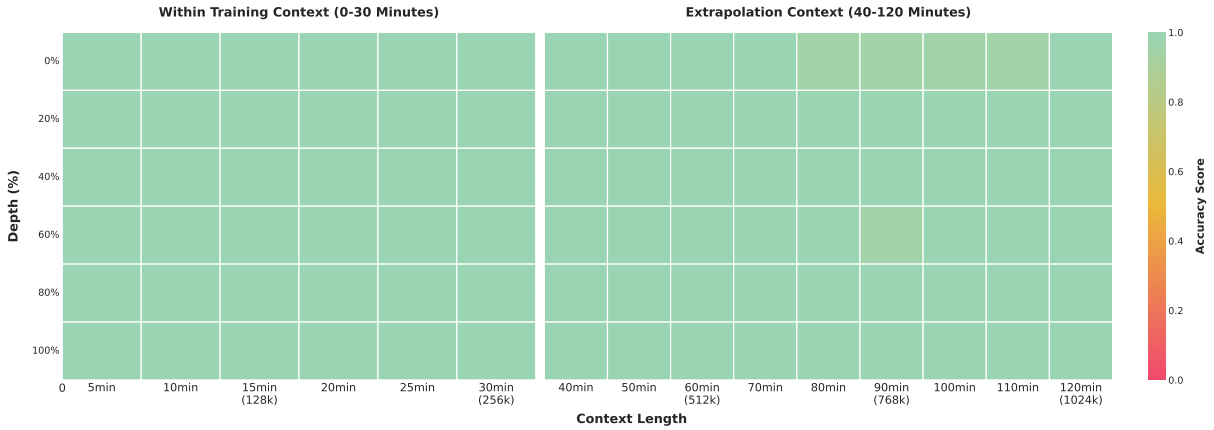


Figure 3: Needle-in-a-Haystack performance heatmap for Qwen3-VL-235B-A22B-Instruct across varying video durations and needle positions. Each cell shows accuracy (%) for locating and answering questions about the inserted "needle" frame.

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.

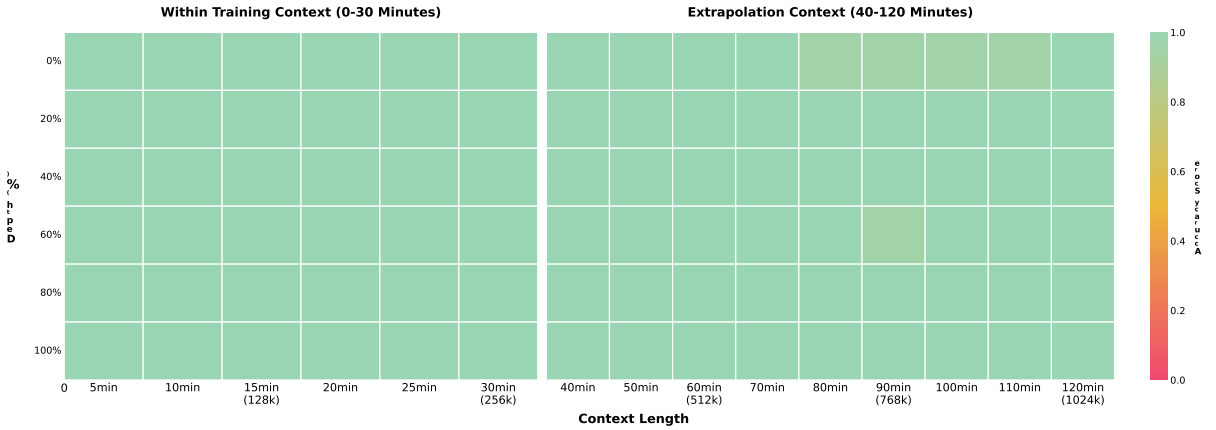


图 3：Qwen3-VL-235B-A22B-Instruct 在不同视频时长和 needle 位置下的 Needle-in-a-Haystack 性能热图。每个单元格显示定位并回答关于插入的“needle”帧的问题的准确率（%）。{v*}

5.12.3 针里寻干草堆

为了评估模型处理长上下文输入的能力，我们对 Qwen3-VL-235B-A22B-Instruct 构建了一个视频“大海捞针”评估。在这个任务中，一个语义上显著的“针”帧——包含关键的视觉证据——被插入到长视频中不同的时间位置。然后，模型需要从长视频中准确地定位目标帧，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，并且帧分辨率被动态调整以保持恒定的视觉 token 预算 {v*}。

如图 3 所示，该模型在时长不超过 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 的位置扩展外推到最多 1M 个 token 的序列（大约 2 小时的视频），该模型仍保持 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

在本文中，我们提出了 Qwen3-VL，这是一个最先进的视觉-语言基础模型系列，它推进了多模态理解和生成的前沿。通过整合高质量的多模态数据迭代和架构创新——例如增强的交错式 MRoPE、DeepStack 视觉-语言对齐和基于文本的时间定位——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。它对 256K-token 交错序列的原生支持能够对长而复杂的文档、图像序列和视频进行稳健的推理，使其非常适合需要高保真跨模态理解的实际应用。密集型和混合专家型变体的可用性确保了跨不同延迟和质量要求的灵活部署，以及我们的后训练策略——包括非思考和思考模式 {v*}。

展望未来，我们设想 Qwen3-VL 成为具身 AI 代理的基础引擎，能够无缝连接数字世界和物理世界。这些代理不仅能够感知和推理丰富的多模态输入，还能够在动态环境中执行果断的、具有上下文意识的动作——与用户互动、操纵数字界面，并通过基于现实的多模态决策来引导机器人系统。未来的工作将侧重于扩展 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，最终目标是使 AI 系统能够在虚拟和物理领域中与人类一起学习、适应和协作。此外，我们正在积极探索统一的理解-生成架构，利用视觉生成能力来进一步提升整体智能。通过在 Apache 2.0 许可下公开整个模型系列，我们旨在促进社区驱动的创新，以实现真正集成、多模态 AI 代理的愿景。

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, Daijun Yu, 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, and Hangdi Xing.

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 的所有贡献者均按姓氏的字母顺序排列。

核心贡献者：白帅，蔡宇轩，陈睿哲，陈科勤，陈雄辉，程泽森，邓良浩，丁伟，高峰，葛春江，葛文斌，郭志芳，黄启东，黄杰，黄飞，惠彬元，姜舒桐，李朝海，李名声，李梅，李凯欣，林子诚，林俊扬，刘雪静，刘家玮，刘成龙，刘洋，刘大一恒，刘世轩，卢敦杰，罗瑞林，吕晨旭，门睿，孟令晨，任轩诚，任星璋，宋思博，孙宇冲，唐骏，涂剑洪，万建成，王鹏，王鹏飞，王秋月，王宇轩，谢天宝，徐一恒，徐海洋，徐进，杨志博，杨明坤，杨建新，杨安，于博文，张飞，张航，张玺，郑博，钟虎门，周靖人，周帆，周静，朱元治，朱科

贡献者：曹益中、陈蓓、程晨、褚云飞、崔泽宇、党凯、邓晓东、范洋、房荣耀、关同坤、贺金政、洪淼、蒋松涛、李政、李晓川、林俊荣、刘玉琼、刘彦涛、倪娜、牛欣瑶、庞雅恬、邱子涵、沈天浩、唐天翼、万宇、魏金熙、吴晨飞、吴步霄、徐骁、薛明峰、闫明、杨雨寰、杨嘉熙、杨可欣、于乐、于浩、于代军、张建伟、张一弛、张振儒、张思琪、张培扬、张北辰、赵洪博、庄贤伟

致谢：我们衷心感谢陈祖龙、邓冰、高峰宇、姜冠君、刘岳和邢航地领导的团队给予的坚定支持。

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, 等. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

AIME。AIME 问题与解答，2025。网址：https://artofproblemsolving.com/wiki/index.php/{v*}AIMEProblemsandSolutions。

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

白帅, 陈克勤, 刘雪静, 王佳林, 葛文斌, 宋思博, 党凯, 王鹏, 王世杰, 唐军, 钟虎门, 朱元治, 杨明坤, 李朝海, 万建成, 王鹏飞, 丁伟, 付哲人, 徐一恒, 叶嘉博, 张熙, 谢天宝, 程泽森, 张航, 杨志博, 徐海阳, 林俊旸. 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: 一个多样化的真实世界数据集，用于使用移动 RGB-D 数据进行 3D 室内场景理解。*arXiv preprint arXiv:2111.08897*, 2021.

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, 和 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.

林晨, 李劲松, 董晓艺, 张盼, 臧宇航, 陈泽辉, 段昊东, 王佳琪, 乔宇, 林达华, 等. 我们在评估大型视觉语言模型的方法上是否正确? *arXiv:2403.20330*, 2024a.

陈世敏, 蓝晓涵, 袁艺天, 解泽群, 和 马琳. Timemarker: 一种通用的视频-LLM，具有卓越的时间定位能力，适用于长视频和短视频理解. *arXiv preprint arXiv:2411.18211*, 2024b.

陈一通, 孟令琛, 彭武剑, 吴祖煊, 和蒋育纲. Comp: 用于视觉基础模型的持续多模态预训练. *arXiv preprint arXiv:2503.18931*, 2025.

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

程显福, 张伟, 张世伟, 杨健, 关向远, 吴先杰, 李祥, 张戈, 刘家恒, 麦玉莹, 等. Simplevqa: 多模态大型语言模型的多模态事实性评估. 见 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 页码 4637–4646, 2025.

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, Dan Zhang, Evan Rosen 等. Gemini 2.5: 通过高级推理、多模态、长上下文和下一代 Agentic 能力推动前沿. *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 等. Molmo and pixmo: 用于最先进多模态模型的开放权重和开放数据。*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 等. Climb: 基于聚类的迭代数据混合引导语言模型预训练。*arXiv preprint arXiv:2504.13161*, 2025.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, 以及 Hervé Jégou. Faiss 库. 2024.

杜梦菲, 吴斌豪, 李泽君, 黄萱菁, 魏忠钰. Embspatial-bench: 使用大型视觉语言模型对具身任务的空间理解能力进行基准测试. *arXiv preprint arXiv:2406.05756*, 2024.

段承祺, 孙凯悦, 方荣耀, 张满园, 冯岩, 罗颖, 刘玉芳, 王珂, 裴鹏, 蔡迅亮, 等. Codeplot-cot: 通过代码驱动图像进行数学视觉推理. *arXiv preprint arXiv:2510.11718*, 2025.

傅超友，戴宇涵，罗涌东，李磊，任书怀，张任锐，王梓涵，周晨宇，沈云航，张梦丹，等. Video-mme: 首个多模态llm在视频分析中的综合评估基准. *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, 和 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, 和 Ranjay Krishna. Blink: 多模态大型语言模型能看但不能感知. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024c.

常高, 郑楚杰, 陈雄辉, 邓凯, 刘世轩, 于 Bowen, 杨安, 白 帅, 周靖人, 和 林俊旸. 软自适应策略优化. *arXiv preprint arXiv:2511.20347*, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, 和 Ram Nevatia. Tall: 通过语言查询进行时间活动定位. 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, 和 Pasquale Minervini. 我们完成mmlu了吗？*CoRR*, abs/2406.04127, 2024. doi: 10.48550/ARXIV.2406.04127. URL https: //doi.org/10.48550/arXiv.2406.04127.

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., & Zhou, T. Hallusionbench: 用于大型视觉语言模型中纠缠的语言幻觉和视觉错觉的高级诊断套件, 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。

胡凯瑞, 吴鹏昊, 蒲凡益, 肖望, 张远瀚, 岳翔, 李勃, 刘自慰. Video-mmmu: 评估从多学科专业视频中获取知识. *arXiv preprint 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.

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 Transac- tions 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.

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.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, J un-Juan Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detect ion. *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 mod el 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 1869-191 9. doi: 10.1007/ s11432-024-4235-6. URL http://dx.doi.org/10.1007/s11432-024-4235-6. Dunjie Lu, Yiheng Xu, J unli Wang, Haoyuan Wu, Xinyuan Wang, Zekun Wang, Junlin Yang, Hongjin Su, Jixuan Chen, Junda Chen, Yuc hen 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, C hunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Eval uating 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, Xiao yi 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, Ale xander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguo us object descriptions. In *CVPR*, 2016. Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Ho que. 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. Mi nesh 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. D eepstack: 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 Zha ng, 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 d iverse pdf document parsing with comprehensive annotations, 2024. URL https://arxiv.org/abs/2412.07626. Samu el J. Paech. Eq-bench: An emotional intelligence benchmark for large language models. *CoRR*, abs/2312.06281, 2 023. doi: 10.48550/ARXIV.2312.06281. URL https://doi.org/10.48550/arXiv.2312. 06281. Roni Paiss, Ariel Ephr at, 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 ber keley 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, P eter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimoda l 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 m ultimodal model achieve human-like mathematical reasoning? *arXiv preprint 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. Vision language mod els 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 bench- marking 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, 和 Samuel R. Bowman. GPQA: A graduate-level google-pr oof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48 550/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 bench mark for contemporary large multimodal models, 2025. URL https://arxiv.org/abs/2502.09696. Mike Roberts, Jas on Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, 和 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 Mahe shwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, et al. INCLUDE: evaluating multilingual languag e understanding with regional knowledge. In *The Thirteenth Interna-tional Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 202 5. Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, 和 Jian Sun. Objects3 65: 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, Zhengy uan Yang, Ruibo Liu, 和 Diyi Yang. Design2code: Bench- marking multimodal code generation for automated fro nt-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 T remblay, Stephen Tyree, Yu Su, 和 Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d visi on-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15768–15780, 2025a. Shuran Song, Samuel P Lichtenberg, 和 Jianxiong Xiao. Sun rgb-d: A rgb-d scene understa nding 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, 和 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-Baptist e Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Mi chiel 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 semanti c 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, 和 Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 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, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, W enbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zho u, 和 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 G u, 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, 和 Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image percept ion 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, Che n 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 mathema tical 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 Chan dra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, et al. MMLU-Pro: A more rob ust 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, 和 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 W ang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhi ka Malladi, Alexis Chevalier, Sanjeev Arora, 和 Danqi Chen. Charxiv: Charting gaps in realistic chart understandi ng in multimodal llms. *arXiv preprint arXiv:2406.18521*, 2024g. Alexander Wettig, Kyle Lo, Sewon Min, Hanna neh Hajishirzi, Danqi Chen, 和 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, Benja min Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, et al. Livebench: A challenging, contamination-free LL M benchmark. *CoRR*, abs/2406.19314, 2024. doi: 10.48550/ARXIV.2406.19314. URL https://doi.org/10.48550/ar Xiv.2406. 19314. Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, 和 Ziwei Liu. Mmsea rch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025a. Penghao Wu 和 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– 13 094, June 2024. Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhan g, Mengyue Wu, Qin Jin, 和 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.2503.05 244. xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/ datasets/xai-or g/RealworldQA, 2024. Accessed: 2025-04-26. Yijia Xiao, Edward Sun, Tianyu Liu, 和 Wei Wang. Logicvista: M ultimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024. Tianbao X ie, Jiaqi Deng, Xiaochuan Li, Junlin Yang, Haoyuan Wu, Jixuan Chen, Wenjing Hu, Xinyuan Wang, Yuhui Xu, Z ekun Wang, Yiheng Xu, Junli Wang, Doyen Sahoo, Tao Yu, 和 Caiming Xiong. Scaling computer-use grounding via user interface decomposition and synthesis, 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, 熊新壮, 沈振楠, 周子龙, 王新元, 陈彦旭, 邓嘉琪, 陈俊达, 王博文, 吴昊远, 陈继轩, 王君力, 卢敦杰, 胡昊, 和于涛. Introducing osworld-verified. *xlang.ai*, 2025b年7月. URL https: //xlang.ai/blog/osworld-verified. 谢天宝, 张丹阳, 陈继轩, 李晓川, 赵思衡, 曹瑞盛, 等. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37 :52040–52094, 2025c. 徐伟业, 王嘉豪, 王维云, 陈哲, 周文刚, 杨爱军, 卢乐伟, 李厚强, 王晓华, 朱犀洲, 等. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models, 2025. URL https://arxiv.org/abs/2504.15279. 杨安, 李安丰, 杨宝松, 张北辰, 惠彬元, 郑博, 于博文, 等. Qwen3 technical report, 2025a. 杨成, 史楚凡, 刘亚欣, 水波, 王俊杰, 景默涵, 徐麟然, 朱新宇, 李思衡, 张宇翔, 等. Chartmimic: Evaluating lmm’s cross-modal reasoning capability via chart-to-code generation. *arXiv preprint arXiv:2406.09961*, 2024a. 杨继涵, 杨书盛, Anjali W Gupta, Rilyn Han, 李飞飞, 和谢赛宁. 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. 杨志博, 唐俊, 李朝海, 王鹏飞, 万建强, 钟虎门, 刘雪静, 杨明坤, 王鹏, 白帅, 靳连文, 和林俊旸. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy, 2024b. URL https://arxiv.org/ abs/2412.02210. 叶嘉博, 张熙, 徐海洋, 刘浩伟, 王俊扬, 朱兆庆, 郑子威, 等. Mobile- agent-v3: Fundamental agents for gui automation. *arXiv preprint arXiv:2508.15144*, 2025. 岳翔, 倪远晟, 张凯, 郑天宇, 刘若琪, 张戈, 等. 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. 岳翔, 郑天宇, 倪远晟, 王宇博, 张凯, 童盛邦, 孙宇轩, 于博涛, 张戈, 孙焕, 等. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024b. 张任瑞, 姜东志, 张益驰, 林浩坤, 郭子瑜, 邱鹏硕, 周傲君, 卢盼, 常凯伟, 乔宇, 等. 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. 赵一伦, 谢陆静, 张浩伟, 甘果, 龙亦韬, 胡志远, 胡桐彦, 陈伟源, 李楚涵, 宋俊扬, 徐智健, 王承业, 等. Mmvu: Measuring expert-level multi-discipline video understanding, 2025. URL https://arxiv.org/abs/2501.12380. 郑子威, Michael Yang, Jack Hong, 赵晨骁, 徐国海, 杨乐, 沈超, 和于兴. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025. 周恩生, 安景坤, 迟骋, 韩毅, 荣山雨, 张驰, 王鹏伟, 王仲远, 黄铁军, 盛律, 等. Roborefer: Towards spatial referring with reasoning in vision- language models for robotics. *arXiv preprint arXiv:2506.04308*, 2025. Jeffrey Zhou, 卢天健, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, 和 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. 周俊杰, 舒炎, 赵博, 吴博雅, 肖诗涛, 杨曦, 熊永平, 张波, 黄铁军, 和刘铮. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 朱婉蓉, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, 和 Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text.
Advances in Neural Information Processing Systems, 36:8958–8974, 2023. 邹承珂, 郭新刚, 杨睿, 张俊宇, 胡斌, 和张欢. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. *arXiv preprint arXiv:2411.00836*, 2024.

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).

A. 基准测试

我们评估了 Qwen3-VL 在各种公共基准上的表现，这些基准涵盖了不同的能力：多模态推理、通用视觉问答、主观体验与指令遵循、文档理解（包括 OCR）、2D/3D 视觉定位和计数、空间推理、视频理解、GUI 代理以及以文本为中心的任务。下面，我们提供了所有使用基准的详细列表。

- 多模态推理：我们在 12 个基准上评估模型，这些基准涵盖了广泛的领域——从数学和 STEM 到视觉推理和解谜任务：MMMU (Yue et al., 2024a), MMMU-Pro (Yue et al., 2024b), MathVision (Wang et al., 2024 b), MathVision-Wild$_{photo}$, 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), VisualPuzzle s (Song et al., 2025b), VLM are Blind (Rahmanzadehgervi et al., 2025), ZeroBench (Main/Subtasks) (Roberts e t al., 2025), 以及 VisuLogic (Xu et al., 2025)。
- 通用视觉问答：我们在 4 个通用 VQA 基准上评估模型：MMBench-V1.1 (Liu et al., 2023b)、RealWorld QA (xAI, 2024)、MMStar (Chen et al., 2024a) 和 Sim- pleVQA Cheng et al. (2025)。
- 主观体验和指令遵循：我们在 3 个基准上评估模型，涵盖主观体验和复杂指令遵循：HallusionBench (G uan et al., 2023)、MM-MT-Bench (Agrawal et al., 2024) 和 MIA-Bench (Qian et al., 2024)。
- 文档理解：我们对 Qwen3-VL 系列在各种 OCR 相关基准测试中的 OCR 和文档理解能力进行了全面评估，这些基准测试包括：DocVQA (Mathew et al., 2021b)、InfoVQA (Mathew et al., 2021a)、AI2D (Kemb havi et al., 2016)、ChartQA (Masry et al., 2022)、OCRBench (Liu et al., 2024)、OCRBench_v2 (Fu et al., 20 24b)、CC-OCR (Yang et al., 2024b)、OmniDocBench (Ouyang et al., 2024)、CharXiv (Wang et al., 2024g) 和 MMLongBench-Doc (Ma et al., 2024)。
- 2D/3D 接地和空间理解：我们在 11 个基准上评估模型，包括 2D 接地、3D 接地和空间理解：RefCOC O/+/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) 和 RoboSpatialHome (Song et al., 2025a)。
- 视频理解：我们使用七个基准来评估模型的视频理解能力：VideoMME (Fu et al., 2024a), MVBench (Li e t al., 2024b), VideoMMMU (Hu et al., 2025), MMVU (Zhao et al., 2025), LVBench (Wang et al., 2024d), ML VU (Zhou et al., 2024), Charades-STA (Gao et al., 2017)。
- 代码能力：我们评估模型的多模态代码能力，特别是在前端重建和SVG生成方面，使用Design2Code (S i et al., 2025)、ChartMimic (Yang et al., 2024a)和UniSVG (Li et al., 2025a)基准测试。
- GUI 代理：我们使用测试感知和决策能力的基准来评估 GUI 代理的能力。对于感知，我们使用 Screen Spot (Cheng et al., 2024)、ScreenSpot Pro (Li et al., 2025b) 和 OSWorldG (Xie et al., 2025a) 来衡量 GUI 接地和跨设备界面布局的理解。对于决策，我们使用 AndroidWorld (Rawles et al., 2024) 和 OSWorld (Xie e t al., 2025c;b) 来评估真实或模拟操作环境中的交互控制、规划和执行。
- 以文本为中心的任务：我们在各种以文本为中心的数据集上评估模型。(1) 知识：MMLU-Pro (Wang et al., 2024f), MMLU-Redux (Gema et al., 2024), GPQA (Rein et al., 2023), SuperGPQA (Team, 2025), (2) 推理：AIME-25 (AIME, 2025), HMMT-25 (HMMT, 2025), LiveBench (2024-11-25) (White et al., 2024), (3) 代码：LiveCodeBench v6 (Jain et al., 2024), CFEval, OJBench (Wang et al., 2025c), (4) 对齐任务：IFEval (Zhou et al., 2023), Arena-Hard v2 (Li et al., 2024d) , Creative Writing v3 (Paech, 2023), WritingBench (Wu et al., 20 25b), (5) 代理：BFCL-v3 (Patil et al., 2024), TAU2-Retail, TAU2-Airline, TAU2-Telecom, (6) 多语言：Mult iIF (He et al., 2024), MMLU-ProX, INCLUDE (Romanou et al., 2025), PolyMATH (Wang et al., 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

为了确保可重复性并方便未来的研究，我们在此提供用于评估我们模型在所有基准测试中的完整提示集。在推理过程中，这些提示被一致地应用，以保持公平性和可比性。

B.1 STEM 与益智游戏

<图像> 问题：{question} 选项：{options} 请从以上选项中选择正确答案。

<图像> {question} {options} 请从选项中选择正确答案。

<图像> 识别问题并解决它。在回答之前，请逐步思考。

<图像> {question}

<图像> 现在，我们需要您解决一道多项选择数学题。请简要描述您的思考过程并提供最终答案（选项）。问题：{question} 选项：{options} 关于格式，请按照以下模板回答，并务必包含两个<>符号：<思考过程>：「您的思考过程」 <答案>：「您的选项」

<图像> {question} 让我们一步一步地思考，并用花括号给出最终答案，像这样：{final answer}

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}
```

<image> ## 问题 {question} ## 回答说明：请提供上述问题的答案。您的回答应遵循以下 JSON 格式，其中包括两个键：“solution”和“short answer”。“solution”键可以包含解决问题所需的详细步骤，“short answer”键应提供简洁的回答。预期 JSON 响应格式示例：{"solution": "[详细的分步解释]", "short answer": "[简洁的答案]" }

<图像> 问题：{question
}

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

<图像> 问题：{question} 选项：{options} 直接用选项的字母回答问题。

<图像> 问题：{question} 选项：{options} 解答选择题，然后从给定的选项中选择一个字母作答。你回复的最后一行应遵循以下格式：'Answer: $LETTER'（不带引号），其中 LETTER 是选项之一。在回答之前，请逐步思考。

B.2 GeneralVQA

<图像> 问题：{question} 选项：{options} 请从以上选项中选择正
确答案。

<图像> {qu
estion}

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

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

B.4 文档理解

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

```
<图像> {question} 用一个词或短语回答问题。
```

```
<图像> 问题：{question} 选项：{options} 请从以上选项中选择正
确答案。
```

```
<图像> {qu
estion}
```

```
<图像>
你是一个 AI 助手，专门用于将 PDF 图像转换为 Markdown 格式。请按照以下说明进行转换：

1. 文本处理：- 准确识别 PDF 图像中的所有文本内容，不进行猜测或推断。- 将识别的文本转换为 Ma
rkdown 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。 - 使用 \( \) 括起行内公式。例如：这是一个
行内公式 \( E = mc2 \) - 使用 \[ \] 括起块公式。例如：\[ \frac{-b \pm \sqrt{b2 - 4ac}}{2a} \]


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

4. 图片处理：- 忽略 PDF 图像中的图片。不要尝试描述或转换图像。

5. 输出格式： - 确保输出的 Markdown 文档具有清晰的结构，元素之间有适当的换行。 - 对于复杂的
布局，尽量保持原始文档的结构和格式。

请严格遵守这些指南，以确保转换的准确性和一致性。你的任务是准确地将PDF图像的内容转换为Mar
kdown格式，不添加任何额外的解释或注释。
```

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 Grounding

<图像>
在图像中找到所有符合描述“{ref_sentence}”的对象。以 JSON 格式报告 bbox 坐标。

<图像> 问题：{question} 选项：{options} 请从以上选项中选择正
确答案。

<图像>
找出所有属于以下类别的实例：{obj_names}。以 JSON 格式报告 bbox 坐标。

<图像>
在提供的图像中定位 {class_name}，并使用 3D 边界框输出它们的位置和尺寸。结果必须是 JSON 格式：["bbox_3d":[x_center, y_center, z_size, x_size, roll, pitch, yaw],"label":"category"]。

B.6 具身/空间理解

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

多项选择：
<视频> 这些是视频的帧。{question} 选项：{options} 直接用给定选项的字
母回答。

开放式的：
<视频> 这些是视频的帧。{question} 请用一个词或短语回答问题
。

<图像> {qu
estion}

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.
```

<图像>，在此图像中定位{object_name}。以JSON格式输出点坐标。例如：[{"point_2d": [x, y
], "label": "point_1"}]

<image> {question} 以 JSON 格式输出点坐标。例如：[{"point_2
d": [x, y], "label": "point_1"}]

B.7 多图像

<图像> 问题：{question} 选项：{options} 请从以上选项中选择正
确答案。

<image_1>text_1>image_2>text_2> ... <image_n>text_n> 直接用给定选项的
字母回答。 < <

B.8 视频理解

<视频> 根据视频选择以下多项选择题的最佳答案。仅以正确选项的字母（A、B、C 或 D）作答
。问题：{question} 可能的答案选项：{options} 最佳答案是：

<视频> 根据视频选择以下多项选择题的最佳答案。仅以正确选项的字母（A、B、C 或 D）作答。问题
：{question} {options} 请逐步推理，识别相关的视觉内容，分析关键的时间戳和线索，然后提供最终答
案。

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.
```

<视频>
给你一段文本查询：{query_text} 视频中描述的内容在什么
时间发生？请以秒为单位返回时间戳。

感知与理解:
<视频> {question} {options} 请忽略视频最后一帧中的测验问题。

适应性多项选择:
<视频>图像> 你应该观看并学习视频内容。然后运用你所学的知识来回答以下多项选择题。此问题的<图像在视频末尾。{question} {options}

适应性-开放性:
<视频>图像> 你应该观看并学习视频内容。然后运用你所学到的知识来回答以下开放式问题。此问<题的图像在视频末尾。{question}

多项选择:
<视频> {question} {options} 视觉信息：已处理的视频 逐步回答给定的多项选择题。首先，清晰地解释你的推理过程。最后，使用以下格式陈述最终答案：“Therefore, the final answer is: $LETTER”（不带引号），其中 $LETTER 是选项之一。在回答之前，请逐步思考。

开放式的:
<视频> {question} 视觉信息：处理后的视频 逐步回答给定的问题。首先清楚地解释你的推理过程。最后，使用以下格式陈述最终答案：“Therefore, the final answer is: "Answer: $ANSWER" (不带引号)，其中 $ANSWER 是问题的最终答案。在回答之前，请一步一步地思考。

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 使用工具的感知

```
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}
```

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:

```
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 编码

```
<图像>
你是一名精通 HTML 和 CSS 的专业 Web 开发者。用户会向你提供一个网页的截图。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。如果涉及到任何图像，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，也请使用 "rick.jpg"。不要虚构任何对外部文件的依赖。你不需要包含用于动态交互的 JavaScript 脚本。请注意所有元素的大小、文本、位置和颜色，以及整体布局。请回复 HTML+CSS 文件的内容:
```

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>

我将给你两张图片。第一张是参考图，第二张是通过代码渲染从第一张生成的。请评价它们的相似度，范围从 0 到 100，其中 0 表示完全不同，100 表示完全相同。请在 LaTeX 中提供分数，并简要解释你的理由。 □

<参考图像>生成图像>

<

B.11 代理

工具

你可以调用一个或多个函数来协助处理用户查询。你在 <tools> . . . </tools> XML 标签中获得了函数签名：<tools> { "name":"computer_use", "description": "使用鼠标与计算机交互。屏幕分辨率为 <display_width_px>x <display_height_px>。" "notes": "点击时，光标尖端应位于目标中心；避免边缘，除非另有要求。不要使用其他工具（type、key、scroll、left_click_drag）。只允许 left_click 和 mouse_move。如果找不到元素，终止并报告失败。", "parameters":{ "type":"object", "required":["action"], "properties":{ "action":{ "type":"string", "enum":["mouse_move","left_click"], "description":"要执行的动作。" }, "coordinate":{ "type":"array", "description":"(x, y)：从左/上角的像素。action=mouse_move 和 action=left_click 需要。" } } } }</tools> 对于每个函数调用，返回一个 JSON 对象，其中包含函数名称和参数，位于 <tool_call> . . . </tool_call> XML 标签中：<tool_call> {{"name": <function-name>, "arguments": <args-json-object>}} </tool_call> 此外，如果你认为任务不可行（例如，任务与图像无关），则返回：<tool_call> {"name": "computer_use", "arguments": {"action": "terminate", "status": "failure"}} </tool_call>