





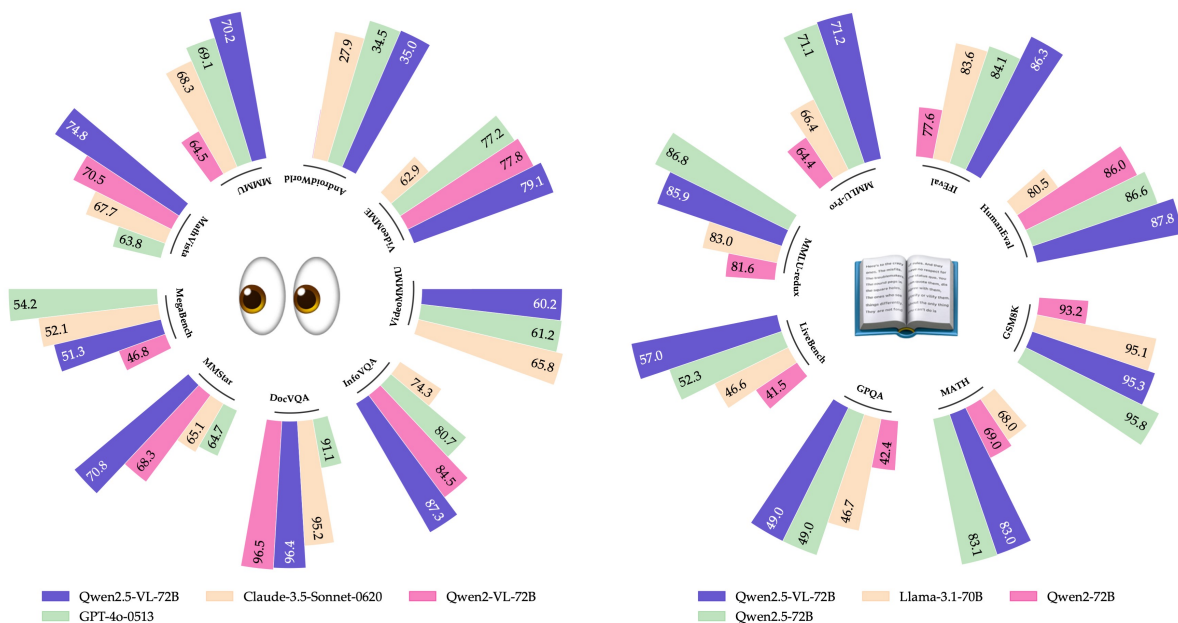
Qwen2.5-VL Technical Report

Qwen Team, Alibaba Group

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

Abstract

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



1 Introduction

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

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

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

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

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

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