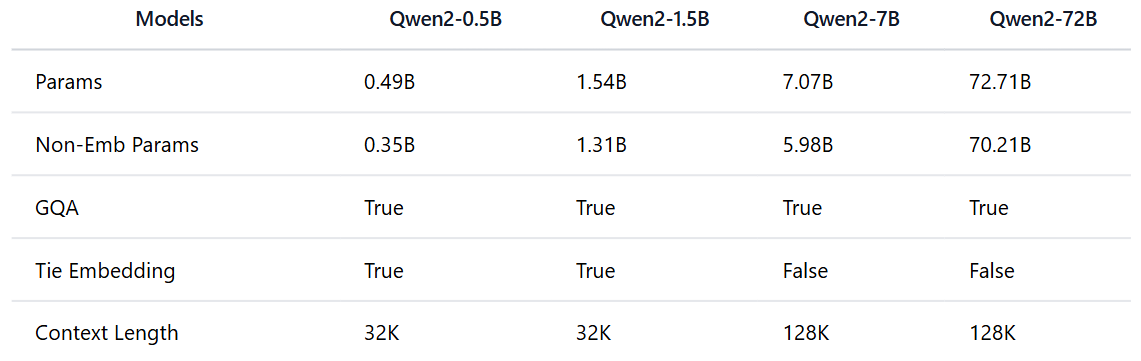
**Qwen2-VL-7B-Instruct**

**Introduction**

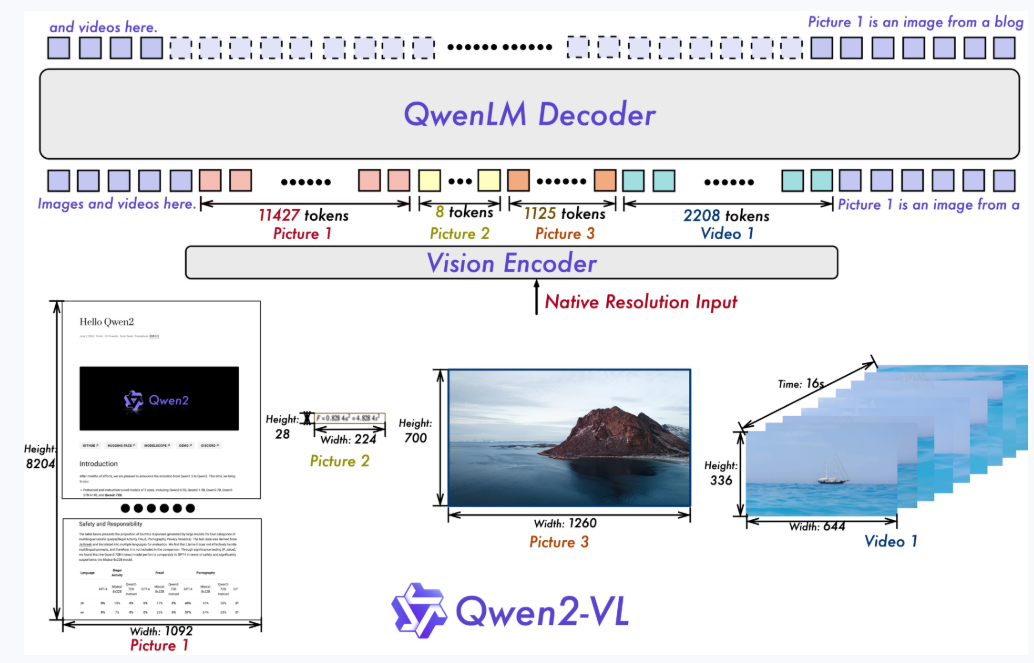
* Qwen2 is a new series of large language models from Alibaba group
* Qwen2 is trained on data in 29 languages, including English and Chinese.
* It is available in 4 parameter sizes: 0.5B, 1.5B, 7B, 72B.
* In the 7B and 72B models, context length has been extended to 128k tokens.
* We can Chat with Images and Video.
* Input: Images, Videos

**Translation into Chinese:**

* Qwen2是阿里巴巴集团新推出的一系列大型语言模型
* Qwen2接受了29种语言的数据培训，包括英语和中文。
* 它有4种参数大小：0.5B、1.5B、7B、72B。
* 在7B和72B模型中，上下文长度已扩展到128k个令牌。
* •我们可以通过图片和视频聊天。
* •输入：图像、视频



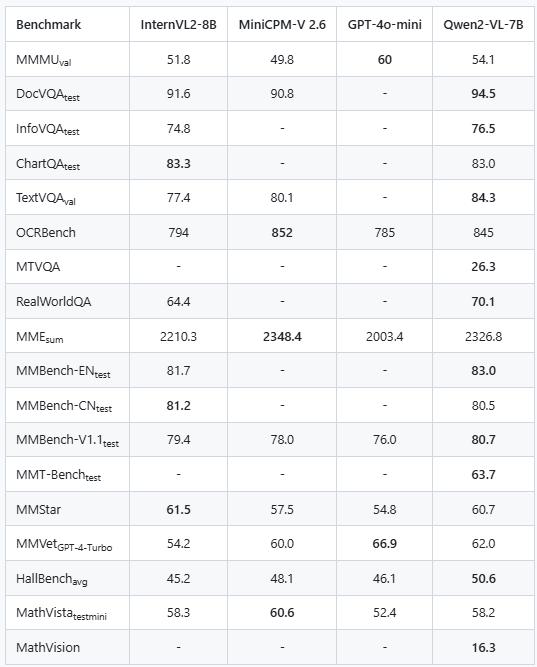
**Model Architecture: It can be used for Images and Videos**



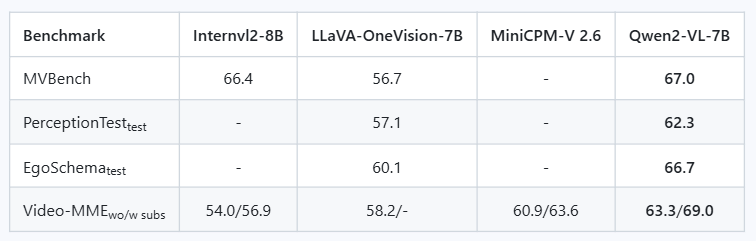
**Benchmark Comparison** (**InternVL2-8B, MiniCPM-V 2.6, GPT-4o-mini, Qwen2-VL-7B**):

**Image Benchmarks:**

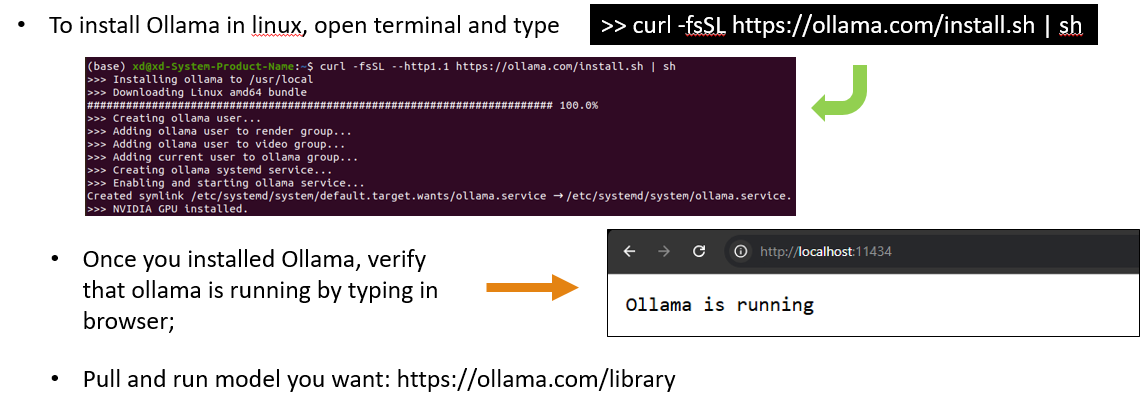
We can see in table that benchmark results of **Qwen2-VL-7B are overall improved as compared to InternVL2-8B, MiniCPM-V 2.6,** and **GPT-4o-mini.**



**Video Benchmarks**

****

**Usage:**

****

ollama pull qwen2:7b-instruct

ollama run qwen2:7b-instruct

**Command Line Terminal**

>> Tell me a story

>> Describe Image image.png

To download model without Ollama:

We can use Modelscope or Huggingface Hub

HuggingFace Hub not work without VPN so we can use Modelscope

from modelscope import snapshot\_download

model\_dir = snapshot\_download("qwen/Qwen2-VL-7B-Instruct")

Install the library (recommendation to use Conda Virtual Environment)

pip install qwen-vl-utils

**Importing Image and describe the image function:**



**Multi-Image:**

****

**Importing Videos and describe the image function:**

****

**Limitations**

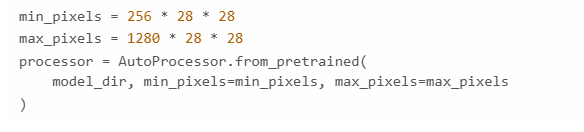
1. Lack of Audio Support: The current model does **not comprehend audio information** within videos.
2. Insufficient Counting Accuracy: Particularly in complex scenes, the accuracy of object counting is not high, necessitating further improvements.

1.缺乏音频支持：当前模型无法理解视频中的音频信息。

2.计数精度不足：特别是在复杂场景中，物体计数的精度不高，需要进一步改进。

**Image Resolution for performance boost**

* model supports a wide range of resolution inputs
* By default, it uses the native resolution for input
* but higher resolutions can enhance performance at the cost of more computation.





* We can set the minimum and maximum number of pixels to achieve an optimal configuration for own needs
* A token count range of 256-1280, to balance speed and memory usage.

