Paper Reading(2025.19-present)

Contents

Qwen2-VL

https://arxiv.org/abs/2409.12191

模型结构

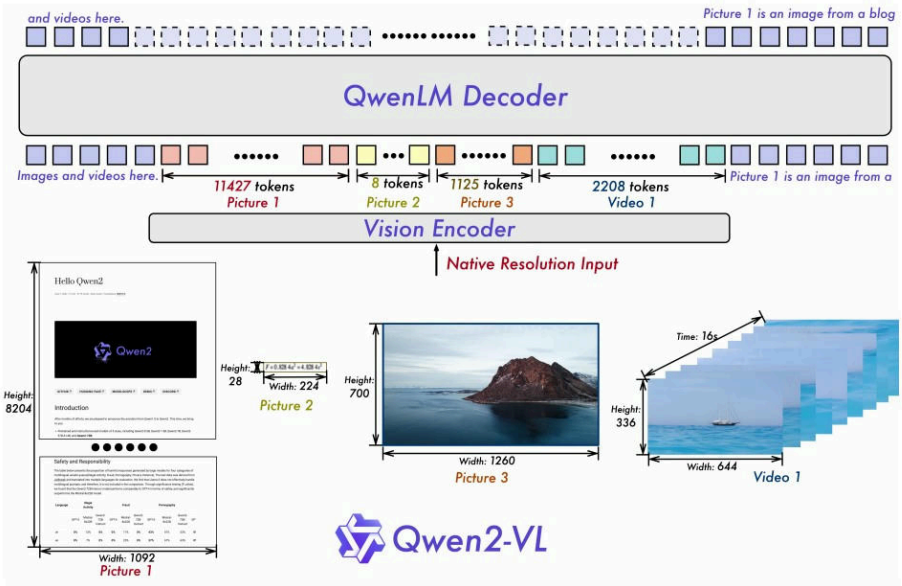

Figure 1: Qwen2-VL structure

- Naive Dynamic Resolution: 可将图片动态转换成若干数量的视觉 tokens，支持任意分辨率。修改 ViT，用 2D-RoPE 代替原本的绝对位置编码嵌入以获取图像的二维信息。在推理阶段，各种分辨率的图像打包成一个序列。