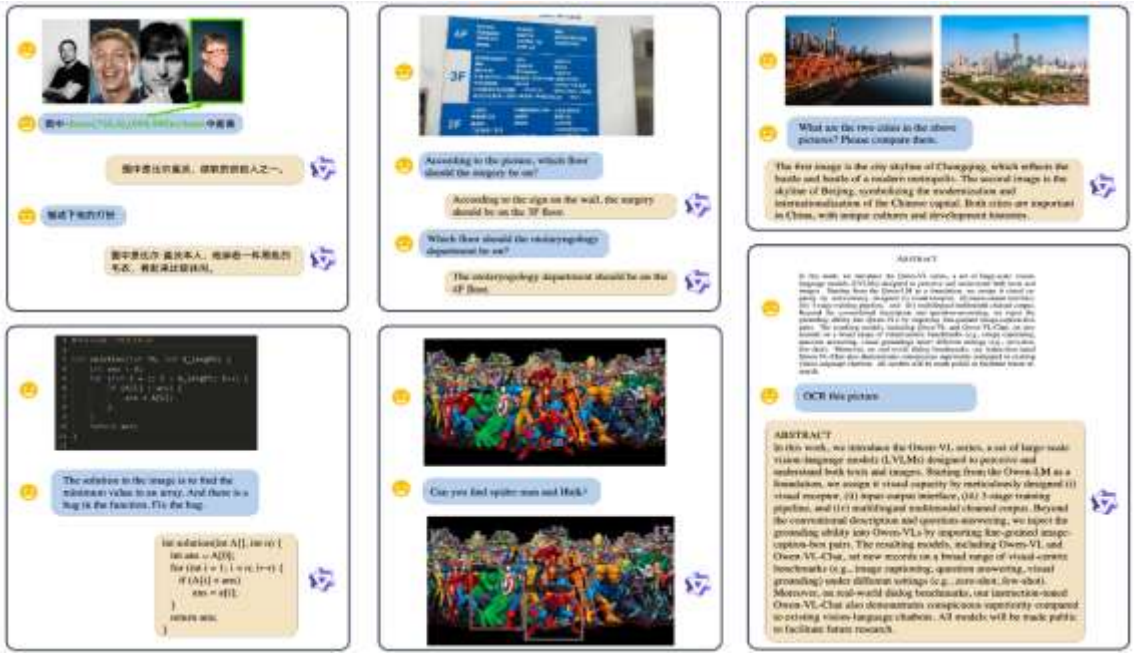
Qwen VL 系列分享

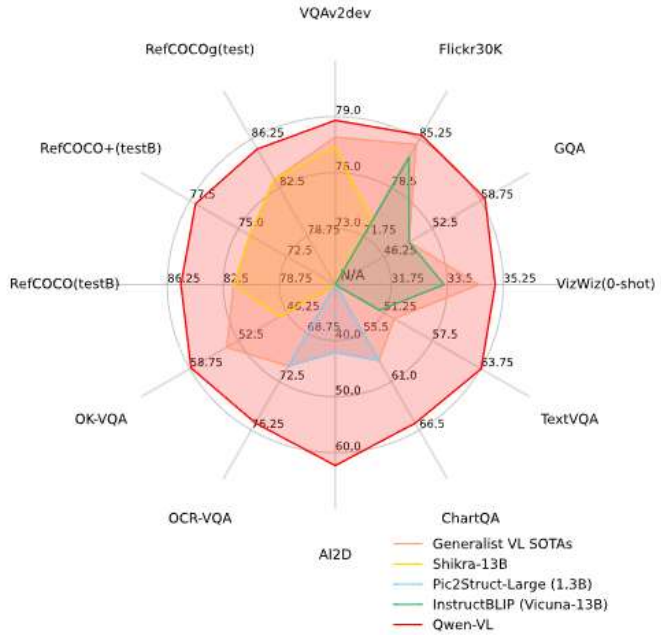
-Qwen VL, Qwen2-VL, Qwen2.5-VL

赖彦涛

2025年9月23日

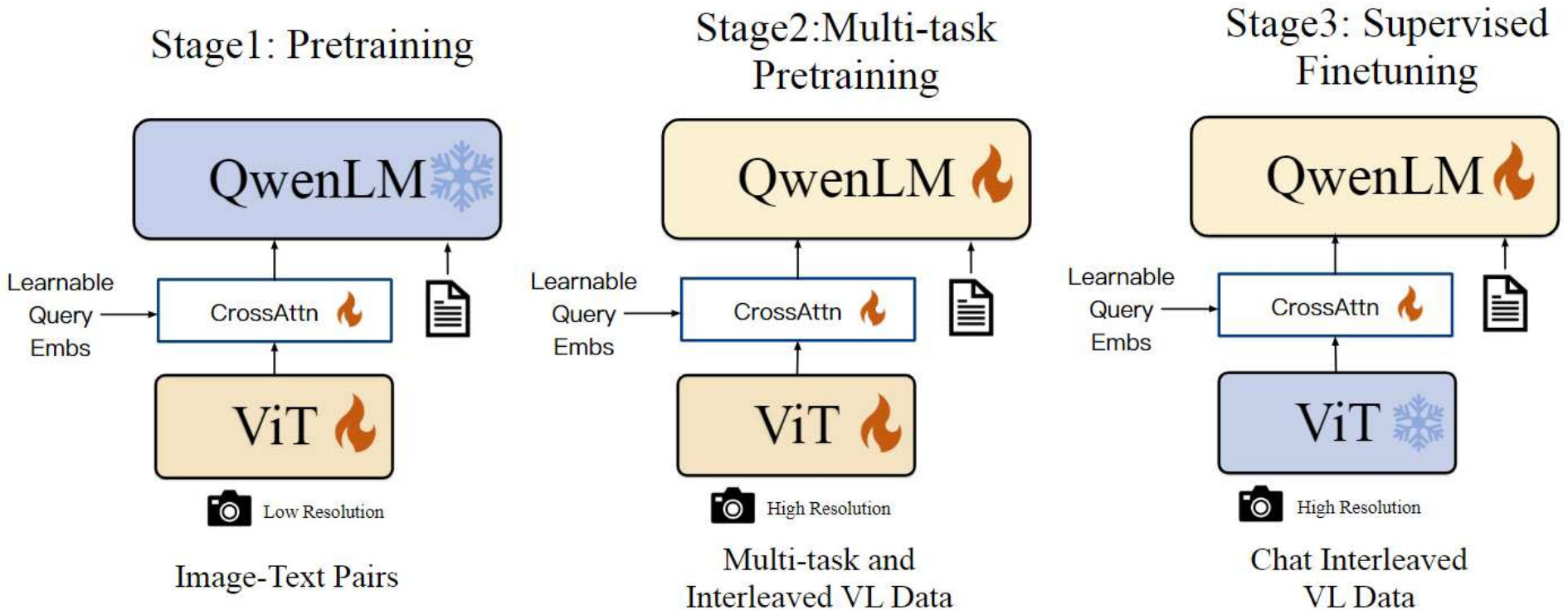
Qwen VL





Qwen-VL的模型效果基本就是中规中矩，和同期的模型对比没有太强的表现

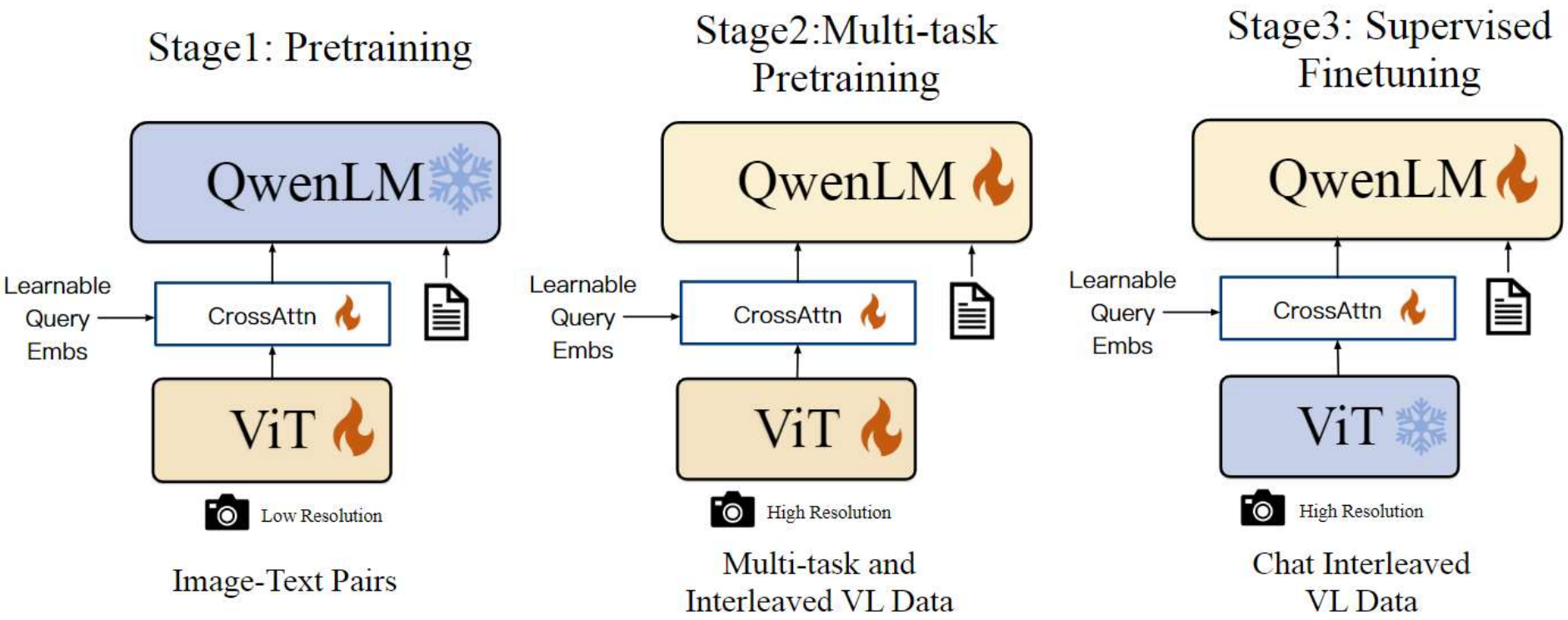
Qwen VL



创新点：

- 视觉编码： 动态高分辨率策略：将图像分割为448x448 patches（最高达1664x1664分辨率）
- 训练框架：三阶段渐进式训练：视觉编码器预训练 → 多模态对齐 → 指令微调
- 跨模态融合：轻量级Adapter设计（仅0.1B参数），通过Q-former连接视觉与语言模型
- 任务支持：支持检测框输出，格式：<box>(x1,y1,x2,y2)</box> 和多轮对话

Qwen VL



训练包括三个阶段：两个pre-training阶段和一个fituning阶段

- 第一阶段（单任务预训练，低分辨率图像224*224，训练ViT和CrossAttntion）
- 第二阶段（多任务预训练，高分辨率图像448*448，全参训练）
- 第三阶段（监督微调，主要提升模型的指令遵循能力和对话能力，全参数训练）

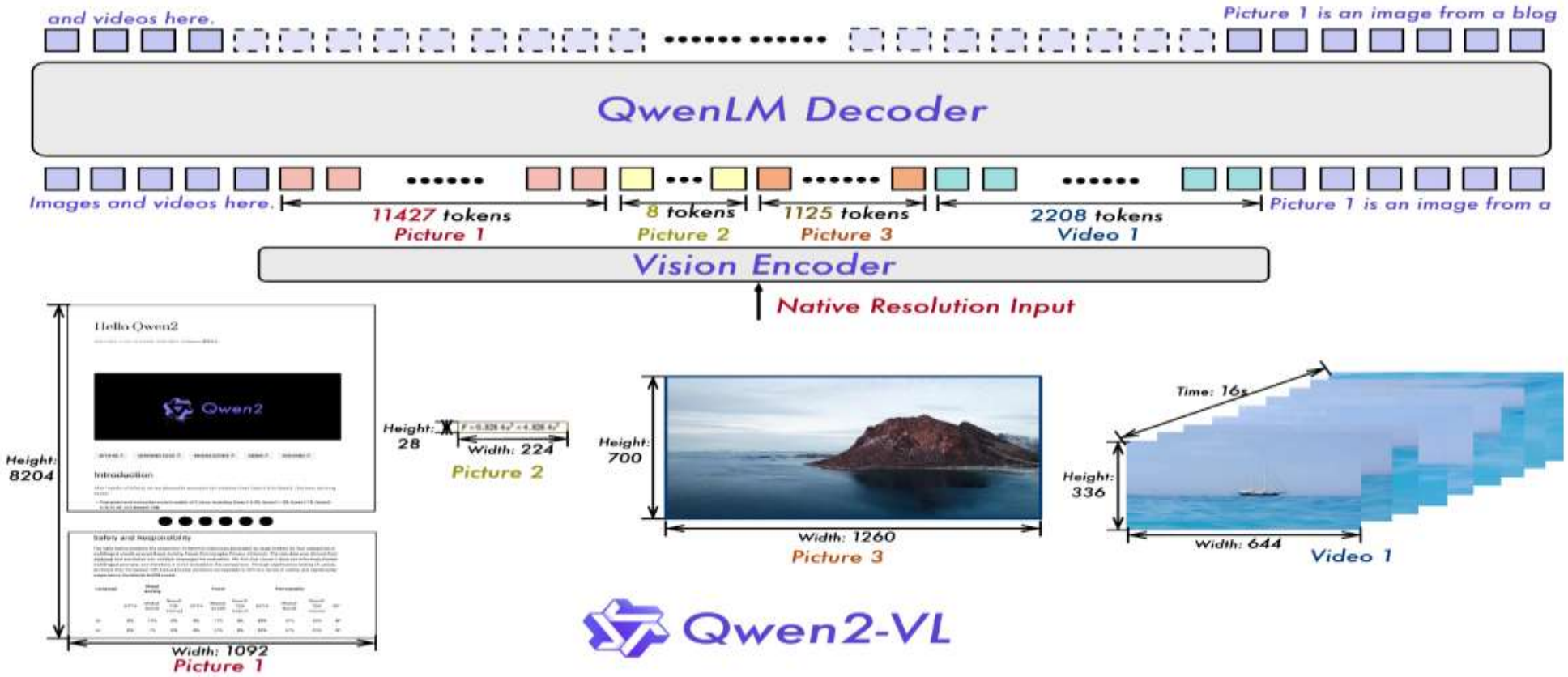
Qwen2-VL

Figure 1: Qwen2-VL capabilities: Multilingual image text understanding, code/math reasoning, video analysis, live chat, agent potential, and more. See Appendix for details.

主要优势：

- 对各种分辨率和比例的图像的先进理解
- 理解超过20分钟的视频
- 可以操作手机、机器人等设备的智能体
- 多语言支持：为了服务全球用户，Qwen2-VL 除了支持英语和中文外，现在还能够理解图像中不同语言的文本

Qwen2-VL



创新点/关键改进

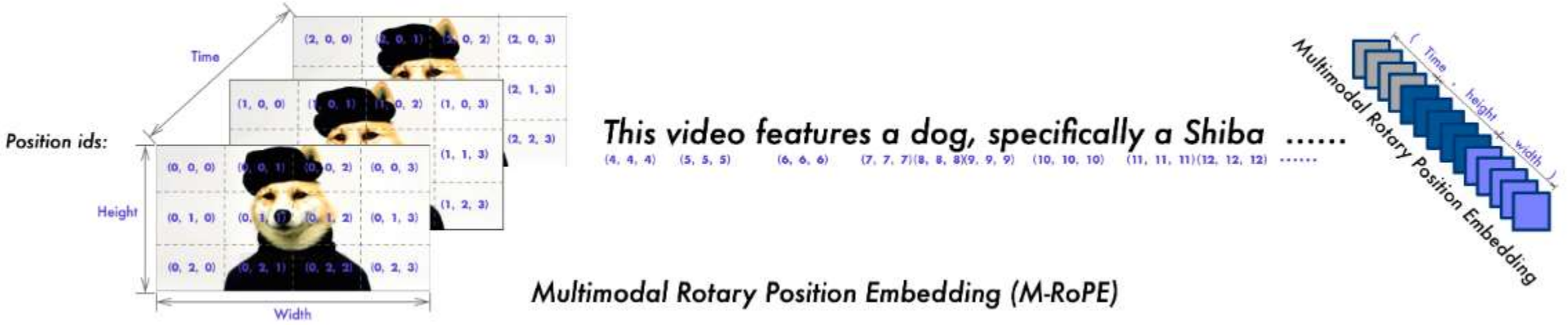
- 原生动态分辨率（Naive Dynamic Resolution）：使用了二位旋转位置嵌入RoPE来增强不同分辨率的适应性。保证模型输入和图像信息的一致性，也更贴近人类感知
- 多模态位置编码 （Mutimodal Rotary Position embedding, MRoPE）：同时捕获和整合一维文本、二维视觉和三维视频位置信息
- 统一图像和视频理解：使用时间轴为2的3d convolutioin编码图像/视频。如果是图像就copy两份，如果是视频，就每秒采样2帧。通过混合训练的方式，结合图像和视频数据，确保在图像理解和视频理解方面具有专业水平

Qwen2-VL

多模态位置编码 （Mutimodal Rotary Position embedding, MRoPE）

为了建模多模态输入，MRoPE将位置编码分解为三个独立的组成部分：时间、高度和宽度

- 文本输入：这三个组件使用相同的位置ID, 是MRoPE在功能上等于传统的1D RoPE
- 图像输入：所有token的时间id一样。根据token在图像中的height和width,对每个图像token分配一个unique id
- 视频输入：时间ID会在每一帧中递增，高度和宽度沿用和静态图像相同的分配方式

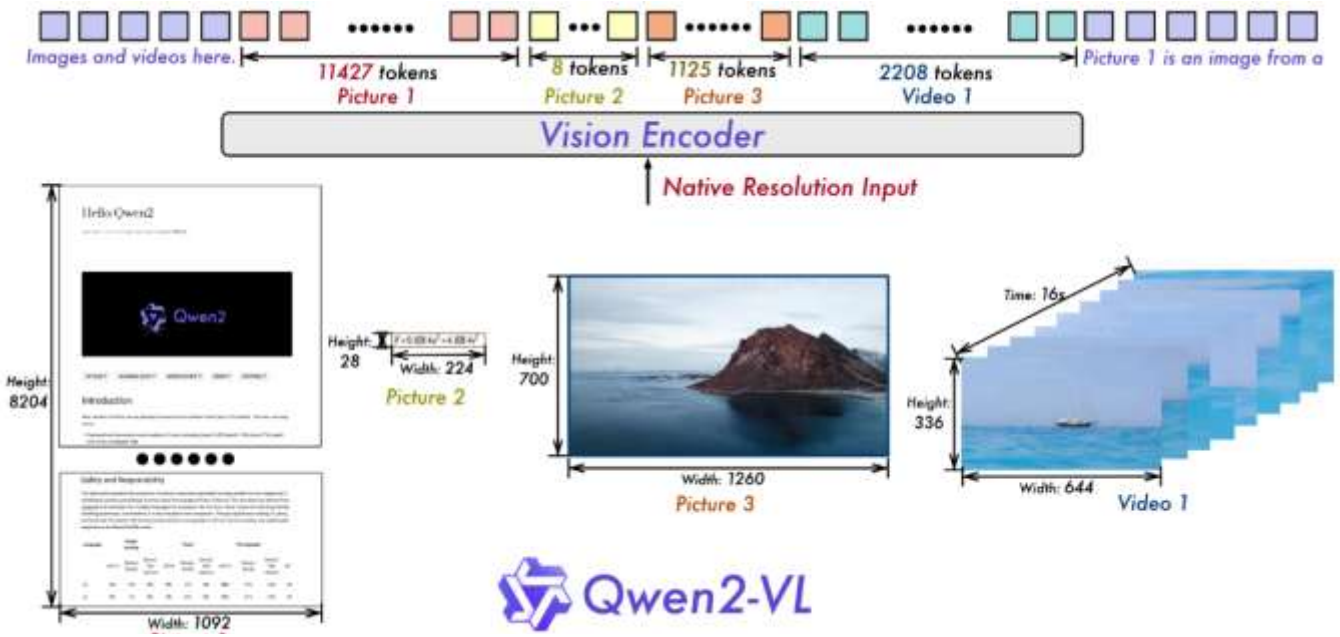


Qwen2-VL

原生动态分辨率（Naive Dynamic Resolution）

与Qwen-VL不同，Qwen2-VL可以处理任意分辨率的图像，将其动态转换为可变数量的视觉标记。

- 为了支持这一功能，Qwen2-VL修改了ViT，删除了原始的绝对位置嵌入，并引入了2D RoPE来捕获图像的二维位置信息
- 压缩视觉令牌：为了减少每个图像的视觉标记，在ViT之后使用一个简单的MLP层将相邻的2×2标记压缩成一个标记
- - 特殊令牌：在压缩的视觉标记的开头和结尾放置特殊的<|vision_start|>和<|visition_end|>标记。



因此，分辨率为28×224的图像，使用patch_size=14的ViT编码，在进入LLM之前将被压缩到10个标记

$$(28 \times 224) \div (14 \times 14) \div (2 \times 2) + 2 = 8 + 2 = 10$$

Qwen2-VL

统一视觉处理方式

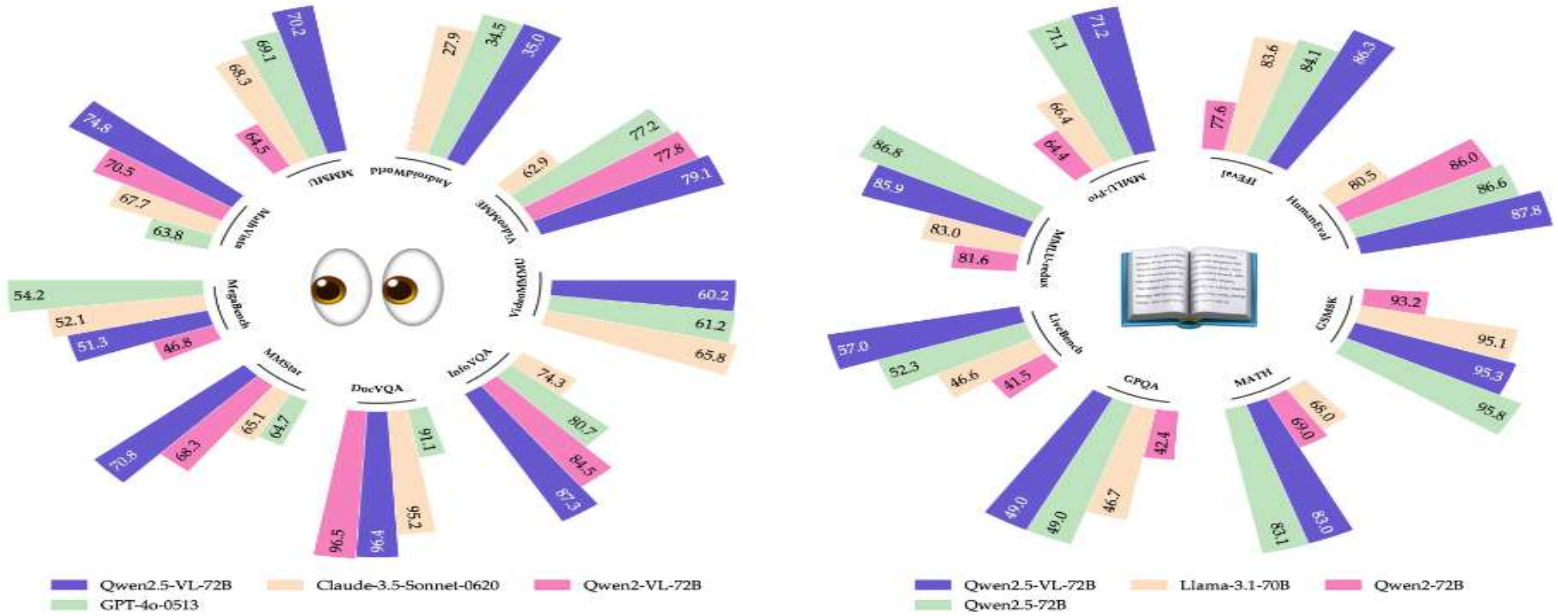
- 为了尽可能完整地保存视频信息，Qwen2-VL以每秒两帧的频率对每个视频进行采样
- 集成了深度为2的**3D卷积**来处理视频输入，使模型能够处理3D tubes 而不是2D patches，从而使其能够在不增加序列长度的情况下处理更多的视频帧
- 为了保持一致性，每个图像都被视为两个相同的帧

Qwen2-VL

➢ 训练流程

- 第一阶段：单任务训练（ViT）：本阶段训练数据由大规模图文对构成，覆盖约 6000 亿个图像与文本的 token，任务涵盖图文对齐、OCR 文本识别与图像分类。训练输入为图像，输出为对应的文本描述
- 第二阶段：多任务训练（全参数训练）：训练任务包括图像描述、视觉问答、多模态推理等多种形式，旨在提升模型对复杂图文交互的理解与表达能力。
- 第三阶段：指令微调（增强指令遵循能力）：在本阶段，模型冻结视觉编码器，仅对语言模型进行指令微调。该阶段使用 ChatML 格式构建的多模态指令数据集，覆盖图像问答、文档解析、多图对比、视频理解与对话、以及具备代理行为的交互场景。

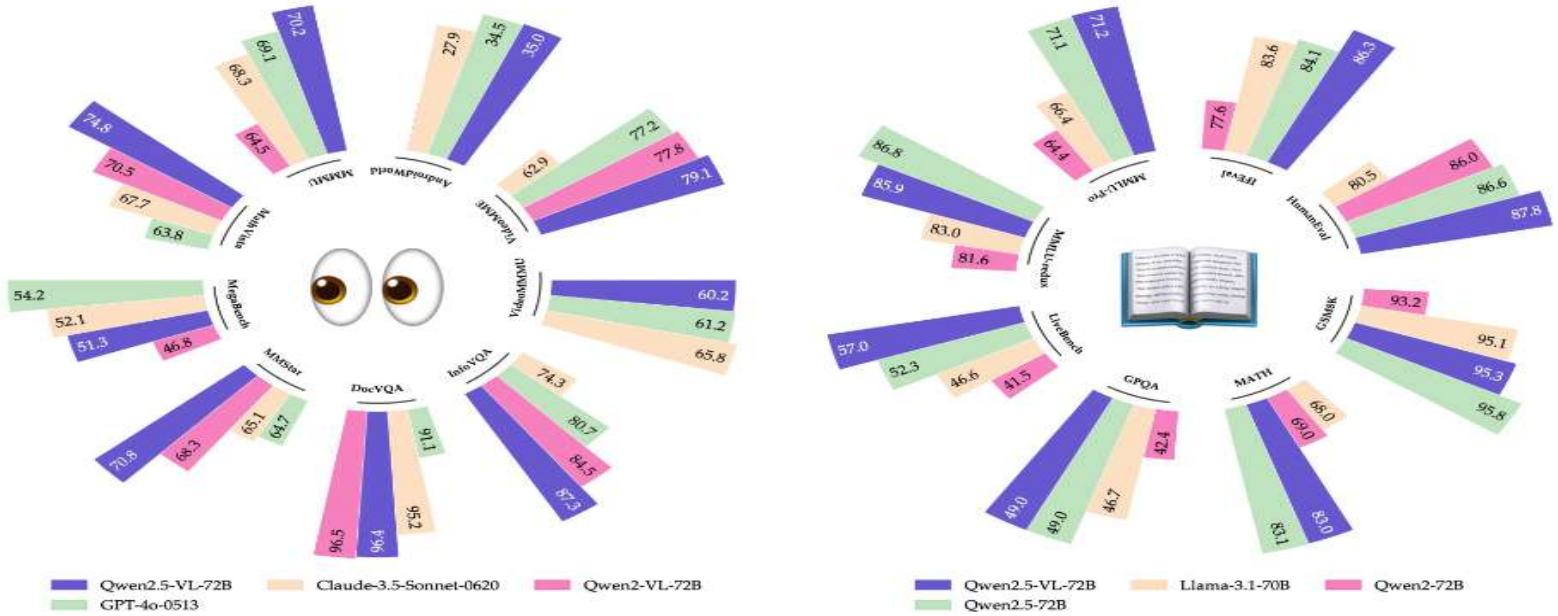
Qwen2.5-VL



主要优势：

- 感知更丰富的世界：Qwen2.5-VL 不仅擅长识别常见物体，如花、鸟、鱼和昆虫，还能够分析图像中的文本、图表、图标、图形和布局。
- Agent: Qwen2.5-VL 直接作为一个视觉 Agent，可以推理并动态地使用工具，初步具备了使用电脑和使用手机的能力。
- 理解长视频和捕捉事件：Qwen2.5-VL 能够理解超过 1 小时的视频，并且这次它具备了通过精准定位相关视频片段来捕捉事件的新能力。
- 视觉定位：Qwen2.5-VL 可以通过生成 bounding boxes 或者 points 来准确定位图像中的物体，并能够为坐标和属性提供稳定的 JSON 输出。
- 结构化输出：对于发票、表单、表格等数据，Qwen2.5-VL 支持其内容的结构化输出。

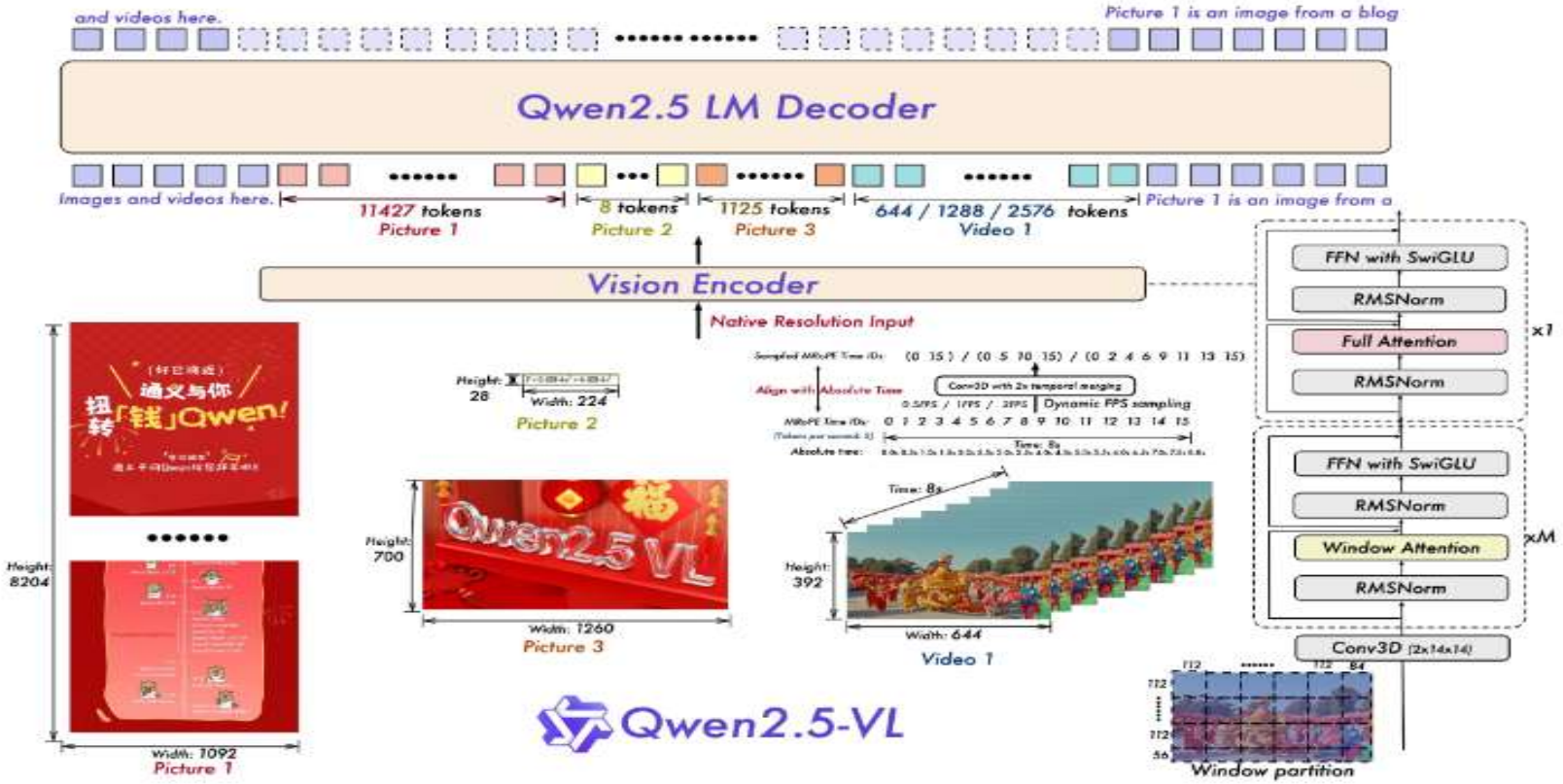
Qwen2.5-VL



创新点：

- 视觉encoder 是重新设计的，引入了window attention
- 动态fps采样：视频理解需要平衡计算效率和信息完整性，不同场景需要不同的采样率（根据内容自动调整视频采样率)
- 在sft阶段后使用了dpo过程
- 高质量数据：从1.2T到4.1T

Qwen2.5-VL



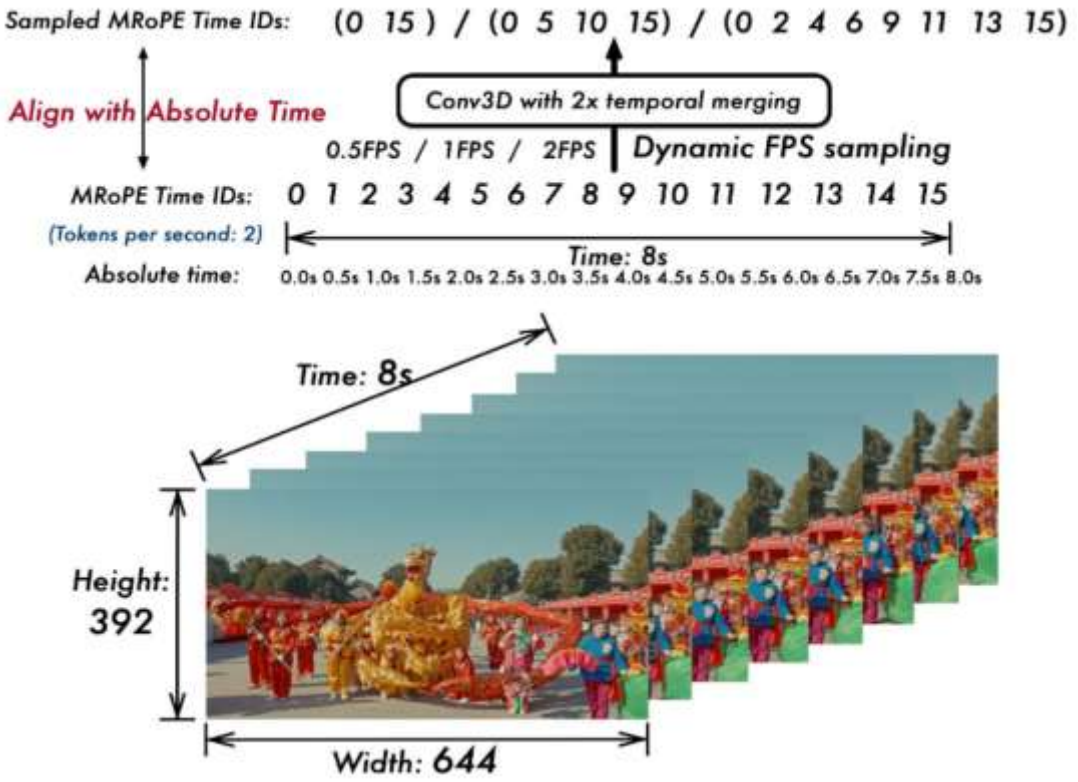
模型架构

- LLM: 使用Qwen2.5 LLM 的预训练权重作为初始化权重。为了更好满足多模态理解的需求，将1D RoPE改成了**MRoPE**
- Vision Encoder:
  - window attention (窗口注意力机制): 确保计算成本随图像快（patch）数量线性增长，而非平方增长。模型中仅有四层使用了完整的自注意力机制（full self-attention）,其余层均采用窗口注意力机制
  - 位置编码： 2D Rotary Positional Embedding (**RoPE**); 层归一化： **RMSNorm**; 激活函数： **SwiGLU**
- MLP-based Vision-Language Merger： 使用两层MLP来将图像embedding 和文本embedding对齐

Qwen2.5-VL

动态帧率（Dynamic Frame Rate）

对于视频输入，引入了动态帧率（Dynamic Frame Rate，FPS）训练和绝对时间编码机制。通过采用不同的帧率，模型能够更好地视频内容中的时间动态信息。



Qwen2.5-VL 训练流程

预训练（Pre-Training）

包含三个阶段，在不同的阶段对这些数据类型的组成和比例进行了精细的调整，以优化模型的学习效果。数据总量从1.2T（Qwen2-VL）增加至约4T（T, trillion,万亿）

- 第一阶段：仅对视觉Transformer(ViT)进行训练，以提升与语言模型的对齐能力。该阶段的数据来源主要包括图像描述、OCR数据
- 第二阶段：全参数微调来增强模型处理复杂视觉信息的能力。该阶段映入了更为复杂且需要推理能力的数据，例如视觉问题（VQA）、多模态数据任务、视频理解任务。
- 第三阶段：全参数微调，扩展序列长度，通过增加long video、long document数据来增加模型的推理能力

后训练（post-training）

包含两个阶段：监督微调SFT和直接偏好优化DPO

- SFT阶段：通过指令微调，来提升指令遵循能力（数据量200w）
- DPO阶段：利用偏好数据将模型输出和人类偏好对齐。

Stages	Visual Pre-Training	Multimodal Pre-Training	Long-Context Pre-Training
Data	Image Caption Knowledge OCR	+ Pure text Interleaved Data VQA, Video Grounding, Agent	+ Long Video Long Agent Long Document
Tokens	1.5T	2T	0.6T
Sequence length	8192	8192	32768
Training	ViT	ViT & LLM	ViT & LLM

谢谢