

Multimodal Large Language Model Meets Remote Sensing

GISLab Short-Term Course 2025 Summer

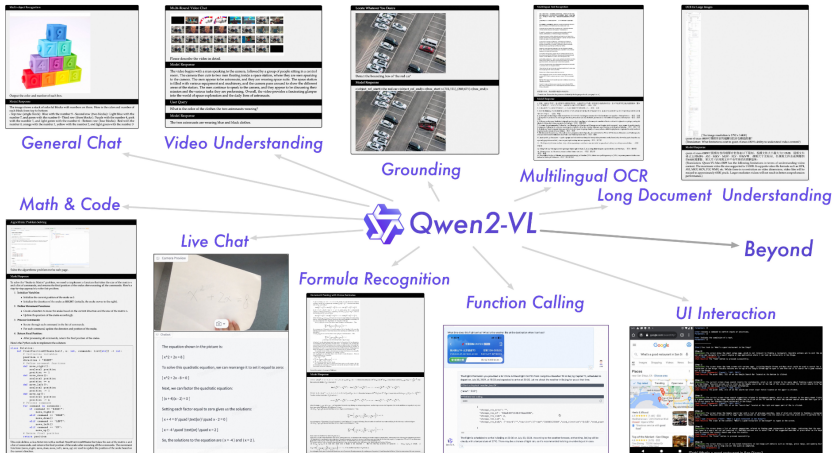
Zhenyuan Chen

School of Earth Science, Zhejiang University

2025

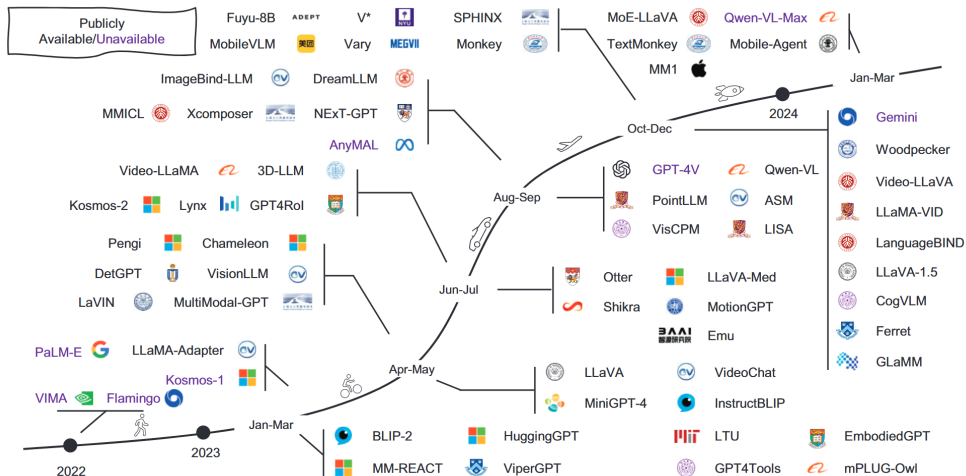
`bili_sakura@zju.edu.cn`

Background: Multimodal Large Language Models (MLLMs)



Qwen2-VL (Wang et al., 2024).

Background



MLLM Timeline (Yin et al., 2024).

What is a Multimodal Large Language Model (MLLM)?

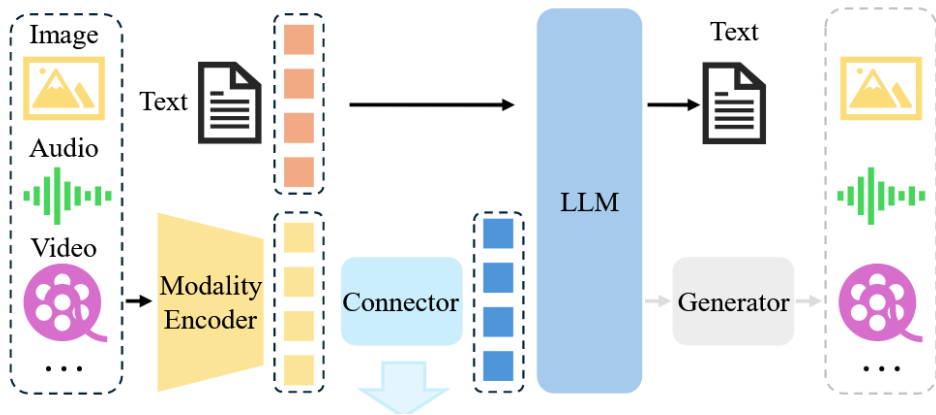
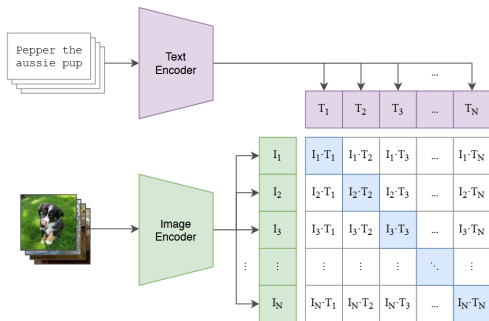


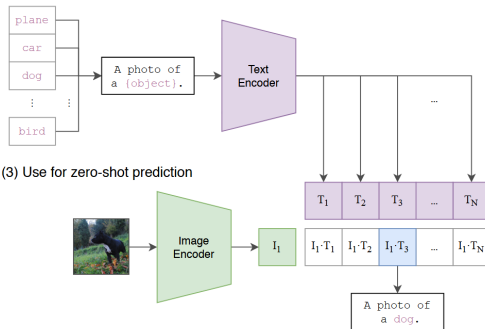
Figure: General architecture of a Multimodal Large Language Model (MLLM) (Yin et al., 2024).

Vision Encoder: CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

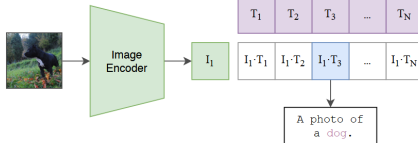
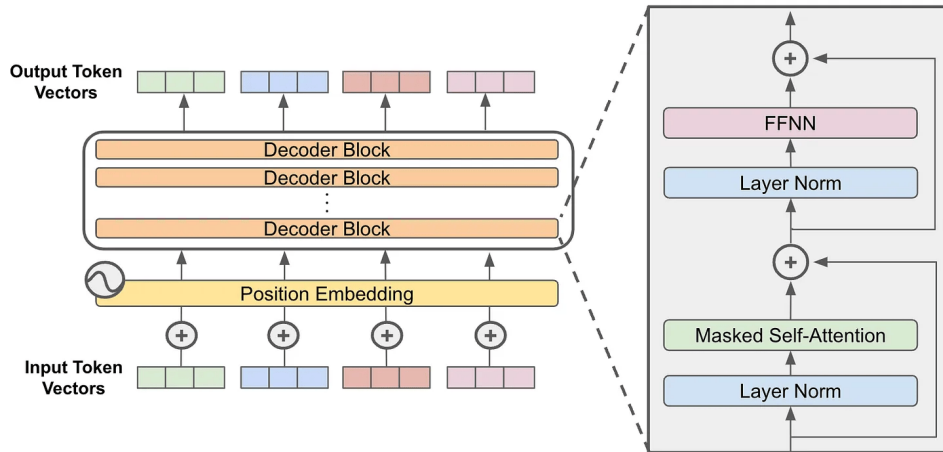


Figure: OpenAI CLIP architecture (Radford et al., 2021).

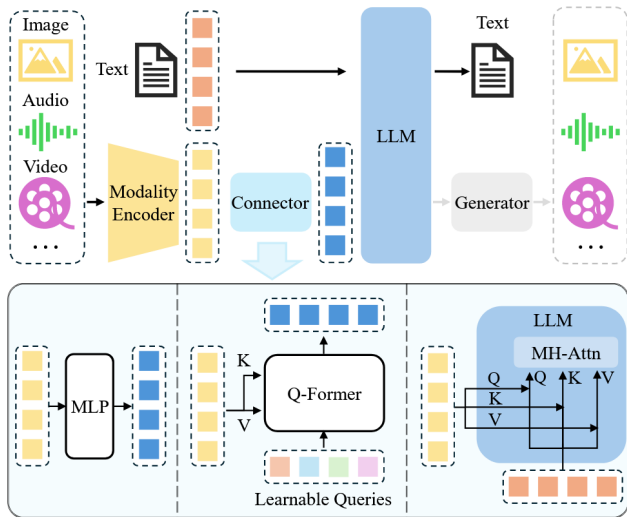
LLM Backbone: Large Language Models



Latest LLMs are generally composed of a series of Transformers decoder blocks (Vaswani et al., 2017).

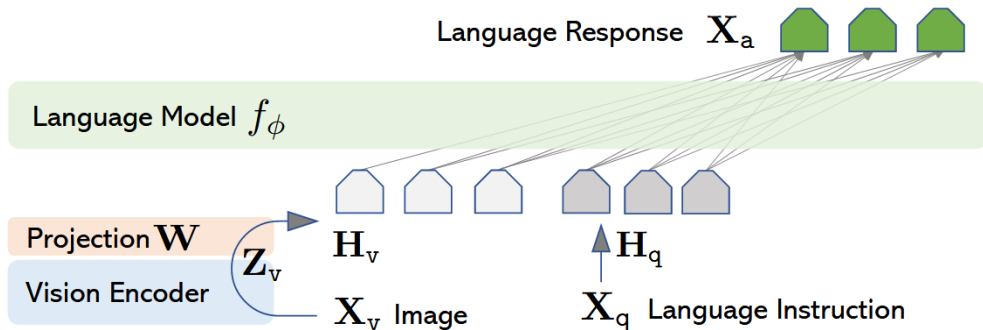
(image source:cameronrwolfe)

Connector: General MLLM



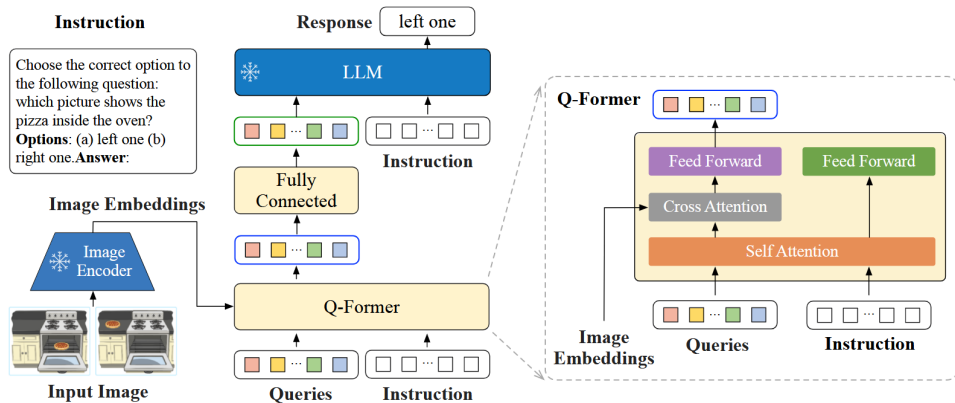
An typical MLLM architecture includes an encoder, a connector, and a LLM. (Yin et al., 2024).

Connector: MLP



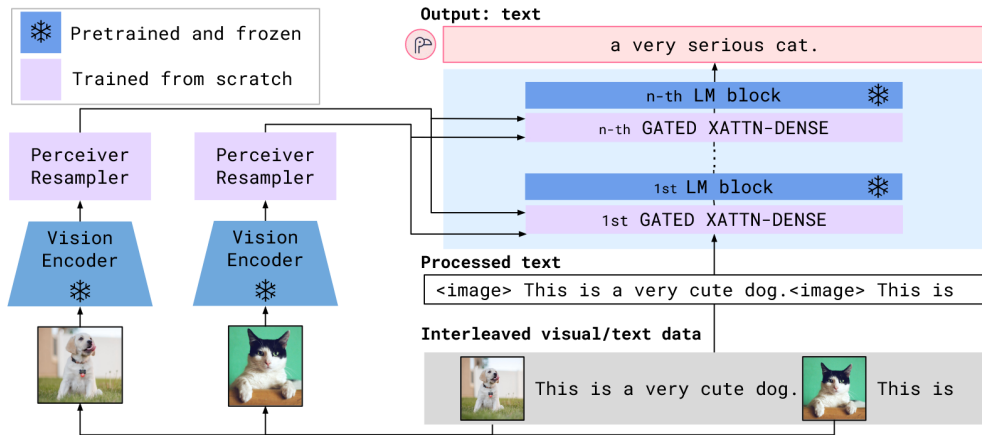
Connector architecture in LLaVA (Liu et al., 2023).

Connector: Q-Former



Connector architecture in InstructBLIP (Dai et al., 2023).

Connector: Feature-level Fusion



Connector architecture in Flamingo (Alayrac et al., 2022).