Multimodal Large Language Model Meets Remote Sensing

GISLab Short-Term Course 2025 Summer

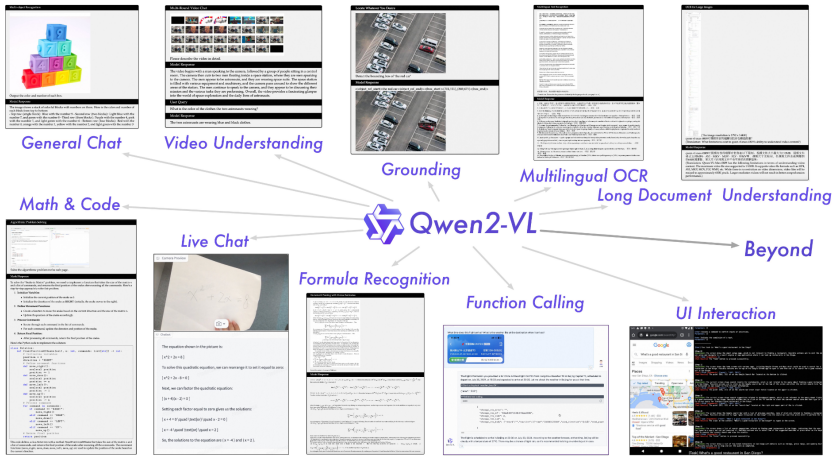
Zhenyuan Chen

School of Earth Science, Zhejiang University

2025

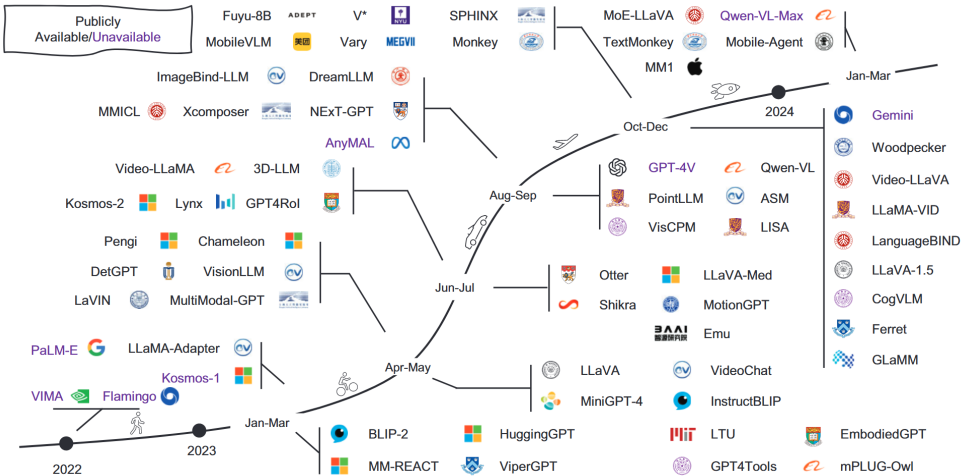
bili_sakura@zju.edu.cn

Background: Multimodal Large Language Models (MLLMs)



Qwen2-VL (**wangQwen2VLEnhancingVisionLanguage2024a**).

Background



MLLM Timeline (**yinSurveyMultimodalLarge2024a**).

What is a Multimodal Large Language Model (MLLM)?

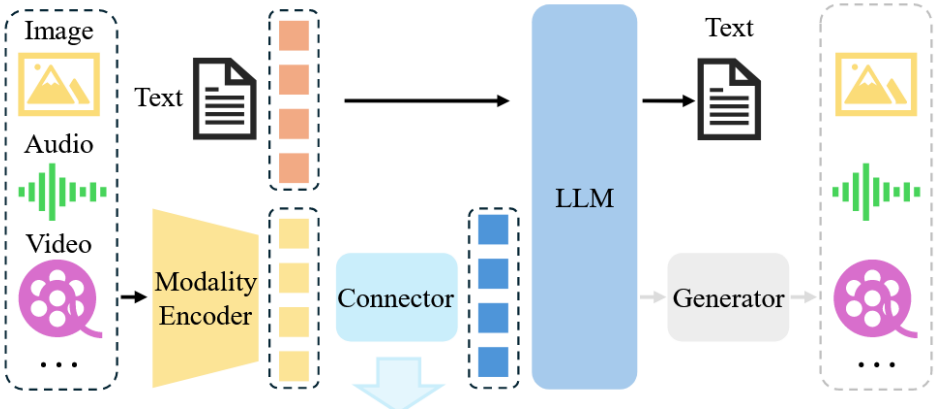


Figure: General architecture of a Multimodal Large Language Model (MLLM) (**yinSurveyMultimodalLarge2024a**).

Vision Encoder: CLIP

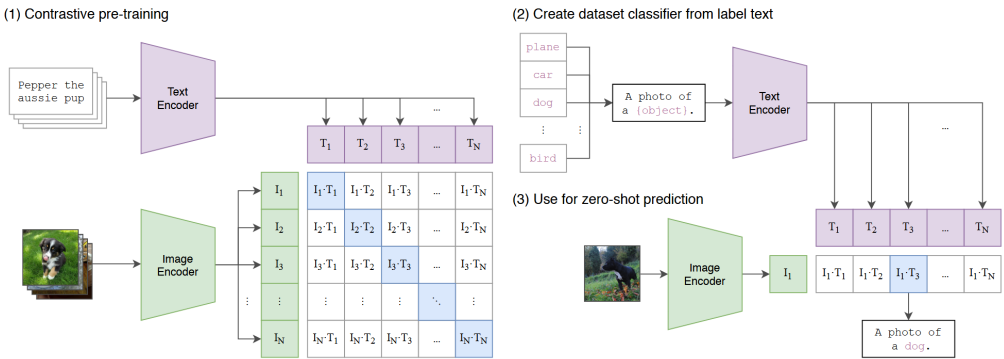
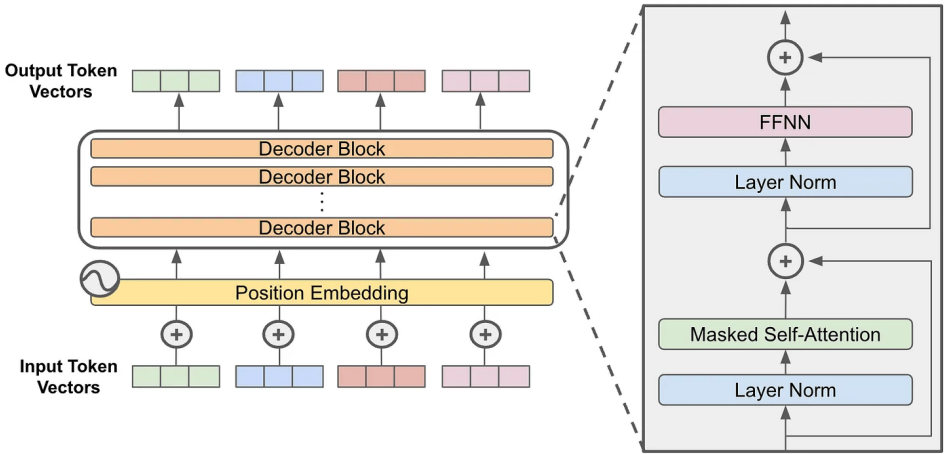


Figure: OpenAI CLIP architecture (**radfordLearningTransferableVisual2021**).

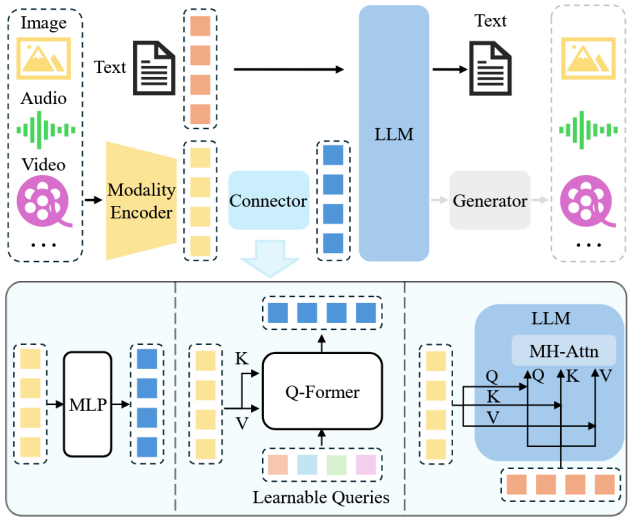
LLM Backbone: Large Language Models



Latest LLMs are generally composed of a series of Transformers decoder blocks (**vaswaniAttentionAllYou2017**).

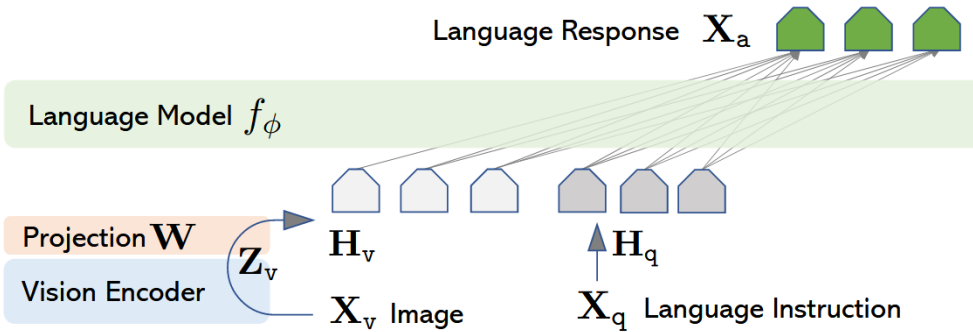
(image source:cameronrwolfe)

Connector: General MLLM



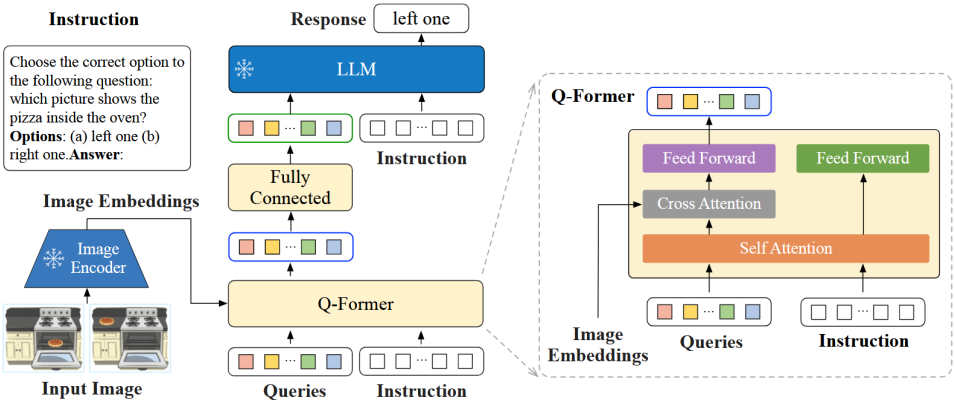
An typical MLLM architecture includes an encoder, a connector, and a LLM. (**yinSurveyMultimodalLarge2024a**).

Connector: MLP



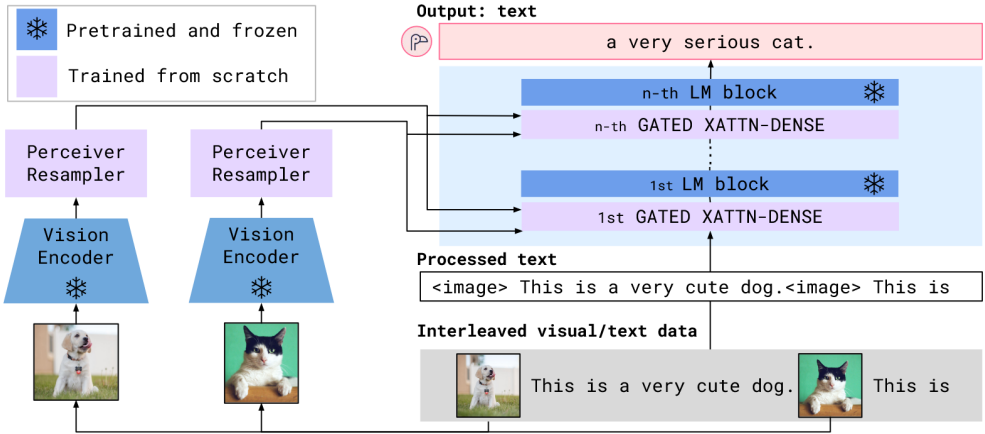
Connector architecture in LLaVA (**liuVisualInstructionTuning2023**).

Connector: Q-Former



Connector architecture in InstructBLIP (**daiInstructBLIPGeneralpurposeVisionLanguage2023a**).

Connector: Feature-level Fusion



Connector architecture in Flamingo (**alayracFlamingoVisualLanguage2022**).