=1920bp =1080bp

Multimodal Large Language Model Meets Remote Sensing

GISLab Short-Term Course 2025 Summer

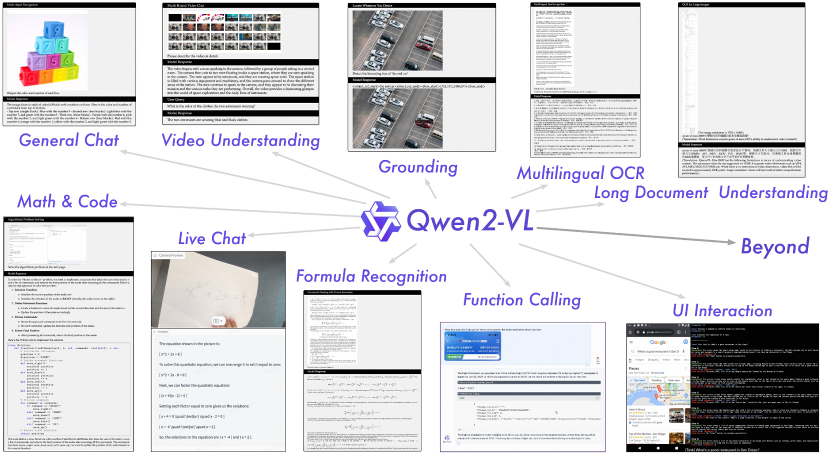
Zhenyuan Chen

School of Earth Science, Zhejiang University

2025

bili_sakura@zju.edu.cn

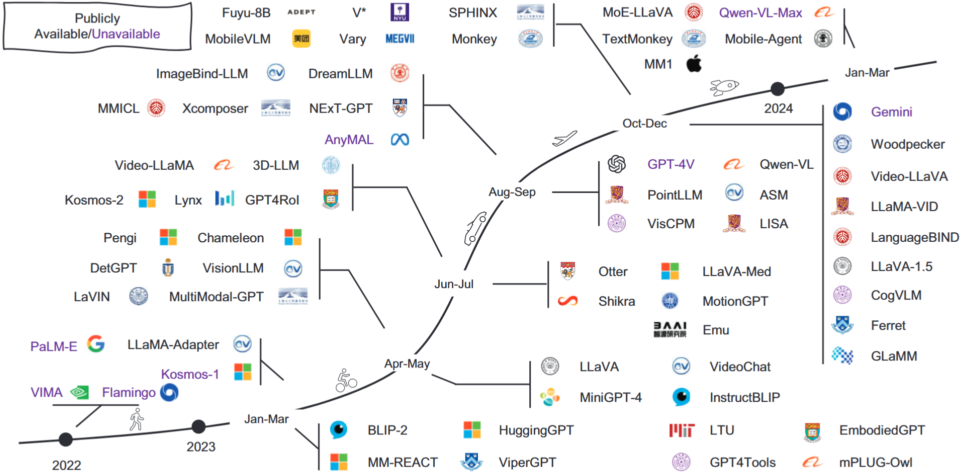
Background: Multimodal Large Language Models (MLLMs)



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

Wang, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution, 2024.

Background



MLLM Timeline (Yin et al., 2024).

Yin, et al. A Survey on Multimodal Large Language Models. National Science Review. 2024.

What is a Multimodal Large Language Model (MLLM)?

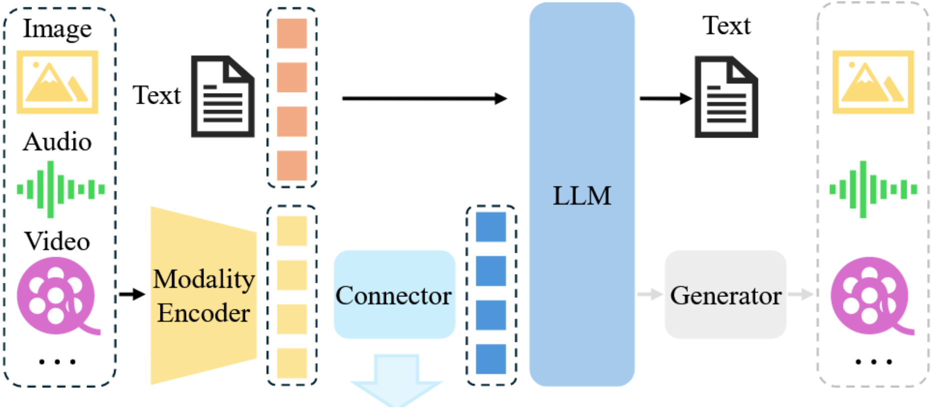


Figure: General architecture of a Multimodal Large Language Model (MLLM) (Yin et al., 2024).

Yin, et al. A Survey on Multimodal Large Language Models. National Science Review. 2024.

Vision Encoder: CLIP

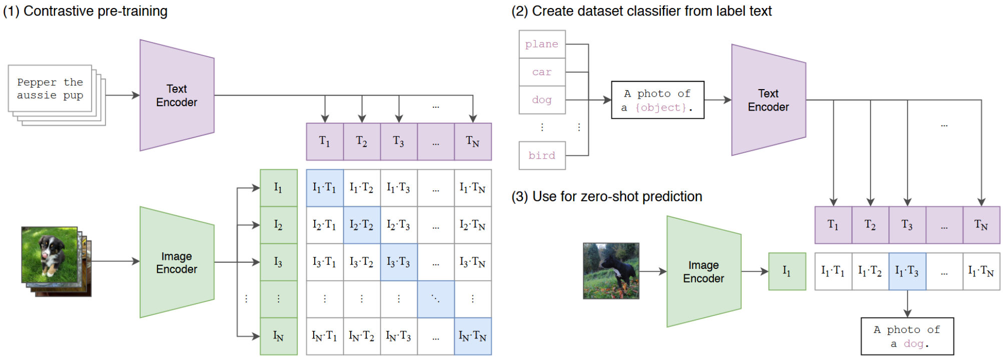
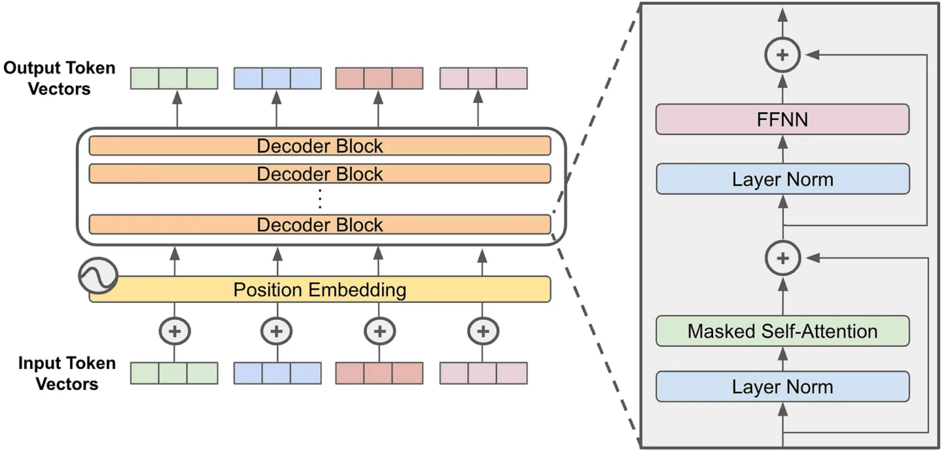

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
T_1 T_2 T_3 ... T_N
Image Encoder
I_1 I_2 I_3 ⋮ I_N
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
$I_2{\cdot}T_1$ $I_2{\cdot}T_2$ $I_2{\cdot}T_3$ ... $I_2{\cdot}T_N$
$I_3{\cdot}T_1$ $I_3{\cdot}T_2$ $I_3{\cdot}T_3$ ... $I_3{\cdot}T_N$
$I_N{\cdot}T_1$ $I_N{\cdot}T_2$ $I_N{\cdot}T_3$ ... $I_N{\cdot}T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1{\cdot}T_1$ $I_1{\cdot}T_2$ $I_1{\cdot}T_3$ ... $I_1{\cdot}T_N$
A photo of a dog.

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

Radford, et al. Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021.

LLM Backbone: Large Language Models

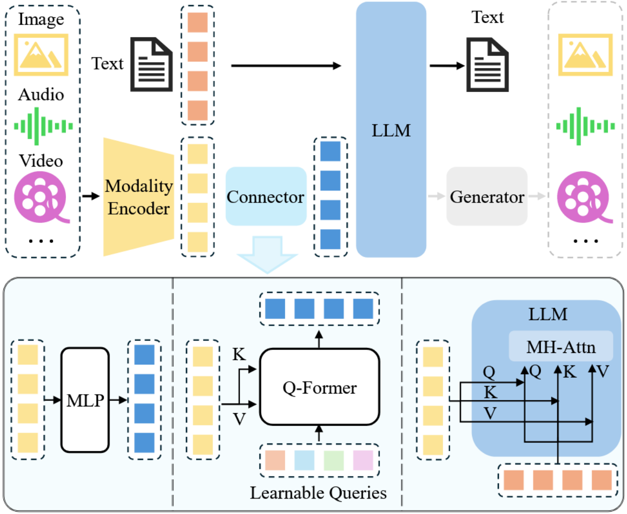


Latest LLMs are generally composed of a series of Transformers decoder blocks (Vaswani et al., 2017).

(image source:cameronrwolfe)

Vaswani, et al. Attention Is All You Need, NeurIPS, 2017.

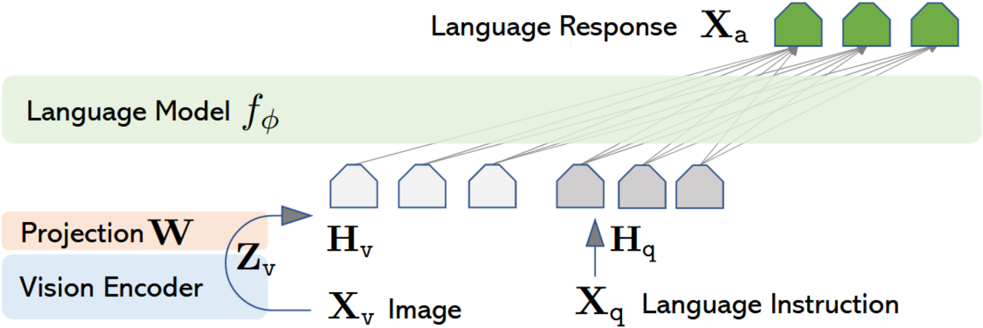
Connector: General MLLM



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

Yin, et al. A Survey on Multimodal Large Language Models. National Science Review. 2024.

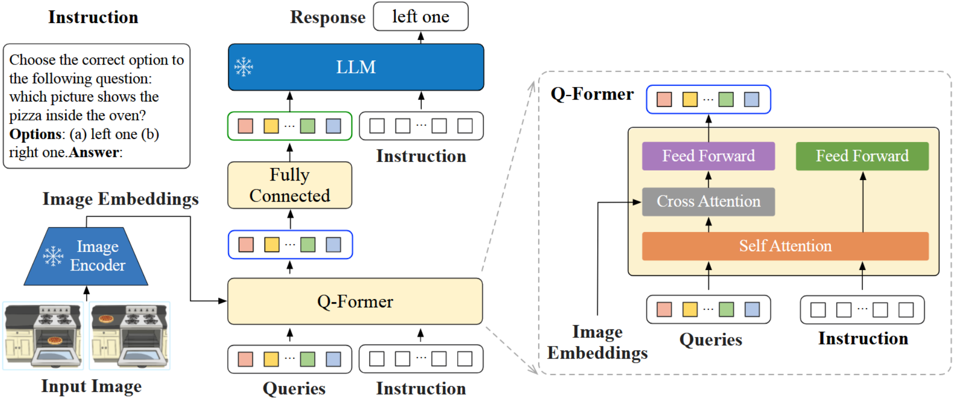
Connector: MLP



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

Liu, et al. Visual Instruction Tuning, NeurIPS, 2023.

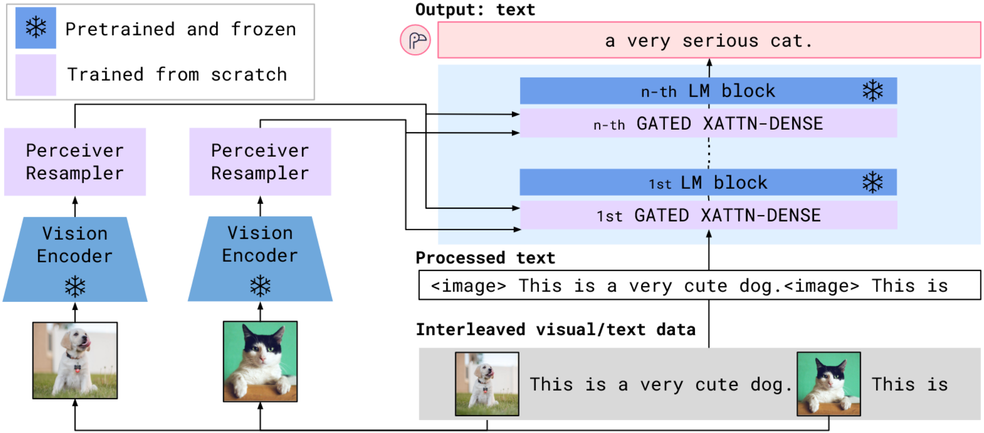
Connector: Q-Former



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

Dai, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, NeurIPS, 2023.

Connector: Feature-level Fusion



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

Alayrac, et al. Flamingo: A Visual Language Model for Few-Shot Learning, NeurIPS, 2022.