



香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

CSC5051/MDS5110

Natural Language Processing

Lecture 10: Multimodal LLMs

Spring 2026
Benyou Wang
School of Data Science

Before the lecture



徐扬生 (Yangsheng Xu)

香港中文大学 (深圳) 博士生陈同学作为 CTO 创办的医疗影像初创公司 Cognita, 以近 1 亿美元的价格被 Mosaic Clinical Technologies™ 收购。Cognita 是专为放射学打造的视觉语言模型领域的先驱, 与仅由放射科医生解读相比, 人工智能生成的结果可减少高达四倍的重大诊断错误, 放射科医生阅片时间最多可节省 76%。

收起



Mosaic Clinical
Technologies™
Acquires Cognita
Imaging Inc.,
Pioneering Next
Phase of Radiology
Intelligence at
Scale

2分钟前

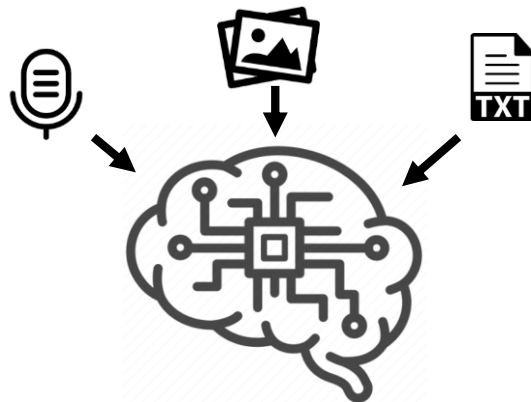
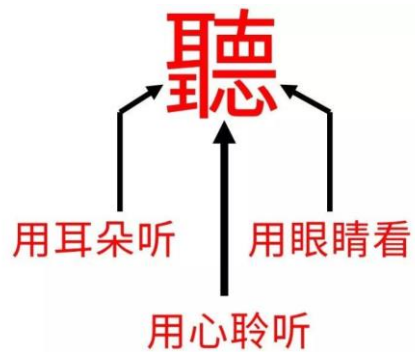


Join the journey on “(multi-modal) LLMs for Society”

Outline

1. Introduce multimodality
 - a. What is Multimodality?
 - b. Why Multimodality?
2. Multimodal Large Language Models
 - a. Early work 1: CLIP: Contrastive Language-Image Pre-training
 - b. Early work 2: Flamingo: the dawns of LMMs (relatively small language models)
 - c. Multimodal LLM
3. Training Multimodal LLMs
 - a. Architecture
 - b. Visual encoders or tokenizer
 - c. Instruction Tuning in Multi-modal LLM
 - d. Multi-stage training
4. Research Directions for Multimodal LMM
5. Our research on Multimodal Large Language Models

Why Multimodality



Human processes **multimodal** infos simultaneously

What is Multimodality?

Multimodal Behaviors and Signals

Language

- **Lexicon**
 - Words
- **Syntax**
 - Part-of-speech
 - Dependencies
- **Pragmatics**
 - Discourse acts

Acoustic

- **Prosody**
 - Intonation
 - Voice quality
- **Vocal expressions**
 - Laughter, moans

Visual

- **Gestures**
 - Head gestures
 - Eye gestures
 - Arm gestures
- **Body language**
 - Body posture
 - Proxemics
- **Eye contact**
 - Head gaze
 - Eye gaze
- **Facial expressions**
 - FACS action units
 - Smile, frowning

Touch

- **Haptics**
- **Motion**

Physiological

- **Skin conductance**
- **Electrocardiogram**

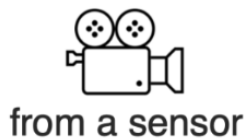
Mobile

- **GPS location**
- **Accelerometer**
- **Light sensors**

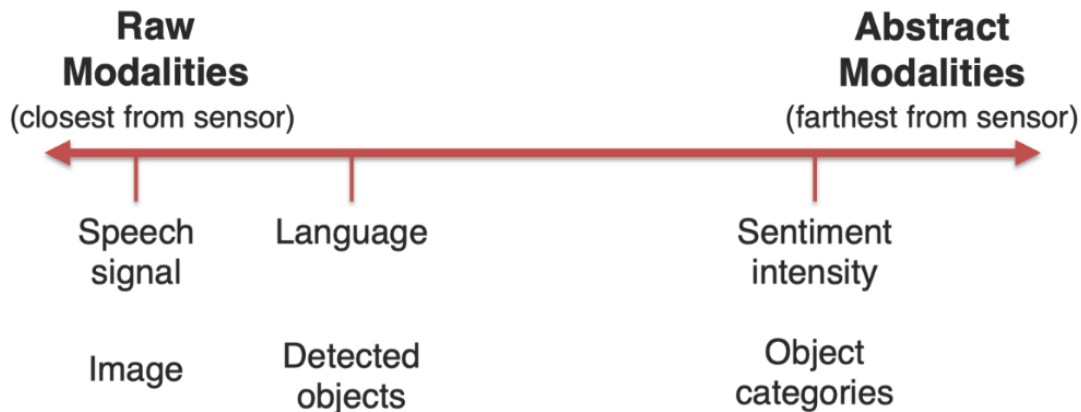
What is Multimodality?

Definition

Modality refers to the way in which something expressed or perceived.



Examples:



What is Multimodality?

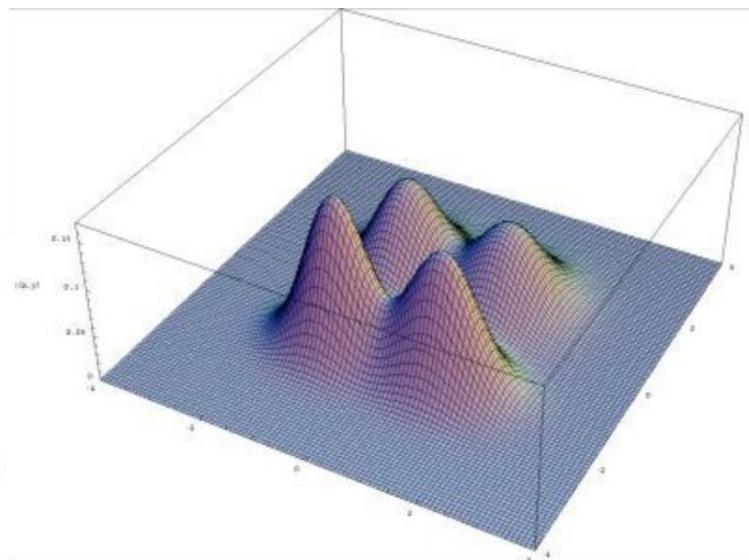
multimodal adjective

mul·ti·mod·al ,məl-tē-'mō-dəl -tī-

: having or involving several modes, modalities, or maxima

| *multimodal* distributions

| *multimodal* therapy

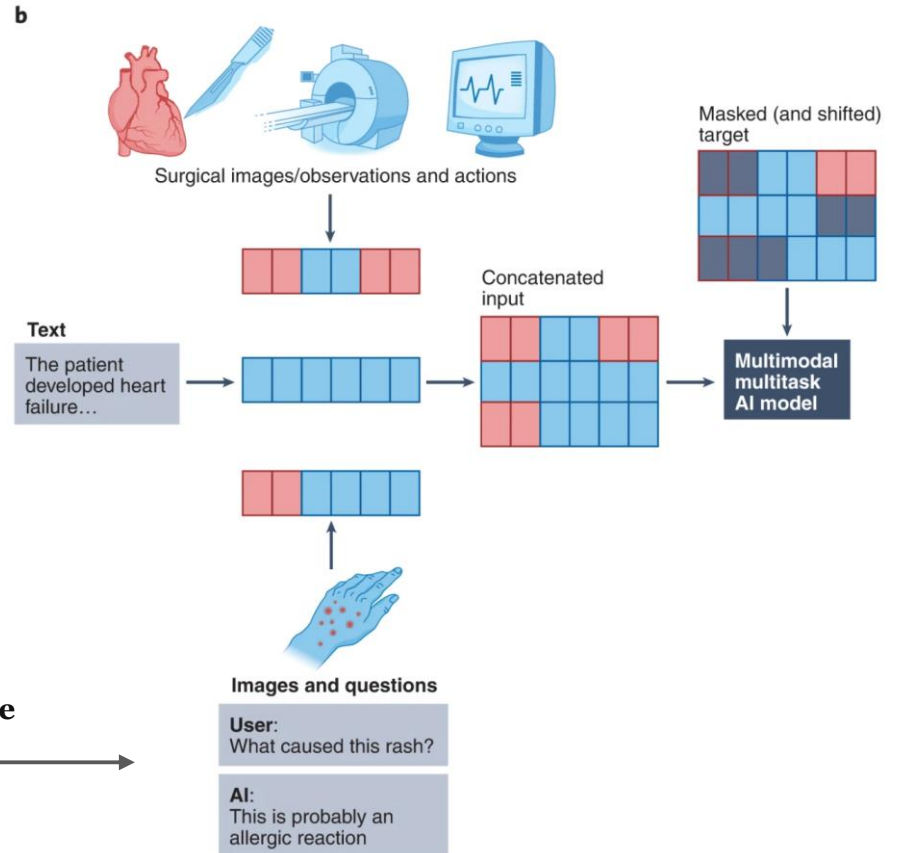


In our case, focusing on NLP: text + one or more other *modality* (images, speech, audio, olfaction, others). We'll mostly focus on images as the other modality.

Why is Multimodality?

Many use cases are impossible without multimodality, especially those in industries that deal with a mixture of data modalities such as healthcare, robotics, e-commerce, retail, gaming, etc.

An example of how multimodality can be used in healthcare. Image from Multimodal biomedical AI



Why is Multimodality?

A range of very good reasons:

- Faithfulness: Human experience is multimodal
- Practical: The internet & many applications are multimodal
- Data efficiency and availability:
 - Efficiency: Multimodal data is rich and “high bandwidth” (compared to language; quoting LeCun, “an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts”), so better for learning?
 - Scaling: More data is better, and we’re running out of high quality text data.



Multimodality is one of the main frontiers of the new foundation model revolution.

Multimodal Large Language Models

— The age of the large model is upon us, so let's just skip the small model and leap directly into the large model era.

Milestones

Given the existence of so many amazing multimodal systems, a challenge of writing this PPT is choosing which systems to focus on. Here, we first focus on two early models: **CLIP (2021)** and **Flamingo (2022)** both for their significance as well as availability and clarity of public details.

- **CLIP** was the first model that could generalize to multiple image classification tasks with zero- and few-shot learning.
- **Flamingo** wasn't the first large multimodal model that could generate open-ended responses. However, Flamingo's strong performance prompted some to consider it the GPT-3 moment in the multimodal domain.
- **Multimodal LLMs**

- [\[CLIP\] Learning Transferable Visual Models From Natural Language Supervision](#) (OpenAI, 2021)
- [Flamingo: a Visual Language Model for Few-Shot Learning](#) (DeepMind, April 29, 2022)

Early Work I: CLIP


Contrastive Language-Image Pre-training

a good practice for alignment: we could “map images to the text space” using a discriminative Language models (encoder-only LLM)

CLIP: Contrastive Language-Image Pre-training


→ CLIP leveraged **natural language supervision** and **contrastive learning**, which allowed CLIP to both **scale up** their data and make training more efficient. We'll go over why/how these two techniques work.

Food101
guacamole (90.1%) Ranked 1 out of 101 labels



- a photo of **guacamole**, a type of food.
- a photo of ceviche, a type of food.
- a photo of edamame, a type of food.
- a photo of tuna tartare, a type of food.
- a photo of hummus, a type of food.

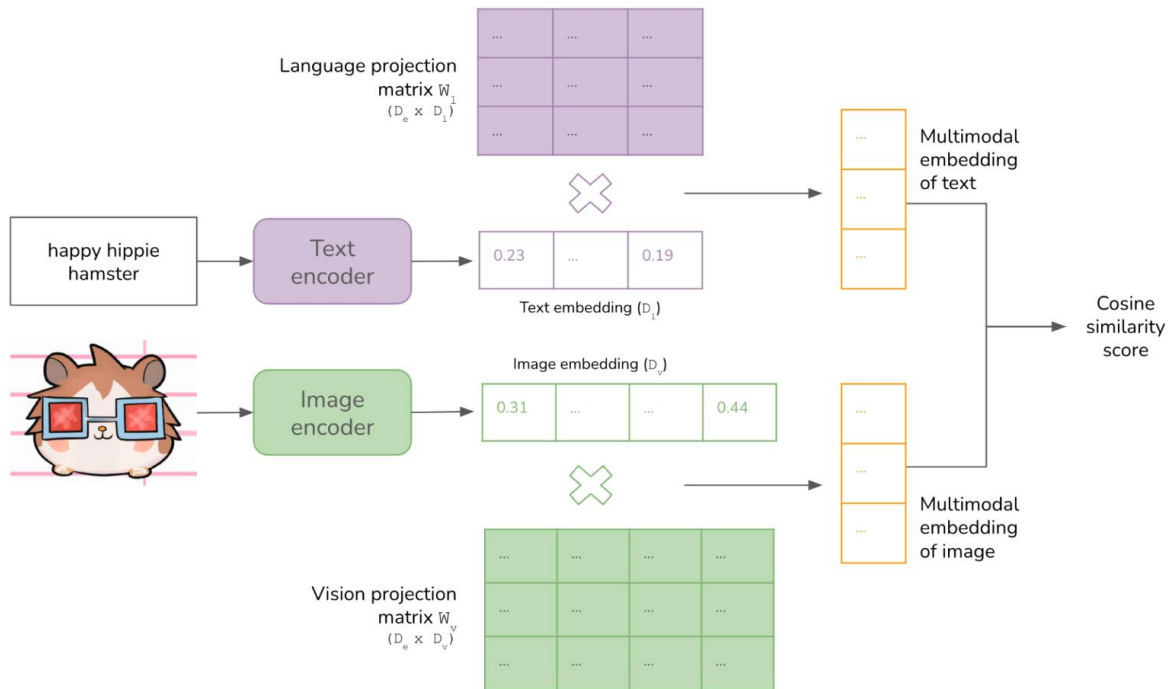
Youtube-BB
airplane, person (89.0%) Ranked 1 out of 23 labels



- a photo of a **airplane**.
- a photo of a bird.
- a photo of a bear.
- a photo of a giraffe.
- a photo of a car.

Zero-shot image classification with CLIP

CLIP's high-level architecture



CLIP's architecture. Both encoders and projection matrices are jointly trained together from scratch. The training goal is to maximize the similarity scores of the right (image, text) pairings while minimizing the similarity scores of the wrong pairings (contrastive learning).

Scalable Data

For many years, image models were trained with manually annotated (image, text) datasets (e.g. ImageNet, MS COCO). This isn't scalable. Manual annotation is time-consuming and expensive.

The CLIP paper noted that none of the then-available (image, text) datasets was big and high quality enough. They created their own dataset – 400M (image, text) pairs – as follows.

1. Construct a list of 500,000 queries. Queries are common words, bigrams, and titles of popular Wikipedia articles.
2. Find images matching these queries (string and substring match). The paper mentioned this search did NOT happen on search engines but didn't specify where. My theory is that since OpenAI already scraped the entire Internet for their GPT models, they probably just queried their internal database.
3. Each image is paired with a text that co-occurs with it (e.g. captions, comments) instead of the query since queries are too short to be descriptive.

Data Example: Wikipedia-based Image Text Dataset

The screenshot shows the Wikipedia article for "Half Dome". Red boxes highlight the following elements:

- Page Title:** "Half Dome" (with a red arrow pointing to the word "Dome")
- Page Description:** The main paragraph describing Half Dome as a granite dome in Yosemite National Park.
- Image:** A photograph of Half Dome with the caption "Sunset over Half Dome from Glacier Point".
- Reference Description:** The metadata section including Elevation (8846 ft), Prominence (1,360 ft), Parent peak (Clouds Rest), and Coordinates (37°44′46″N 119°31′59″W).

The article content includes:

Half Dome is a granite dome at the eastern end of Yosemite Valley in Yosemite National Park, California. It is a well-known rock formation in the park, named for its distinct shape. One side is a sheer face while the other three sides are smooth and round, making it appear like a dome cut in half.^[c] The granite crest rises more than 4,737 ft (1,444 m) above the valley floor.

Contents [hide]

- 1 Geology
- 2 Ascents
- 3 Hiking the Cable Route
- 4 Notable ascents
- 5 Notable free climbs
- 6 In culture
- 7 See also
- 8 References
- 9 External links

Geology [edit]

Main article: Geology of the Yosemite area

The impression from the valley floor that this is a round dome that has lost its northwest half, is just an illusion. From Washburn Point, Half Dome can be seen as a thin ridge of rock, an arête, that is oriented northeast-southwest, with its southeast side almost as steep as its northwest side except for the very top. Although the trend of this ridge, as well as that of Tenaya Canyon, is probably controlled by master joints, 80 percent of the northwest "half" of the original dome may well still be there.

Table of Metadata:

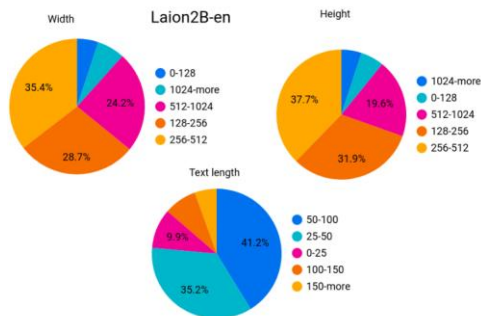
Reference Description	Highest point
Elevation	8846 ft (2696 m) NAVD 88 ^[1]
Prominence	1,360 ft (410 m) ^[1]
Parent peak	Clouds Rest ^[1]
Coordinates	37°44′46″N 119°31′59″W﻿ / ﻿37.74611°N 119.53306°W﻿ / 37.74611; -119.53306 ^[2]

Geography

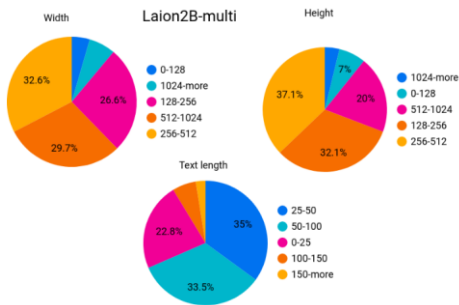
Even a lager one: Laion-5B

● Dataset columns

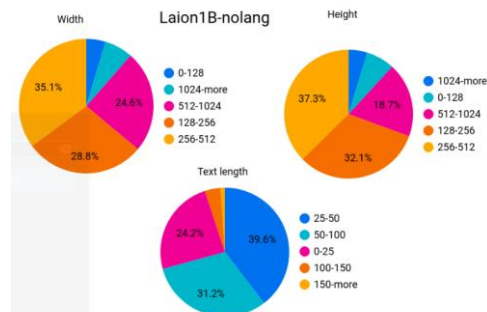
- **URL**: the image url, millions of domains are covered
- **TEXT**: captions, in english for en, other languages for multi and nolang
- **WIDTH**: picture width
- **HEIGHT**: picture height
- **LANGUAGE**: the language of the sample, only for laion2B-multi, computed using [cld3](#)
- **similarity**: cosine between text and image ViT-B/32 embeddings, clip for en, mclip for multi and nolang
- **pwatermark**: probability of being a watermarked image, computed using our [watermark detector](#)
- **punsafe**: probability of being an unsafe image, computed using our [clip based detector](#)



Average text length: 67



Average text length: 52

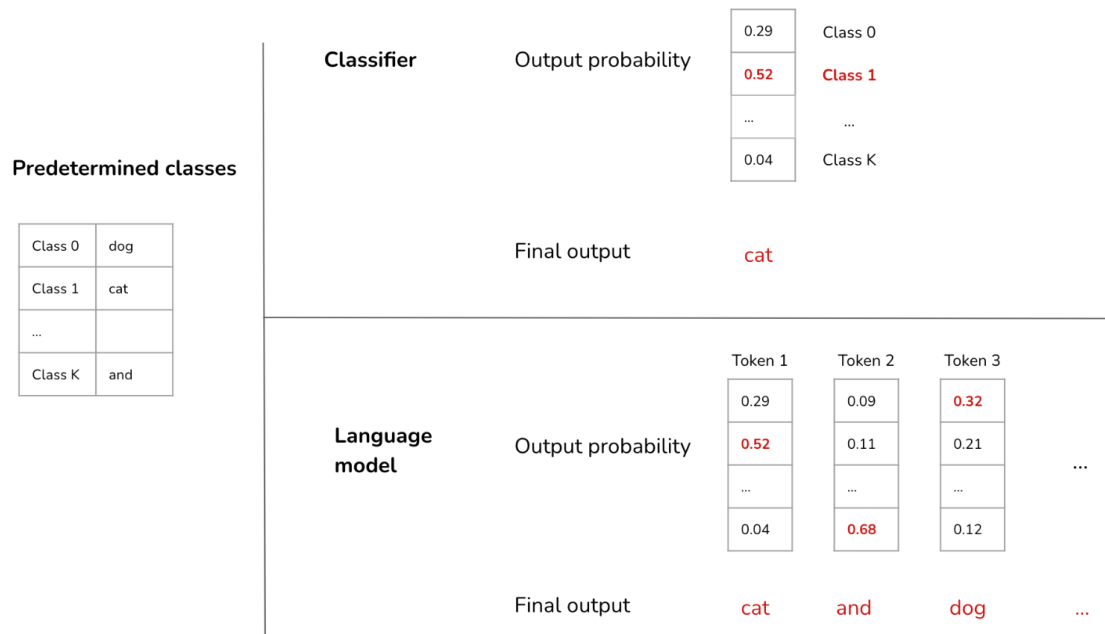


Average text length: 46

Existing Data

数据类型	数据种类	数据数量	压缩后数据大小	压缩格式	采集时间	数据说明
开源图像数据集	LAION-5B	50亿图文对	240TB	tar	2023年6月	/
	Laion400	4亿图文对	12TB	tar	2023年6月	采集地址: https://laion.ai/laion-400-open-dataset/ 图片大小256*256
	OBELICS	3亿图片文对	26TB	tar	2024年2月	采集地址: https://huggingface.co/datasets/HuggingFaceM4/OBELICS
	MMC4	3.2亿张图文对	26TB	tar	2024年1月	fewer faces v1.1版本 https://github.com/allenai/mmc4/blob/main/README.md
	Coyo-700m	4.1亿图文对	50TB	tar	2024年10月	分辨率两边不小于300像素
	MINT-1T	26亿图文对	272TB	tar	2024年8月	采集地址: https://huggingface.co/datasets/mlfoundations/MINT-1T-HTML
	BLIP3-KALE	1.83亿图文对	12.5TB	tar	2024年12月	使用网络采集文本来增强合成的图像captions, 以生成基于事实的图像captions, 可用于训练更强大、知识更丰富的多模态模型。数据集全量 2.18 亿个图文对, 因下载时部分链接失效, 实际落盘1.83亿图文对。
	CapsFusion-120M	1.2亿图文对	4TB	tar	2024年8月	采集地址: https://huggingface.co/datasets/BAAI/CapsFusion-120M
	Pixelprose	1500万图文对	3TB	tar	2024年8月	采集地址: https://huggingface.co/datasets/tomg-group-umd/pixelprose
商用图像数据集	谷歌艺术	400万图片	13.9TB	图片格式	2025年6月	最短边长大于1024P, 以高清晰度图像为主
	LFI-Gallery	100万图片	5TB	图片格式	2025年6月	Leica Fotografie International的gallery (画廊) 模块的所有图片。图片分辨率大于1080P比率较高
	马蜂窝 (游记图片)	1.5亿张游记图片	110TB	图片格式	2024年7月	游记图片, 以及图片和原始html游记 对应关系的meta信息。 另有游记文本数据集 (详见 社交互动sheet页): HTML格式, 保留所有原始信息, 纯文本, 含图片URL
	下厨房 (封面图+步骤图)	348万张封面图片 1881万张步骤图	7.5TB	图片格式	2025年6月	菜谱图片: 为解析处理后的数据, 提取了菜谱标题、详细烹饪步骤等信息。包括封面图和做菜步骤图, 且图片已下载。 另有菜谱文本数据集 (详见 文本垂类sheet页): 353个csv文件保存了, 菜谱的原始HTML信息, HTML包含 用料、详细烹饪步骤及配图URL等信息
	Sogou图库	280万张图文对	269GB	qz	/	包含原图, 缩略图, 链接, 文字说明
	生物百科	2200万张图片	12TB	zip	2025年6月	采集地址: https://www.inaturalist.org/ 含50万余个物种, 每个物种图片含meta信息, 并配有中英文维基百科、抖音百科、especies百科, 以及各个百科页面的图片
	植物百科	1015万张图片	3.5TB	zip	2025年9月	植物图像数据库, 知识图谱训练语料

Discriminative vs. generative for image classifications



If a classifier outputs only one class for each input, a language model outputs a sequence of classes. Each generated class is called a token. Each token is from a predetermined list, the vocabulary, of the language model.

Contrastive objective (CLIP)



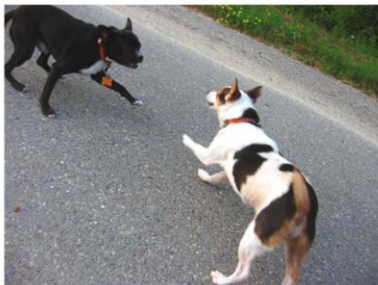
Several men in hard hats are operating a giant pulley system.

Workers look down from up above on a piece of equipment.

Two men working on a machine wearing hard hats.

Four men on top of a tall structure.

Three men on a large rig.



A black dog and a white dog with brown spots are staring at each other in the street.

A black dog and a tri-colored dog playing with each other on the road.

Two dogs of different breeds looking at each other on the road.

Two dogs on pavement moving toward each other.

A black dog and a spotted dog are fighting.

Texts for an image is **diverse**

While the language model objective allows for vastly more flexible outputs, CLIP authors noted this objective made the training difficult. They hypothesized that this is because the model tries to generate exactly the text accompanying each image, while many possible texts can accompany an image: alt-text, caption, comments, etc.

Contrastive objective (CLIP)

Contrastive learning is to overcome this challenge. Instead of predicting the exact text of each image, CLIP was trained to predict whether a text is more likely to accompany an image than other texts.

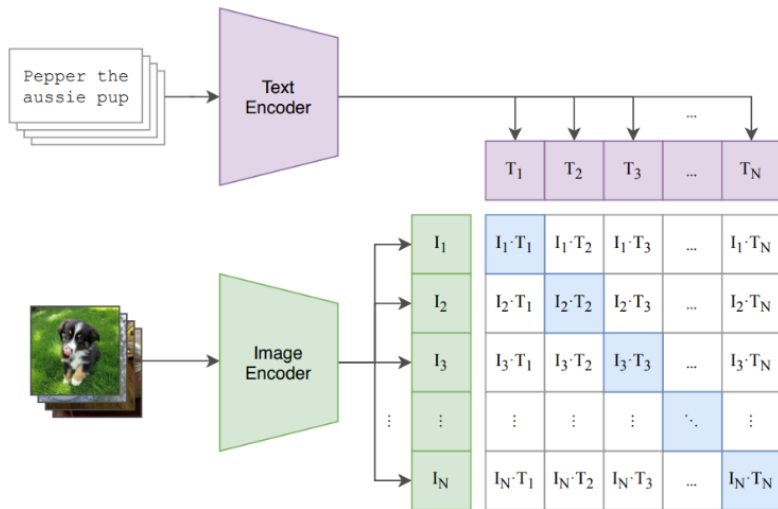
For each batch of N (image, text) pairs, the model generates N text embeddings and N image embeddings.

- Let V_1, V_2, \dots, V_n be the embeddings for the N images.
- Let L_1, L_2, \dots, L_n be the embeddings for the N texts.

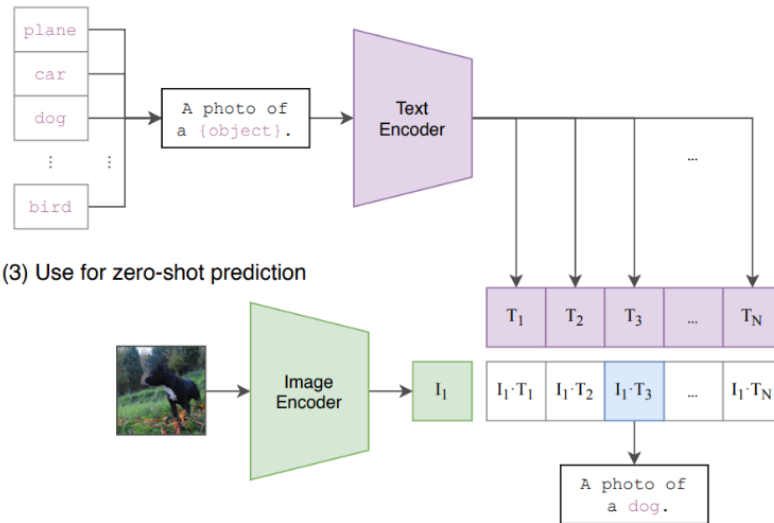
CLIP computes the cosine similarity scores of the N^2 possible (V_i, L_j) pairings. The model is trained to maximize the similarity scores of the N correct pairings while minimizing the scores of the $N^2 - N$ incorrect pairings. For CLIP, $N = 32,768$.

CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

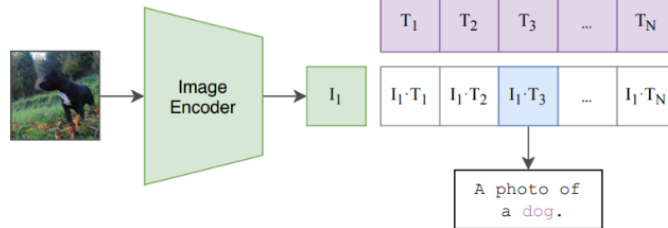
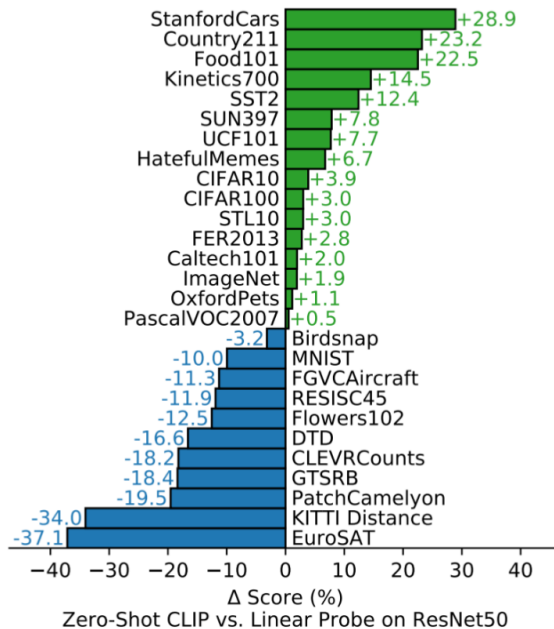


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

CLIP applications: classification

Today, for many image classification tasks, CLIP is still a strong out-of-the-box baseline to be used as-is or fine-tuned.



CLIP applications: text-based image retrieval

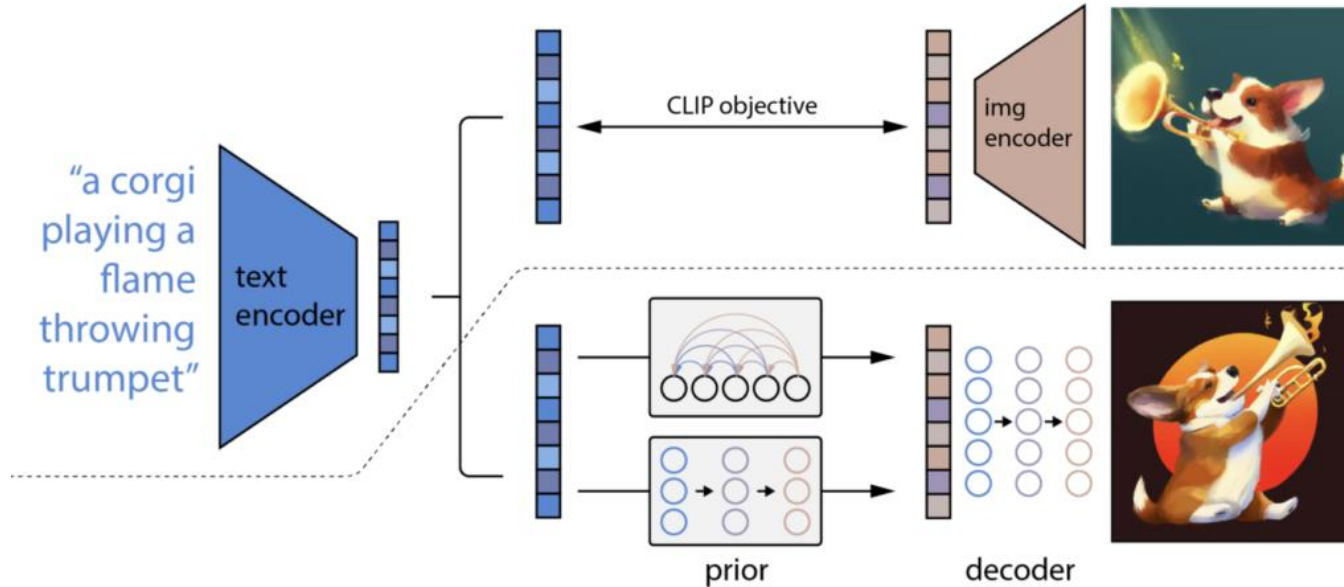
Text-based image retrieval

Since CLIP's training process was conceptually similar to image-to-text retrieval and text-to-image retrieval, CLIP "*displays significant promise for widely-applicable tasks like image retrieval or search.*" However, "*on image retrieval, CLIP's performance relative to the overall state of the art is noticeably lower.*"

There are attempts to use CLIP for image retrieval. For example, [clip-retrieval](#) package works as follows:

1. Generate CLIP embeddings for all your images and store them in a vector database.
2. For each text query, generate a CLIP embedding for this text.
3. Query in the vector database for all images whose embeddings are close to this text query embedding.

CLIP applications: image generation



- [\[CLIP\] Learning Transferable Visual Models From Natural Language Supervision](#) (OpenAI, 2021)
- https://tryolabs.com/blog/2022/08/31/from-dalle-to-stable-diffusion?utm_source=blog&utm_medium=edgeAlliance&utm_campaign=edgeAlliance&utm_id=edgeAlliance

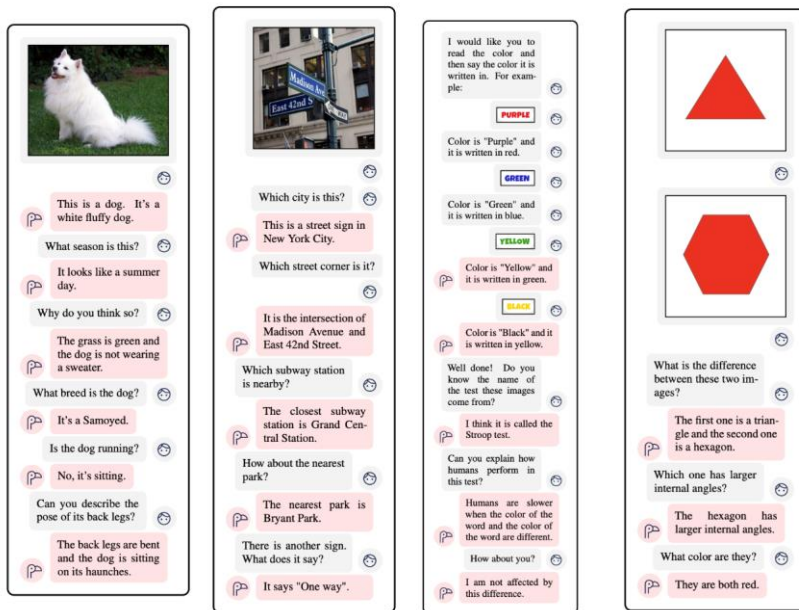
Early Work II: Flamingo, the dawns of LMMs

a good practice to enable “talking on images” with generative language models

Flamingo: the dawns of LMMs (多模态大模型的曙光)

In a reductive view, **Flamingo** = **CLIP** + **An autoregressive language model**

- Adding a decoder LM. Unlike CLIP, Flamingo can generate text responses.
- Attention interaction between image encoder and LM



Flamingo can generate text responses conditioned on both text and images

Flamingo's high-level architecture

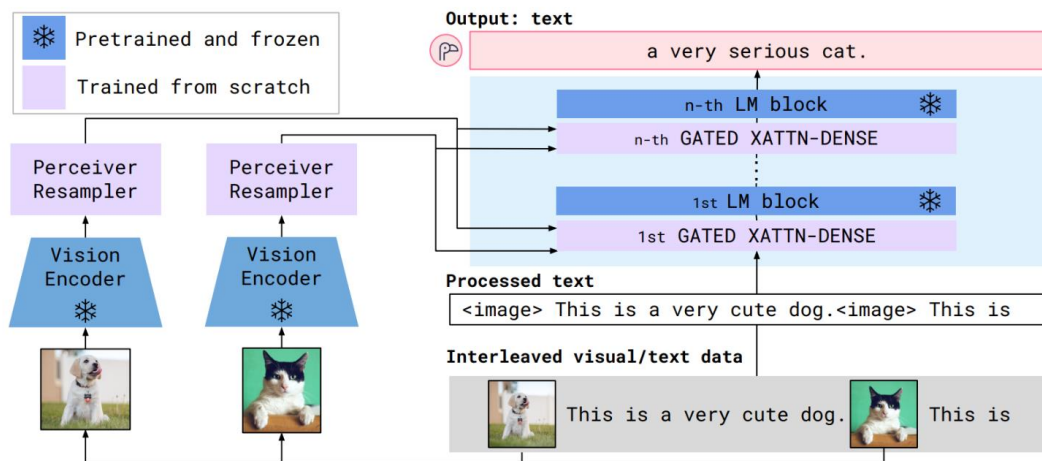


Figure 3: **Flamingo architecture overview.** Flamingo is a family of visual language models (VLMs) that take as input visual data interleaved with text and produce free-form text as output.

At a high level, Flamingo consists of **2 parts**:

- Vision encoder: a CLIP-like model is trained using contrastive learning. The text encoder of this model is then discarded. The vision encoder is frozen to be used in the main model.
- Language model: Flamingo finetunes Chinchilla to generate text tokens, conditioned on visuals and text, using language model loss, with two additional components Perceiver Resampler and GATED XATTN-DENSE layers.

Data

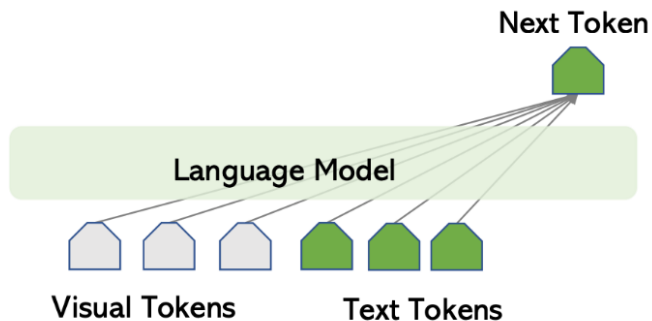
Flamingo used 4 datasets: 2 (image, text) pair datasets, 1 (video, text) pair dataset, and 1 interleaved image and text dataset.



Dataset	Type	Size	How	Training weight
M3W	Interleaved image and text dataset	43M webpages	For each webpage, they sample a random subsequence of 256 tokens and take up to the first 5 images included in the sampled sequence.	1.0
ALIGN	(Image, text) pairs	1.8B pairs	Texts are alt-texts, averaging 12 tokens/text.	0.2
LTIP	(Image, text) pairs	312M pairs	Texts are long descriptions, averaging 20.5 tokens/text.	0.2
VTP	(Video, text) pairs	27M short videos	~22 seconds/video on average	0.03

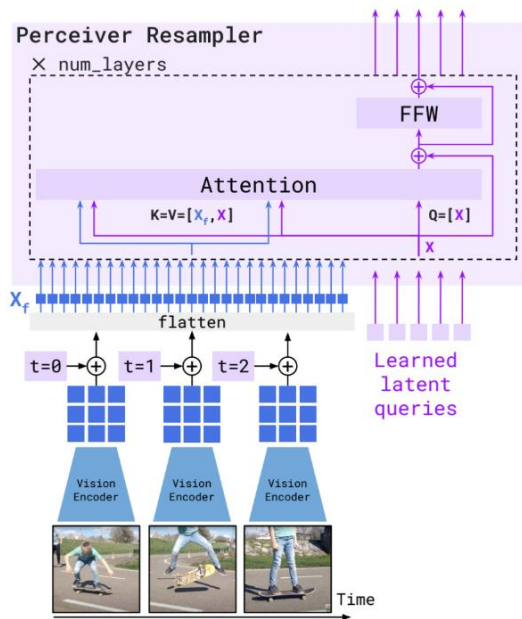
Flamingo's language model

Flamingo uses Chinchilla as their language model. More specifically, they freeze the 9 pretrained Chinchilla LM layers. A traditional language model predicts the next text token based on the preceding text tokens. Flamingo predicts the next text token based on both the preceding text and visual tokens.



Perceiver Resampler

As the visual inputs can be both images and videos, the vision encoder can produce a variable number of image or video features. Perceiver Resampler converts these variable features into a consistent 64 visual outputs.



```
def perceiver_resampler(  
    x_f, # The [T, S, d] visual features (T=time, S=space)  
    time_embeddings, # The [T, 1, d] time pos embeddings.  
    x, # R learned latents of shape [R, d]  
    num_layers, # Number of layers  
):  
    """The Perceiver Resampler model."""  
  
    # Add the time position embeddings and flatten.  
    x_f = x_f + time_embeddings  
    x_f = flatten(x_f) # [T, S, d] -> [T * S, d]  
    # Apply the Perceiver Resampler layers.  
    for i in range(num_layers):  
        # Attention.  
        x = x + attention_i(q=x, kv=concat([x_f, x]))  
        # Feed forward.  
        x = x + ffw_i(x)  
    return x
```

GATED XATTN-DENSE layers

GATED XATTN-DENSE layers are inserted between existing and frozen LM layers to allow the language model to attend more efficiently to the visual tokens when generating text tokens. Without these layers, Flamingo authors noted a drop of 4.2% in the overall score.

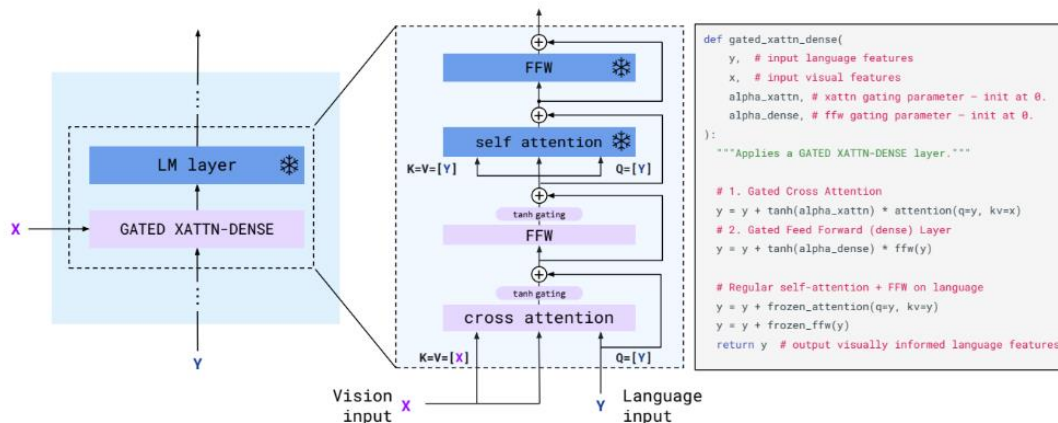


Figure 4: **GATED XATTN-DENSE layers.** To condition the LM on visual inputs, we insert new cross-attention layers between existing pretrained and frozen LM layers. The keys and values in these layers are obtained from the vision features while the queries are derived from the language inputs. They are followed by dense feed-forward layers. These layers are *gated* so that the LM is kept intact at initialization for improved stability and performance.

Loss function: autoregressive language models

Flamingo computes the likelihood of text y conditioned on the interleaved images and videos x .

$$p(y|x) = \prod_{l=1}^N p(y_l | y_{<l}, x_{\leq l})$$

The training loss function was a weighted sum of expected negative log-likelihoods of generated text across all 4 datasets, with λ_m being the training weight of dataset m .

$$\sum_{m=1}^M \lambda_m E_{(x,y) \sim D_m} \left[- \sum_{l=1}^L \log p(y_l | x) \right]$$

Flamingo Training

While Flamingo isn't open-sourced, there are many open-source replications of Flamingo.

- [IDEFICS](#) (HuggingFace)
- [mlfoundations/open_flamingo](#)

CLIP vs. Flamingo

		CLIP (2021)	Flamingo (2022)
Data		<p>- WIT (WebImageText): 400M (image, text) pairs scraped from the Internet.</p> <p>500K queries Up to 20K (image, text) per query</p>	<p>4 datasets:</p> <ul style="list-style-type: none"> - ALIGN: 1.8B (image, text) pairs. Avg. 12 tokens/text. - LTIP (Long Text & Image Pairs): 312M pairs. Avg. 20.5 tokens/text. - VTP (Video & Text Pairs): 27M (short video, text) pairs. Avg. 22 seconds/video. - M3W (MultiModal MassiveWeb): 43M webpages with interleaved images and text. Up to 256 tokens and 5 images per page.
Natural language supervision for vision encoder	Text encoder	CBOW or text transformer (a smaller version of GPT-2)	BERT. It's only used to train the vision encoder and discarded after.
	Vision encoder	ResNet or Vision Transformer (ViT)	NormalizerFree ResNet (NFNet). Once trained, it's frozen before being used in Flamingo's main model.
	Training	<ul style="list-style-type: none"> - Text and image embeddings are projected into a common multimodal embedding space using linear projections - Text and image encoders are jointly trained from scratch using contrastive objectives. 	Same contrastive objectives as CLIP
Language model		X	<ul style="list-style-type: none"> - Chinchilla (9 layers) - Perceiver Resampler maps from varying-sized visual features to a fixed number of visual tokens (64). - GATED XATTN-DENSE layers are inserted between Chinchilla LM layers to allow the LM to attend to visual tokens when generating text tokens. - Chinchilla is pretrained and frozen. Perceiver Resampler and GATED XATTN-DENSE layers are trained from scratch.

Multimodal LLMs


— Welcome to the Era of Large Language Models.

(Large) language models really improve the **cognition**-level abilities (e.g., reasoning and planning), but it also needs **perceiving** multi-modal input.

Multimodal LLMs

LLaVA Chatbot

what is written? who could be interested in the content of the screenshot?



The screenshot displays a cityscape with a large city skyline and a sunset over a body of water. The image also contains text, which says "recent advances in vision foundation models." This content might be of interest to researchers, academics, or professionals in the fields of computer vision, artificial intelligence, or urban planning, as they could be looking to gain insights or stay updated on recent advancements in these areas. The combination of the visually appealing cityscape and the technical information suggests that the image might be related to a conference or a presentation on the topic.

Enter text and press ENTER

Submit

Visual chat with LMM. Generated by LLaVA: <https://llava-vl.github.io/>

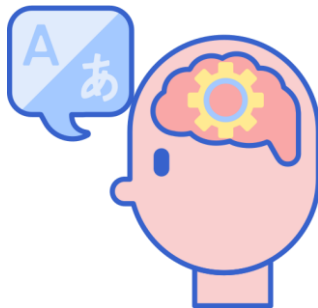
Preliminary Idea: Intelligence over Language

Emergent phenomena have extensively already occurred in language-based LLMs.



These LLMs now generally possess very powerful **semantic understanding capabilities**.

This also implies that **language is a crucial modality for carrying intelligence**.



language

Preliminary Idea: Language Intelligence as Pivot



Given this premise, **nearly all CURRENT MLLMs are built based on language-based LLMs** as the core decision-making module (i.e., the brain or central processor).



By adding additional external non-textual modality modules, LLMs are enabled with multimodal abilities.

- Extend the capability boundary, next milestone towards more advanced intelligence
- More applications



Language for communication and also **thought**



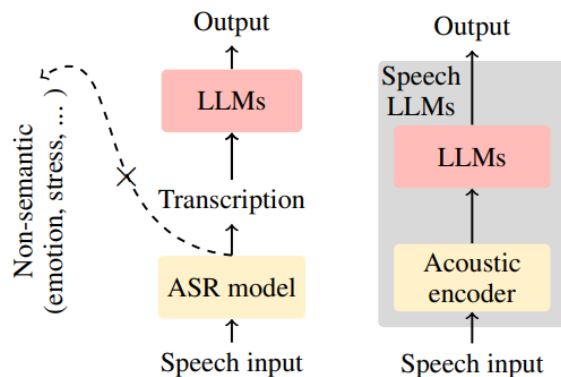
“There is no reason to believe ... that the “essential purpose” of language is “communication”. Language can be used to transmit information, but it also serves many other purposes: to establish relations among people, to express or clarify thought, for play, for creative mental activity, to gain understanding, and so on. In my opinion, there is no reason to accord privileged status to one or the other of these modes.”

~ NOAM CHOMSKY

A recent perspective paper from Nature: Language is primarily a tool for communication rather than thought. <https://www.nature.com/articles/s41586-024-07522-w>

An interesting question:

- Can Speech2speech models works generally without involving language/words!



Some people argues that speech-to-speech is cool; but without language, it probably cannot think (reason and plan)!

An interesting question:

- Can a speech2speech end2end model answer this question well?

Count from one to one hundred, pausing whenever you encounter a prime number.

Output 1, 2 [pause], 3 [pause], 4, 5 [pause], 6, 7 [pause], 8, 9, 10, 11 [pause], 12, 13 [pause], 14, 15, 16, 17 [pause], 18, 19 [pause], 20, 21, 22, 23 [pause], 24, 25, 26, 27, 28, 29 [pause], 30, 31 [pause], 32, 33, 34, 35, 36, 37 [pause], 38, 39, 40, 41 [pause], 42, 43 [pause], 44, 45, 46, 47 [pause], 48, 49, 50, 51, 52, 53 [pause], 54, 55, 56, 57, 58, 59 [pause], 60, 61 [pause], 62, 63, 64, 65, 66, 67 [pause], 68, 69, 70, 71 [pause], 72, 73 [pause], 74, 75, 76, 77, 78, 79 [pause], 80, 81, 82, 83 [pause], 84, 85, 86, 87, 88, 89 [pause], 90, 91, 92, 93, 94, 95, 96, 97 [pause], 98, 99, 100

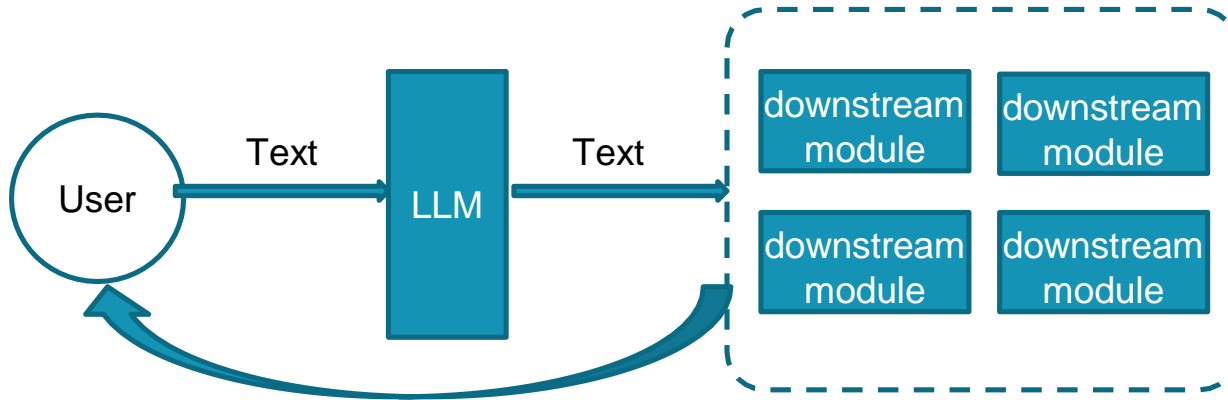
Without textual planning, speech2speech might not solve this problem.

Type-I: LLM as Discrete Scheduler/Controller

The role of the LLM is to **receive textual signals** and **instruct textual commands to call downstream modules**.

+ Key feature:

All message passing within the system, such as “multimodal encoder to the LLM” or “LLM to downstream modules”, is facilitated through **pure textual** commands as the medium.



Type-I : LLM as Discrete Scheduler/Controller

+ Representative MLLMs:

+ Visual-ChatGPT

+ HuggingGPT

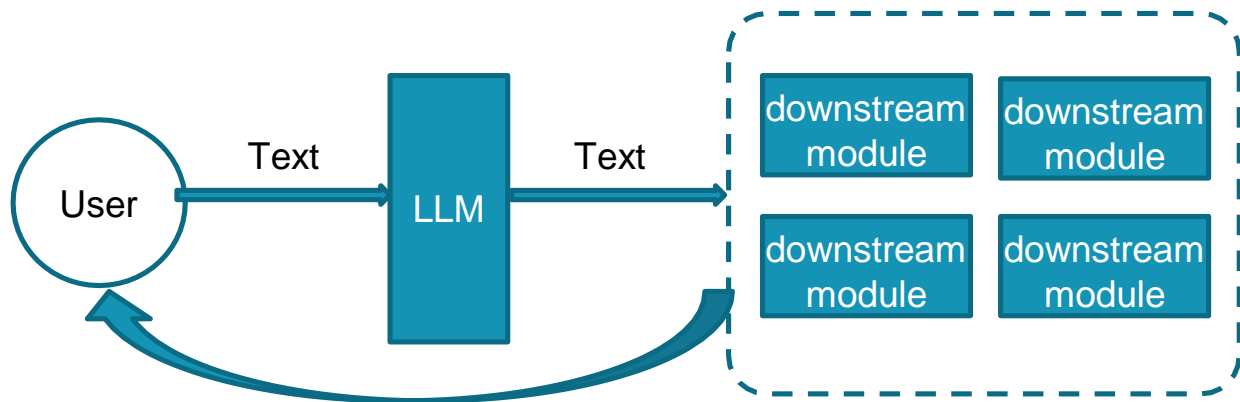
+ MM-REACT

+ ViperGPT

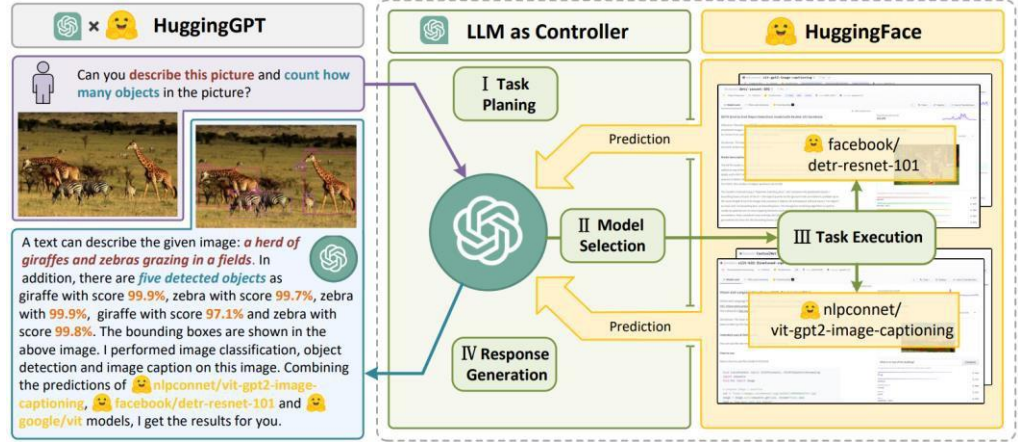
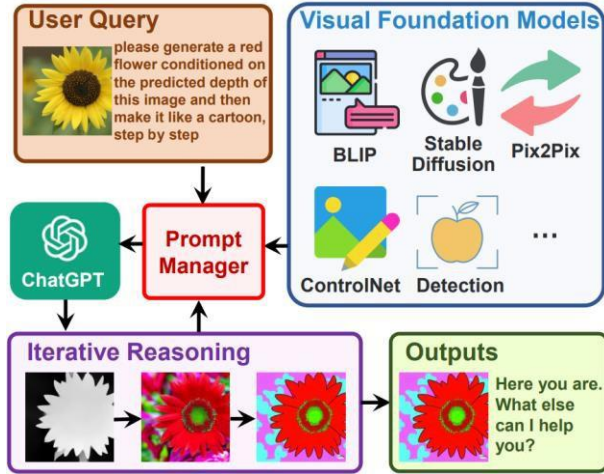
+ AudioGPT

+ LLaVA-Plus

+ ...



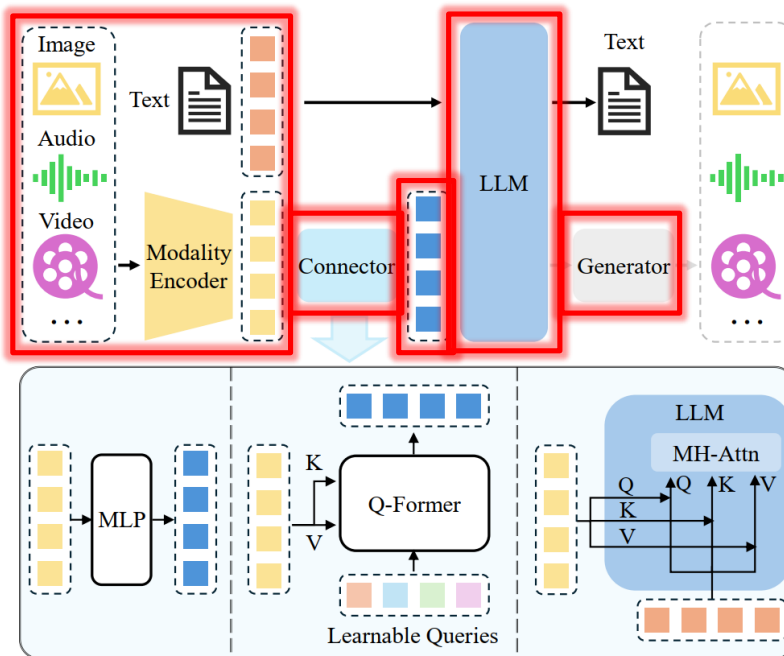
Type-I : LLM as Discrete Scheduler/Controller



- + Quick to build (without training), flexible extension to many tool features
- + Information loss in text medium, the bottle-neck

Type-II: LLM as Joint Part of System

- + > 90% MLLMs belong to this category.
- + Higher upper-bound, better integrated into a unified model



[1] A Survey on Multimodal Large Language Models.
https://github.com/BradyFU/A_wesome-Multimodal-Large-Language-Models, 2023.

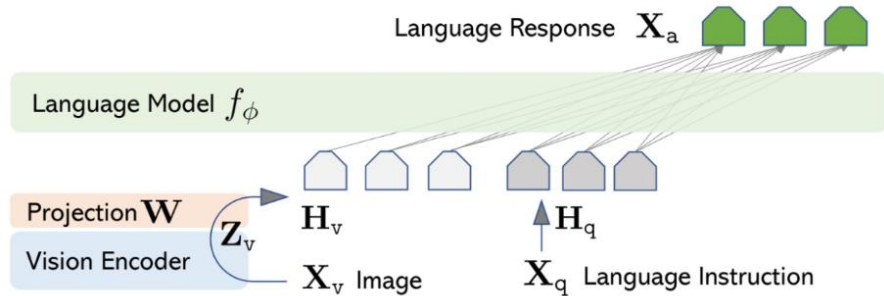
Training LLMs

- Architecture
- Visual encoders or tokenizer
- Multi-modal instruction tuning
- Multi-stage training

Architecture of LLMs

LLaVA: Large Language-and-Vision Assistant

□ Architecture



□ Two-stage Training

•Stage 1: Pre-training for Feature Alignment.

Only the projection matrix is updated, based on a subset of CC3M.

•Stage 2: Fine-tuning End-to-End. Both the projection matrix and LLM are updated

•**Visual Chat**: Our generated multimodal instruction data for daily user-oriented applications.

•**Science QA**: Multimodal reasoning dataset for the science domain.

Other Multimodal LLMs: Qwen-VL

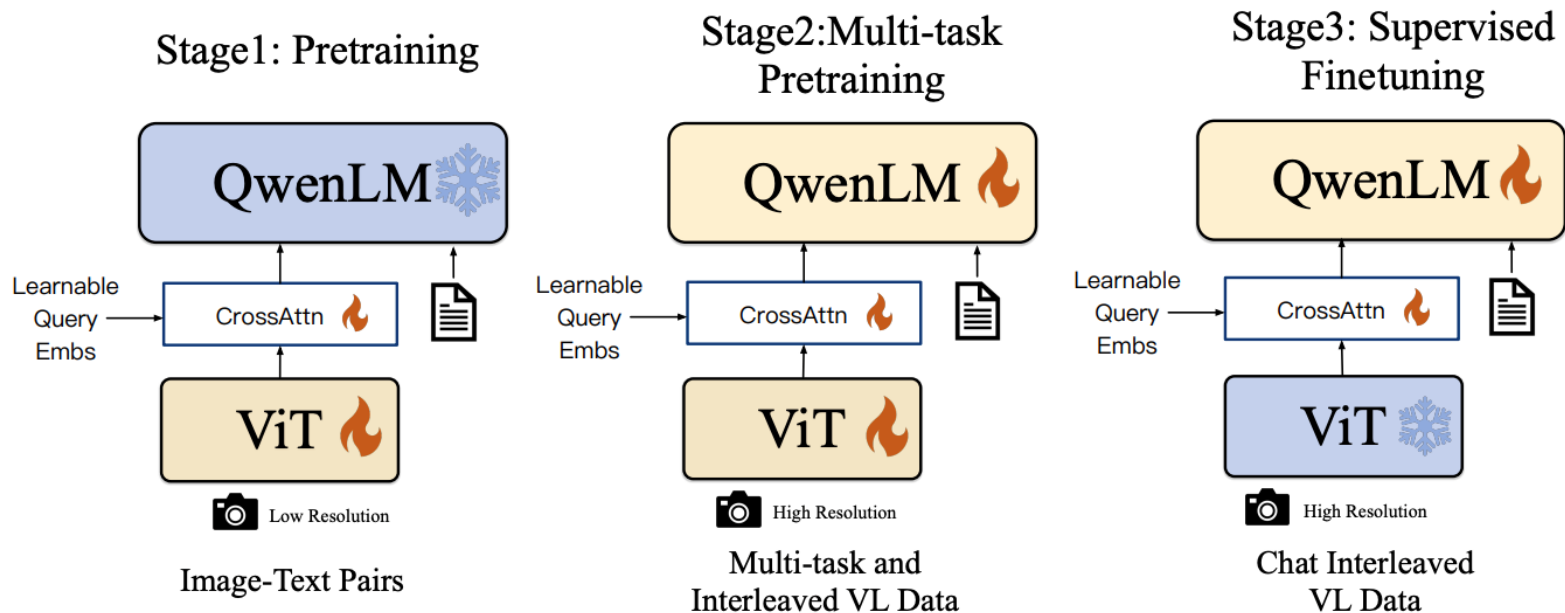


Figure 3: The training pipeline of the Qwen-VL series.

Other Multimodal LLMs: CogVLM-Chat

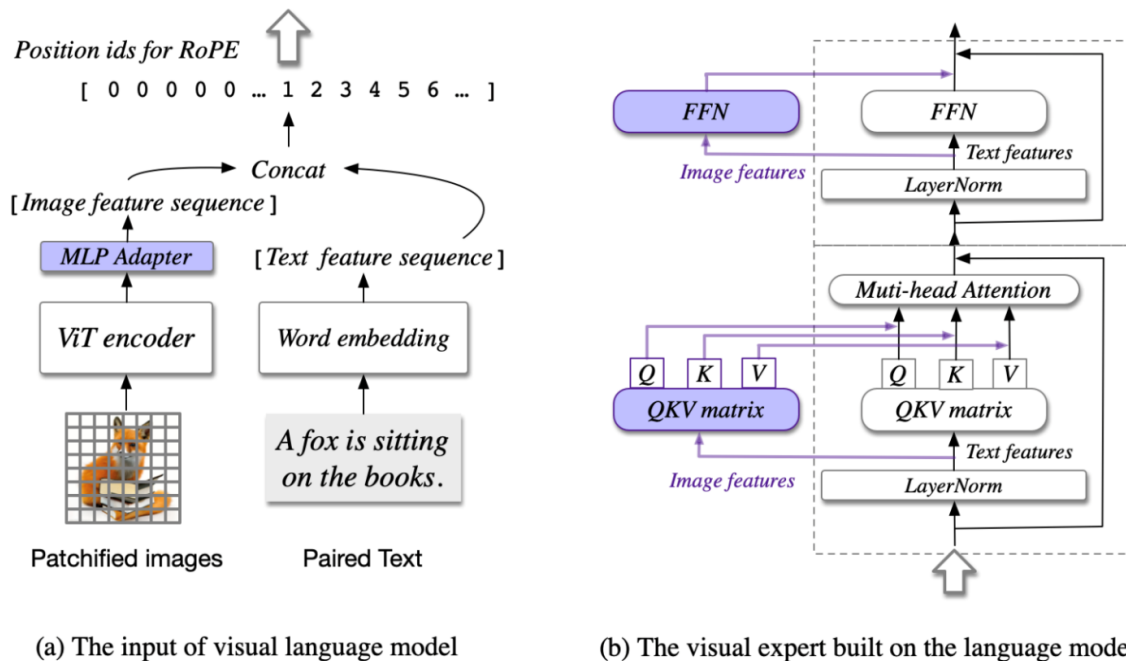
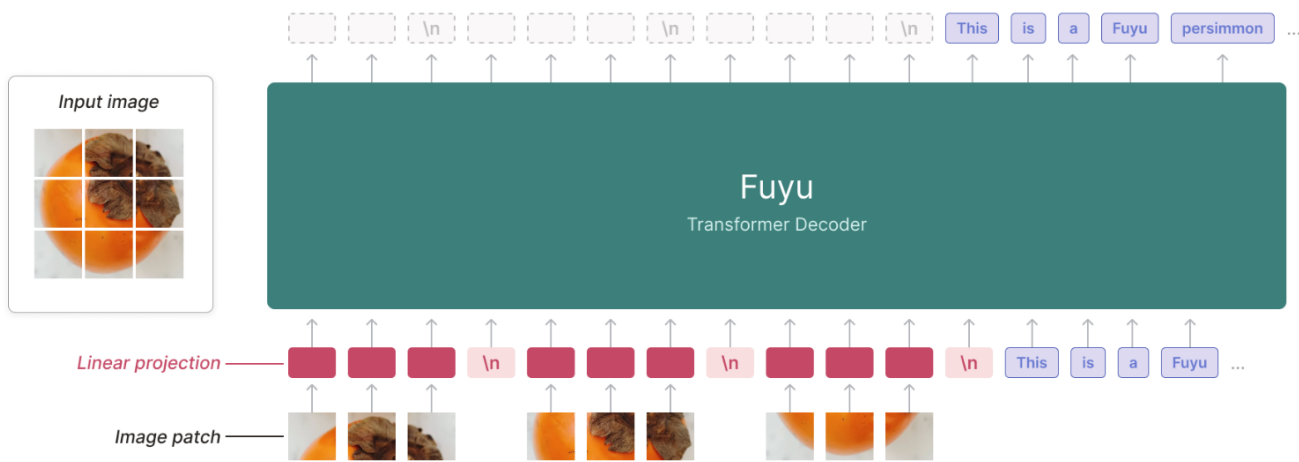


Figure 3: The architecture of CogVLM. (a) The illustration about the input, where an image is processed by a pretrained ViT and mapped into the same space as the text features. (b) The Transformer block in the language model. The image features have a different QKV matrix and FFN. Only the purple parts are trainable.

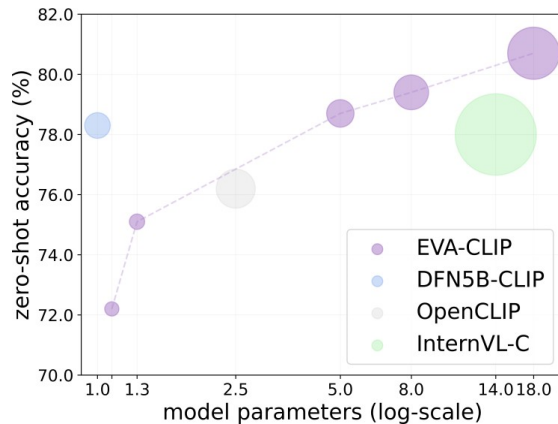
Other Multimodal LLMs: Fuyu



Visual encoders or tokenizer

Visual Encoder

- + **CLIP-ViT** is the most popular choice for vision-language models.
 - × Providing image representations well aligned with text space.
 - × Scale well with respect to parameters and data.
- + **SigLIP** is gaining increasing popularity (smaller and stronger)



Visual Encoder

- + Limitations of existing pretrained ViTs:
 - × Fixed low-resolution (224x224 or 336x336) in square shape
- + High-resolution perception is essential, especially for OCR capability!



Low resolution encoding misses fine-grained visual details!

Visual Encoder

+ High-resolution Multimodal LLMs

× Image slice-based: Split high-resolution images into slices

× Representatives:

- ◆ GPT-4V, LLaVA-NeXT, MiniCPM-V 2.0/2.5, LLaVA-UHD, mPLUG- DocOwl 1.5, SPHINX, InternLM-XComposer2-4KHD, Monkey

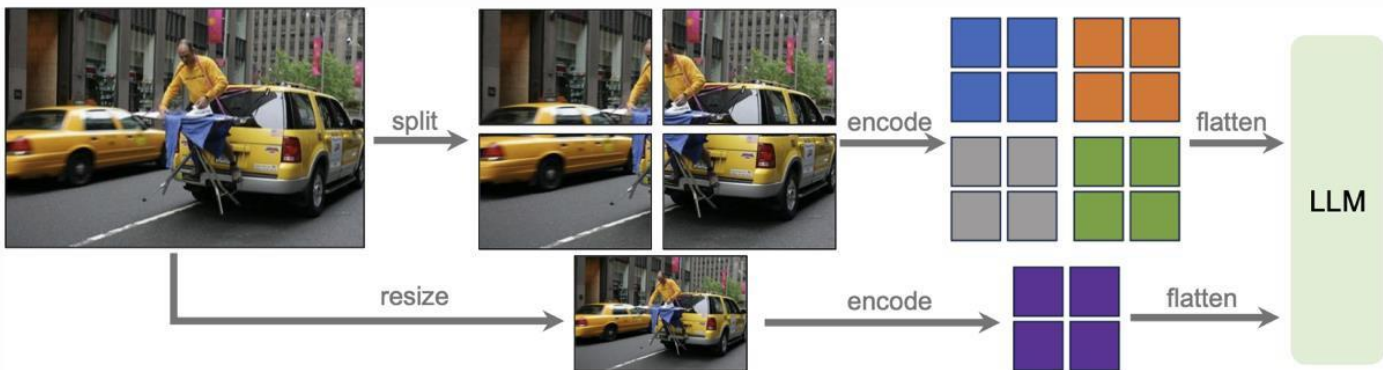


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

Non-Visual Encoder

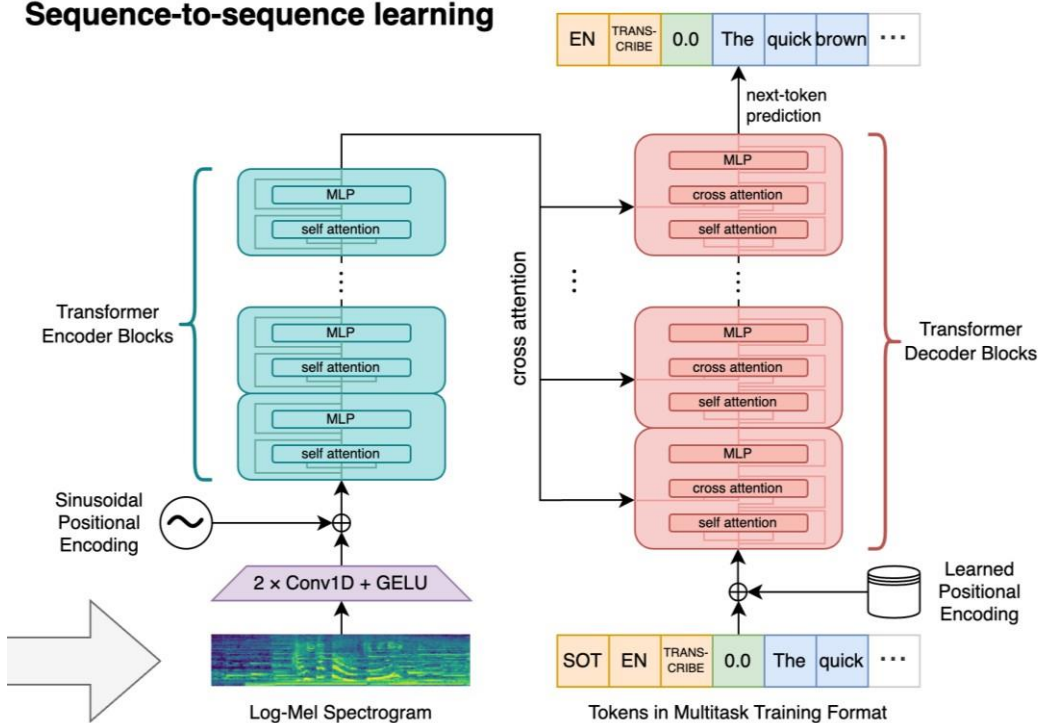
+ Audio:

- × Whisper
- × AudioCLIP
- × HuBERT
- × BEATs

+ 3D Point:

- × Point-BERT

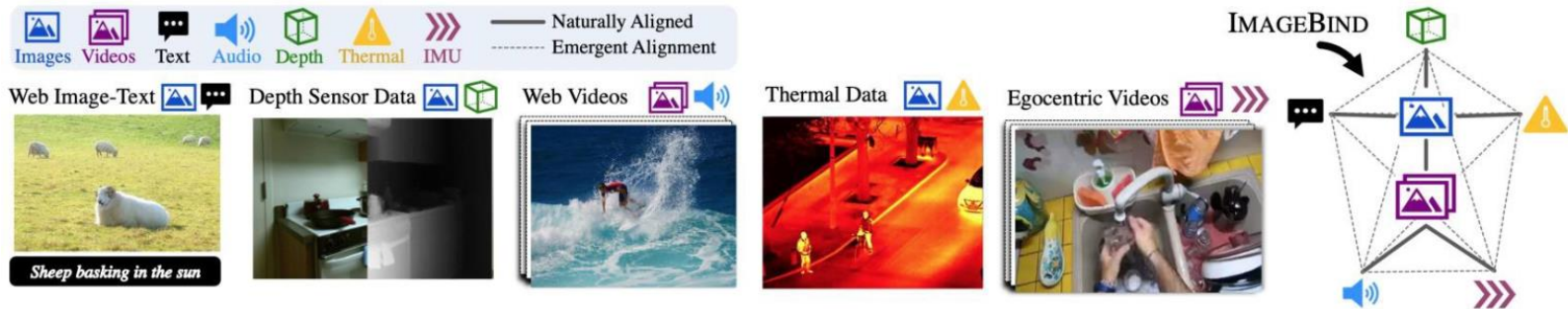
Sequence-to-sequence learning



Unified Multimodal Encoder

+ ImageBind:

- × Embedding all modalities into a joint representation space of **Image**.
- × Well aligned modality representations can benefit LLM understanding

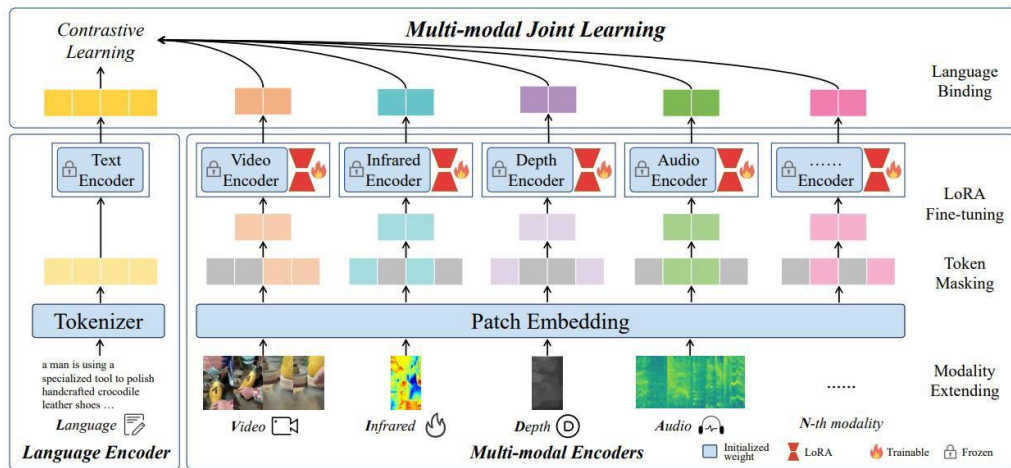


[1] ImageBind: One Embedding Space To Bind Them All. 2023

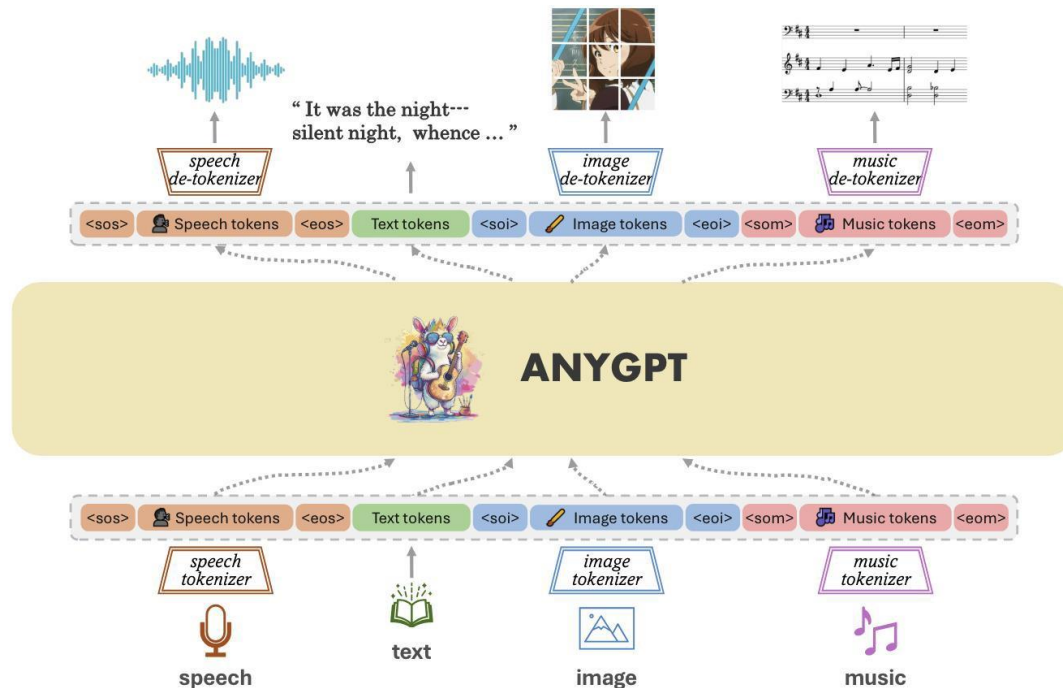
Unified Multimodal Encoder

+ LanguageBind:

- × Embedding all modalities into a joint representation space of **Language**.
- × Well aligned modality representations can benefit LLM understanding



Tokenization

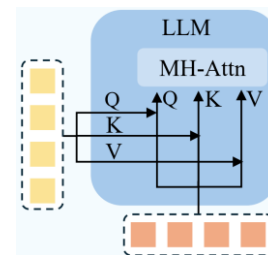
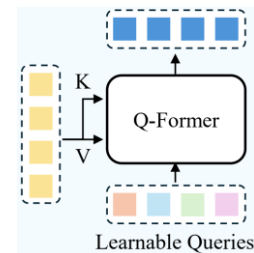
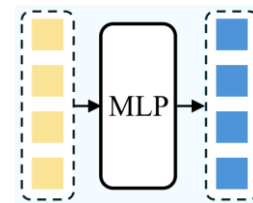


✦ Represent multimodal signals as discrete tokens in a codebook

✦ **Advantages:** support **unified** multimodal signal **understanding** and **generation** in an auto-regressive next-token prediction framework

Tokenization

- **Methods to Connect Multimodal Representation with LLM**
 - + Projecting multimodal (e.g., image) representations into LLM semantic space
 - × Q-Former: BLIP-2, InstructBLIP, VisCPM, VisualGLM
 - × Linear projection: LLaVA, MiniGPT-4, NExT-GPT
 - × Two-layer MLP: LLaVA-1.5/NeXT, CogVLM, DeepSeek-VL, Yi-VL
 - × Perceiver Resampler: Flamingo, Qwen-VL, MiniCPM-V, LLaVA-UHD
 - × C-Abstractor: HoneyBee, MM1



Instruction Tuning in Multi-modal LLM

Recap on Language Modeling



GPT-2



GPT-3



ChatGPT
InstructGPT



GPT-4

What's new?

In-context-learning
Chain-of-thoughts (CoT)

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following
Multimodal Input with image

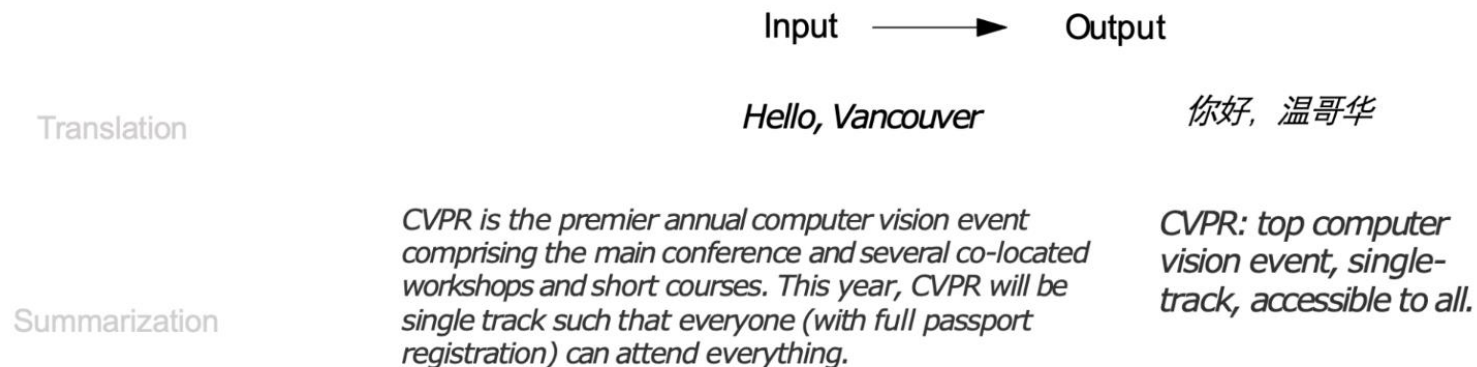
Multimodal
Space

CLIP
Flamingo

Gap?
Instruction-Following
→ **Alignment Research**

Multimodal GPT-4

Instruction Tuning in LLM



- Task instructions are implicit.
- Individual models are trained, or multi-tasking without specifying the instructions
- Hard to generalize to new tasks in zero-shot

Instruction Tuning in LLM



GPT-2



GPT-3



ChatGPT
InstructGPT



GPT-4

What's new?

In-context-learning
Chain-of-thoughts (CoT)

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following

In-context-learning
Chain-of-thoughts (CoT)
Instruction-Following
Multimodal Input with image

**Open Source
Community**

LLaMA



Alpaca



GPT4-Alpaca








Vicuna



Tulu



Instruction Tuning in LLM

	LLaMA 	Alpaca 	Vicuna 	GPT4-Alpaca 	...	Tulu 
Data Source		GPT-3.5	ShareGPT (Human & GPT)	GPT-4 (text-only)	...	Mixed Data
Instruction- following Data (#Turns)	None	52K	500K (~150K conversions)	52K	...	

Self-Instruct with Strong Teacher LLMs & Mixed Human Data

Instruction Tuning in Multimodal LLM

Visual Instruction Tuning with GPT-4

<https://l1ava-vl.github.io/>

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee (* Equal contribution)

Self-Instruct with Strong Teacher LLMs

But No Teacher is available on multiGPT4?

	LLaMA	Alpaca	Vicuna
Teacher			
		GPT-3.5	ShareGPT (Human & GPT)
Instruction-following Data	None	52K	700K (70 conversions)

GPT-4-LLM



GPT-4
(text-only)

LLaVA



GPT-4
(text-only)

- 158K multimodal instruction following data (First & High Quality)

—————▶ Multimodal Chatbot

GPT-assisted Visual Instruction Data Generation

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples

→ Text-only GPT-4

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.

Luggage surrounds a vehicle in an underground parking area

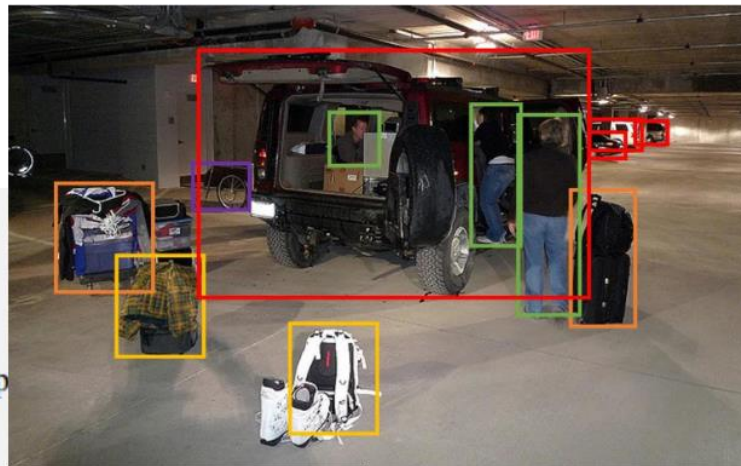
People try to fit all of their luggage in an SUV.

The sport utility vehicle is parked in the public garage, being packed for a trip

Some people with luggage near a van that is transporting it.

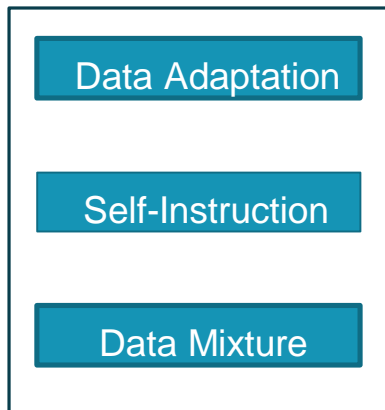
Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

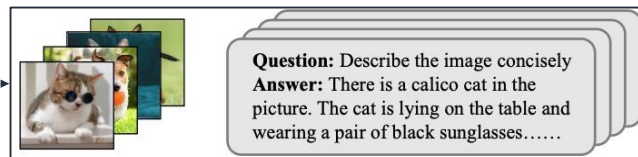


MLLM Instruction Tuning Framework

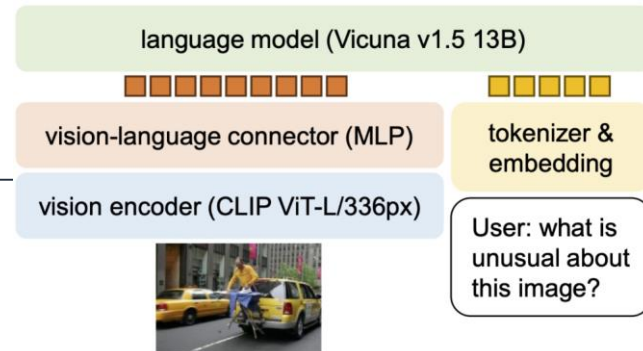
Data Construction



Visual Instruction-following data



Visual Instruction Tuning Framework Example: LLaVA-1.5



Multi-stage training in Multimodal LLMs

Training Data

- Image-text paired data (i.e. captions)
- Visual instruction tuning data (Visual QA)

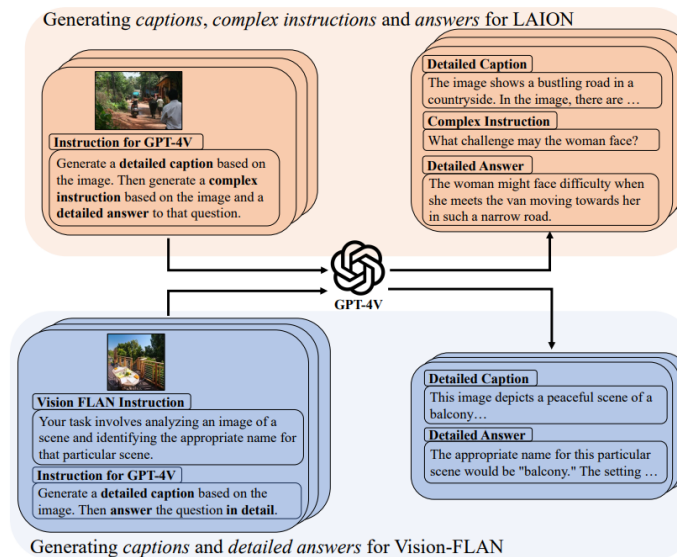


Figure 1: Pipeline for scaling up high-quality data. Prompts in the figure are shown for demonstration purpose. See the detailed prompt in Appendix A.2 and A.3.

Pretraining Alignment Data (i.e. image caption)



COCO: Young children standing on a platform waiting for a train to arrive. Adults and children watching a train slowly leave. A family near a railroad track watching the train pass. People waiting on a platform as a train pulls up. A train station with a green train on the tracks and children waiting for it to go by.

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

A red signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machines that frequent these tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.

The sky above is a clear blue, providing a serene backdrop to the bustling activity of the station. In the distance, trees can be seen, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

Coarse-gained Image-text

Coarse-gained Image-text

Fine-gained Image-text

Pretraining Alignment Data (i.e. image caption)

- **Coarse-gained Image-text**
Data volume is **large**, the captions are **shorts** and **noisy**.

- **Fine-gained Image-Text**
High quality, **longer** and **more accurate descriptions**, fine-gained alignment between different modalities.

Dataset	Samples	Taxonomies
CC-3M	3.3M	Coarse-grained Image-Text
CC-12M	12.4M	Coarse-grained Image-Text
LAION-5B	5.9B	Coarse-grained Image-Text
SBU-Captions	1M	Coarse-grained Image-Text
ShareGPT4V-PT	1.2M	Fine-grained Image-Text
LVIS-Instruct4V	111k	Fine-grained Image-Text
<u>ALLaVA</u>	709k	Fine-grained Image-Text
MSR-VTT	200k	Video-Text
<u>WavCaps</u>	24k	Audio-Text
MMC-Instruction	600k	Chart-Text

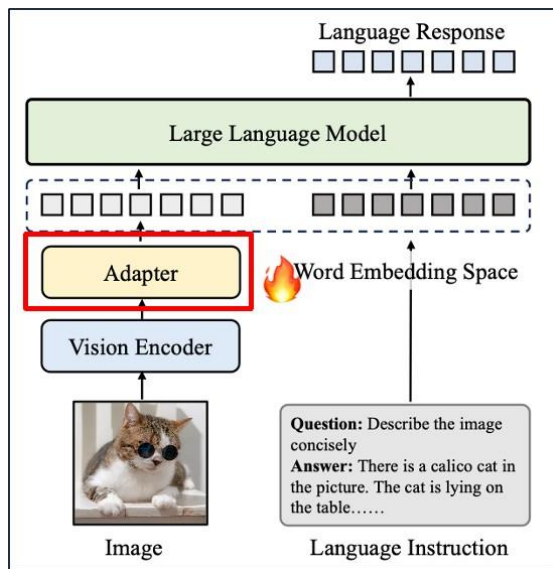
Instruction Tuning Data (visual QA)

Dataset	Size	Modalities	Constructions
LLaVA-Instruct-158k	158k	Image, Text	ChatGPT-generated
LRV-Instruction	400k	Image, Text	GPT4-generated
MMC-Instruction	600k	Chart, Text	GPT4-generated/adapted
Clotho-Detail	3.9k	Text, Audio	GPT4-generated
MACAW-LLM	119k	Image, Video, Text	GPT-3.5-turbo-generated
MIMIC-IT	2.8M	Image, Video, Text	ChatGPT-generated
StableLLaVA	126k	Image, Text	StableDiffusion & ChatGPT-generated
LAMM	196k	Image, PointCloud, Text	GPT4-generated
VIGC-LLaVA	1.8M	Image, Text	Model-generated
ALLaVA-4V	1.3M	Image, Text	GPT-4V

The Typical Training Paradigm

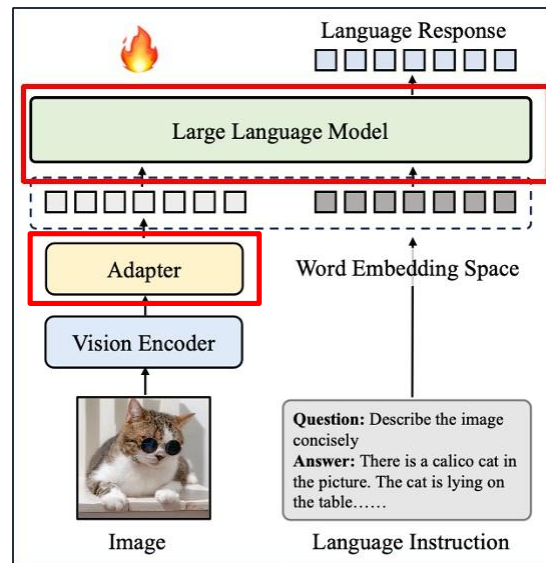
Stage1: Pretraining Stage

+Align different modalities, provide world knowledge



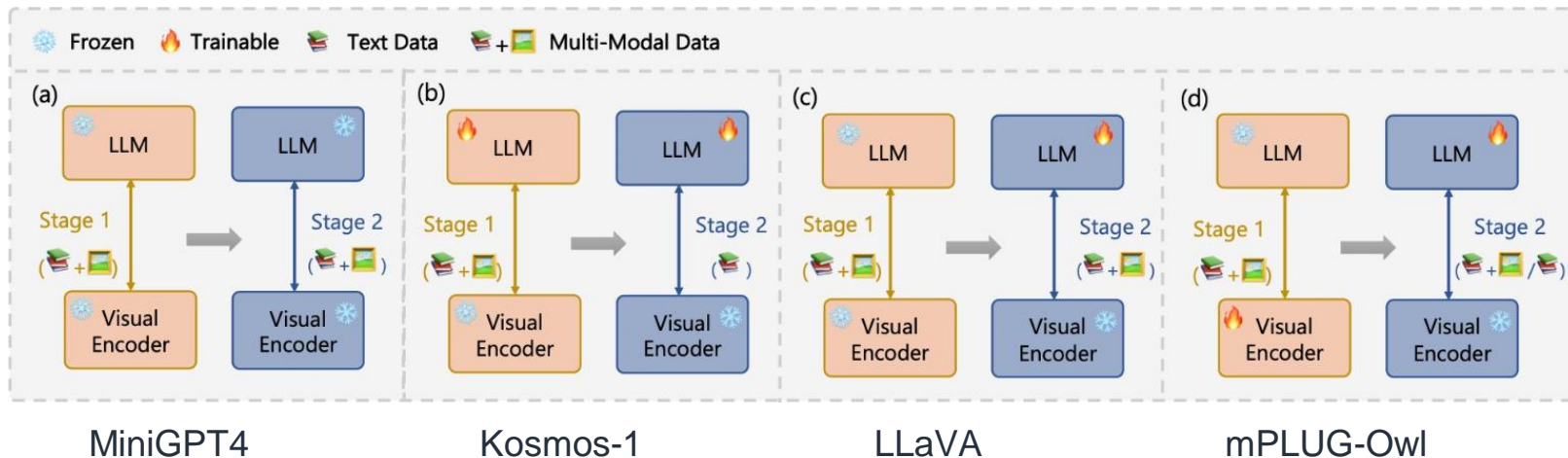
Stage2: Instruction Tuning Stage

+ Teach models to better understand the instructions from users and fulfill the demanded tasks.



Training Paradigms

👉 Training paradigms of popular multimodal large language models.



1 *mPLUG-Owl: Language Models with Multimodality. 2023.*

2 *Visual Instruction Tuning. NeurIPS 2023.*

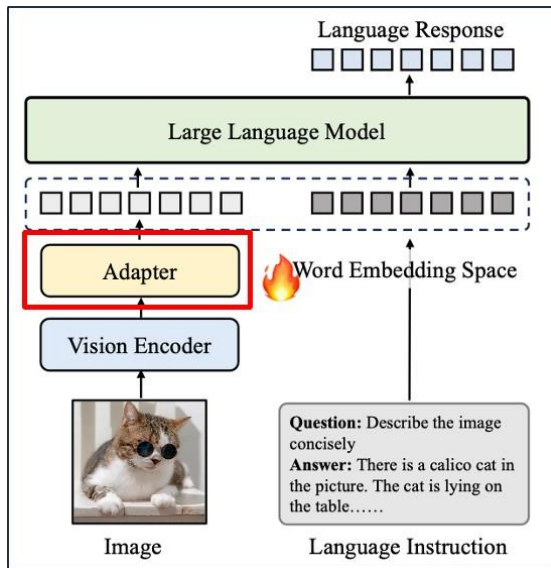
3 *MINIGPT-4: ENHANCING VISION-LANGUAGE UNDERSTANDING WITH ADVANCED LARGE LANGUAGE MODELS. 2023.*

4 *Language Is Not All You Need: Aligning Perception with Language Models. 2023.*

Pretraining Paradigms

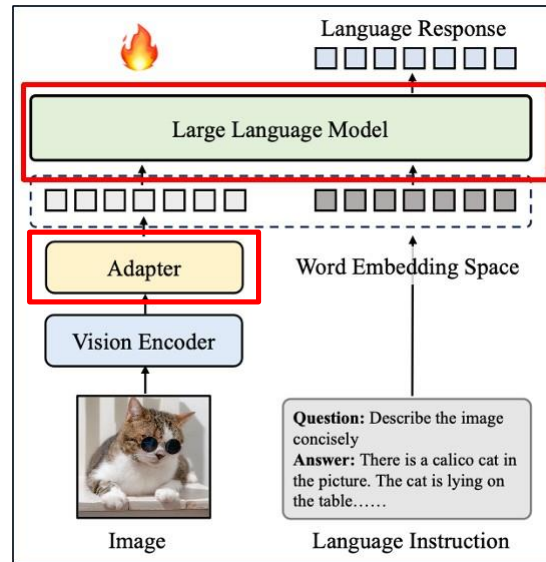
Stage1: Pretraining Stage

+Align different modalities, provide world knowledge



Stage2: Instruction Tuning Stage

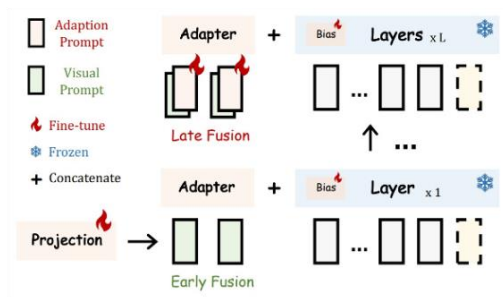
+ Teach models to better understand the instructions from users and fulfill the demanded tasks.



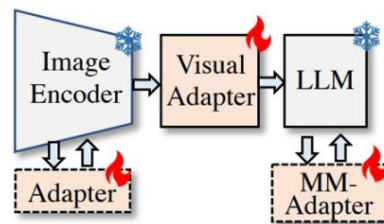
Research Directions for Multimodal LMMs

(1) Efficient multimodal LLMs:

- **Projection** (see Fuyu) or **token reductions** [1]
- LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model
- Cheap and Quick: Efficient Vision-Language Instruction Tuning for Large Language Models



LLaMA-Adapter V2: 14M parameters



Mixture-of-Modality Adaptation

LAVIN: 3.8M parameters

- QLoRA: Efficient Finetuning of Quantized LLMs

Finetuning 65B LLaMA for 24 hours on a single GPU, reaching 99.3% of the performance level of ChatGPT

(2) Generating multimodal outputs (vision, action, etc.)

Generating Images with Multimodal Language Models (Koh et al., Jun 2023) is an awesome paper that shows how LMMs can generate and retrieve images together with generating texts.

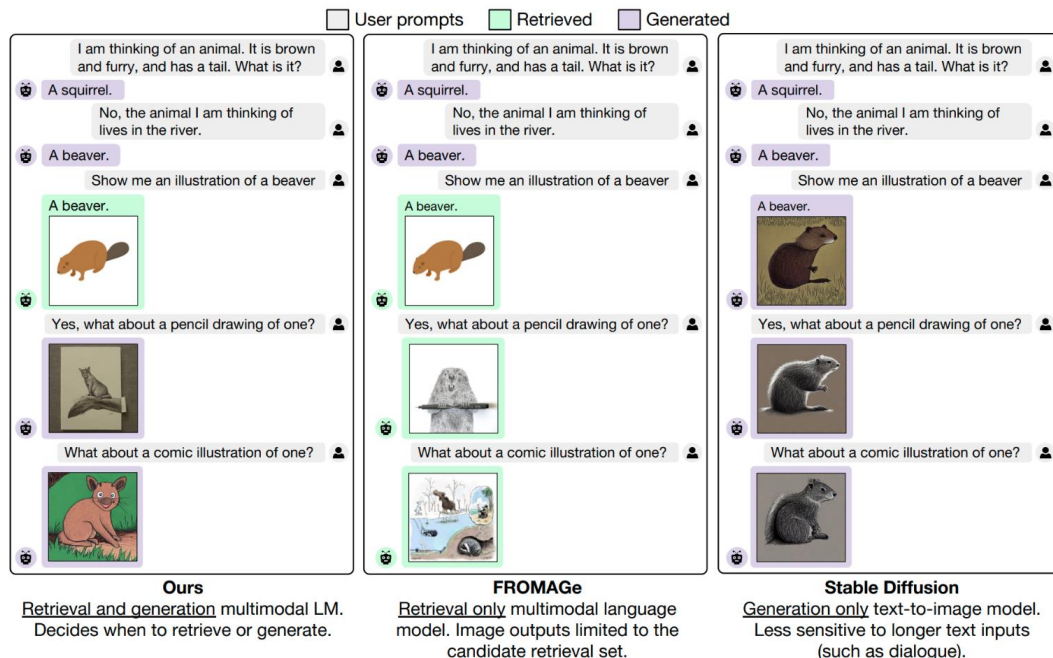
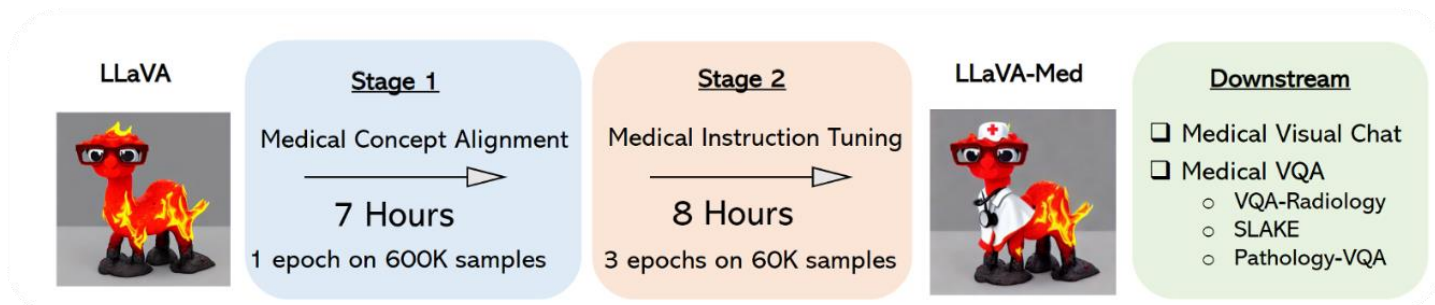


Figure 1: Our model is capable of generating text, retrieving images, generating novel images, and interleaving results into coherent multimodal dialogue.

(3) Vertical Applications (e.g. medical applications)

- PathAsst: Redefining Pathology through Generative Foundation AI Assistant for Pathology
- PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering
- LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day



User: What is this image about?
LLaVA-Med: The image is a coronal view of a computed tomography (CT) scan of the abdomen and pelvis. It provides a detailed view of the internal organs and structures within the abdominal and pelvic cavities.

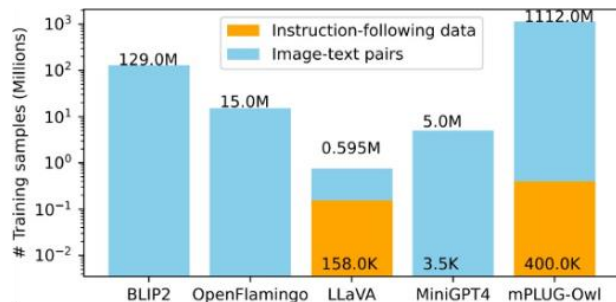
User: What is abnormal about this image?
LLaVA-Med: The abnormality in this image is the presence of a large left-sided pelvic mass. This mass could be due to various reasons, such as a tumor, cyst, or other growth. Further evaluation and tests would be needed to determine the exact cause and appropriate treatment for this abnormality.

User: What is this image about?
LLaVA: The image is a medical illustration showing a person's internal anatomy, especially focusing on their large and small intestines, stomach, liver and kidneys. **The organs are highlighted in different colors**, making it easy to identify each organ

User: What is abnormal about this image?
LLaVA: **As a visual AI assistant, I cannot determine the specific abnormality on the chest X-ray based on the provided image description. The image description mainly focuses on the organs and their colors...**

(4) Evaluations

- Evaluating Object Hallucination in Large Vision-Language Models
- On Evaluating Adversarial Robustness of Large Vision-Language Models
- On the Hidden Mystery of OCR in Large Multimodal Models
- LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark



14 tasks: 10 tasks:
OCR OCR + Reasoning

BLIP-2 OPT _{6.7b}	62.12	7.31
BLIP-2 FlanT5 _{XXL}	63.74	13.59
OpenFlamingo	56.68	11.28
LLaVA	53.30	10.37
MiniGPT4	40.12	7.82
mPLUG-Owl	61.79	16.03
Supervised-SoTA	84.83	72.75



(5) Incorporating more data modalities

Today, most multimodal systems work with text and images. It's only a matter of time before we need systems that can incorporate other modalities such as **videos, music, and 3D**. Wouldn't it be amazing to have one shared embedding space for ALL data modalities?

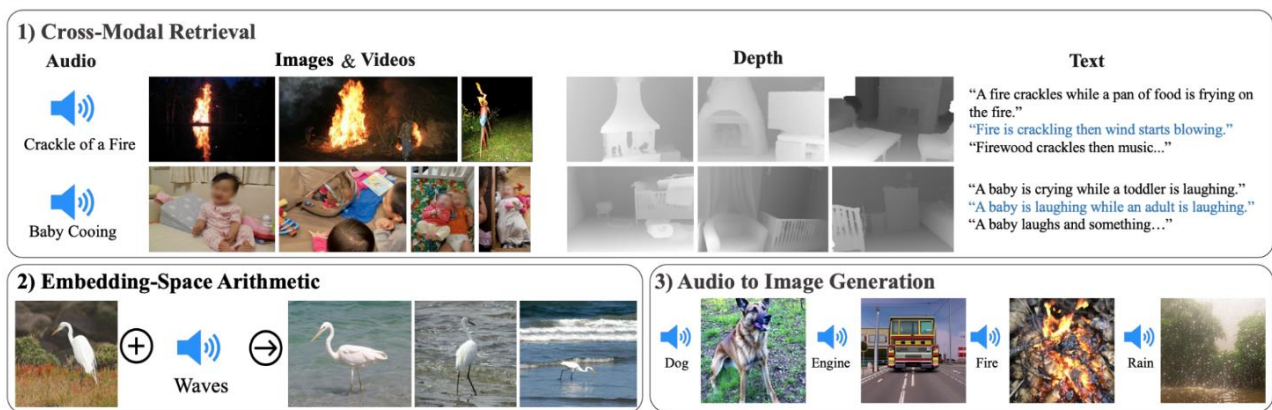
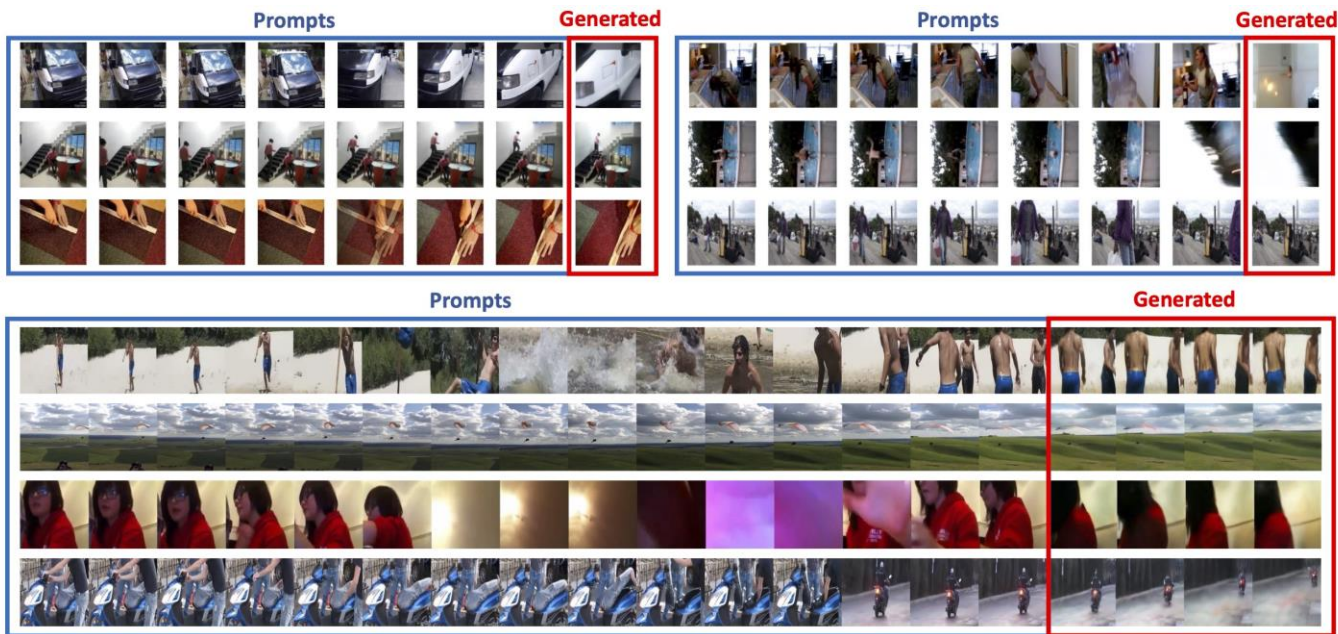


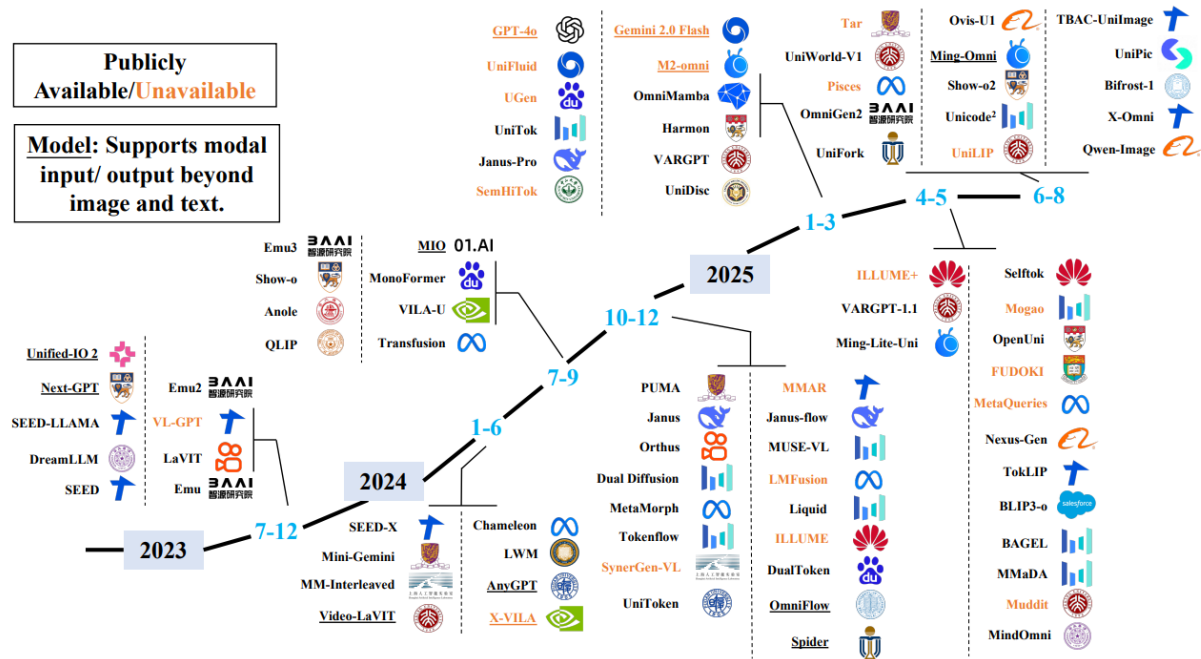
Figure 1. IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth and text, that aren't observed together. **2)** Adding embeddings from different modalities naturally composes their semantics. And **3)** Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 [61] decoder designed to work with CLIP text embeddings.

(6) Large vision models using next-frame prediction



Frame predictions. LVM predicts the next frame (marked in red) given previous video frames as prompt. The results reveal the LVM can predict the video frames while considering dynamic objects and camera motion.

(7) Unify multi-modal understanding and generation



Applications of multi-modality generations

Example:



中国科学院、中国工程院公布了2025年院士增选结果
生成一张封面， 21:9



显示思路 (Nano Banana Pro) ▾



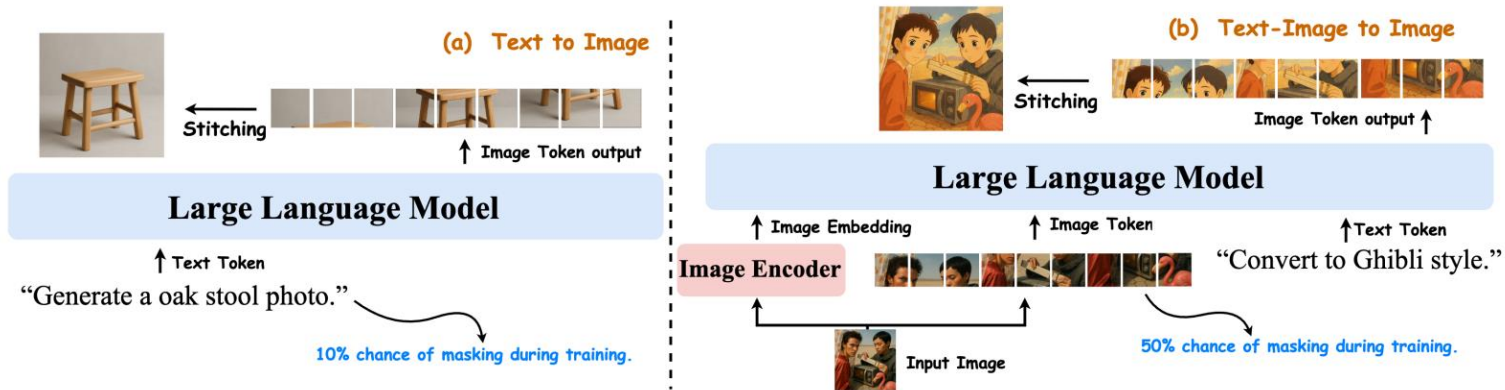
[2]: 院士增选 - 中国工程院 [🔗](#)



Generated by **Nana Banana**

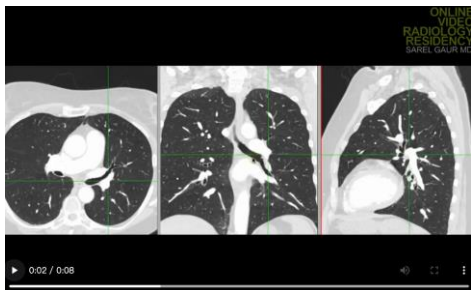


Janus



Junyong Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, Benyou Wang. ShareGPT-4o-Image: Aligning Multimodal Models with GPT-4o-Level Image Generation.

Medical Sora: Towards a Medical World Simulator



Medical image analysis



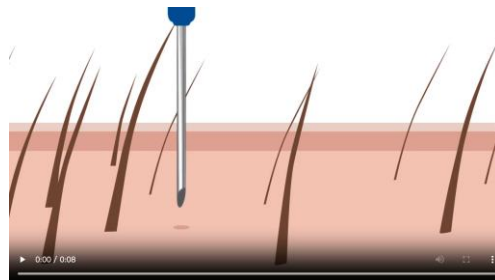
medical care



Instrument Usage



Operation



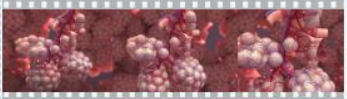
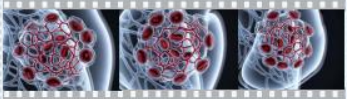




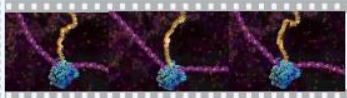


Medical Popularization



Medical Scenario

A medical world simulator offers immense application potential.

Micro world simulator

Alveoli Demonstration			
Create a medical animation of human lung alveoli with detailed capillaries and blood flow.			
Cell Division			
Create a simplified video of cell division, from one to two to four			
DNA Replication			
Create a simple animation of DNA turning into an RNA strand through polymerase.			
Prompt	Real Video	Sora	Veo3

Edge-device Multi-modal LLMs

MLLMs for Edge devices

在端侧设备运行，甚至可以流式地运行



A naïve demo



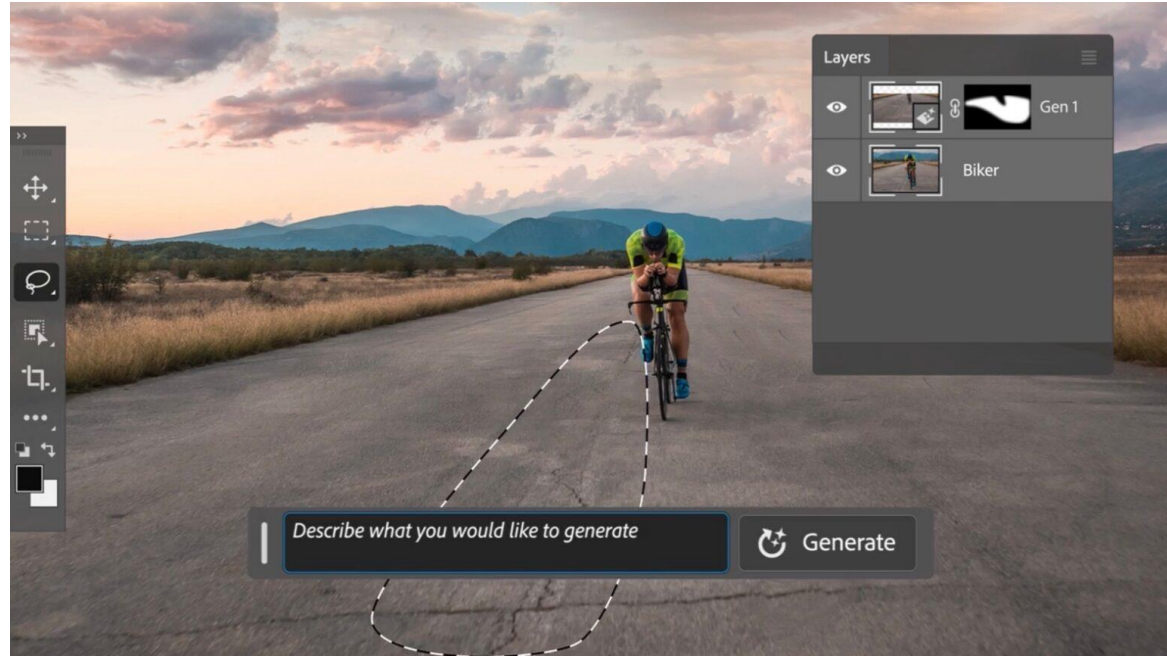
MLLMs running on edge devices

Final remarks from multi-modal interactions

The necessity to Multi-modal LLLs

- Speech is more convenient for interaction;
 - See HuatuoGPT applications
- Video is language-agnostic
 - We could also enjoy the fun from TikTok in India without language barrier
- Data with other modality needs a **language interface** to interact;
 - Draw with language interface-based AI tools

Adobe Firefly turns Photoshop into effortless AI image generator



<https://www.cultofmac.com/news/adobe-firefly-photoshop-ai-image-generator>

Final conclusion of this course



Human



Agent (generated)

Hope it could finally pass tuning test in a video interaction !

Final projects

- **Starbuck time every noon next week.**
 - Tentative time: **12:30 -13:00 pm from 24-25 Dec.**



AIMO 3 is coming

AI|MO · FEATURED CODE COMPETITION · 5 MONTHS TO GO

AI Mathematical Olympiad - Progress Prize 3

Solve international-level math challenges using artificial intelligence models

[Join Competition](#) ...

AI|MO
ARTIFICIAL INTELLIGENCE
MATHEMATICAL OLYMPIAD

[Overview](#) [Data](#) [Code](#) [Models](#) [Discussion](#) [Leaderboard](#) [Rules](#)

Overview

The goal of this competition is to create open-source algorithms and models that can solve olympiad-level math problems written in LaTeX format. Your participation will help to advance AI models' mathematical reasoning skills and drive frontier knowledge.

Start
15 hours ago

Close
5 months to go

Merger & Entry

Competition Host
AI|MO

Prizes & Awards
\$2,207,152 ⓘ
Awards Points & Medals

Participation
1,953 Entrants
15 Participants
15 Teams
15 Submissions

Contact me if you want to work on it — (full-time in five months)

<https://www.kaggle.com/competitions/ai-mathematical-olympiad-progress-prize-3>

Acknowledgement

- <https://huyenchip.com/2023/10/10/multimodal.html>
- <https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models>
- <https://arxiv.org/pdf/2306.14895.pdf>
- <https://web.stanford.edu/class/cs224n/>
- <http://www.boyangli.org/slides/Multimodal+LLM-UBC-June-26-2023.pdf>
- <https://huyenchip.com/2023/10/10/multimodal.html>
- <https://cmu-multicomp-lab.github.io/mmmml-tutorial/icml2023/>
- <https://mllm2024.github.io/CVPR2024/>