

多模态大语言模型

Outline

1. 多模态大语言模型（MLLM）的定义
2. （图-文）多模态模型的发展历程
3. GPT4v实战
4. 图文对话系统的搭建
5. Beyond VL：支持更多模态输入的大语言模型
6. 使用多模态大语言模型完成更多任务

Multimodal Large Language Models

- 模态的定义

模态 (modal) 是事情经历和发生的方式，我们生活在一个由多种模态 (Multimodal) 信息构成的世界，包括视觉信息、听觉信息、文本信息、嗅觉信息等等。

➤ 描述同一对象的多模态数据：



图像



视频



音频

下雨天，淅淅沥沥
地，街道上布满了
雨水，路边的树木
倒影在路面上。

文本

- MLLMs 的定义



雨天时出门不带伞的用户接收到一份处理多模态信息能力的模型。[1]

“听雨声，观察路面积水情况。判断现在是否适合出门？”

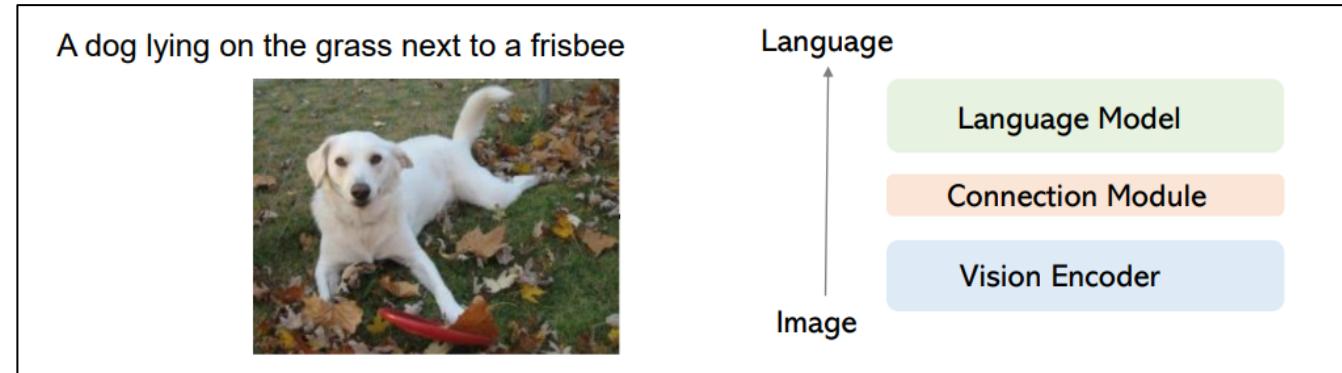
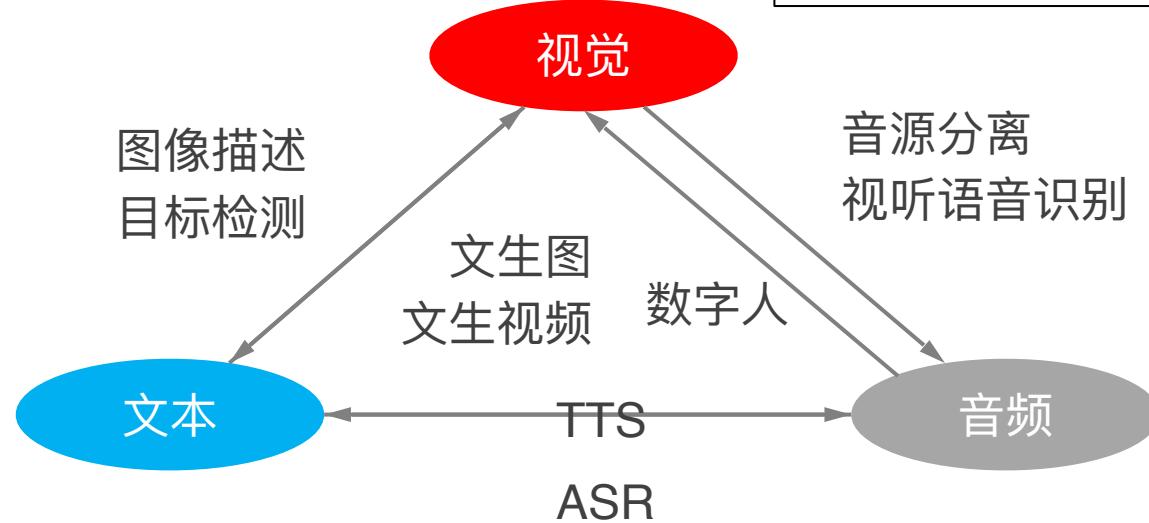
Multimodal Large Language Models

- 几个概念的区分
 - 跨模态模型
 - 单模态大模型
 - 多模态模型
 - 多模态语言大模型

Multimodal Large Language Models

- 跨模态模型

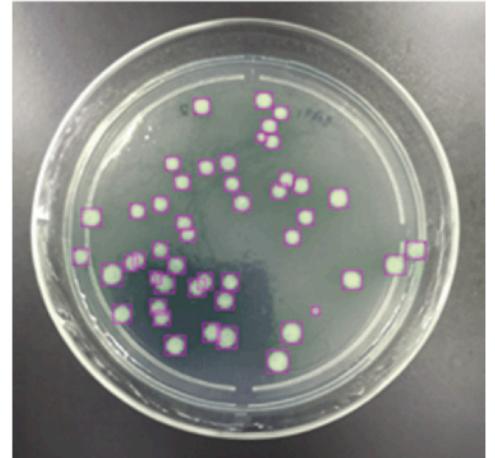
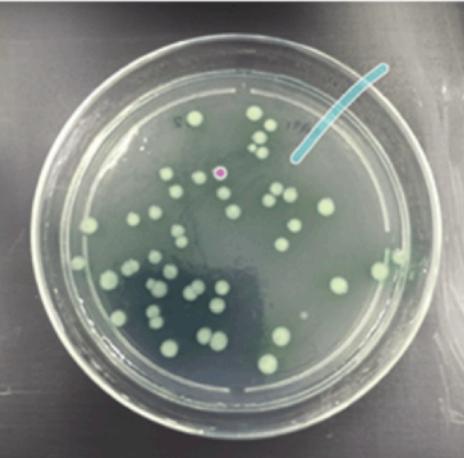
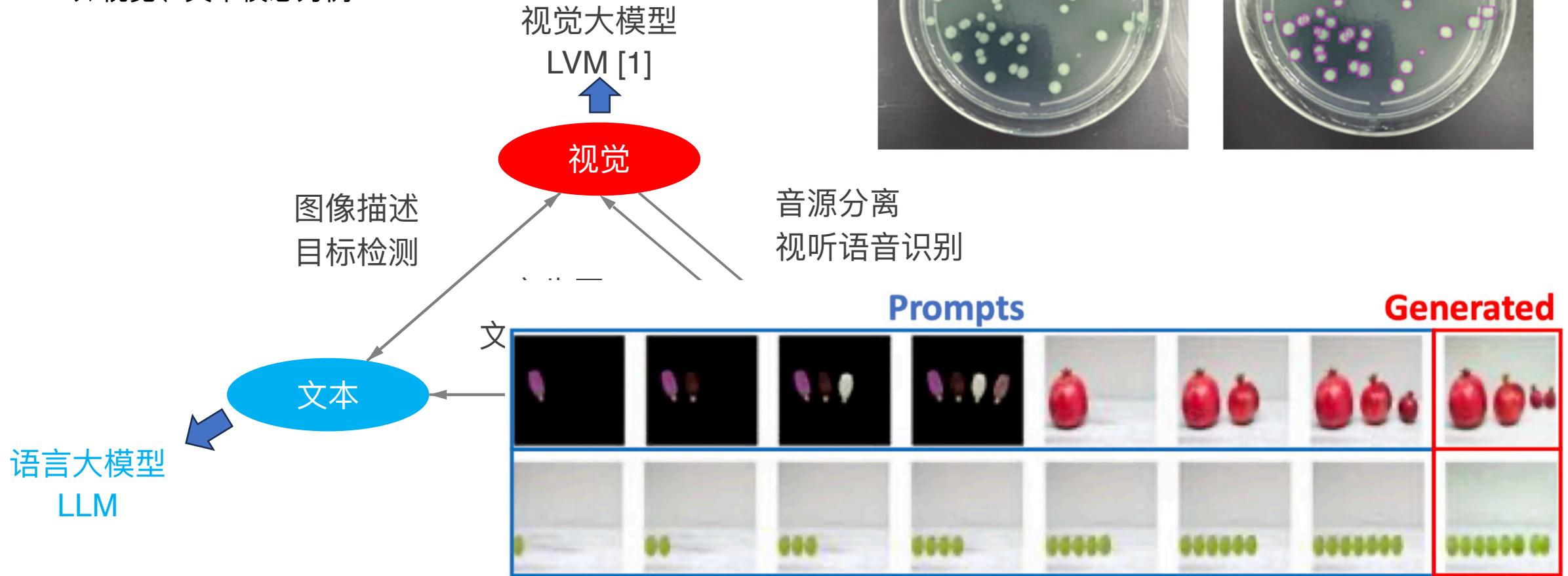
以 视觉、文本、音频三种模态为例：



Multimodal Large Language Models

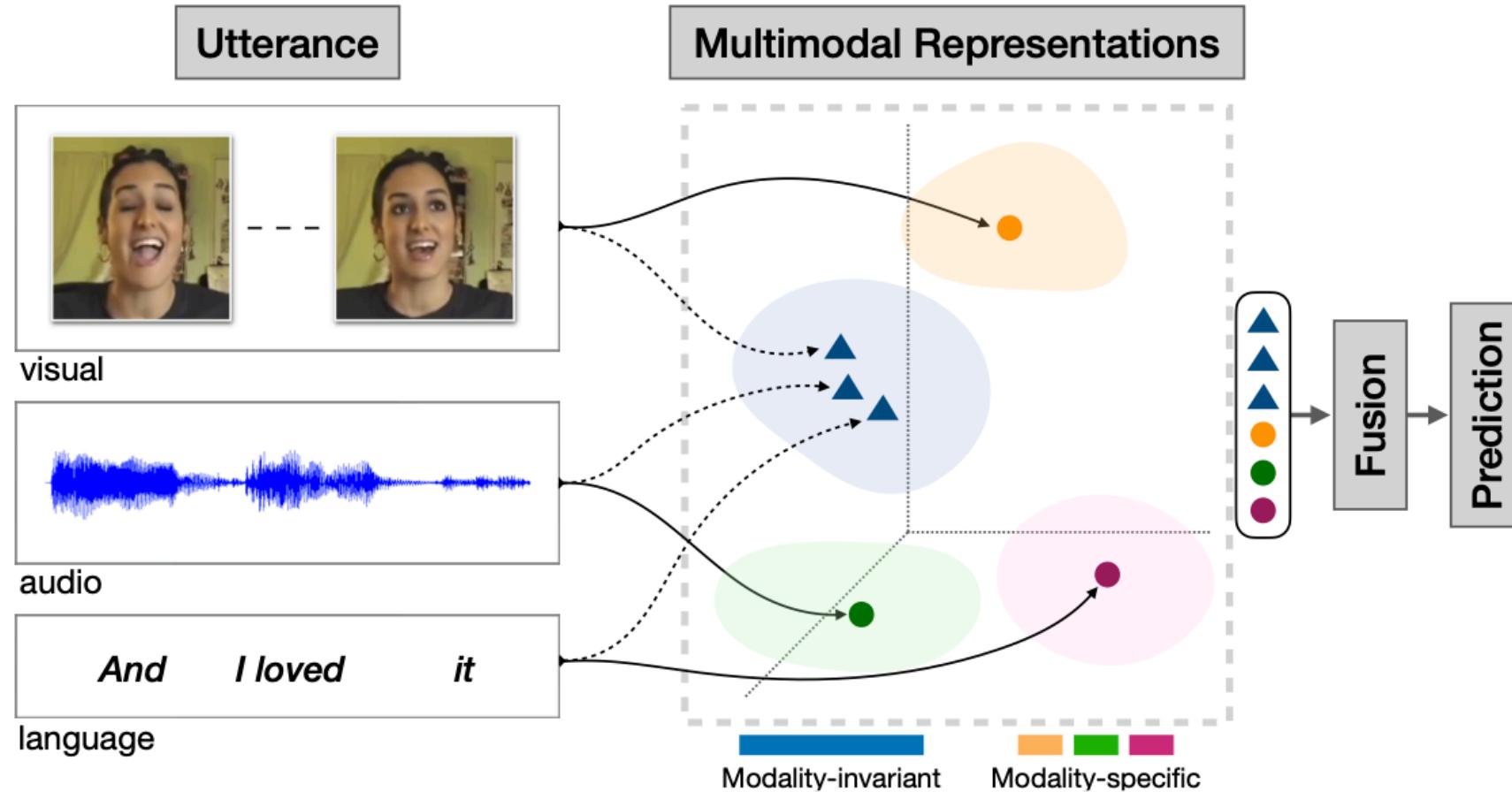
- 单模态大模型

以 视觉、文本模态为例：



Multimodal Large Language Models

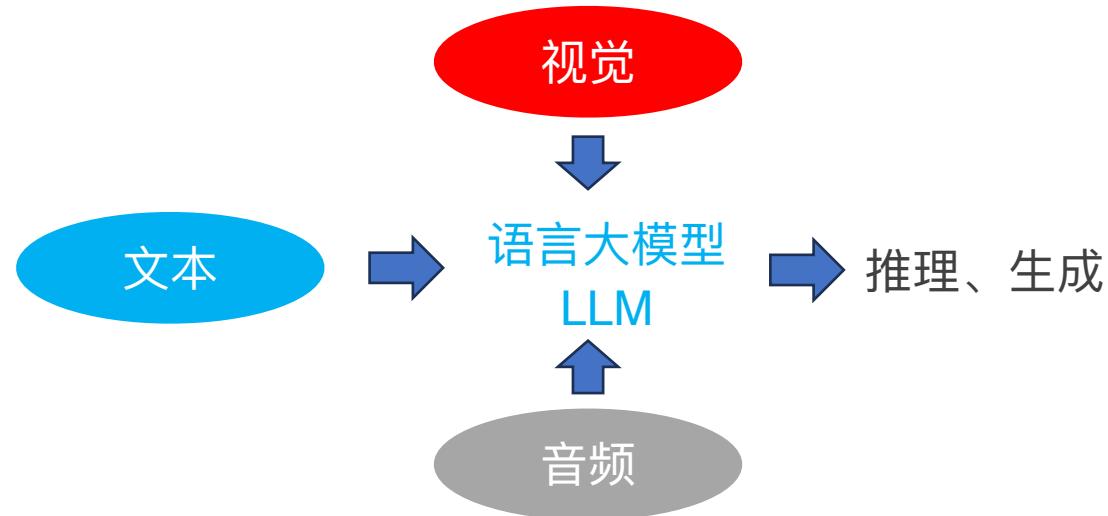
- 多模态模型



Multimodal Large Language Models

- 多模态大语言模型

以 视觉、文本、音频三种模态为例：



“由LLM扩展而来的具有接收与推理多模态信息能力的模型。”

(图-文) 多模态模型的发展历程

- 以 视觉-文本 多模态模型为例，四个关键里程碑：
 1. Vision Transformer (ViT) 模型[1] (图像表示的token化)

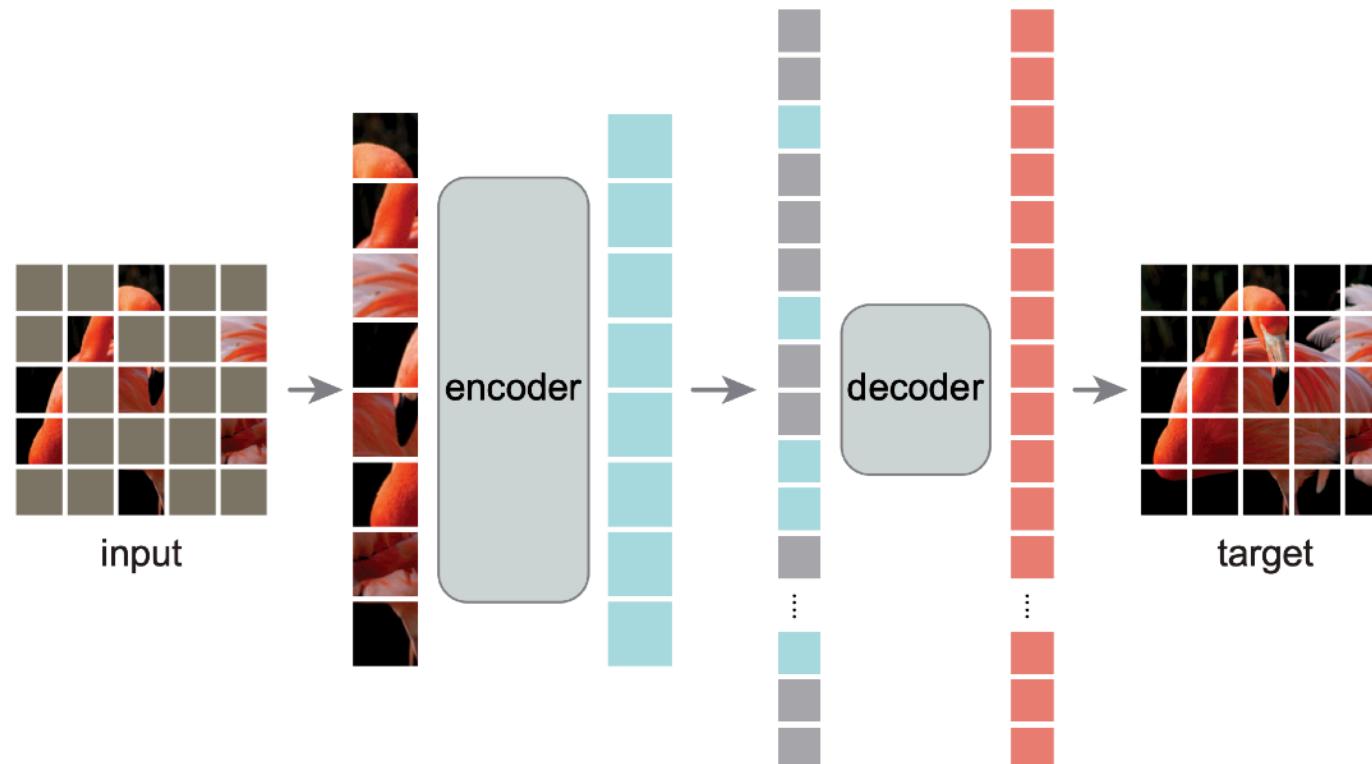


(图-文) 多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

1. Vision Transformer (ViT) 模型[1] (图像表示的token化)

➤ Mask Image Modeling 无监督图像特征学习



[1] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

多模态模型的发展历程

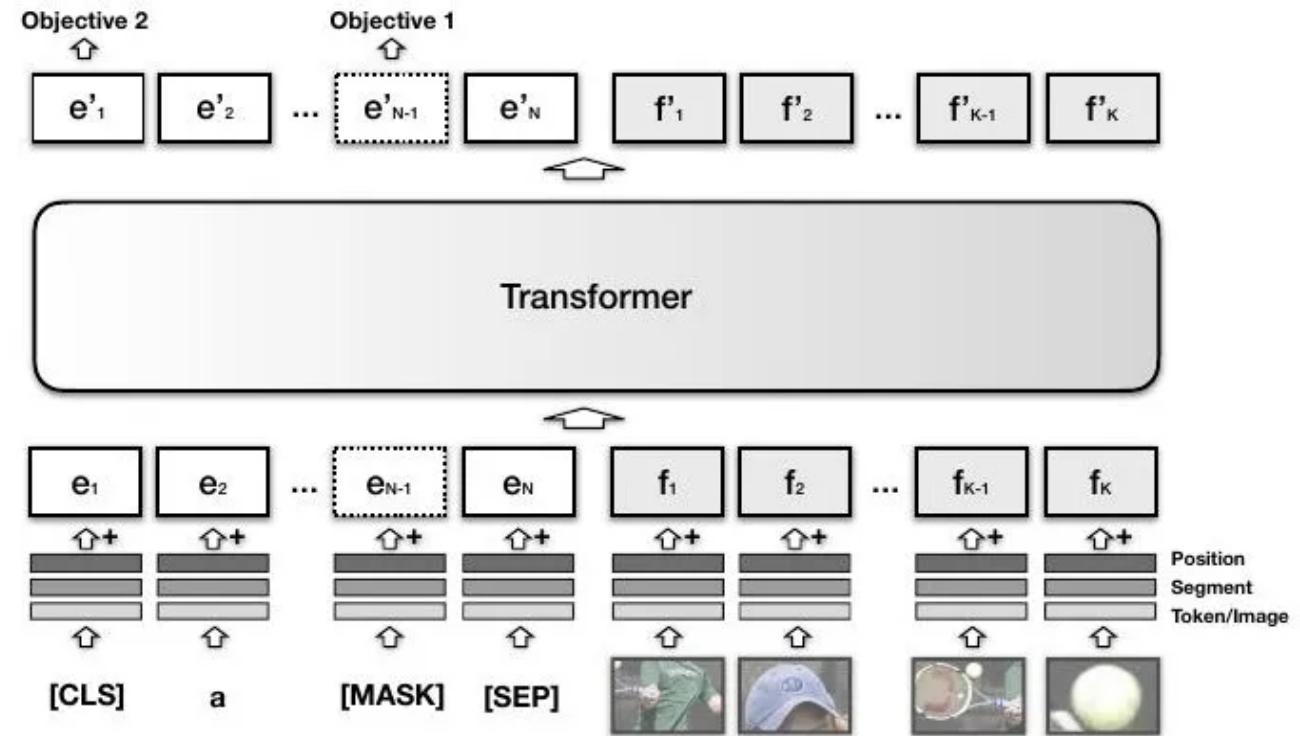
- 以视觉-文本 多模态模型为例，四个关键里程碑：

2. 基于Transformer架构的图像-文本联合建模

➢ 以VisualBert [1]为例：



A person hits a ball with a tennis racket



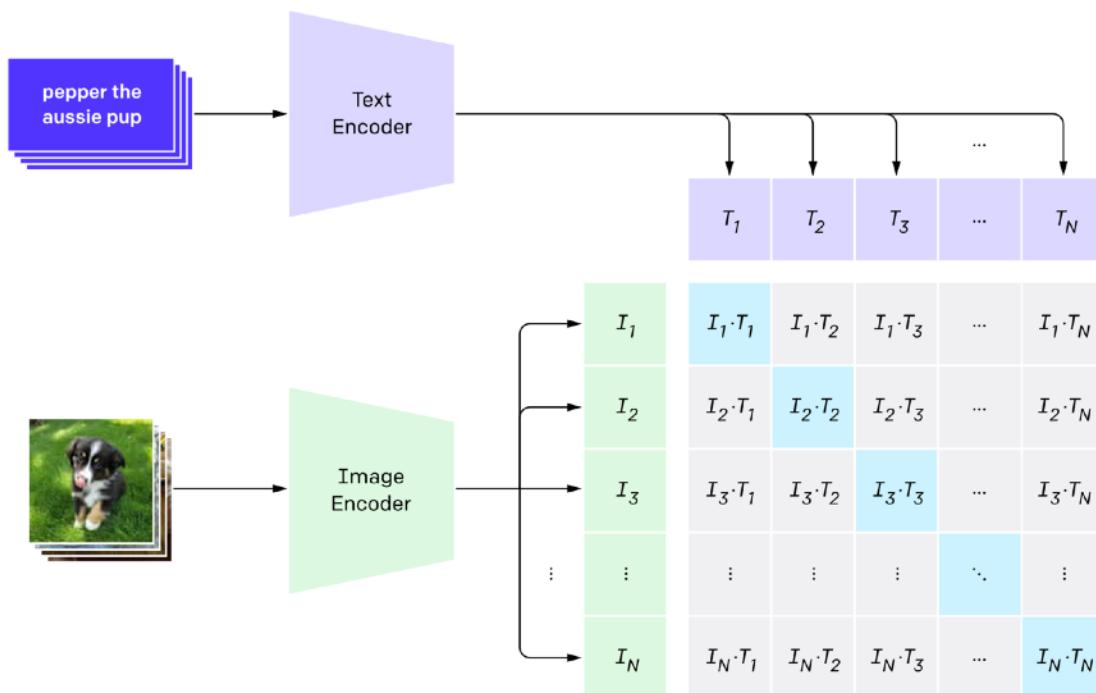
[1] Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." arXiv preprint arXiv:1908.03557 (2019).

多模态模型的发展历程

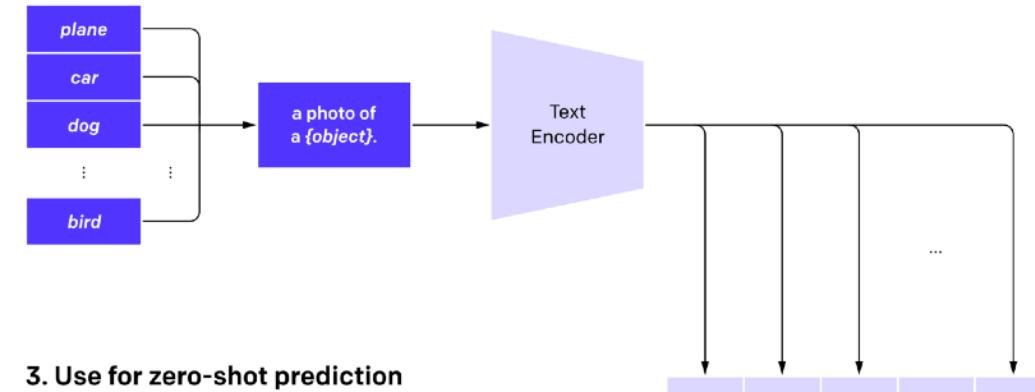
- 以视觉-文本 多模态模型为例，四个关键里程碑：

3. 大规模 图-文 Token对齐模型 CLIP [1]

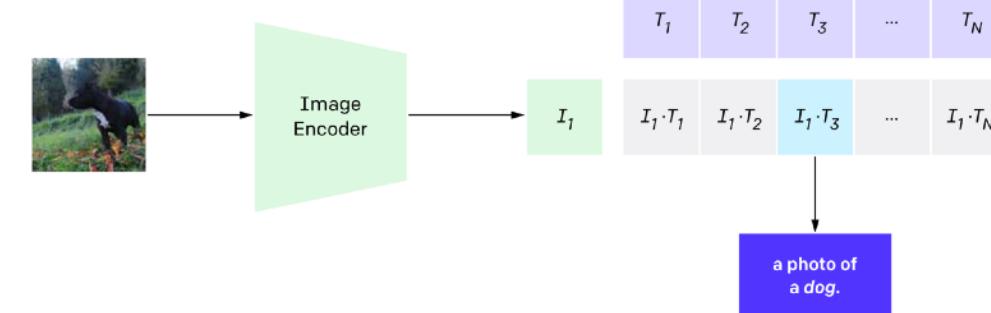
1. Contrastive pre-training



2. Create dataset classifier from label text



3. Use for zero-shot prediction



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

多模态模型的发展历程

- 以 视觉-文本 多模态模型为例，四个关键里程碑：

- 3. 大规模 图-文 Token对齐模型 CLIP [1]

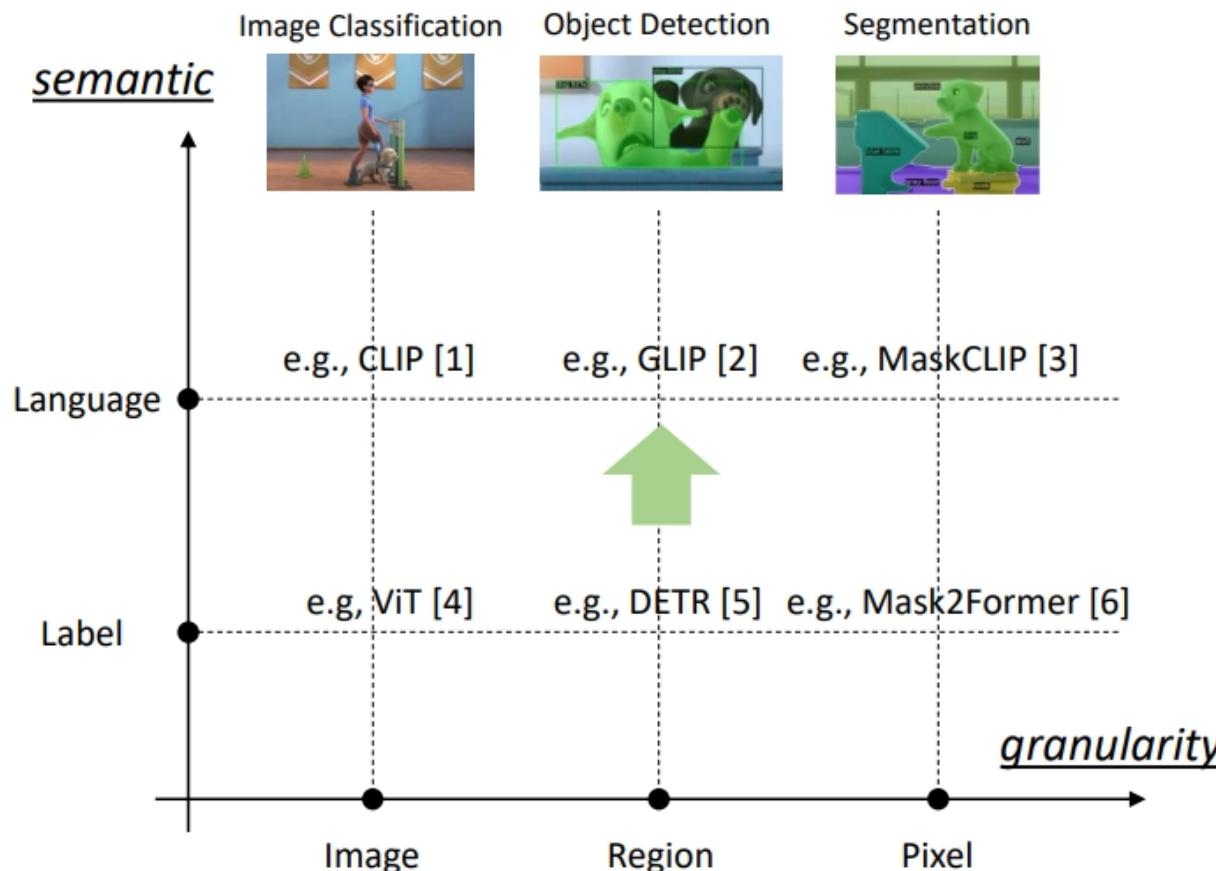
模型实践：

参考实验室notebook: `clip.ipynb`

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

3. 大规模 图-文 Token对齐模型 -> **开域**下的图像分类-目标检测-图像分割



[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

[2] Li, Liunian Harold, et al. "Grounded language-image pre-training." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

[3] Zhou, Chong, Chen Change Loy, and Bo Dai. "Extract free dense labels from clip." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.

[4] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

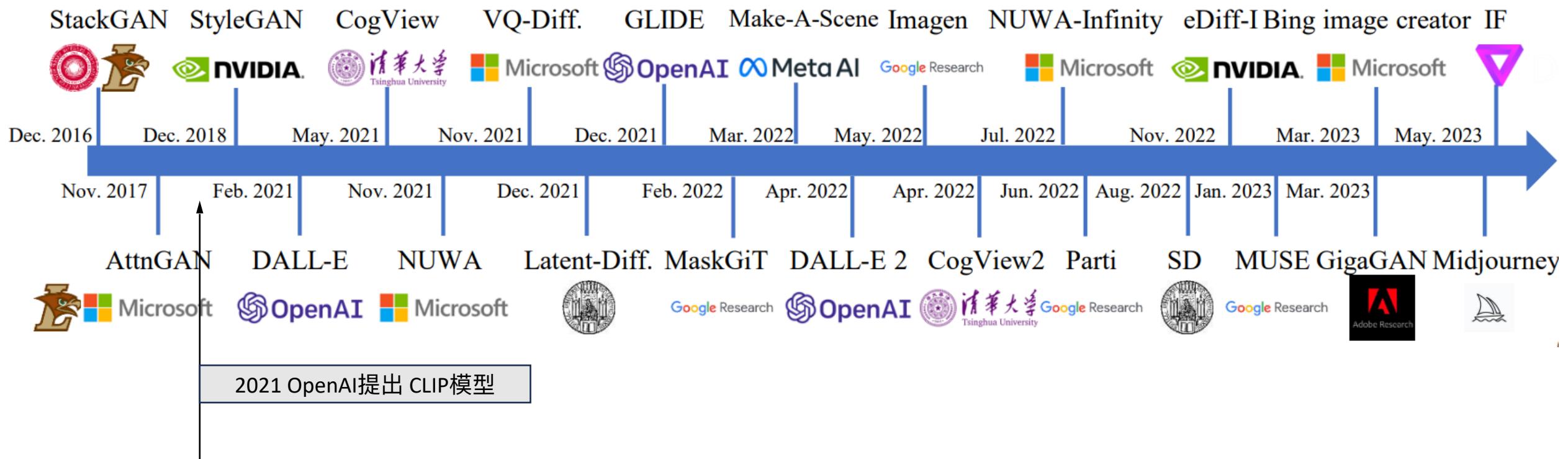
[5] Carion, Nicolas, et al. "End-to-end object detection with transformers." *European conference on computer vision*. Cham: Springer International Publishing, 2020.

[6] Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

3. 大规模 图-文 Token对齐模型 -> 文生图任务的复兴



多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➤ OpenAI: GPT-4 with Vision (GPT4v)

Interleaved Image-text Pairs

Prompt:

How much did I pay for tax?
Where is the information?



GPT-4V:

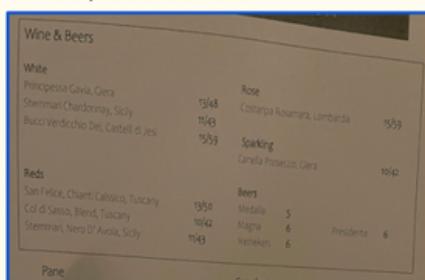
In the first receipt, you paid \$3.72 for tax. The information is located at the bottom of the receipt. In the second receipt, you paid \$4.23 for tax. The information is located near the bottom of the receipt, under the "Sales Tax" line. In the third receipt, you paid \$4.58 for tax. The information is located at the bottom of the receipt, under the "TAX" line. In total, you paid \$50.53 in tax across all three receipts.

Prompt:

How much should I pay for the beer on the table according to the price on the menu?



Here is a photo of the menu.



GPT-4V:

According to the menu, the price for a Magna beer is \$6. Since there are two Magna beers on the table, the total cost for the beers would be \$12.

支持图-文交替输出

- 输入：可以接收文本、图像信息输入
- 输出：自然语言文本

不支持视频，但支持含多张图像的序列输入

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➤ OpenAI: GPT-4 with Vision (GPT4v)

特性一：遵循文字提示

Prompt:

Describe the image with a sentence consisting of **three/six/nine** words.



GPT-4V:

Three: Seaside dining table

Six: Dining table overlooking ocean with beer

Nine: Outdoor restaurant table with ocean view and a beer

Prompt:

Please read the text in this image and return the information in the following JSON format (note xxx is placeholder, if the information is not available in the image, put "N/A" instead).

```
{"Surname": xxx, "Given Name": xxx, "USCIS #": xxx, "Category": xxx, "Country of Birth": xxx, "Date of Birth": xxx, "SEX": xxx, "Card Expires": xxx, "Resident Since": xxx}
```



GPT-4V:

```
{"Surname": "STEVENS", "Given Name": "ICHIKA", "USCIS #": "000-000-000", "Category": "IR1", "Country of Birth": "Japan", "Date of Birth": "18 AUG 1988", "SEX": "F", "Card Expires": "07/04/2027", "Resident Since": "07/05/2017"}
```

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➤ OpenAI: GPT-4 with Vision (GPT4v)

特性二：理解视觉指向和参考

Describe the pointed region in the image.



GPT-4V:

The pointed region in the image is a row of hanging lights on a wire. The lights are small and round, and they are evenly spaced along the wire. The wire is stretched between two poles and is located above a patio area with tables and umbrellas. The lights appear to be decorative and are likely used to provide ambient lighting in the evening.

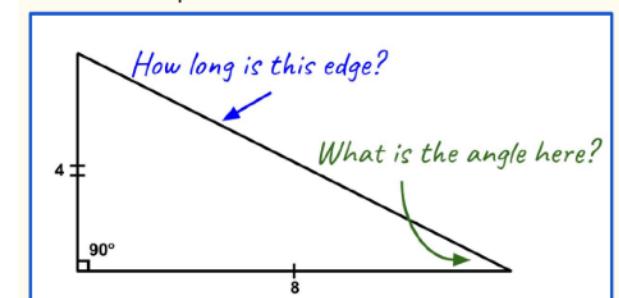
Describe the pointed region in the image.

Method	Validation set								Test set							
	in.		near.		out.		overall		in.		near.		out.		overall	
	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C	S
OSCAR	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4	84.8	12.1	82.1	11.5	73.8	9.7	80.9	11.3
Human	84.4	14.3	85.0	14.3	95.7	14.0	87.1	14.2	80.6	15.0	84.6	14.7	91.6	14.2	85.3	14.6
VIVO	92.2	12.9	87.8	12.6	87.5	11.5	88.3	12.4	89.0	12.9	87.8	12.6	80.1	11.1	86.6	12.4
VinVL	103.7	13.7	95.6	13.4	83.8	11.9	94.3	13.1	98.0	13.6	95.2	13.4	78.0	11.5	92.5	13.1
UFO	103.9	14.5	95.5	13.8	83.5	12.3	94.3	13.6	98.9	14.3	94.7	13.9	77.9	12.1	92.3	13.6
mPLUG	-	-	-	-	-	-	114.8	14.8	-	-	-	-	-	-	-	-
SimVLM	113.7	-	110.9	-	115.2	-	115.2	-	113.7	-	110.9	-	115.2	-	115.2	-
LEMON	118.0	15.4	116.3	15.1	120.2	14.5	117.3	15.0	112.8	15.2	115.5	15.1	110.1	13.7	114.3	14.9
UniversalCap	123.2	15.0	121.5	15.3	123.4	14.4	122.1	15.0	118.9	15.4	120.6	15.3	114.3	14.1	119.3	15.1
CoCa	-	-	-	-	-	-	122.4	15.5	-	-	-	-	-	-	120.6	15.5
GIT _B	100.7	13.8	97.7	13.5	89.6	12.5	96.6	13.4	-	-	-	-	-	-	-	-
GIT _L	107.7	14.9	107.8	14.5	102.5	13.7	106.9	14.4	-	-	-	-	-	-	-	-
GIT	129.8	16.3	124.1	16.0	127.1	15.7	125.5	16.0	122.4	16.2	123.9	16.0	122.0	15.7	123.4	15.9
GIT2	126.9	16.1	125.8	16.2	130.6	15.8	126.9	16.1	124.2	16.4	125.5	16.1	122.3	15.6	124.8	16.1

GPT-4V:

The highlighted red region in the image is a table cell that contains a numerical value of 122.3. It is located in the "Test set" section of the table, under the "out" column and the "GIT2" row. This value represents the overall performance metric for the GIT2 method on the test set, in the "out" category.

Answer the question in the following figure. Provide intermediate steps.



GPT-4V:

To find the length of the blue arrowed edge, we can use the Pythagorean theorem since we have a right triangle. The theorem states that $a^2 + b^2 = c^2$, where a and b are the two legs of the right triangle and c is the hypotenuse. Plugging in the values we have, we get:

$$4^2 + 8^2 = c^2 \quad 16 + 64 = c^2 \quad 80 = c^2 \quad \sqrt{80} = c \quad 8.94 \approx c$$

Therefore, the length of the blue arrowed edge is approximately 8.94.

To find the angle at the green arrow, we can use the tangent function since we have the opposite and adjacent sides of the angle. The tangent function states that $\tan\theta = \text{opposite}/\text{adjacent}$. Plugging in the values we have, we get:

$$\tan\theta = 4/8 \quad \tan\theta = 1/2 \quad \theta = \tan^{-1}(1/2) \quad \theta \approx 26.57^\circ$$

Therefore, the angle at the green arrow is approximately 26.57°.

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

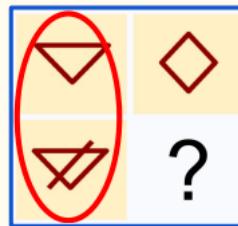
4. 多模态大语言模型的出现：

➢ OpenAI: GPT-4 with Vision (GPT4v)

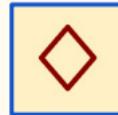
特性三：支持视觉+文本联合提示

(1) Prompt:

Find out the pattern in the first column, and use it to infer the missing figure in the second column.



For example, the circled region in the first left column shows the pattern of adding a line through it. Therefore, the answer is a figure of a diamond



with a line through it, looking like



Answer the question in the following figure.

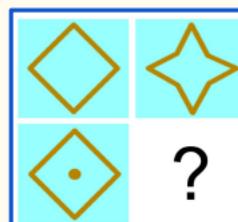


Figure out which figure of a to f that should logically replace the missing figure.

What would be the next image to complete the matrix?

- a. 
- b. 
- c. 
- d. 
- e. 
- f. 

GPT-4V:

The pattern in the first column is that a dot is added in the center of the figure. Therefore, the missing figure in the second column would be a star with a dot in the center, which is option C.

多模态模型的发展历程

- 以 视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➢ OpenAI: GPT-4 with Vision (GPT4v)

特性四：少样本上下文学习

In-context Few-shot (2-shot)

Prompt:

What is the read of the speed meter?



The yellow pointer is roughly at the middle between 80 and 100 mph. The middle between 80 and 100 is 90. The speed is around 91 mph, as the pointer is just passing 90 mph.



The yellow needle is roughly at the middle between 20 and 40 mph. The middle between 20 and 40 is 30. The speed is around 29 mph, as the needle is close to 30 mph but not quite 30 yet.



GPT-4V:

The yellow pointer is at the long tick between 0 and 20 mph. The long tick at the middle between 0 and 20 is 10. The speed is around 9 mph, as the pointer is close to 10 mph but not quite 10 yet.

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➤ OpenAI: GPT-4 with Vision (GPT4v)

特性五：强大的视觉认知能力

Prompt:

识人

Who is the person in the image and what is the person doing?



GPT-4V:

The person in the image is Jensen Huang, the CEO and co-founder of NVIDIA. He is holding and presenting a product from NVIDIA, likely a graphics processing unit (GPU).

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the famous Lombard Street in San Francisco, California, known for its steep hairpin turns. The street is lined with colorful flowers and shrubbery, and there are houses and buildings on both sides. In the background, you can see the city skyline and the Coit Tower on Telegraph Hill. This street is a popular tourist attraction, and it is often considered one of the most crooked streets in the world.

Prompt:

Describe the name of the dish.



识菜

GPT-4V:

The dish in the image is a bowl of Mapo Tofu, a popular Chinese dish from the Sichuan province. It consists of soft tofu cubes in a rich and spicy sauce made from chili peppers, fermented bean paste, and Sichuan peppercorns. The dish is garnished with thin slices of green onions.

Prompt:

Describe both the image and logo in details



Logo

GPT-4V:

The image shows a close-up of a person's foot wearing a white Nike Air Force 1 sneaker. The sneaker is a low-top style with a white rubber sole and white laces. The shoe's upper is made of a textured white leather, giving it a premium look. The logo shown on the sneaker is the iconic Nike Swoosh. It is displayed on the side of the shoe in a silver color. The Swoosh is a curved, elongated checkmark shape with a pointed tip at the end. It represents speed and motion, embodying the brand's athletic and sporty aesthetic.

Prompt:

Describe the image

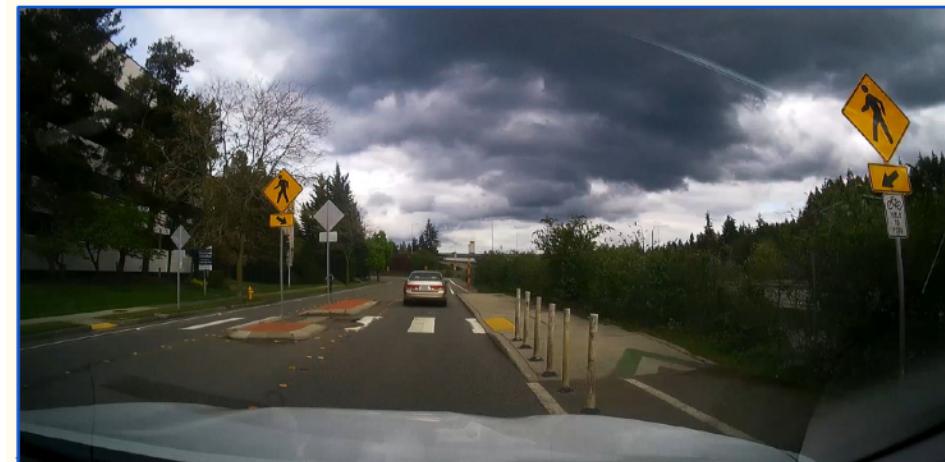


GPT-4V:

The image is a black and white X-ray of a human skull, viewed from the front. The top and bottom of the skull are visible. The teeth and jaw bones are clearly visible. The nasal cavity and eye sockets are also visible but not as clear. The image is labeled with an "L" on the bottom right corner, indicating that it is the left side of the skull.

Prompt:

Describe the image



GPT-4V:

The image is a view of a road from the perspective of a car dashboard camera. The road has two lanes, and the car is in the right lane behind a silver vehicle. There are two yellow signs on the side of the road warning drivers of an upcoming pedestrian crossing and a curve to the left. The sky is cloudy and gray, and trees and bushes line the sides of the road.

通用场景分析

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➢ OpenAI: GPT-4 with Vision (GPT4v)

特性五：强大的视觉认知能力

Prompt: 文字识别

What are all the scene text in the image?

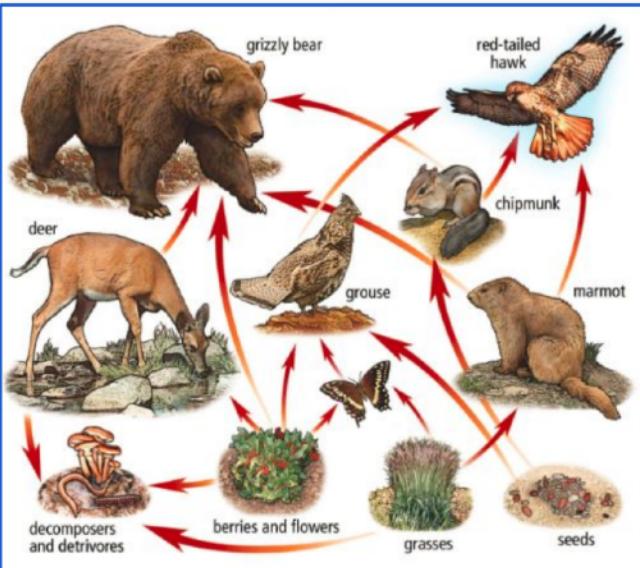


GPT-4V:

The scene text in the image are:

- Vacation
- Beach
- Smile
- Summer
- Watermelon
- Flip flops
- Lemonade

Which of the organisms is the producer in this food web?



GPT-4V:

The producers in this food web are the berries and flowers, grasses, and seeds. These are the organisms that use energy from the sun to produce their own food through photosynthesis.

图表文档理解

Prompt:

Count the number of apples in the image.

GPT-4V:

There are 16 apples in the image.

Prompt:

Localize each car in the image using bounding box. What is the image size of the input image?



GPT-4V:

The image size of the input image is 500 x 281 pixels.

Bounding box for the black car:

- Top left corner: (63, 114)
- Bottom right corner: (196, 220)

Bounding box for the white car:

- Top left corner: (287, 96)
- Bottom right corner: (474, 220)

Visualization of Detected Bbox:



目标定位

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➢ OpenAI: GPT-4 with Vision (GPT4v)

特性六：时序视觉信号理解

Prompt:

In which frame does the player kick the ball? Did the goalkeeper block the ball?

a



b



c



Prompt Continued

d



e



f



GPT-4V:

The player kicks the ball at frame c. The ball was not blocked by the goalkeeper as it can be seen in frame f that the ball is in the net and the goalkeeper is on the ground.

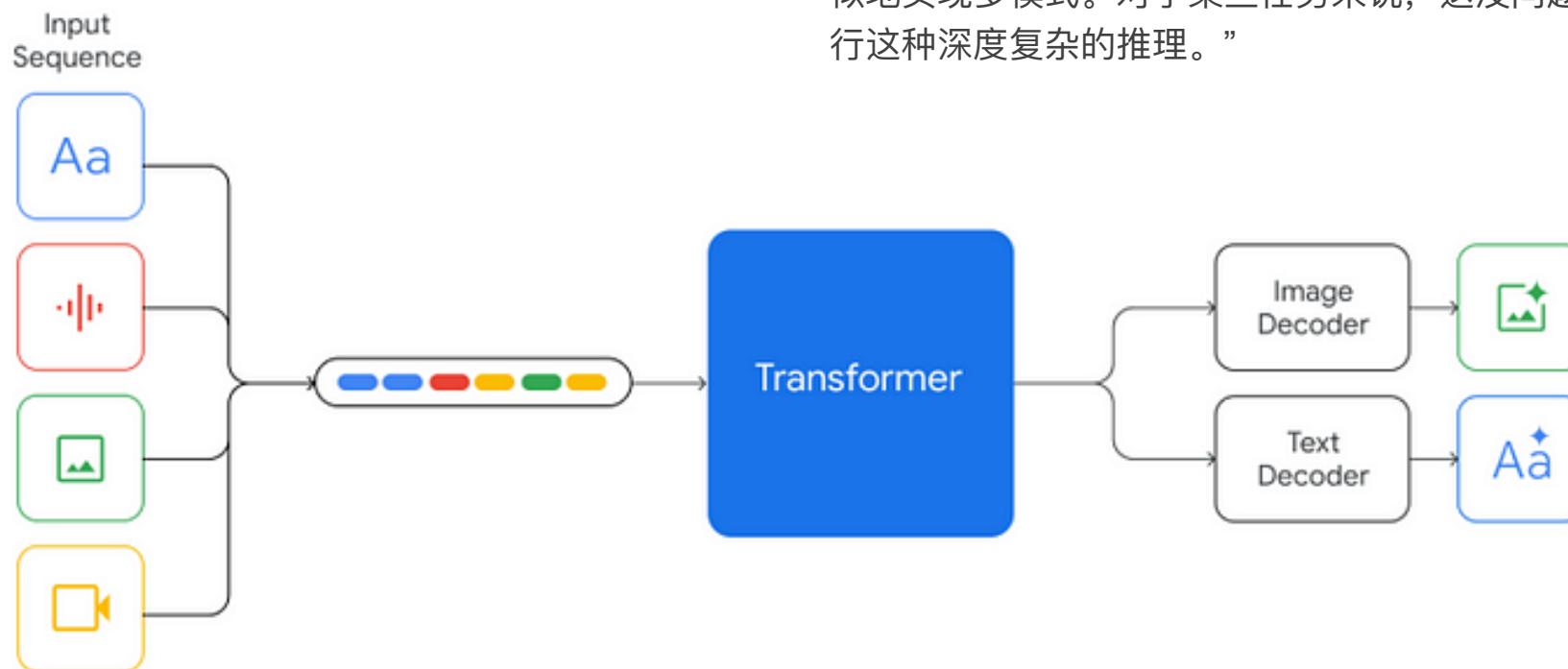
多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➢ Google: Gemini

- 输入：可以接收文本、语音、图像、视频信息输入
- 输出：自然语言文本、图像



“伪”多模态 vs. “原生”多模态

GPT-4v

Gemini

“到目前为止，大多数模型都是通过训练单独的模块，然后将它们拼接在一起实现多模态。对于某些任务来说，这没问题，但在多模态空间中，你无法进行这种深度复杂的推理。”

——DeepMind CEO哈萨比斯

多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➤ Google: Gemini

特性一：支持多模态内容输出

Prompt	Model Response
Please create a blog post about a trip to New York, where a dog and his owner had lots of fun. Include a few pictures of the dog posing happily at different landmarks.	Hi, I'm Cooper. I'm a 2 year old Goldendoodle who loves to travel. Last month, my family and I went to New York City. It was my first time in the Big Apple, and I had a blast! We stayed in a hotel near Times Square. It was so much fun to see all the bright lights and people. We went to see a Broadway show, and I even got to meet some of the actors. One of my favorite parts of New York City was Central Park. It was so big and beautiful. I loved running around and playing fetch. We also went to the Statue of Liberty and the Empire State Building. I was so excited to see all the famous landmarks. I had so much fun in New York City. It's definitely one of my favorite places to travel. I can't wait to go back someday!



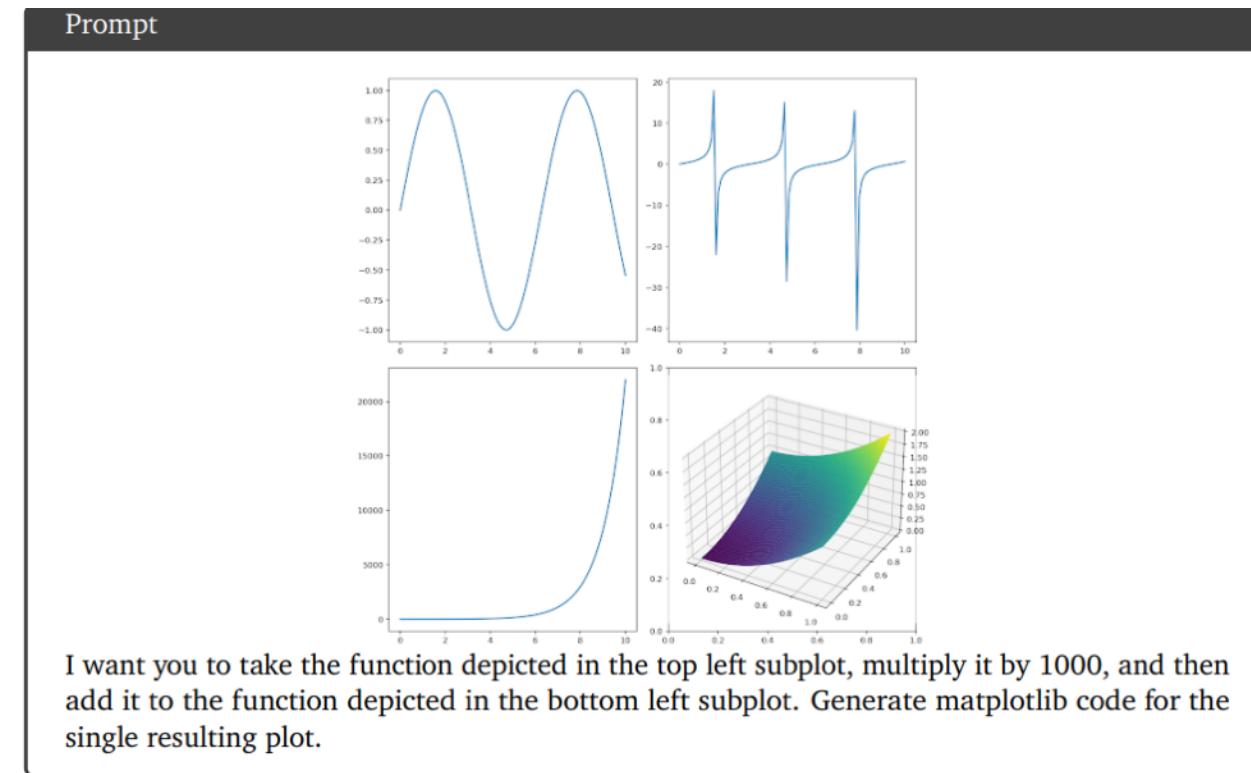
多模态模型的发展历程

- 以视觉-文本 多模态模型为例，四个关键里程碑：

4. 多模态大语言模型的出现：

➤ Google: Gemini

特性二：复杂图像理解与代码生成



GPT4v实战

API调用、使用场景实践：

参考实验室notebook: gpt4_gradio.ipynb

从零开始的图文对话系统搭建之旅

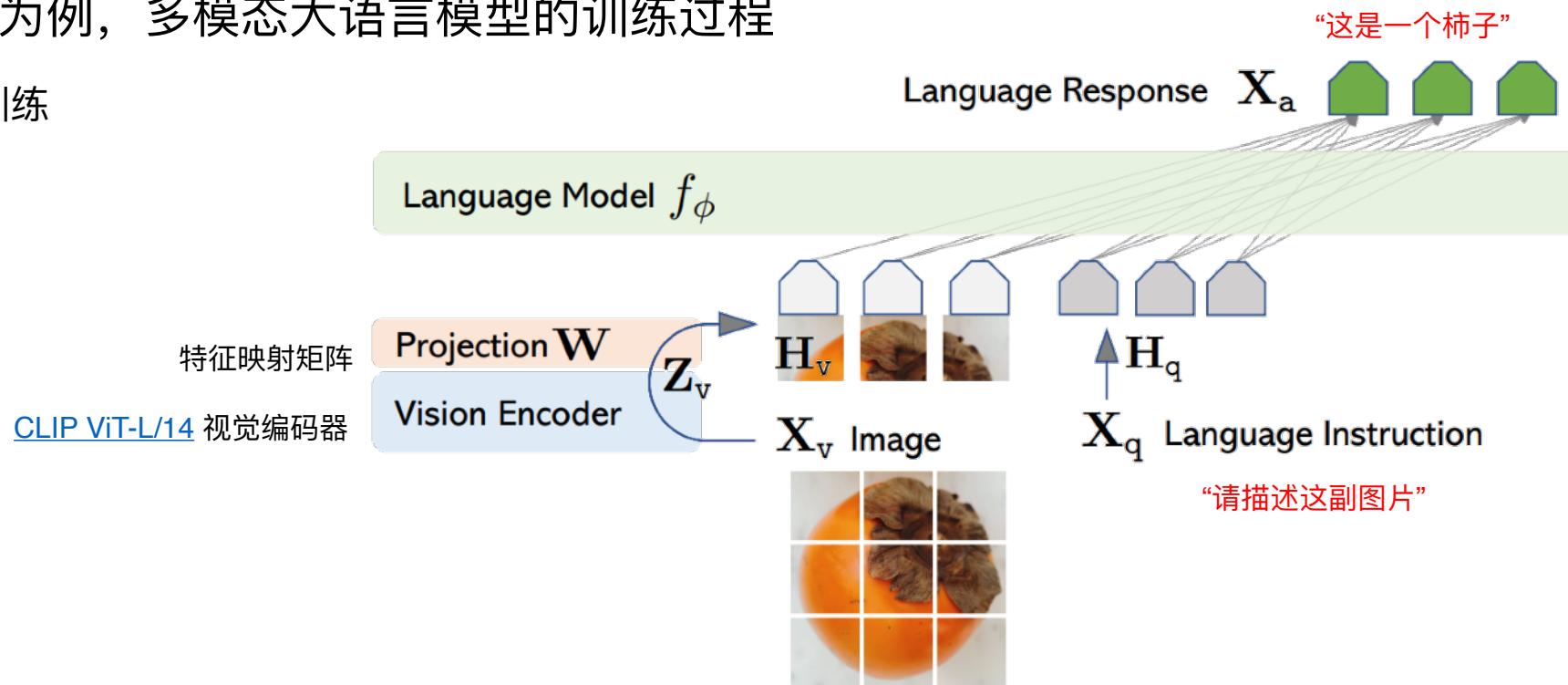
- 开源社区多模态大语言模型一览

多模态大语言模型	LLM基础模型	MME Benchmark得分
ShareGPT4V	Vicuna-13B	1618.70
LVIS-INSTRUCT4V	Vicuna-13B	1574.89
LLaVA	Vicuna-13B	1531.31
InternLM-XComposer-VL	InternLM-7B	1528.45
Qwen-VL-Chat	Qwen-7B	1487.58
GPT-4V	-	1409.43
LLaMA-Adapter V2	LLaMA-Adapter-v2.1-7B	1328.39
BLIP-2	Flant5xxl	1293.84
InstructBLIP	FlanT5xxl	1212.82
VisualGLM-6B	VisualGLM-6B	705.31
Multimodal-GPT	Multimodal-GPT-9B	654.72
MiniGPT-4	Vicuna-13B	581.66

从零开始的图文对话系统搭建之旅

- 以LLaVA为例，多模态大语言模型的训练过程

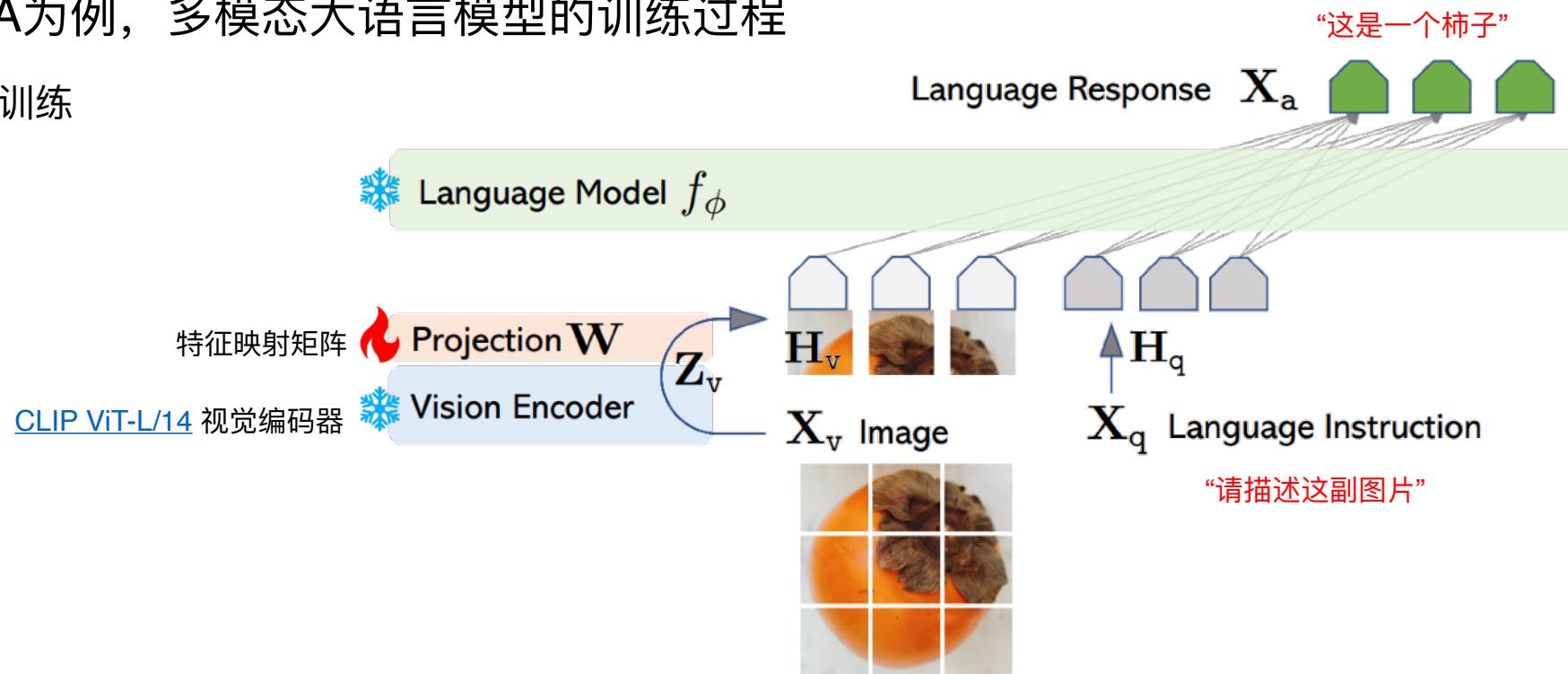
➤ 模型训练



LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 模型训练



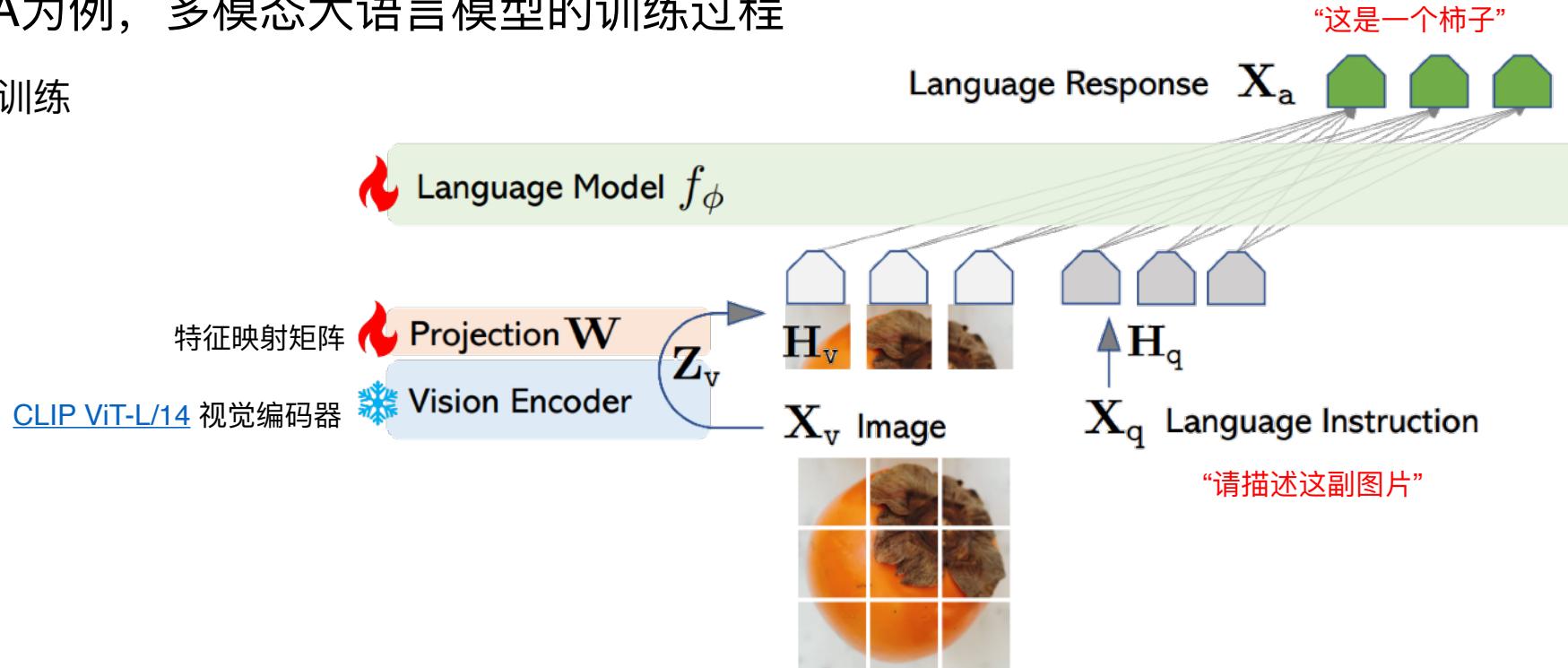
两阶段训练过程：

➤ 阶段一：特征对齐的预训练。只更新特征映射矩阵

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 模型训练



两阶段训练过程：

- 阶段一：特征对齐的预训练。只更新特征映射矩阵
- 阶段二：端到端微调。特征投影矩阵和LLM都进行更新

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

- 数据准备

图-文对齐数据

图像数据来源：

- LAION Dataset
- Conceptual Captions Dataset
- SBU Captions Dataset

共计558K条数据

conversations list	image string	id string
[{ "from": "human", "value": "Render a clear and concise summary of the photo.\n<image>" }, { "from": "gpt", "value": "select luxury furniture 3 - inch gel memory foam mattress topper" }]	00453/004539375.jpg	004539375
[{ "from": "human", "value": "Write a terse but informative summary of the picture.\n<image>" }, { "from": "gpt", "value": "a grey watch with an army style..." }	00223/002239345.jpg	002239345
[{ "from": "human", "value": "<image>\nWhat is this?", { "from": "gpt", "value": "a dragon kite flying in the blue sky stock images" }]	00694/006947502.jpg	006947502
[{ "from": "human", "value": "<image>\nRender a clear and concise summary of the photo." }, { "from": "gpt", "value": "\$ 10 - cute cheap printed mini dress -..." }	00511/005116462.jpg	005116462
[{ "from": "human", "value": "What is in the photo?\n<image>" }, { "from": "gpt", "value": "augmented reality using aruco markers in opencv" }]	00201/002017886.jpg	002017886
[{ "from": "human", "value": "Describe the image concisely.\n<image>" }, { "from": "gpt", "value": "a pair of childrens boots in the shape of knitted shoes..." }	00181/001819539.jpg	001819539

基于BLIP短文本描述打标



图-文指令数据

图像数据来源：

coco	gqa	ocr_vqa	textvqa	VG_100K	VG_100K_2	Total
118287	148854	207572	25119	64346	43903	608081

共计665K条数据

[{ "from": "human", "value": "<image>\nWhat skill set might someone need to perform such a frisbee trick?", { "from": "gpt", "value": "To perform the frisbee trick shown in the image, where the man is passing a frisbee between or underneath his legs, a person would need a combination of skills. These skills include good hand-eye coordination, agility, balance, flexibility, and dexterity. Additionally, the ability to throw and catch the frisbee accurately while maintaining control of bodily movements would also be essential. To perfect the trick, practicing these skills and building up muscle memory through repetition would be beneficial." }]

GPT4生成数据

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 数据准备

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]



图像精细描述

Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle. Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.

对话

Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV).

Question: Where is the vehicle parked?

Answer: The vehicle is parked in an underground parking area, likely in a public garage.

Question: What are the people in the image doing?

Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.

复杂推理

Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 数据准备

图-文对齐数据

```
[  
 {  
   "id": "004539375",  
   "image": "00453/004539375.jpg",  
   "conversations": [  
     {  
       "from": "human",  
       "value": "Render a clear and concise summary of the photo.\n<image>"  
     },  
     {  
       "from": "gpt",  
       "value": "select luxury furniture 3 - inch gel memory foam mattress  
topper"  
     }  
   ]  
},
```

图-文指令数据

```
[  
 {  
   "id": "000000033471",  
   "image": "coco/train2017/000000033471.jpg",  
   "conversations": [  
     {  
       "from": "human",  
       "value": "<image>\nWhat are the colors of the bus in the image?"  
     },  
     {  
       "from": "gpt",  
       "value": "The bus in the image is white and red."  
     },  
     {  
       "from": "human",  
       "value": "What feature can be seen on the back of the bus?"  
     },  
     {  
       "from": "gpt",  
       "value": "The back of the bus features an advertisement."  
     },  
     {  
       "from": "human",  
       "value": "Is the bus driving down the street or pulled off to the side?"  
     },  
     {  
       "from": "gpt",  
       "value": "The bus is driving down the street, which is crowded with people and other  
vehicles."  
     }  
   ]  
},
```

LLaVA图文对话系统搭建

图-文指令数据

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 自定义数据准备

图-文对齐数据

```
[  
{  
  "id": "任意一个非重复id，例如 0888888",  
  "image": "图像文件路径，例如： my_data/hongyadong.jpg"  
  "conversations": [  
    {  
      "from": "human",  
      "value": "请简要描述这幅图像。\n<image>"  
    },  
    {  
      "from": "gpt",  
      "value": "这是一张关于重庆洪崖洞的外景照片。"  
    }  
  ]  
},
```

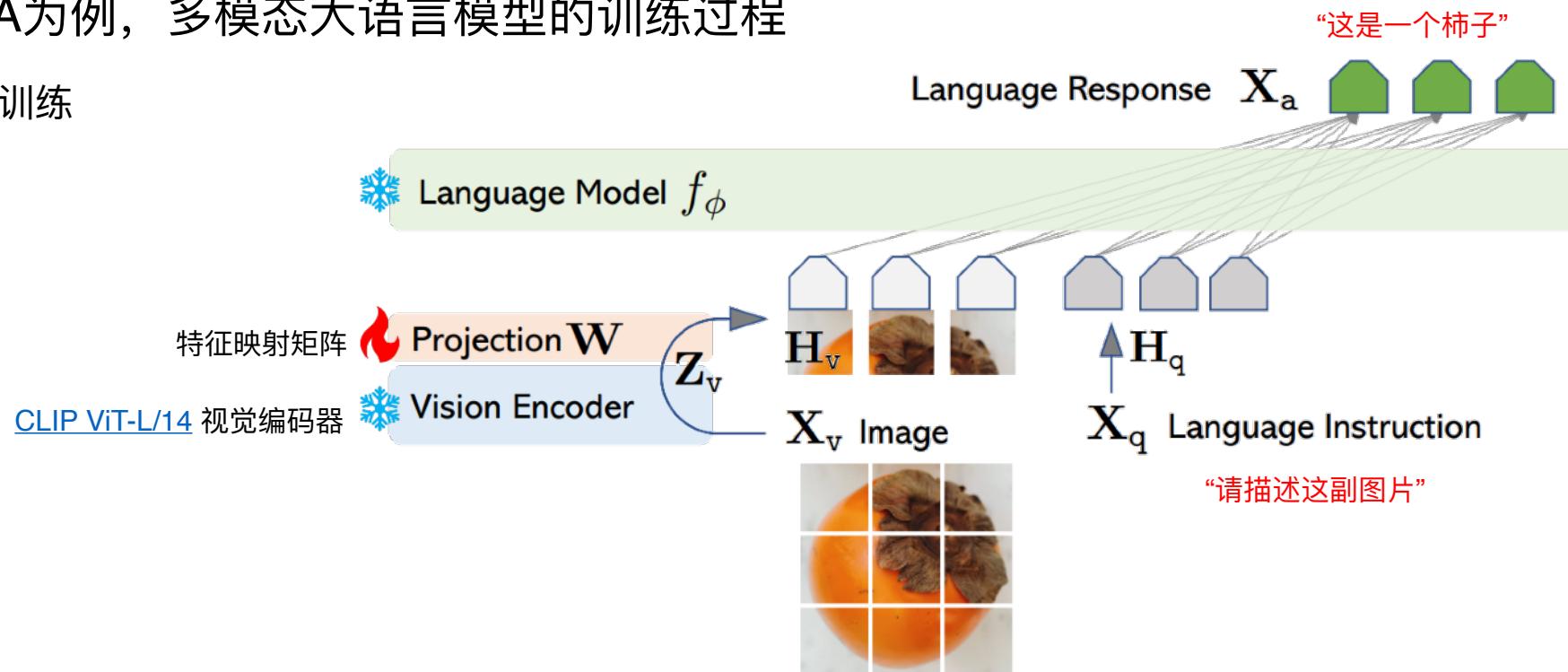


```
[  
{  
  "id": "任意一个非重复id，例如 0999999",  
  "image": "图像文件路径，例如： my_data/hongyadong.jpg",  
  "conversations": [  
    {  
      "from": "human",  
      "value": "<image>\n这张照片拍摄于哪里?"  
    },  
    {  
      "from": "gpt",  
      "value": "这是一张拍摄于重庆市洪崖洞景区的照片."  
    },  
    {  
      "from": "human",  
      "value": "图中有什么可以识别的文字内容?"  
    },  
    {  
      "from": "gpt",  
      "value": "图片右下角墙体上写着\"洪崖洞\"三个字."  
    }  
  ]  
},
```

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 模型训练



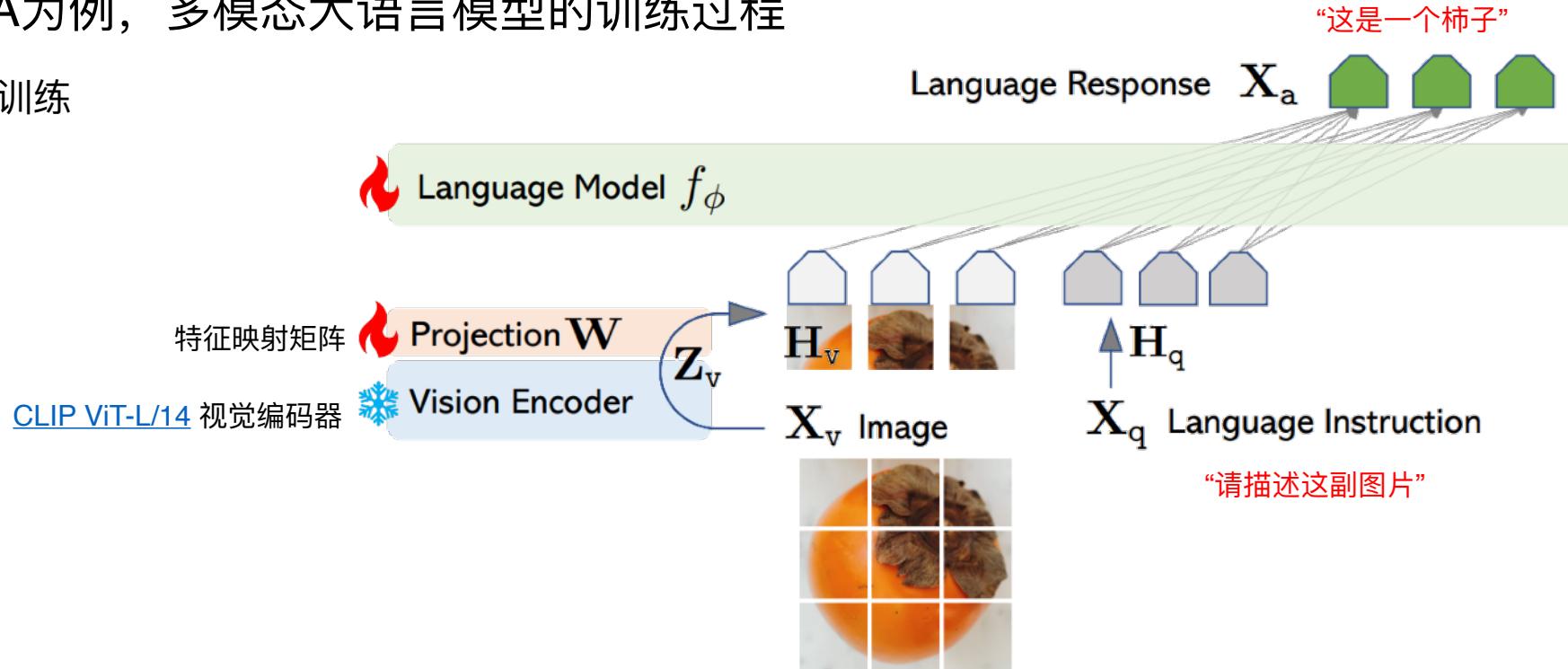
两阶段训练过程：

➤ 阶段一：特征对齐的预训练。尽更新特征映射矩阵

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

➤ 模型训练



两阶段训练过程：

- 阶段一：特征对齐的预训练。仅更新特征映射矩阵
- 阶段二：端到端微调。特征投影矩阵和LLM都进行更新

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的训练过程

- 模型训练

特征对齐的预训练

Hyperparameter	Global Batch Size	Learning rate	Epochs	Max length	Weight decay
LLaVA-v1.5-13B	256	1e-3	1	2048	0

- 8x A100 (80GB) 耗时5.5h
- 基于DeepSpeed ZeRO-2
- 输入图像分辨率336px
- 训练参数：特征映射层结构（2层全连接层）
- 训练脚本：

[LLaVA/scripts/v1_5/fine_tune.sh at main · haotian-liu/LLaVA · GitHub](https://github.com/haotian-liu/LLaVA/blob/main/LLaVA/scripts/v1_5/fine_tune.sh)

端到端微调

Hyperparameter	Global Batch Size	Learning rate	Epochs	Max length	Weight decay
LLaVA-v1.5-13B	128	2e-5	1	2048	0

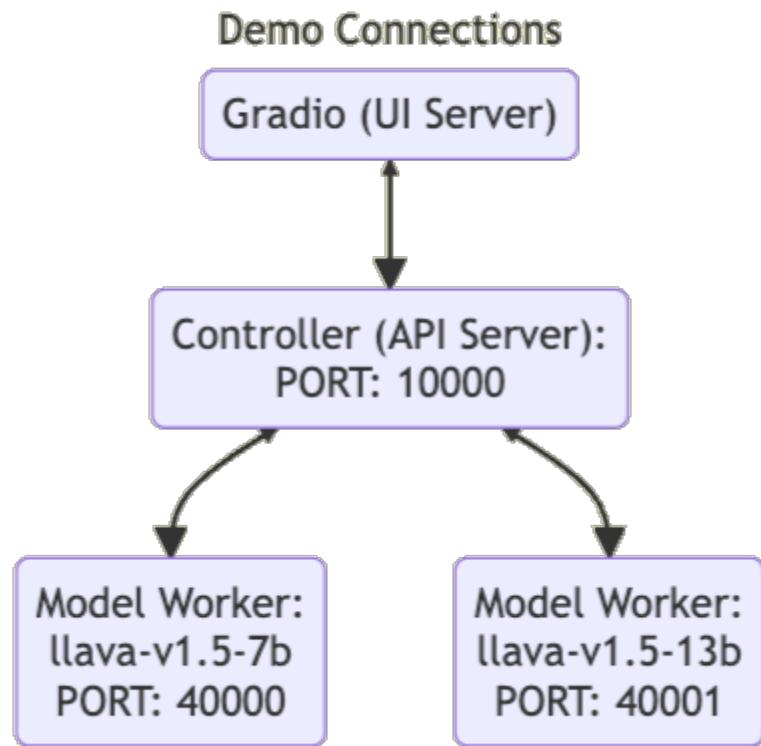
- 8x A100 (80GB) 耗时20h
- 基于DeepSpeed ZeRO-3
- 输入图像分辨率336px
- 训练参数：特征映射层结构（2层全连接层）以及LLM
- 训练脚本：

[LLaVA/scripts/v1_5/fine_tune.sh at main · haotian-liu/LLaVA · GitHub](https://github.com/haotian-liu/LLaVA/blob/main/LLaVA/scripts/v1_5/fine_tune.sh)

LLaVA图文对话系统搭建

- 以LLaVA为例，多模态大语言模型的部署过程

➤ 模型部署



1. 启动API server:

```
python -m llava.serve.controller --host 0.0.0.0 --port 10000
```

2. 启动WebUI:

```
python -m llava.serve.model_worker --host 0.0.0.0 --controller http://localhost:10000 --port 40000 --worker http://localhost:40000 --model-path liuhaotian/llava-v1.5-13b
```

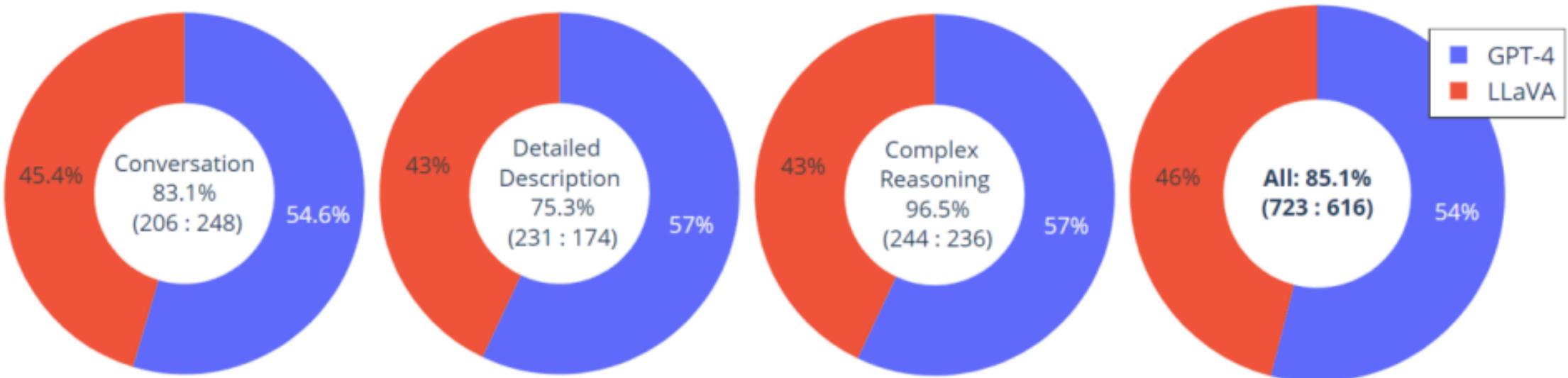
3. 启动Worker:

```
python -m llava.serve.model_worker --host 0.0.0.0 --controller http://localhost:10000 --port <different from 40000, say 40001> --worker http://localhost:<change accordingly, i.e. 40001> --model-path <ckpt2> --load-4bit
```

LLaVA图文对话系统搭建

- LLaVA模型效果评测

➤ 在线可试用demo：[LLaVA \(hliu.cc\)](https://llava.hliu.cc/) <https://llava.hliu.cc/>



- 基于30张未见图片做评测，每张图片对应3种instructions：对话、详细描述、推理。共计90组图文指令
- 使用GPT-4对LLaVA和GPT-4输出结果进行1-10的打分
- LLaVA 生成结果与GPT-4生成结果的平均相关程度达85.1%

图文多模态大词典

• MME评测集

[BradyFU/Awesome-Multimodal](#)

<https://github.com/BradyFU/Awesome-Multimodal>

Perception (Coarse-Grained Tasks)

Existence

[Y] Is there a **elephant** in this image?
[N] Is there a **hair drier** in this image?

[Y] Is there a **refrigerator** in this image?
[N] Is there a **donut** in this image?

Count

[Y] Is there a total of **two** person appear in the image?
[N] Is there only **one** person appear in the image?

[Y] Are there **two** pieces of pizza in this image?
[N] Is there only **one** piece of pizza in this image?

Position

[Y] Is the motorcycle on the **right** side of the bus?
[N] Is the motorcycle on the **left** side of the bus.

[Y] Is the baby on the **right** of the dog in the image?
[N] Is the baby on the **left** of the dog in the image?

Color

[Y] Is there a **red** coat in the image?
[N] Is there a **yellow** coat in the image?

[Y] Is there a **red** couch in the image?
[N] Is there a **black** couch in the image?

Perception (OCR Task)

OCR

[Y] Is the phone number in the picture "**0131 555 6363**"?
[N] Is the phone number in the picture "**0137 556 6363**"?

[Y] Is the word in the logo "**high time coffee shop**"?
[N] Is the word in the logo "**high tite cofeee shop**"?

Commonsense Reasoning

[Y] Should I **stop** when I'm about to **cross** the street?
[N] When I see the sign in the picture, can I **cross** the street?

[Y] Is there **one** real cat in this picture?
[N] Is there **two** real cats in this picture?

Numerical Calculation

[Y] Is the answer to the arithmetic question in the image **65**?
[N] Is the answer to the arithmetic question in the image **56**?

[Y] Should the value of "a" in the picture equal **3**?
[N] Should the value of "a" in the picture equal **2**?

Cognition (Reasoning Tasks)

Text Translation

老味道

共同努力

[Y] Appropriate to translate into English '**classic taste**'?
[N] Appropriate to translate into English '**strawberry flavor**'?

Code Reasoning

```
x = 10
if x < 20:
    print("Hello")
else:
    print("World")
```

[Y] Python code. Is the output of the code '**Hello**'?
[N] Python code. Is the output of the code '**World**'?

```
a = '12345'
b = a.find('1')
print(b)
```

[Y] Python code. Is the output of the code '**0**'?
[N] Python code. Is the output of the code '**1**'?

Perception (Fine-Grained Tasks)

Poster

[Y] Is this movie directed by **francis ford coppola**?
[N] Is this movie directed by **franklin j. schaffner**?

[Y] Is this movie titled **twilight (2008)**?
[N] Is this movie titled the **horse whisperer (1998)**?

[Y] Is the actor inside the red box called **Audrey Hepburn**?
[N] Is the actor inside the red box called **Chris April**?

[Y] Is the actor inside the red box named **Jim Carrey**?
[N] Is the actor inside the red box named **Jari Kinnunen**?

Scene

[Y] Does this image describe a place of **moat water**?
[N] Does this image describe a place of **marsh**?

[Y] Is this picture captured in a place of **galley**?
[N] Is this picture captured in a place of **physics laboratory**?

[Y] Is this an image of **Beijing Guozijian**?
[N] Is this an image of **Klinikkirche (Pfafferode)**?

[Y] Is this a picture of **Church of Saint Giles in Prague**?
[N] Is this a picture of **Pfarrkirche St. Martin an der Raab**?

Landmark

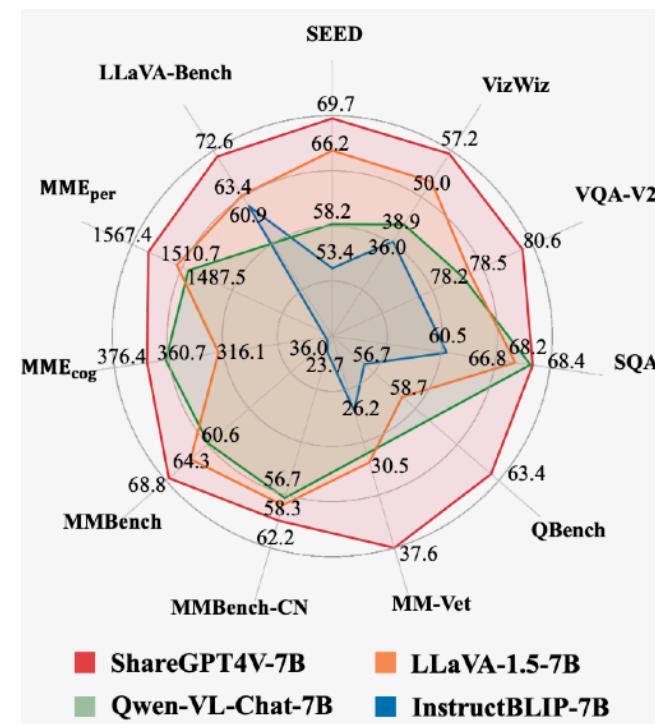
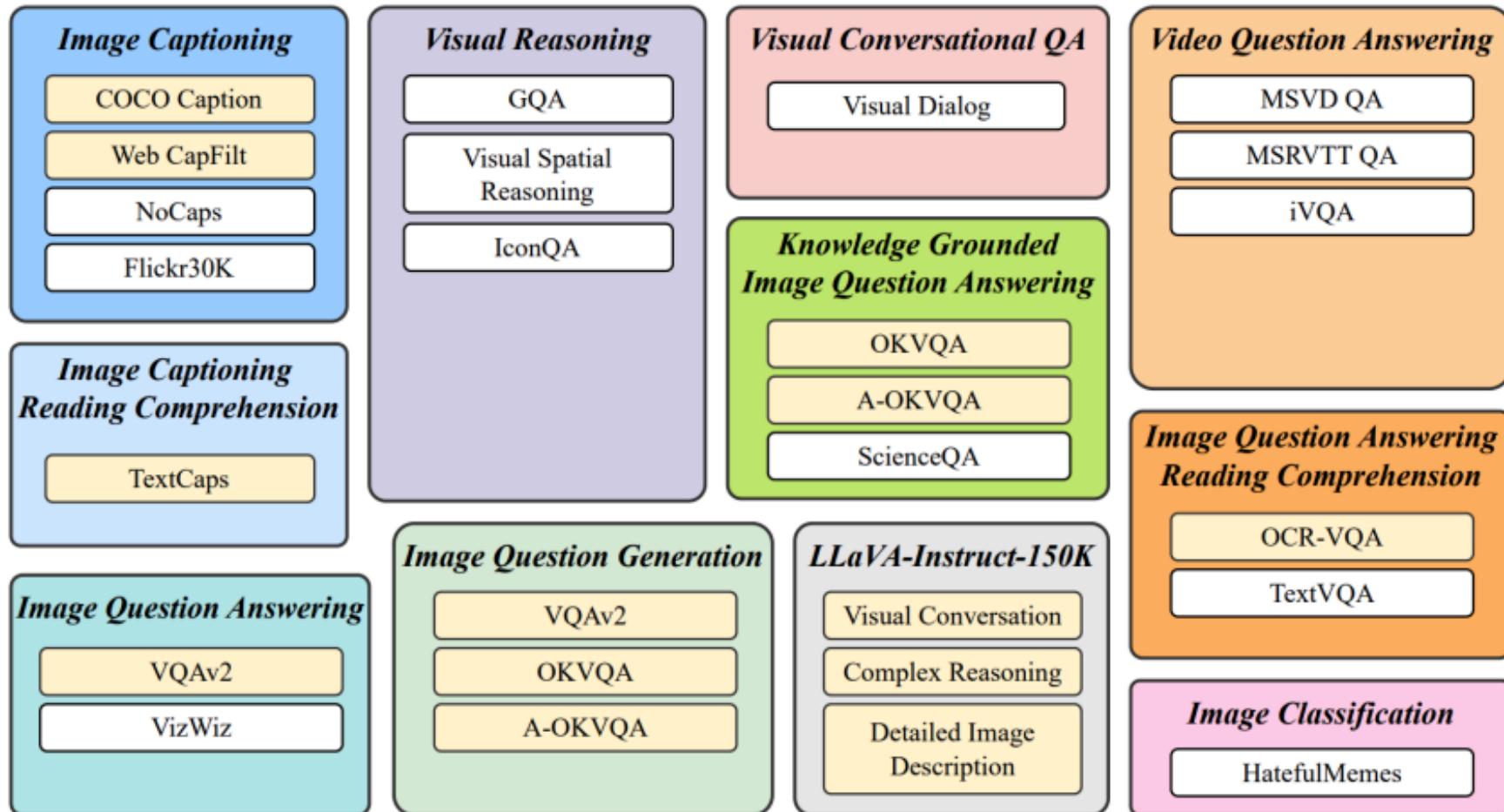
[Y] Is this artwork displayed in **musée du louvre**?
[N] Is this artwork displayed in **galleria nazionale d'arte moderna e contemporanea**?

[Y] Does this artwork belong to the type of **still-life**?
[N] Does this artwork belong to the type of **mythological**?

[Y] Is this artwork displayed in **musée du louvre**?
[N] Is this artwork displayed in **galleria nazionale d'arte moderna e contemporanea**?

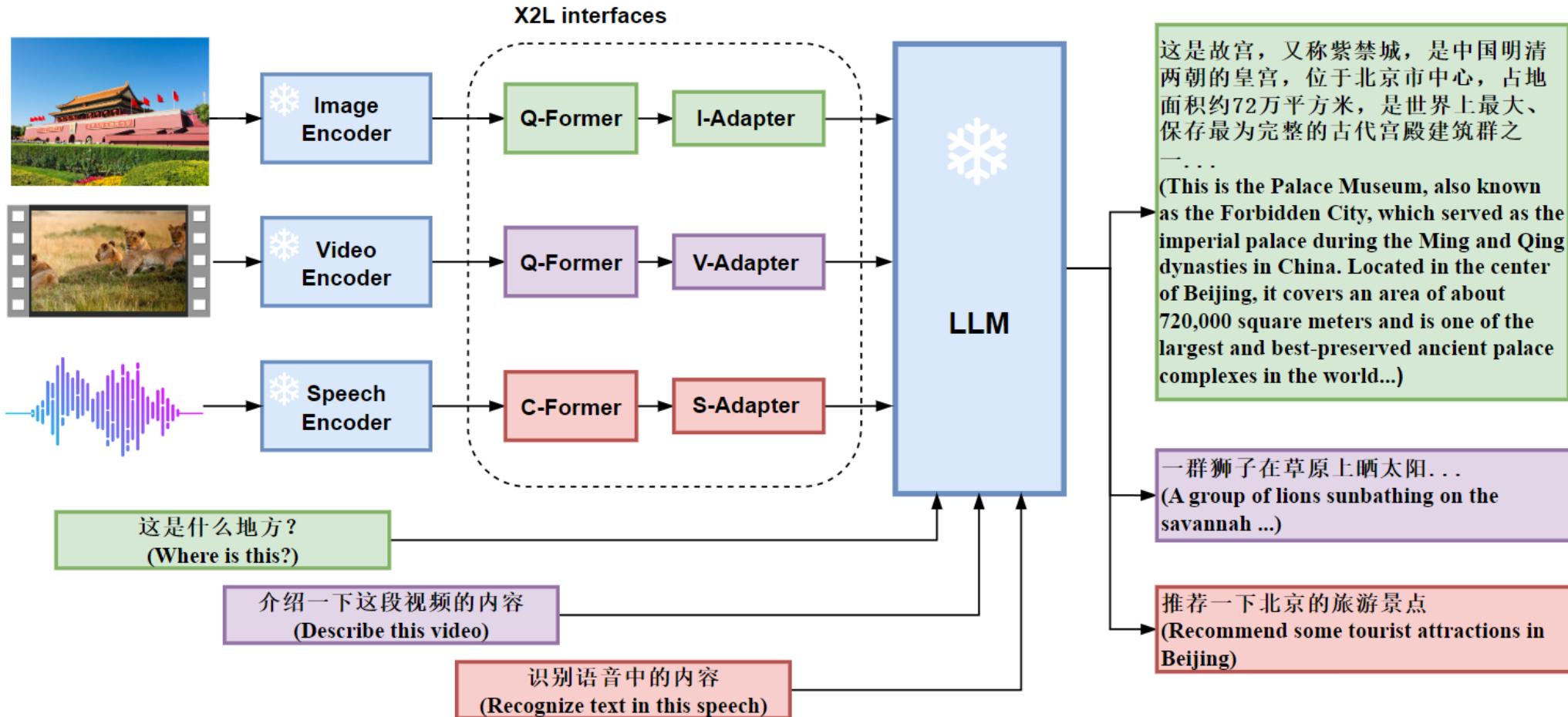
图文多模态大语言模型的评测

- 各类benchmark上的综合测试



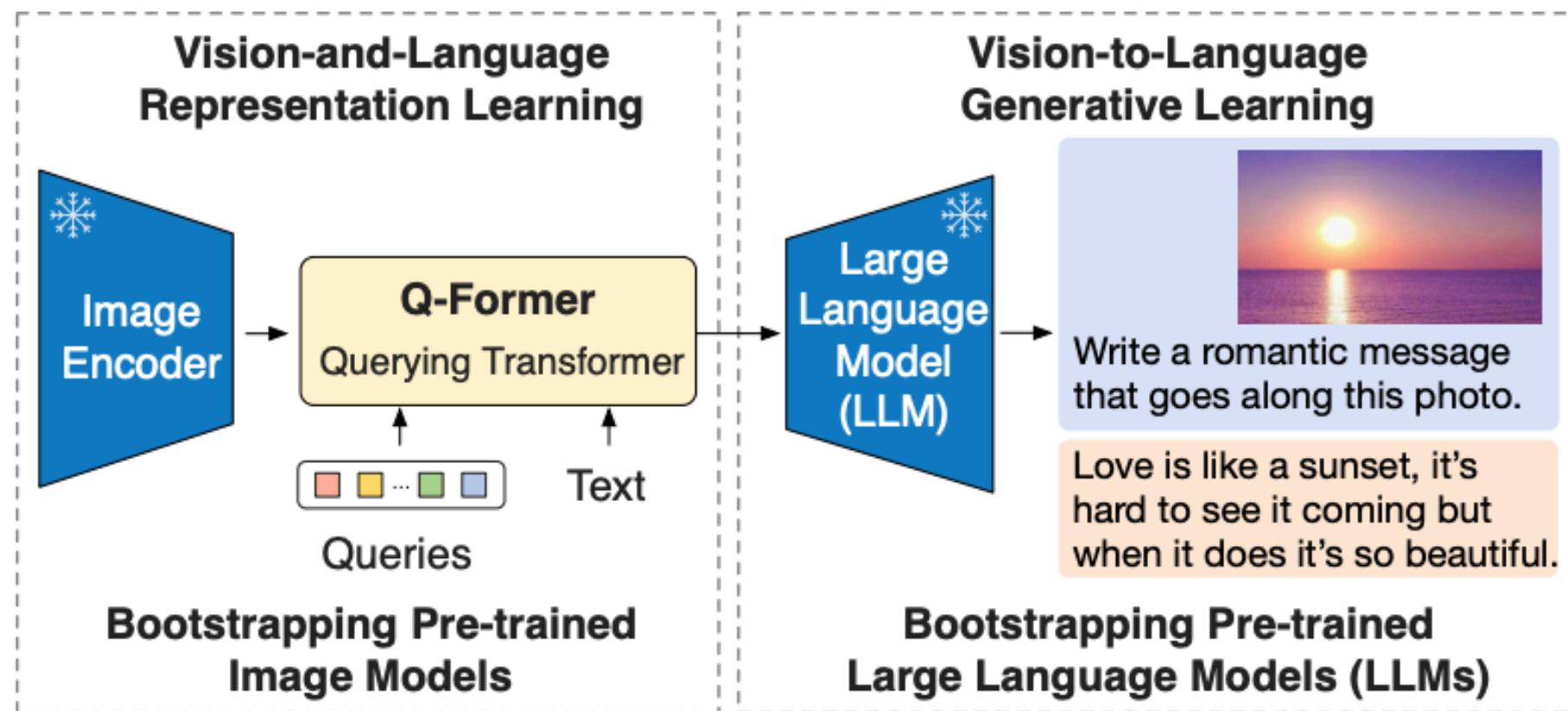
Beyond VL: 支持更多模态输入的大语言模型

- X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages



Beyond VL: 支持更多模态输入的大语言模型

- Q-former结构
- BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models



Beyond VL: 支持更多模态输入的大语言模型

- X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages

图像interface训练数据

Dataset	Data	Lang.
CC3M	Data Json	ZH
MSCOCO	Data Json	ZH
Visual Genome	Data Json	ZH
Flickr30k	Data Json	ZH
SBU	Data Json	ZH
AI Challenger	Data Json	ZH
Wukong captions	Data Json	ZH

语音interface训练数据

Dataset	Data	Lang.
AISHELL-2	Data Json	ZH
VSDial-CN	Data Json	ZH

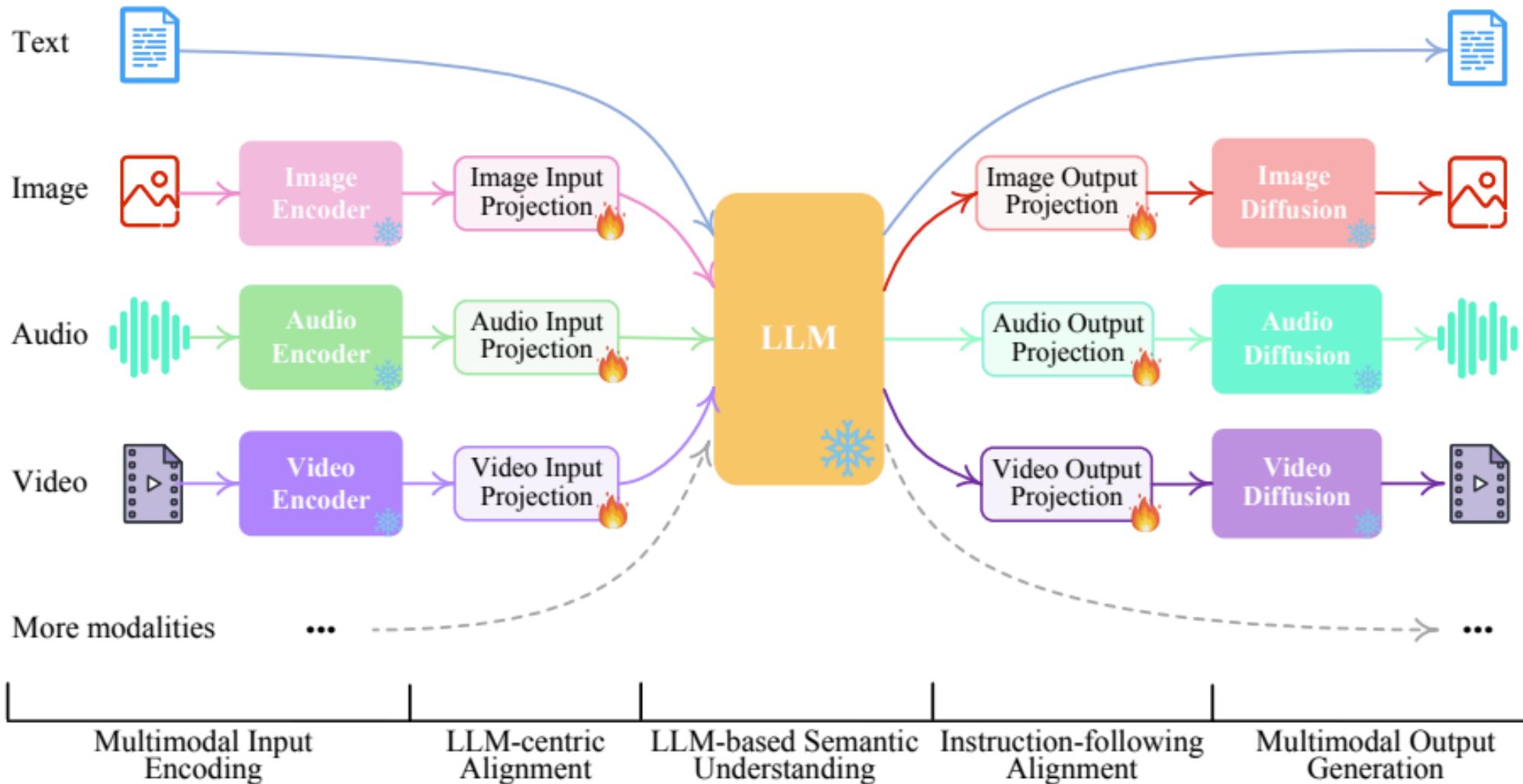
视频interface训练数据

Dataset	Video	Data
MSRVTT	Video Url	Data Json
ActivityNet	Video Url	Data Json

- Instruction Tuning阶段的数据未公开
- Q-former初始化自BLIP-2；
- C-former初始化自ASR模型
- 训练过程分为两个阶段：
 1. Adapter特征映射矩阵训练
 2. Instruction tuning训练

Beyond VL: 支持更多模态输入及输出的大语言模型

- NExT-GPT: Any-to-Any Multimodal LLM



Beyond VL: 支持更多模态输入及输出的大语言模型

- NExT-GPT: Any-to-Any Multimodal LLM

- 网络结构配置

Encoder		Input Projection		LLM		Output Projection		Diffusion	
Name	Param	Name	Param	Name	Param	Name	Param	Name	Param
Text	—	—	—	—	—	—	—	—	—
Image				Vicuna [12] 7B	Transformer 31M	SD [68]	1.3B		
Audio	ImageBind [25] 1.2B	Linear	4M	(LoRA 33M <td>Transformer 31M </td> <td>AudioLDM [51] 975M </td> <td></td> <td></td> <td></td>	Transformer 31M	AudioLDM [51] 975M			
Video					Transformer 32M	Zeroscope [8] 1.8B			

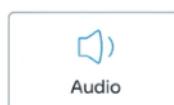
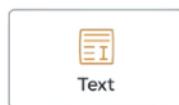


Encoder:

- ✓ ImageBind: One Embedding Space To Bind Them All

Decoder:

- ✓ SD: Stable Diffusion Model v1.5
- ✓ AudioLDM: TTS Diffusion Model [AudioLDM: Text-to-
Audio Generation with Latent Diffusion Models - Speech](#)

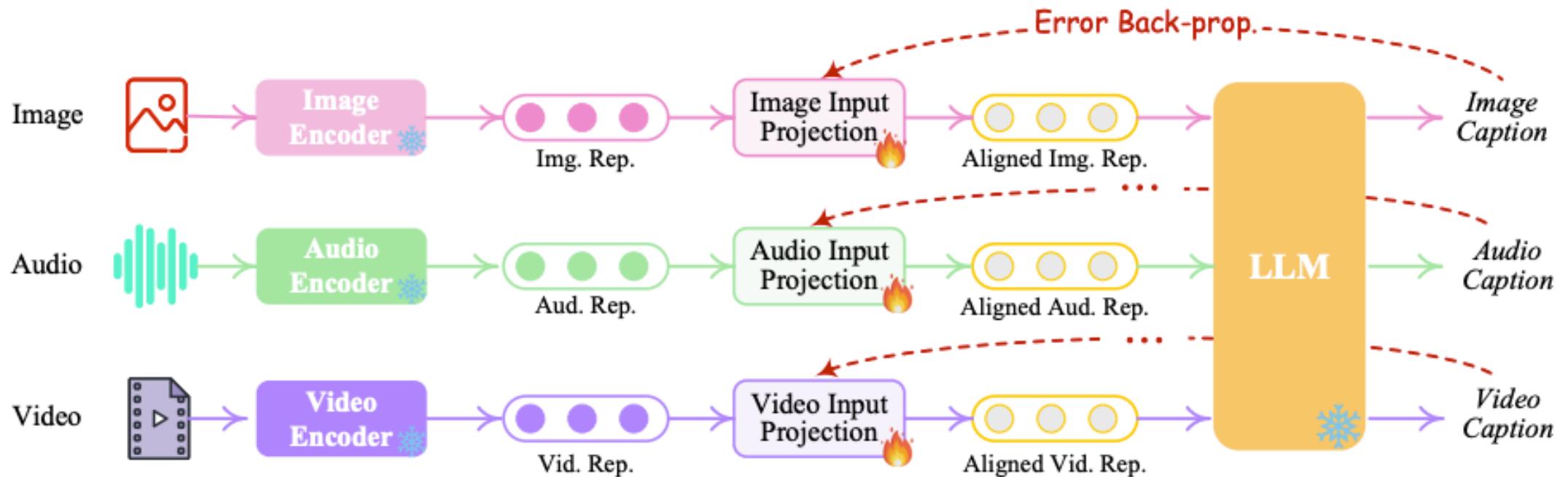


Beyond VL: 支持更多模态输入及输出的大语言模型

- NExT-GPT: Any-to-Any Multimodal LLM

三阶段训练过程：

➤ 阶段一：以LLM为中心的Encoder特征对齐：只更新input projection layer

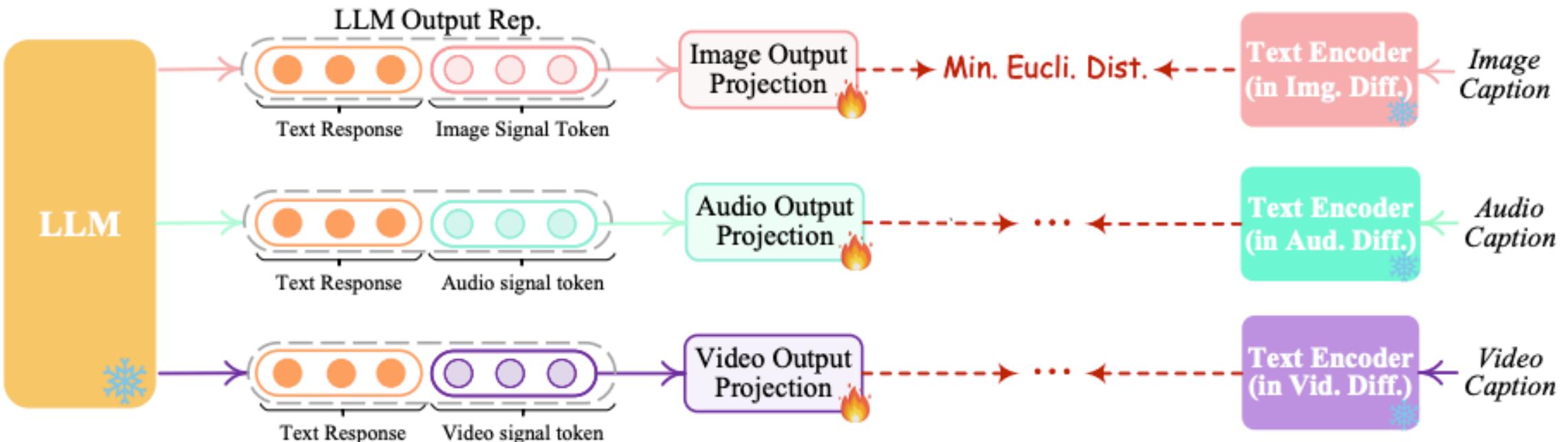


Beyond VL: 支持更多模态输入及输出的大语言模型

- NExT-GPT: Any-to-Any Multimodal LLM

三阶段训练过程：

- 阶段一：以LLM为中心的Encoder特征对齐：只更新input projection layer
- 阶段二：Decoder段输出结果与指令对齐：只更新output projection layer



本质上是在做text conditional input的对齐，所使用训练数据还是文本数据

Beyond VL: 支持更多模态输入及输出的大语言模型

- NExT-GPT: Any-to-Any Multimodal LLM

三阶段训练过程：

- 阶段一：以LLM为中心的Encoder特征对齐：只更新input projection layer
- 阶段二：Decoder段输出结果与指令对齐：只更新output projection layer
- 阶段三：指令微调：LLM Lora, input以及output projection layers

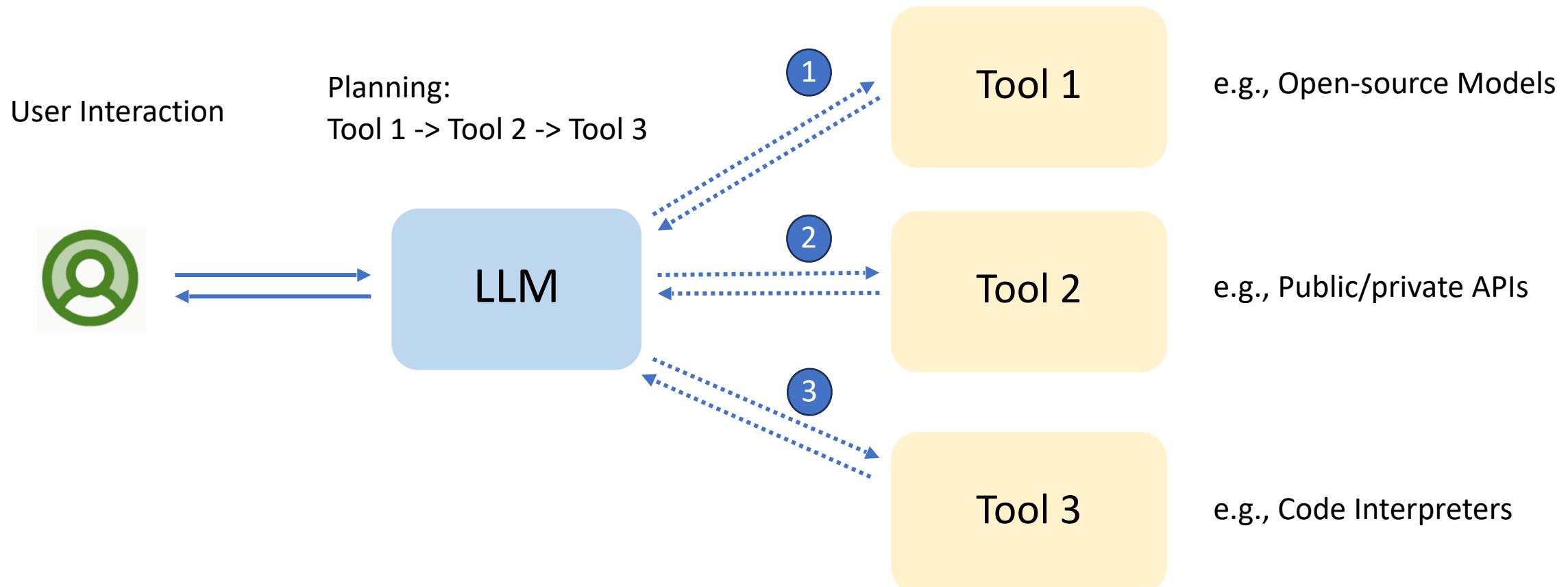
Beyond VL: 支持更多模态输入及输出的大语言模型

- NExT-GPT: Any-to-Any Multimodal LLM

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
► Existing data							
MiniGPT-4 [109]	CC [10], CC3M [71]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [47]	SD [68]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [104]	COCO [50]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [106]	MS-COCO [50], VG [41]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [104]	COCO [50], CC3M [71], LAION [70]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [44]	WebVid [5]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [54]	ActivityNet [28]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [103]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [15]	Multiple	T+I/V→T	Auto	✗	-	-	~ 1.6M
MIMIC-IT [42]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [77]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [107]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M ³ IT [45]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [97]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [108]	Clotho [20], VGGSS [11]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [96]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
► In this work							
T2M	Webvid [5], CC3M [71], AudioCap [38]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

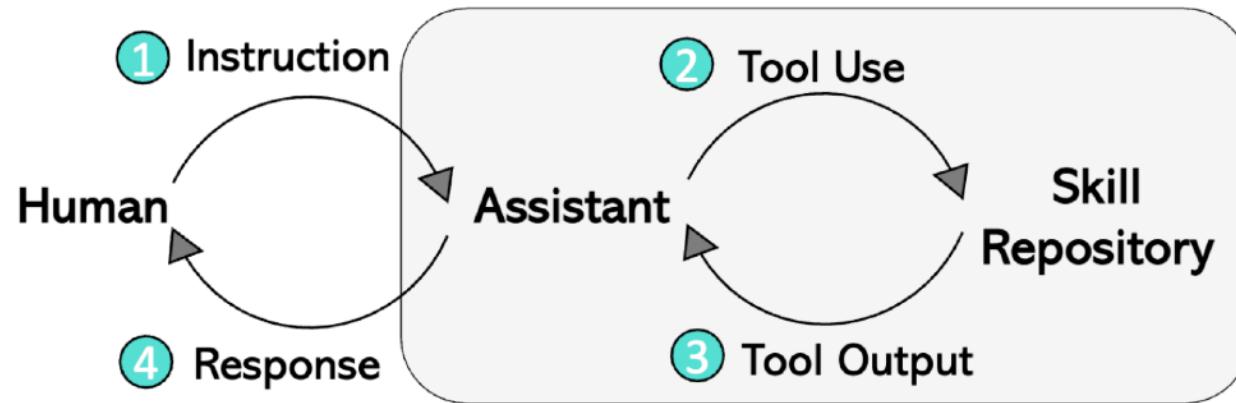
使用多模态大语言模型完成更多任务

- Multimodal Agents - Chaining Multimodal Experts with LLMs



LLaVA-Plus

- LLaVA-Plus: Large Language and Vision Assistants that Plug and Learn to Use Skills



LLaVA Plus的四步过程:

- ① 用户提供任务文本指令以及相关图片输入
- ② LLM分析指令和图像，选择是否调用额外的工具完成指令，如果是，则生成调用工具所需要的 prompt
- ③ 运行工具，返回工具输出结果给LLM
- ④ LLM汇总工具所输出结果，根据用户事先提供的文本指令和图片输入，生成相应的回复。

LLaVA-Plus

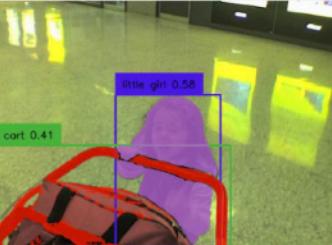
- LLaVA-Plus: Large Language and Vision Assistants that Plug and Learn to Use Skills



LLaVA-Plus

- LLaVA-Plus: Large Language and Vision Assistants that Plug and Learn to Use Skills

➤ 训练样本

Human ($\mathbf{X}_q, \mathbf{I}_q$)	<p>Can you segment the girl and the cart in the image and describe their relationship? <i>(The source image from COCO with ID 569536)</i></p> 
Assistant ($\mathbf{X}_{\text{skill_use}}$)	<p>“thoughts” To accurately segment and identify the girl and the cart in the image, we can leverage the grounding_dino model for open-set detection and the segmentation model SAM. Integrating these models will enable us to generate a comprehensive summary of the segmented subjects. “actions” [“API_name”: “grounding_dino+sam”, “API_params”: “caption”: “little girl and cart .”] “value” I will use grounding_dino+sam to help to answer the question. Please wait for a moment.</p> <p><i>Tool execution: Grounding DINO + SAM</i></p> <p>grounding_dino+sam model outputs: ‘boxes’: [[0.35, 0.37, 0.66, 0.97], [0.0, 0.57, 0.69, 1.0]], ‘logits’: [0.58, 0.41], ‘phrases’: [‘little girl’, ‘cart’]</p>
Skill Repo. Human ($\mathbf{X}_{\text{skill_result}}$)	<p>Can you segment the girl and the cart in the image and describe their relationship?</p>
Assistant ($\mathbf{X}_{\text{answer}}$)	<p>“thoughts” Thanks to the output of grounding_dino+sam. I can answer the question better. “actions” [] “value” Sure! I segment and identify the little girl and the cart in the image.</p> <ul style="list-style-type: none">- Centered in the image, a little girl is discernible with bounding box coordinates [0.35, 0.37, 0.66, 0.97]. She’s seen actively maneuvering a cart within the airport’s ambiance.- To the bottom left of the image, a cart can be pinpointed, bounded by the coordinates [0.0, 0.57, 0.69, 1.0]. It’s laden with various luggage items, indicating travel preparations or arrivals. <p>The interaction between the girl and the cart epitomizes a typical airport scenario: a traveler utilizing the cart to facilitate her movement with luggage. This tableau, whether signifying an arrival or departure, encapsulates the vibrancy and expectation inherent to air travel.</p> 

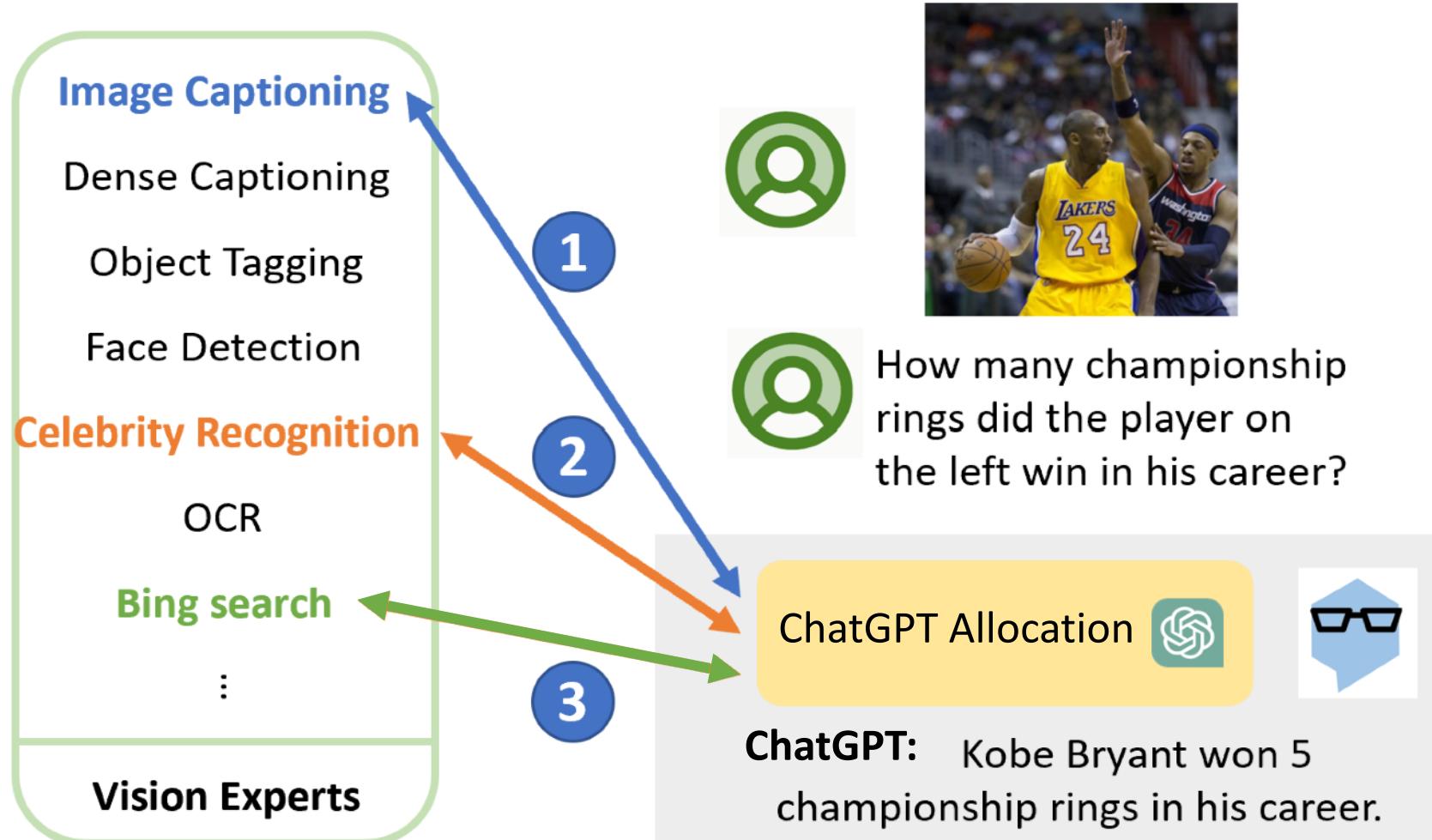
Thought

Action

Value

使用纯语言模型完成视觉推理任务

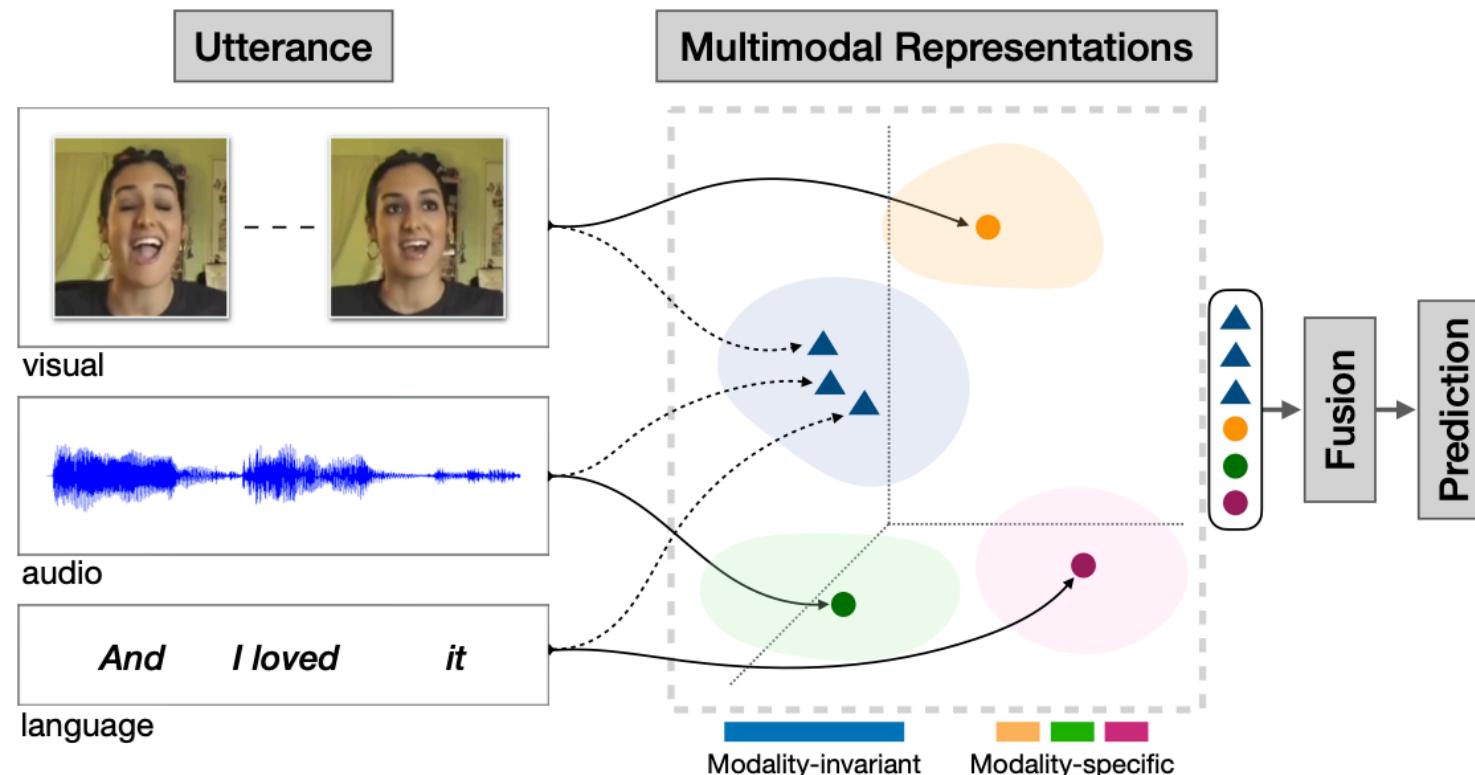
- MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action



Multimodal Models

- 多模态学习一般范式

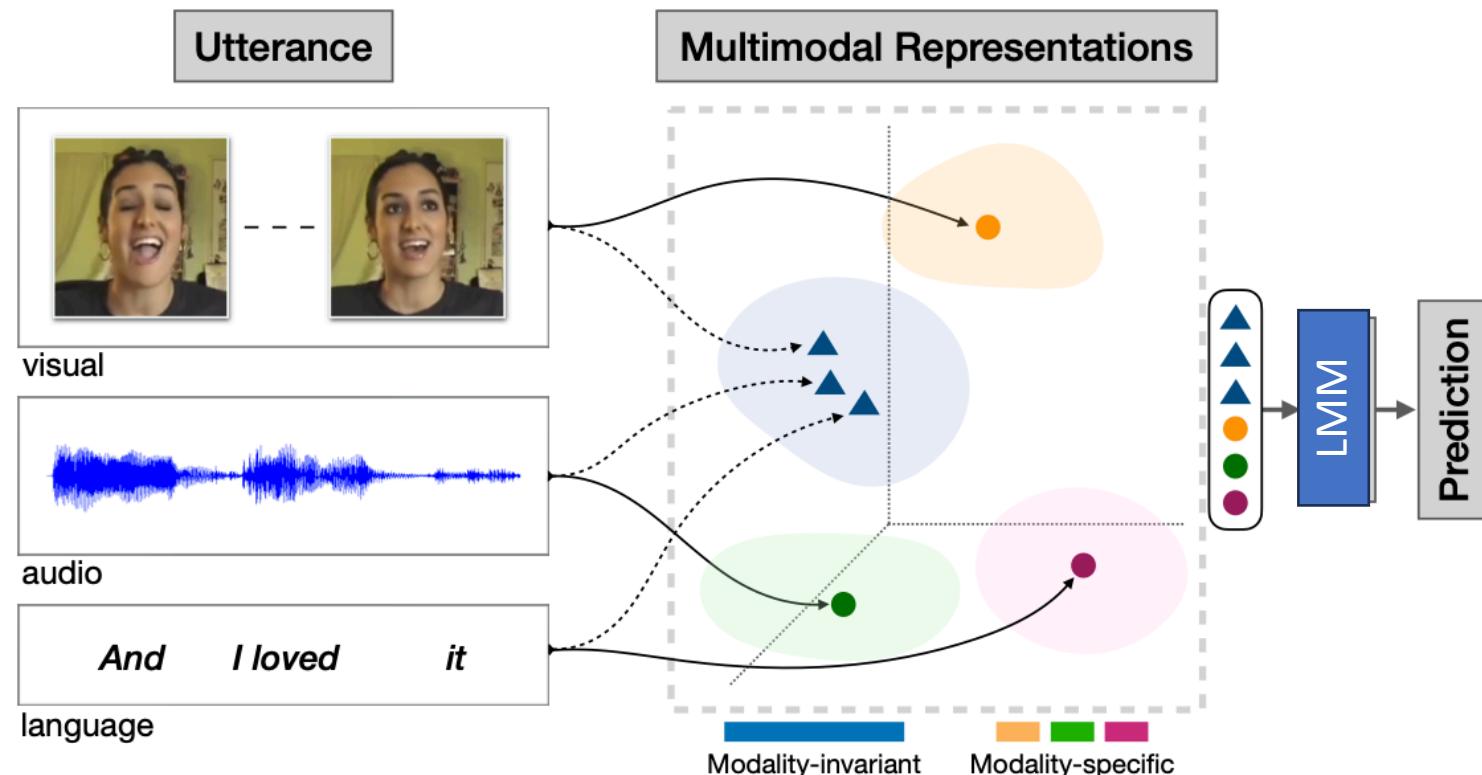
以 视觉、文本、音频三种模态为例：



Multimodal Large Language Models

- 多模态大语言模型学习一般范式

以 视觉、文本、音频三种模态为例：



- 多模态学习的核心在于特征对齐；
- 多模态大语言模型的本质在于All-to-one (LMM) 的特征对齐范式；
- 多模态大语言模型正处于飞速发展阶段

相比于LLaMA与textonly GPT4之间的差距，开源多模态大语言模型与GPT4V之间的差距更小。

- 多模态大模型领域的发展现状：
技术发展先于应用，存在更多想象空间。

- 多模态大模型最全收录：
包括多模态指令微调、多模态上下文学习、多模态Agent、多模态数据集等各类信息：
<https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models>

- 多模态大模型榜单：
<https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation>