



多模态大语言模型领域进展分享

殷述康 / 博士在读

DataFunSummit # 2024

DataFun.



- 背景介绍
- 多模态大语言模型介绍
- 多模态大语言模型演进
- 团队相关工作介绍
- 未来展望



01

背景介绍



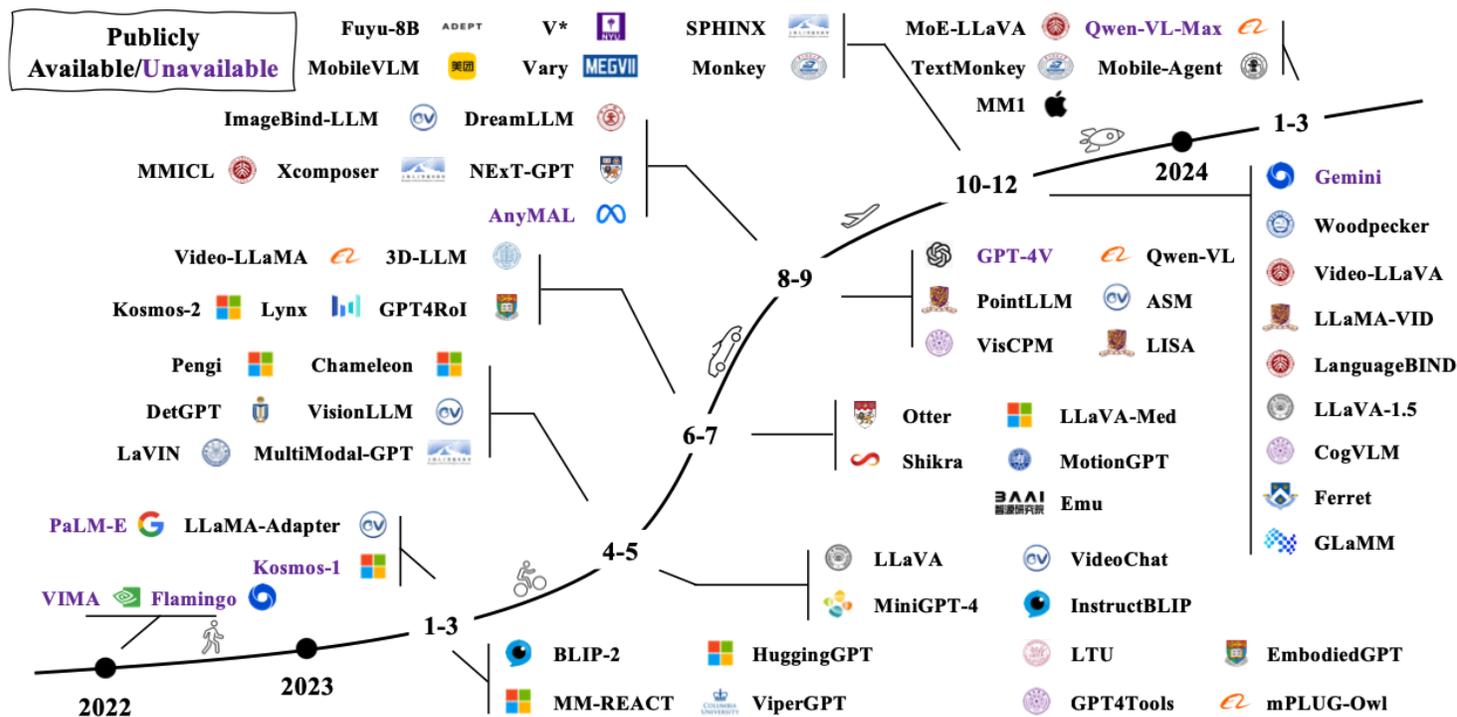
背景-LLM正走向多模态

- 大语言模型(LLM)是近几年来最火热的方向之一
 - 可以解决各种传统NLP任务, 如文本分类、命名实体识别等
 - 可以做更高级的任务
 - 作为聊天机器人, 按照要求扮演某个角色 (强大的指令遵循能力)
 - 做高阶的推理任务, 如写代码、解数学问题等 (强大的推理能力, CoT进一步增强)
- 然而LLM存在固有的限制
 - 无法处理多模态的输入, 导致有些任务无法做或者很难做, 如根据网站截图给出源代码、理解一张表情包的含义
 - 无法获取更多的多模态的世界知识, 如名画、名人等



背景-LLM正走向多模态

- 多模态大语言模型(MLLM)的兴起
 - 就在近两年，工业和学术界都在积极转向多模态模型，比如OpenAI的GPT-4V、GPT-4o谷歌的Gemini-Pro



在短短的两年间，已有上百个模型涌现，包括大企业的闭源模型和学术社区的开源模型探索。

背景-MLLM能做什么

- 能做传统视觉/多模态任务

Prompt:

描述這張圖片



GPT-4V:

這張圖片是一隻小貓在草地上走路的畫面。小貓是橘色的，有藍色的眼睛，看起來很可愛。它的身體很小，四肢纖細，正在小心翼翼地草地上行走。

Caption任务

Prompt:

Count the number of people in the image.



GPT-4V:

There are 4 people in the image.

计数任务

Prompt:

Localize each person in the image using bounding box. The image size is (width, height) = (998,786)



GPT-4V:

There is only one person in the image, so the bounding box for this person would be:

- Top left: (254, 118)
- Bottom right: (731, 693)

Explanation: The top left point of the bounding box is at $(x, y) = (254, 118)$, which is the top left corner of the person in the image. The bottom right point of the bounding box is at $(x, y) = (731, 693)$, which is the bottom right corner of the person in the image. This bounding box encompasses the entire person in the image.

Visualization of Detected Bbox:



定位任务

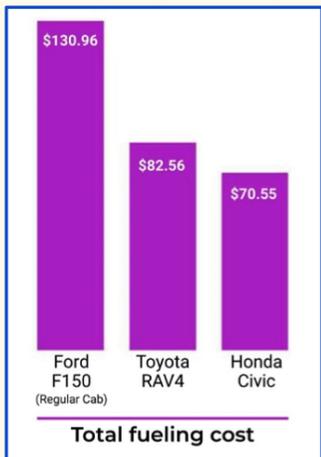
背景-MLLM能做什么

- 能做更复杂的复合型任务, 比如基于视觉的感知和理解任务;

Chart Understanding and Reasoning

Prompt:

What is the average total fueling cost excluding the Ford F150?



GPT-4V:

To find the average total fueling cost excluding the Ford F150, we need to add the total fueling costs of the Toyota RAV4 and Honda Civic and divide by 2 (since there are 2 vehicles in the calculation).

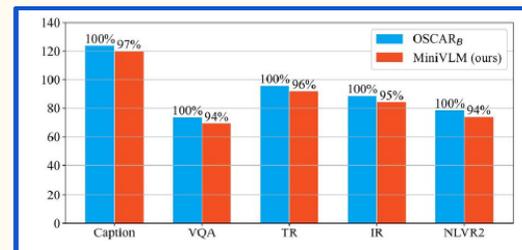
$$(82.56 + 70.55) \div 2 = 76.55$$

Therefore, the average total fueling cost excluding the Ford F150 is \$76.55.

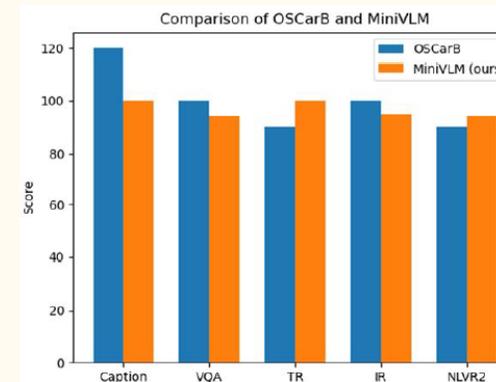
图表推理

Prompt:

Write python codes to generate similar figures.



GPT-4V:



根据图表写代码



02

多模态大语言模型介绍



介绍-MLLM的基本方面

- 由于大企业的模型是闭源的，学术界正积极研究探索开源的模型。

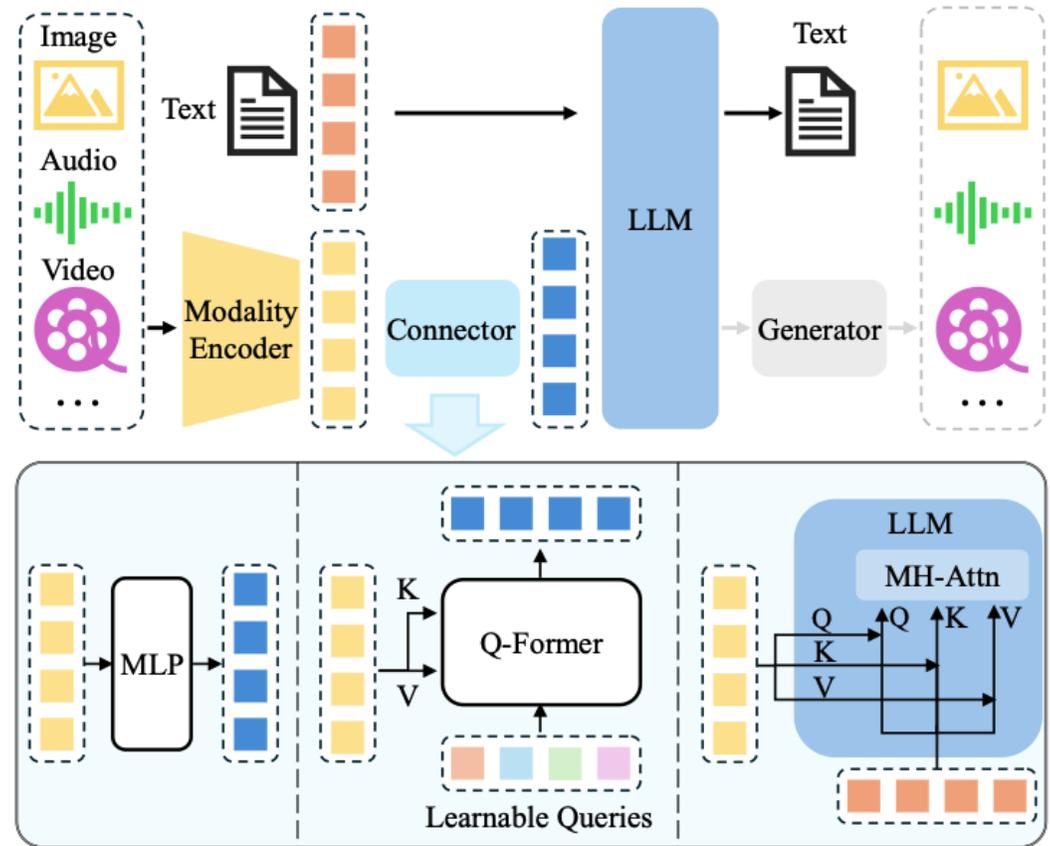
模型架构

数据与训练方法

模型评估

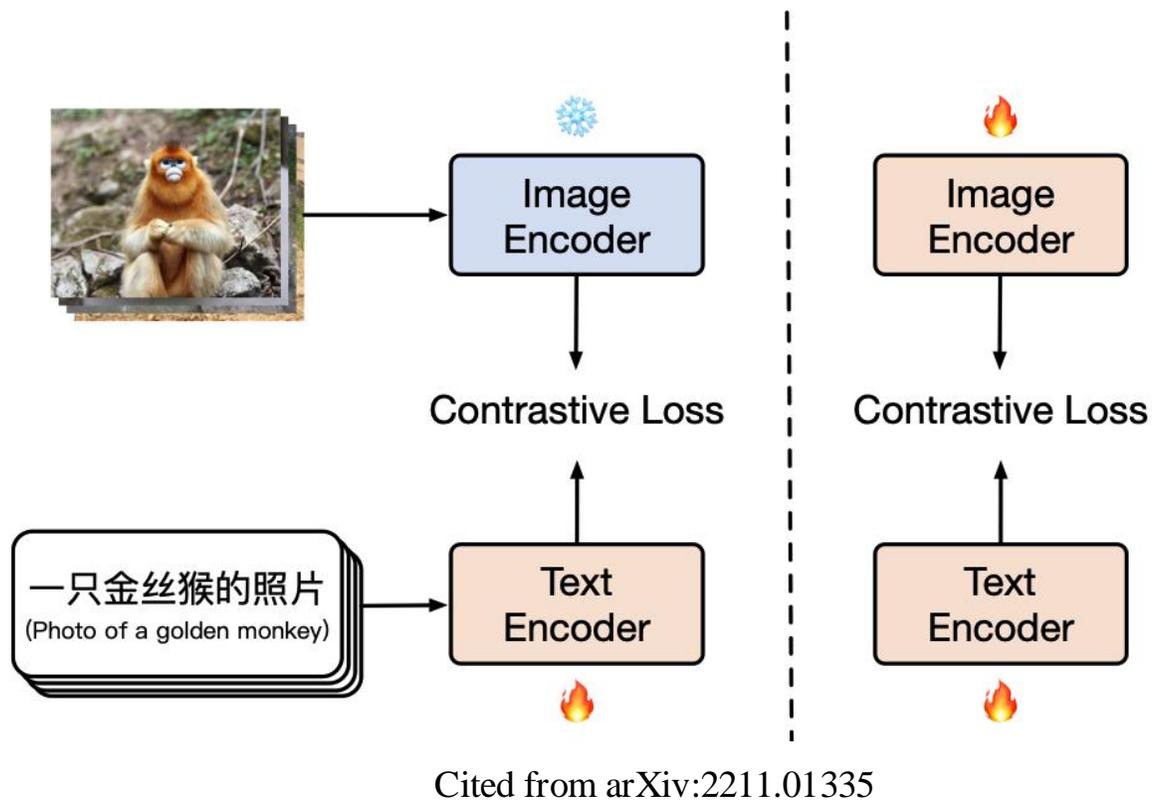
介绍-MLLM的架构

- 常用的架构一般包含三个部分：
 - 编码器
 - 连接器
 - 大语言模型



介绍-MLLM的架构

- 视觉编码器
 - 常用的是基于CLIP预训练的ViT
 - 对于常见的224x224分辨率图片，patch大小为14，最后共得到 $14 \times 14 = 256$ 个tokens



介绍-MLLM的架构

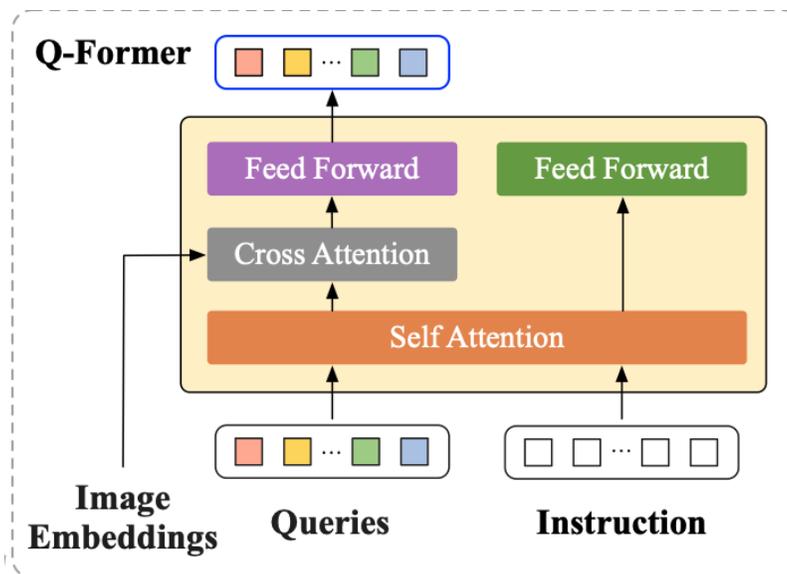
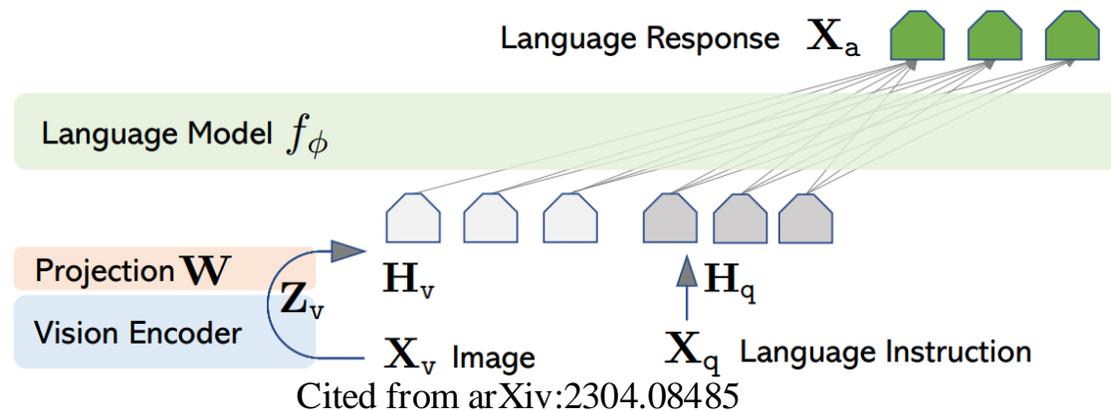
- 连接器

- MLP结构

- 不改变视觉token的数量，使用线性层或者多层感知机做投影。

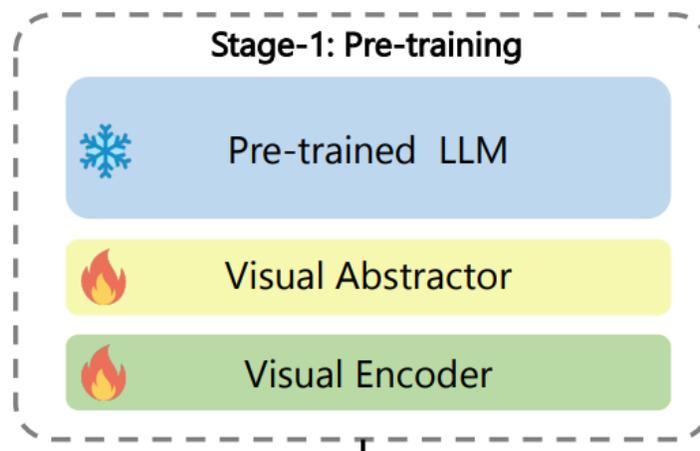
- Q-Former

- 压缩图片token至固定的数量，提高运算效率
 - Q指query，使用一组可学习的query向量从视觉token中抽取更紧凑的表征信息



介绍-MLLM的数据和训练方法

- 第一阶段：模态对齐训练
 - 将视觉的语义空间与文本空间对齐
 - 一种做法是冻结LLM，训练视觉编码器和连接器
 - 通常使用大量的图文配对数据训练，如caption数据。输入图片，预测图片的文本描述



A horse carrying a large load of hay and two people sitting on it.



Bunk bed with a narrow shelf sitting underneath it.



The man at bat readies to swing at the pitch while the umpire looks on.

介绍-MLLM的数据和训练方法

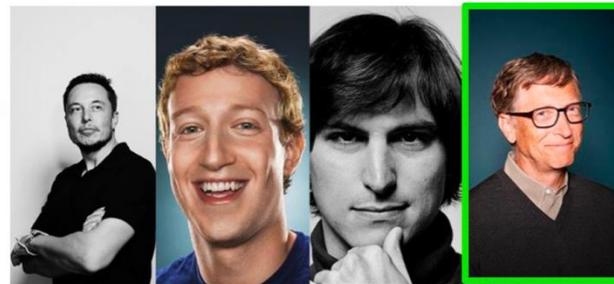
- 第二阶段：指令微调训练
 - 提升模型的指令遵循能力，学习泛化到各种任务
 - 通常使用各种任务的数据，如VQA数据、检测数据等，并改造为指令格式（多轮对话形式）
 - 输入图片和指令，预测回答



右边的男人是谁？



The man on the right in the image is Yao Ming.



图中<box>(750,0),(999,999)</box>中是谁

图中是比尔盖茨，微软的创始人之一。



描述下他的打扮

图中是比尔·盖茨本人，他穿着一件黑色的毛衣，看起来比较休闲。



介绍-MLLM的评测

- 常规任务Benchmark

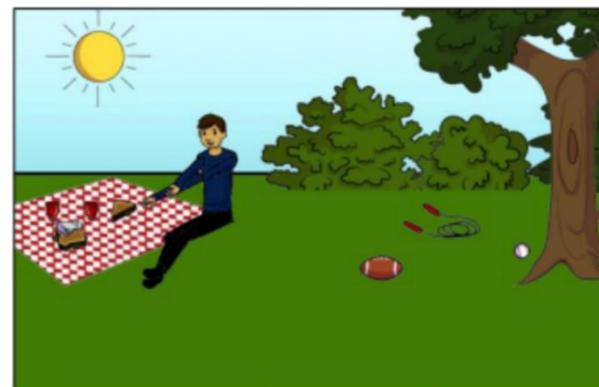
- 一般聚焦某个具体的特定任务，关注特定指标
- VQA任务：要求模型基于图片内容进行回答，常关注浅层元素如物体类别、属性，有时涉及浅层推理，回答一般比较简单。
 - 一般使用准确率作为评估指标



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

介绍-MLLM的评测

- 专门Benchmark
 - 除了基本的感知能力，还关注推理等能力，任务一般更为复杂、困难

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?

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'*?



03

多模态大语言模型演进

演进-更高的分辨率

- 更高的分辨率能看见更多的信息，更准确地回答问题



酒精度: $\geq 15.5\%$ vol



What is the alcohol content of the product in the picture?

15.5%



..... of the product in the picture is 15.5% vol.



The alcohol content of the product in the picture is 40%.



The alcohol content of the product in the picture is 30%.



The alcohol content of the product in the picture is 15%.



保质期: 90天



What is the shelf life of the product in the picture?

90 days



The shelf life of the product shown in the picture is 9 months.



The shelf life of the product in the picture is 18 months.



The shelf life of the product in the picture is 12 months.



The shelf life of the product in the picture is 2 years.



净含量: 125ml



What is the net weight of the product in the picture?

896_Resolution: 125 ml

784_Resolution: 250 ml



672_Resolution: 50 g

演进-更高的分辨率

- 如何提高视觉编码器的分辨率？
 - 思路一：使用更大的分辨率进行微调 (Qwen-VL: 224 -> 448, arXiv:2308.12966)

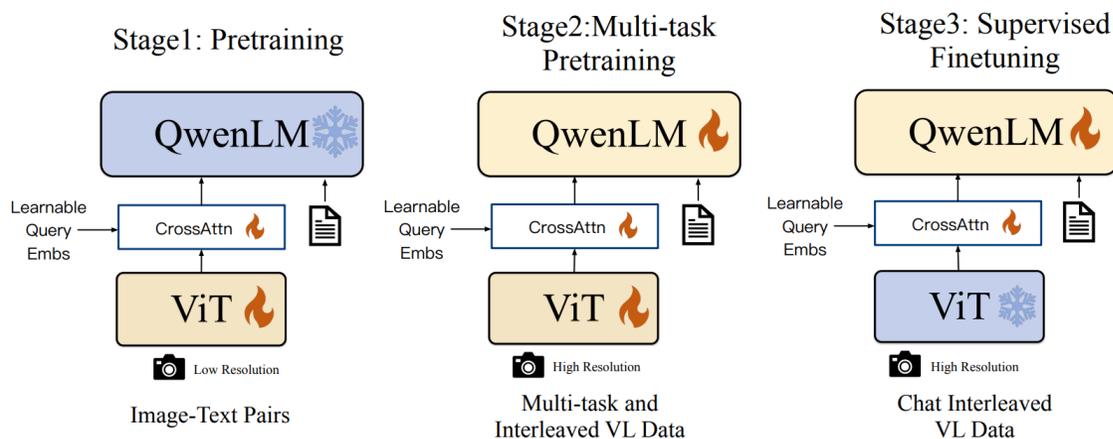


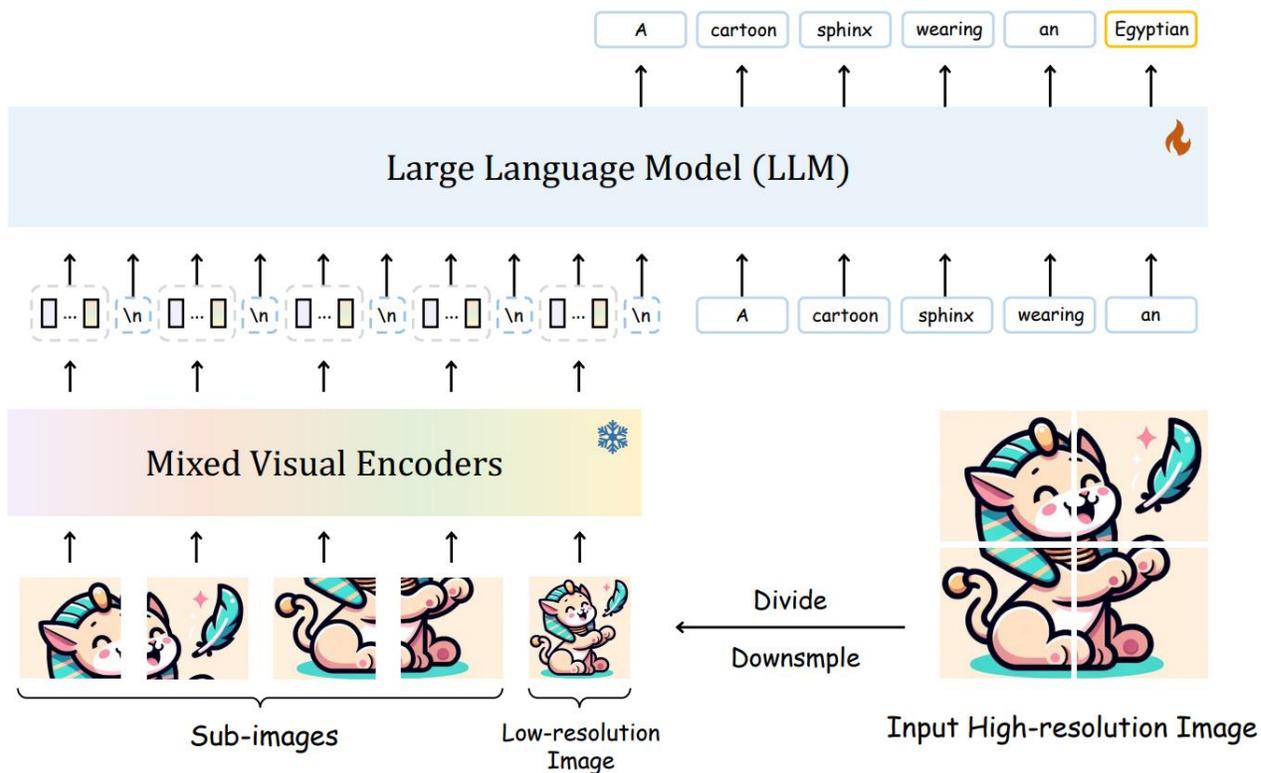
Figure 3: The training pipeline of the Qwen-VL series.

Table 8: Training hyperparameters of Qwen-VL

Configuration	Pre-training	Multi-task Pre-training	Supervised Fine-tuning
ViT init.	Open-CLIP-bigG	Qwen-VL 1st-stage	Qwen-VL 2nd-stage
LLM init.	Qwen-7B	Qwen-7B	Qwen-VL 2nd-stage
VL Adapter init.	random	Owen-VL 1st-stage	Owen-VL 2nd-stage
Image resolution	224 ²	448 ²	448 ²

演进-更高的分辨率

- 如何提高视觉编码器的分辨率？
 - 思路二：将大分辨率的图片切成多块，每块依然是原来的分辨率
 - 一般保留一张低分辨率的完整图片作为全局特征



演进-更丰富的输入形式

- 从只支持单图输入到支持多图、视频输入

Counting

Question: How many cat are in these two images?

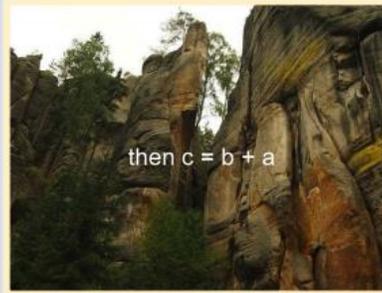
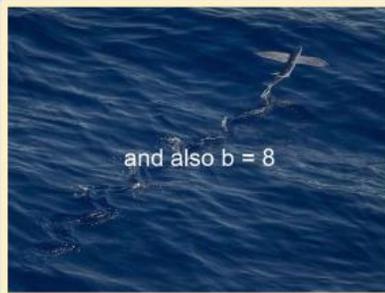
Answer: 2



Synthetic Visual Logic Chain

Question: What's the value of c ?

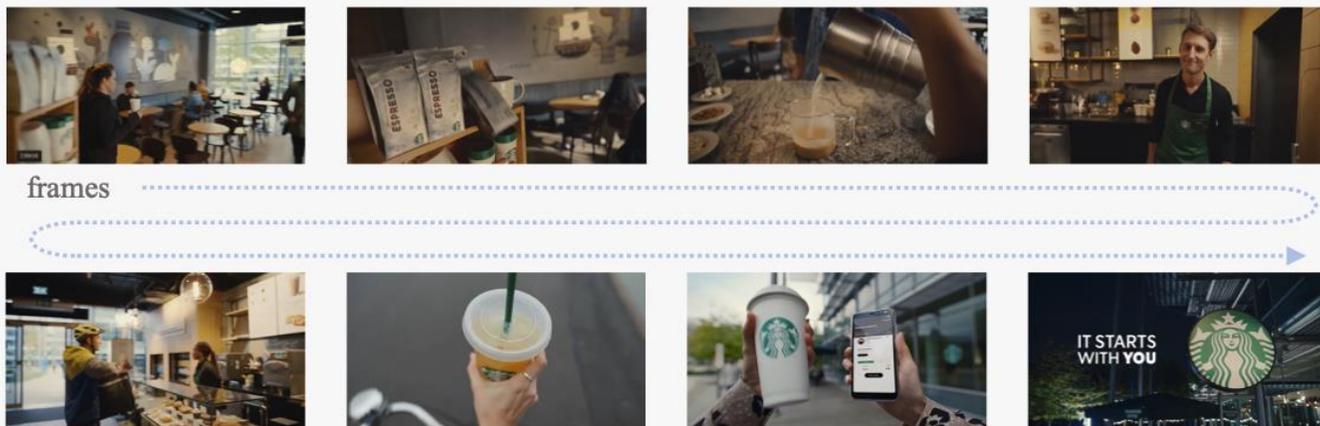
Answer: 29



演进-更丰富的输入形式

- 从只支持单图输入到支持多图、视频输入

Chat on Video



frames

What is the most likely use of this video?



The most likely use of this video is as an advertisement for Starbucks. The sequence of events depicted in the video, such as the preparation of a coffee drink, the interaction between customers and baristas, and the closing shot of the Starbucks logo, are characteristic of a promotional piece designed to showcase the brand's products and services. The emphasis on the coffee-making process, the use of the Starbucks app, and the friendly interactions between customers and staff all serve to create a positive image of the brand. Additionally, the final shot of the Starbucks logo with the text "IT STARTS WITH YOU" suggests a call to action for viewers to engage with the brand, further supporting the idea that this video is an advertisement aimed at attracting new customers and encouraging existing ones to continue patronizing Starbucks.

演进-更丰富的I/O模态支持

- 输出更多的模态？如图片、音频、视频…
- 输出图文交错的内容，给文字配上图，使内容更生动

请根据给定标题：“熊猫：自然界的温柔使者”，写一篇长文章。首先，详细介绍熊猫的基本特征，如外貌，习性等。然后，探讨熊猫的自然栖息地及其面临的环境挑战，包括栖息地丧失和气候变化等问题。之后，详细讨论熊猫的保护工作现状，包括建立自然保护区，大熊猫的人工繁育等。最后，分析熊猫在中国和世界范围内的文化影响。字数不少于800字。

熊猫：自然界的温柔使者

大熊猫，熊科哺乳动物，体肥头大、圆脸圆耳，幼年体色为灰白色，因其独立特征和憨态可掬的样子被人们誉为“活化石”和“中国国宝”。



从外表来看，大熊猫似乎是自然界最“温柔”的生物：黑白相间的花纹让它的外形看起来更像是一种卡通形象；圆滚滚的身躯和大大的头部让人忍不住想要抱一抱，即使不熟悉的人也会因为其憨态可掬的外表而心生好感……

但就是这样一种温顺的生物，却也正面临着种种危机——栖息地丧失、气候变化等因素都对其生存构成威胁。



此外，气候变暖也是大熊猫目前所面临的重要问题之一。据研究，自20世纪70年代以来，我国东部沿海地区出现了持续增温现象，这种趋势仍在继续发展。高温会导致竹林生长缓慢，从而降低了竹子作为大熊猫主要食物的供应量。除此之外，夏季降雨量增多还会造成泥石流、滑坡等灾害，进一步破坏大熊猫的栖息地。这些因素共同作用，使得大熊猫的种群数量日益减少。



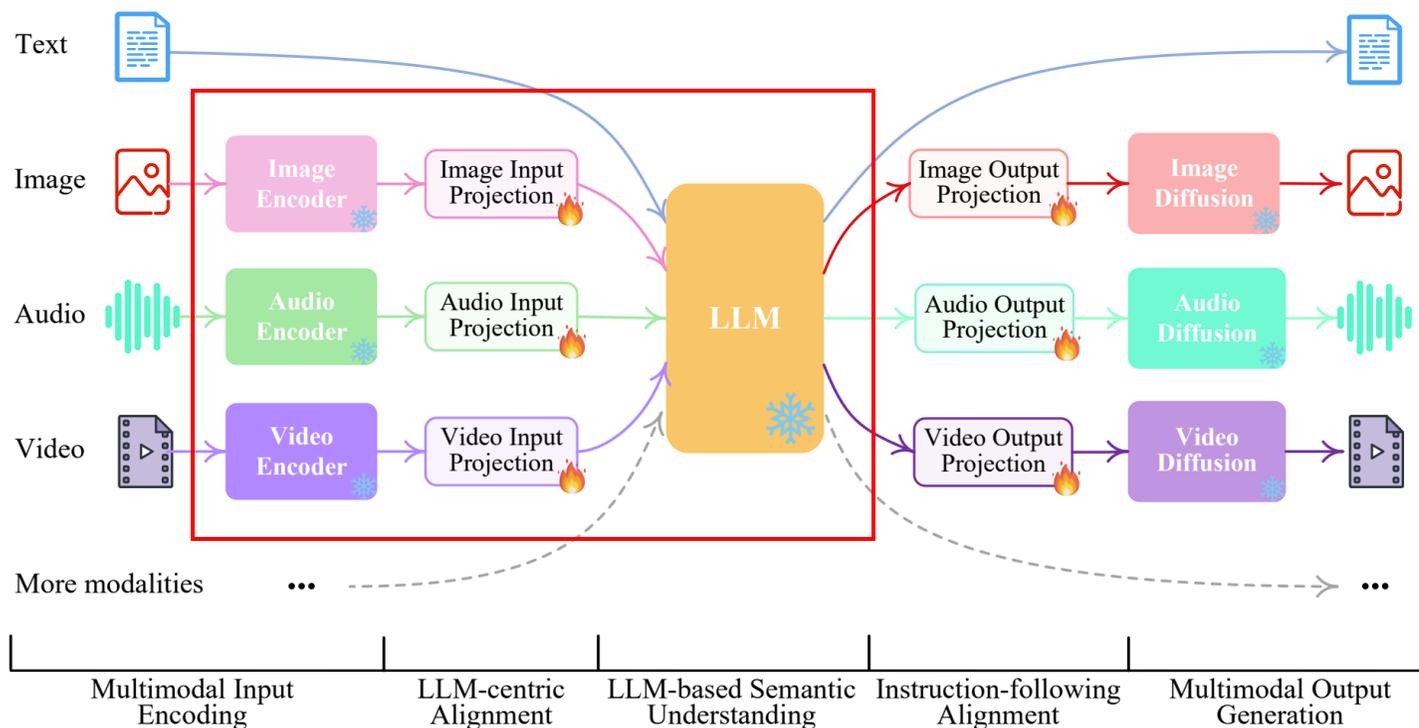
熊猫的保护现状如何？

针对上述问题，政府和企业相继采取了一系列措施保护大熊猫及其栖息地。例如，建立自然保护区和森林公园，限制人类活动范围，改善大熊猫的食物来源等。同时，还通过科学管理、合理开发利用等方式推动了当地经济发展。尽管如此，由于历史原因和自然环境的恶化等原因，野生大熊猫仍处于濒危状态。根据国际自然保护联盟最新发布的红色名录，目前全球现存的大熊猫约有1864只，其中约90%以上为中国境内的大熊猫。



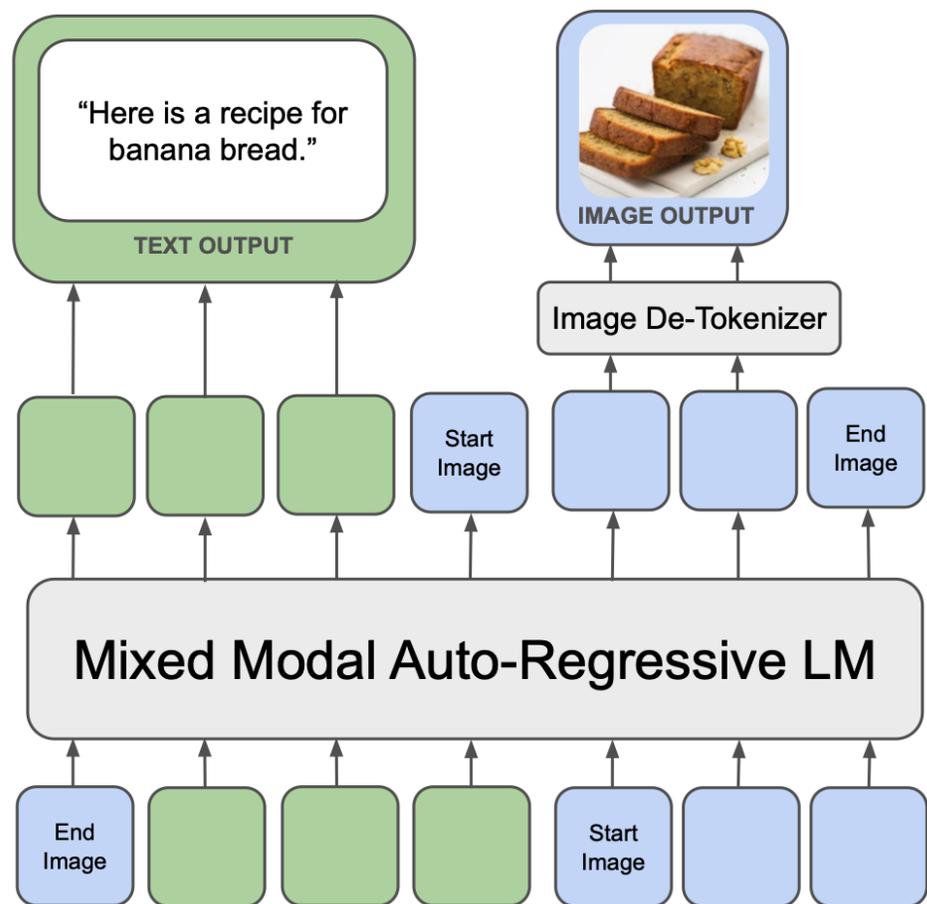
演进-更丰富的I/O模态支持

- 输出更多的模态？如图片、音频、视频…
 - 基于基础的MLLM框架，后接各种生成模型如扩散模型，生成除文本外其他模态的信息。



演进-更丰富的I/O模态支持

- 输出更多的模态?
 - 统一的多模态模型—扩充语言模型的词表，将连续信号离散化
 - 加入Tokenizer，训练时将图片转化为离散token进行预测
 - 推理时统一预测多模态token，图片token经过De-tokenizer转化为图片输出
 - 天然支持图文交错的训练数据





04

团队相关工作介绍

相关工作-幻觉缓解

- 背景：在开源模型探索的早期，幻觉问题还比较严重
 - 幻觉: 大模型在描述**对象属性**、**对象数量**等方面不准确的现象。样例如下图所示。
 - 减少模型输出的幻觉，提升准确性和用户体验: [Woodpecker](#)



请描述这张图片。



这张图中出现了一只**红色**的狗，它在草坪上奔跑。在这只狗的周围，还可以看见**其它几只狗**。有**几只离这只红色的狗更近**，**其它的距离更远一些**。

Yin, Shukang, et al. "Woodpecker: Hallucination correction for multimodal large language models." arXiv:2310.16045

相关工作-幻觉缓解

- 能否集成专家模型缓解幻觉?
 - 利用外部反馈获取更准确的认知，从而增强模型，减少幻觉
 - 基础的感知能力：object类别与数量、object的属性信息(颜色、位置等)



大模型生成的带有
幻觉的描述



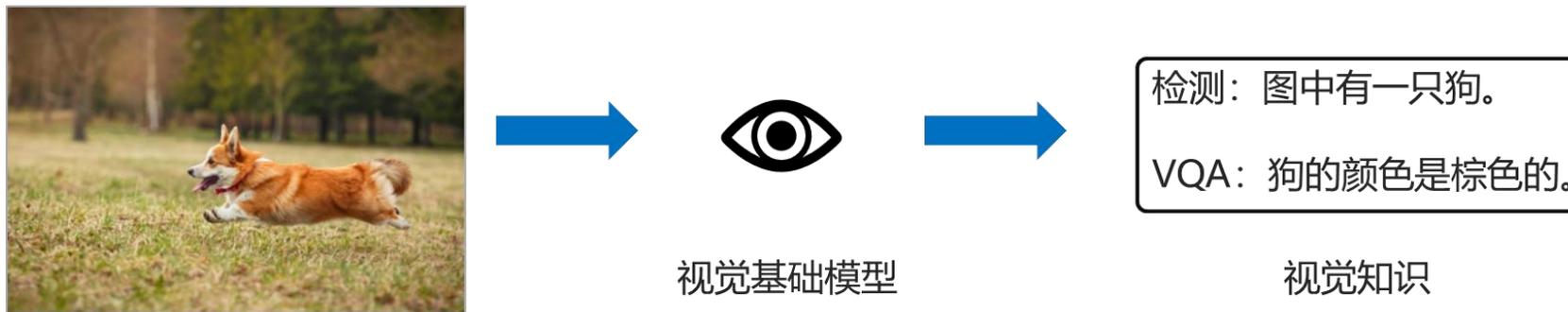
基于知识的
幻觉修正框架



可靠的描述

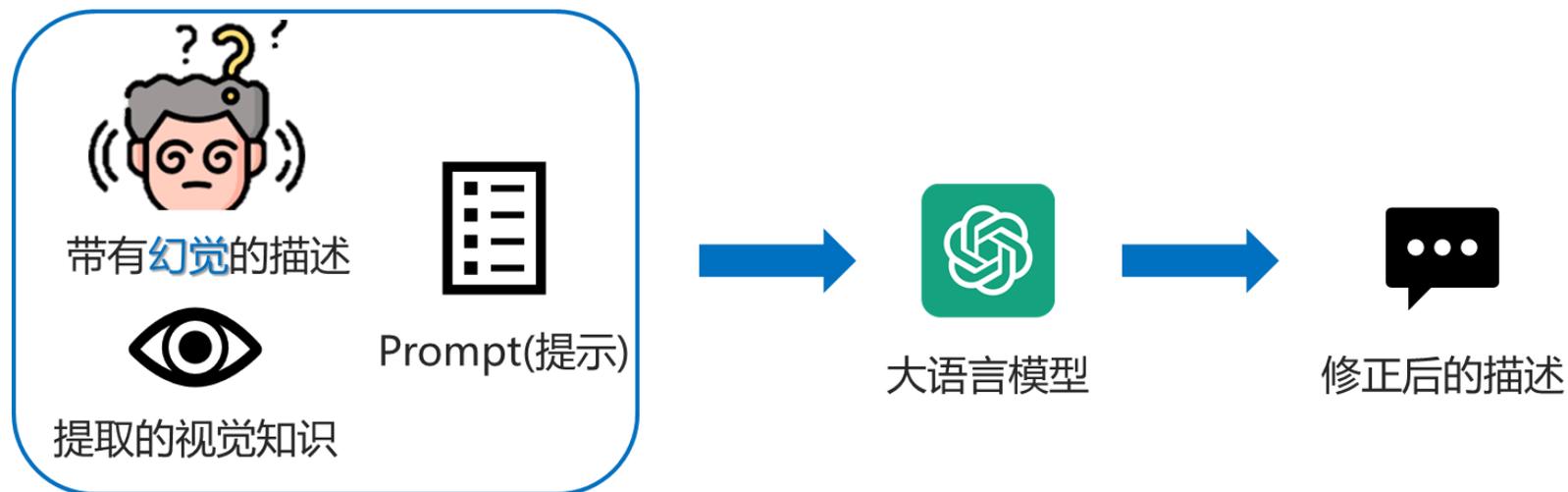
相关工作-幻觉缓解

- 实现:使用传统视觉模型提升感知能力
 - 视觉基础模型具有强大的感知能力。因此,我们利用预训练的视觉基础模型从图片中提取可靠的视觉信息。这些视觉信息为幻觉修正提供了参考。
 - 使用目标检测模型从图片中提取目标的数量信息,使用VQA(视觉问答)模型提取目标的属性信息。



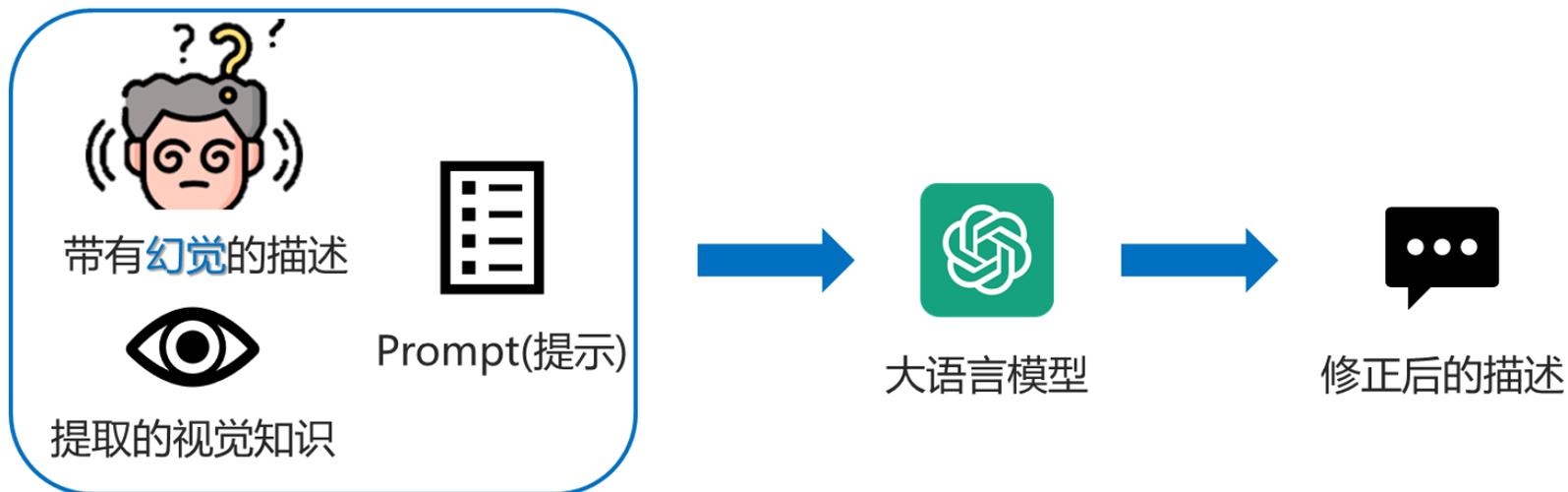
相关工作-幻觉缓解

- 实现:利用视觉信息辅助大语言模型做幻觉修正
 - **大语言模型**具有强大的**推理**能力与丰富的语言学知识。因此,大语言模型能够基于已知信息进行推理,对幻觉做修正,同时保持语义连贯性、完整性
 - 使用设计好的Prompt(提示)模板将原描述、视觉知识组织起来,送入大语言模型中。大语言模型基于视觉知识、文本语义进行推理,将描述中带有幻觉的部分进行修正。



相关工作-幻觉缓解

- 实现:利用视觉信息辅助大语言模型做幻觉修正
 - **大语言模型**具有强大的**推理**能力与丰富的语言学知识。因此,大语言模型能够基于已知信息进行推理,对幻觉做修正,同时保持语义连贯性、完整性
 - 使用设计好的Prompt(提示)模板将原描述、视觉知识组织起来,送入大语言模型中。大语言模型基于视觉知识、文本语义进行推理,将描述中带有幻觉的部分进行修正。



相关工作-幻觉缓解

- 实验结果:利用纠正框架可以显著减少幻觉
 - 受益于基础视觉模型的知识提取,几个典型的多模态大模型在感知能力上均有提升
 - 为了构建更强大的系统,使用外置专家模块作为信息补充仍然是有效的手段

Method	w/Ours	Object-level		Attribute-level		Total
		Existence	Count	Position	Color	
LLaVA [21]	✗	195.00	95.00	53.33	78.33	421.67
	✓	195.00	160.00	55.00	155.00	565.00
MiniGPT-4 [47]	✗	100.00	61.67	53.33	65.00	280.00
	✓	183.33	163.33	60.00	121.67	528.33
mPLUG-Owl [40]	✗	101.67	73.33	58.33	66.67	300.00
	✓	200.00	131.67	78.33	145.00	555.00
Otter [15]	✗	185.00	95.00	50.00	118.33	448.33
	✓	195.00	160.00	51.67	165.00	571.67

上表为开源多模态大模型在MME的存在、计数、位置和颜色子集上的测评结果, w/Ours表示是否加入我们提出的修正框架



这张图中出现了一只棕色的狗,它在草坪上奔跑。

这张图中出现了一只红色的狗,它在草坪上奔跑。在这只狗的周围,还可以看见其它几只狗。有几只离这只红色的狗更近,其它的距离更远一些。

案例:大语言模型基于提取的视觉信息进行推理,修正了其中的属性方面(颜色)与数量方面的幻觉。



相关工作-长视频理解测评

- 背景:现存测试集无法充分评估多模态大模型在理解长视频全局性内容的能力
 - 【长视频理解】绝大多数现存评测基准平均时长不超过1分钟
 - 【通用模态理解】现存评测所提供的模态信息（如文本、音频）均不丰富
 - 【全局信息利用】现存评测只利用局部信息即可得到答案，难以评估上下文理解、整合能力

Benchmarks	#Videos	#Clips	Len.(s)	#QA Pairs	Anno.	QA Tokens	Sub. Tokens	Multi-level	Open-domain	Sub.&Aud.
MSRVTT-QA [60]	2,990	2,990	15.2	72,821	A	8.4	X	X	✓	X
MSVD-QA [60]	504	504	9.8	13,157	A	7.6	X	X	✓	X
TGIF-QA [18]	9,575	9,575	3.0	8,506	A&M	20.5	X	X	✓	X
ActivityNet-QA [65]	800	800	111.4	8,000	M	10.2	X	X	✓	X
TVQA [21]	2,179	15,253	11.2	15,253	M	27.8	159.8	X	X	X
How2QA [25]	1,166	2,852	15.3	2,852	M	16.9	31.1	X	✓	X
STAR [58]	914	7,098	11.9	7,098	A	19.5	X	X	✓	X
NExT-QA [59]	1,000	1,000	39.5	8,564	A	25.3	X	X	✓	X
MVBench [24]	3,641	3,641	16.0	4,000	A	27.3	X	X	✓	X
Video-Bench [45]	5,917	5,917	56.0	17,036	A&M	21.3	X	X	✓	X
EgoSchema [43]	5,063	5,063	180.0	5,063	A&M	126.8	X	X	X	X
AutoEval-Video [8]	327	327	14.6	327	M	11.9	X	X	✓	X
TempCompass [37]	410	500	11.4	7,540	A&M	49.2	X	X	✓	X

平均时长较短

多模态信息不够丰富

相关工作-长视频理解测评

- 人工采集三种长度的视频各300条，每条人工标注3个问答对，提供字幕以及音频
- 问题设计均考虑视觉相关性，并且具有足够高的有效时长（38%）
- 需要更充分地利用前后视频信息，对视频建立更深入的全局理解



Video-MME

视频中，那个戴着绷带、拿着信封的男人是如何受伤的？

- A. 他在放烟花时，一只手被烟花击中了。
- B. 他在试图扑灭一栋着火的房子时，手臂受伤了。
- C. 他在追赶Wayne的摩托车时，从地上摔下来，手受伤了。
- D. 他在侮辱Wayne的父亲时，被Wayne用食物引诱的狗拖下了一只胳膊。

被狗拖下
[选项D]



戴着绷带的男人
拿着一个信封



追赶Wayne的摩托车
[选项C]



一栋着火的房子
[选项B]



被烟花击中
[选项A]

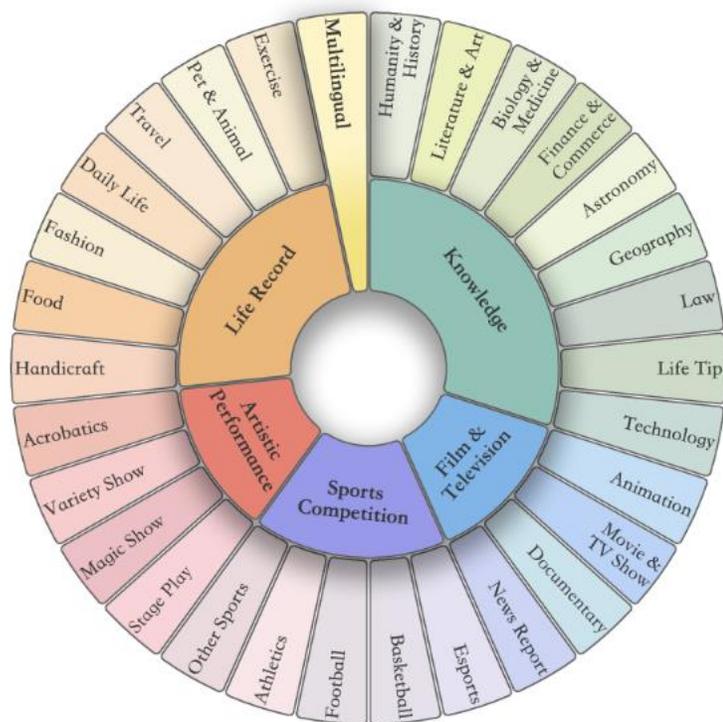


选项在视频中均出现

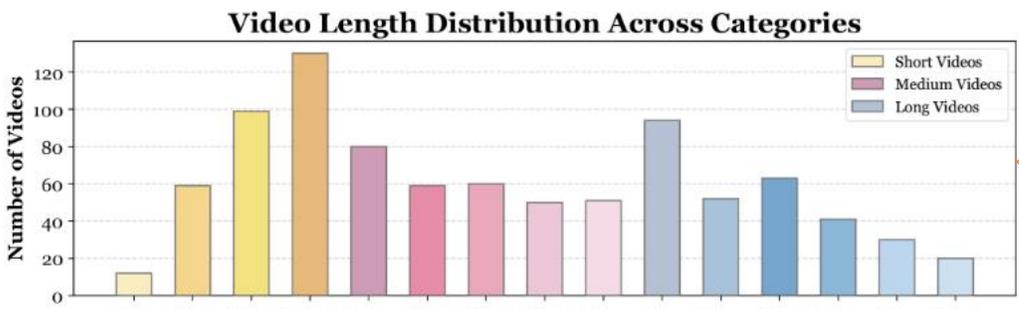
模型做出正确选择需对问题和视频有全局理解

相关工作-长视频理解测评

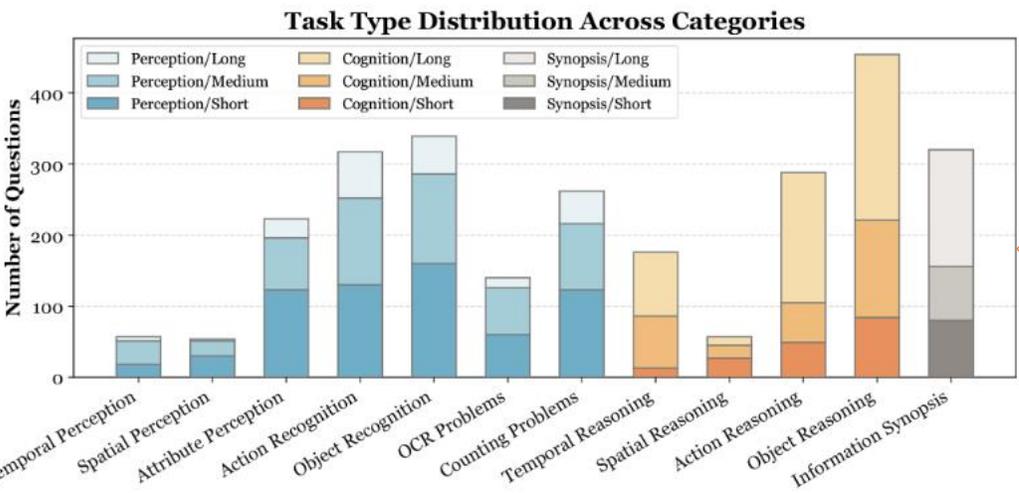
- 视频类别、视频时长、问题类型的多样性



视频涵盖各个领域



视频时长覆盖广
中、长视频分布均匀



问题类别多
短、中视频注重感知
长视频注重推理

相关工作-长视频理解测评

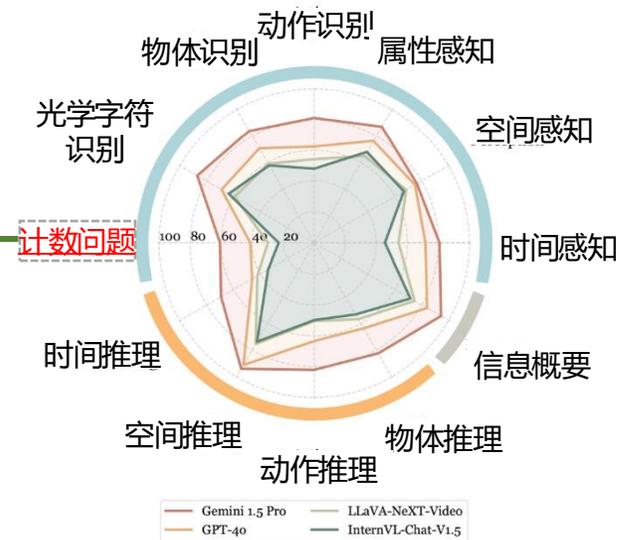
- 对现存的开源与商业大模型进行了综合评测

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs						
<i>Open & Closed-source Image MLLMs</i>									
Qwen-VL-Chat [5]	7B	46.4	47.1	38.1	39.8	38.0	38.3	40.9	41.7
Qwen-VL-Max [5]	-	56.5	58.3	49.9	49.8	49.0	46.9	51.8	51.7
InternVL-Chat-V1.5 [9]	20B	61.2	62.4	47.3	50.0	46.0	47.0	51.5	53.2
<i>Open-source Video MLLMs</i>									
Video-LLaVA [28]	7B	45.9	47.1	38.1	40.2	37.3	39.6	40.4	42.3
VideoChat2 [24]	7B	38.2	41.6	33.2	34.3	29.7	31.9	33.7	35.9
ST-LLM [33]	7B	47.0	49.9	36.9	42.2	31.8	37.3	38.6	43.2
Chat-UniVi-V1.5 [19]	7B	46.3	51.4	40.3	45.2	36.9	42.3	41.2	46.3
LLaVA-NeXT-Video [68]	34B	63.1	66.4	51.1	53.2	44.6	48.7	52.5	56.0
<i>Closed-source MLLMs</i>									
GPT-4V [45]	-	71.4	74.5	56.5	59.3	54.2	57.2	60.7	63.7
GPT-4o [46]	-	77.1	77.5	62.1	63.0	59.2	56.7	66.2	65.8
Gemini 1.5 Pro [51]	-	82.3	84.7	75.3	82.6	67.5	76.3	75.7	81.6

字幕对多模态大模型的视频理解具有正向作用

开源模型最优结果 < 闭源模型最差结果
开源模型仍具有较大提升空间

模型在12种任务类型上的表现雷达图



上表中为多模态大模型在Video-MME中短、中、长三种视频上的表现
其中w/ subs代表使用字幕信息, w/o subs代表不使用字幕信息

- 图片大模型:** QwenVL系列、InternVL-Chat-V1.5
- 视频大模型:** Video-LLaVA、Video-Chat2等, GPT4V/4o以及Gemini 1.5Pro

实体感知问题仍是现存多模态大模型的共同瓶颈

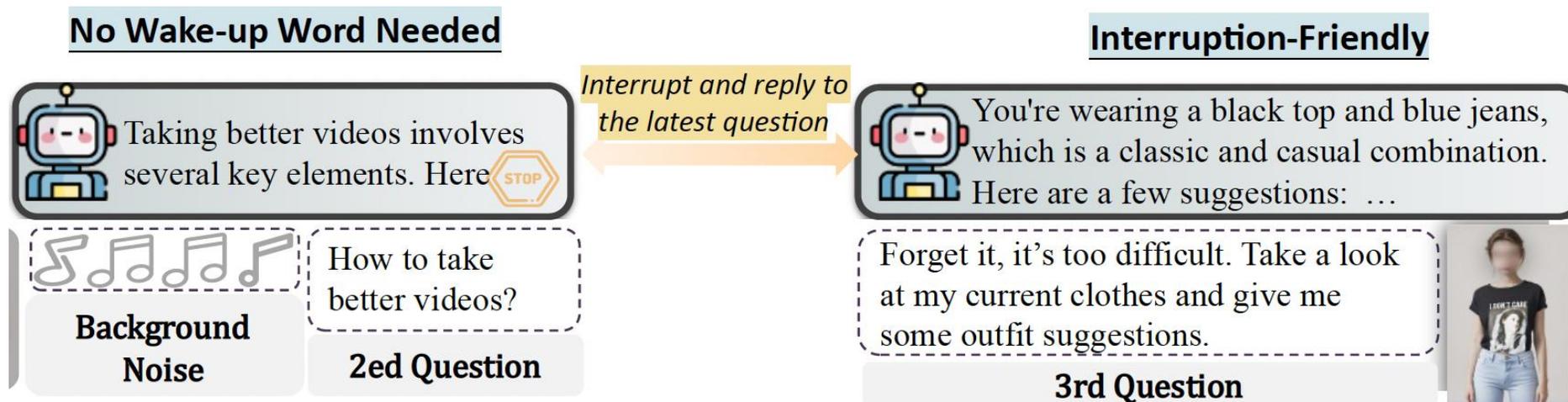
相关工作-多模态交互体验提升

- 背景:GPT-4o的发布展现了多模态实时对话交互体验的新可能
 - 可以随时打断模型的输出,而不需要等待输出完再进行下一轮交互/提问
 - 实时响应速度,模型及时回复,而不需要等待几秒钟的延迟
 - 模型能够捕捉人类情绪(如说话人是失落/兴奋的),也能够以不同的情绪(比如激昂、戏剧化的)、风格(模仿机器人、歌唱式的)回答



相关工作-多模态交互体验提升

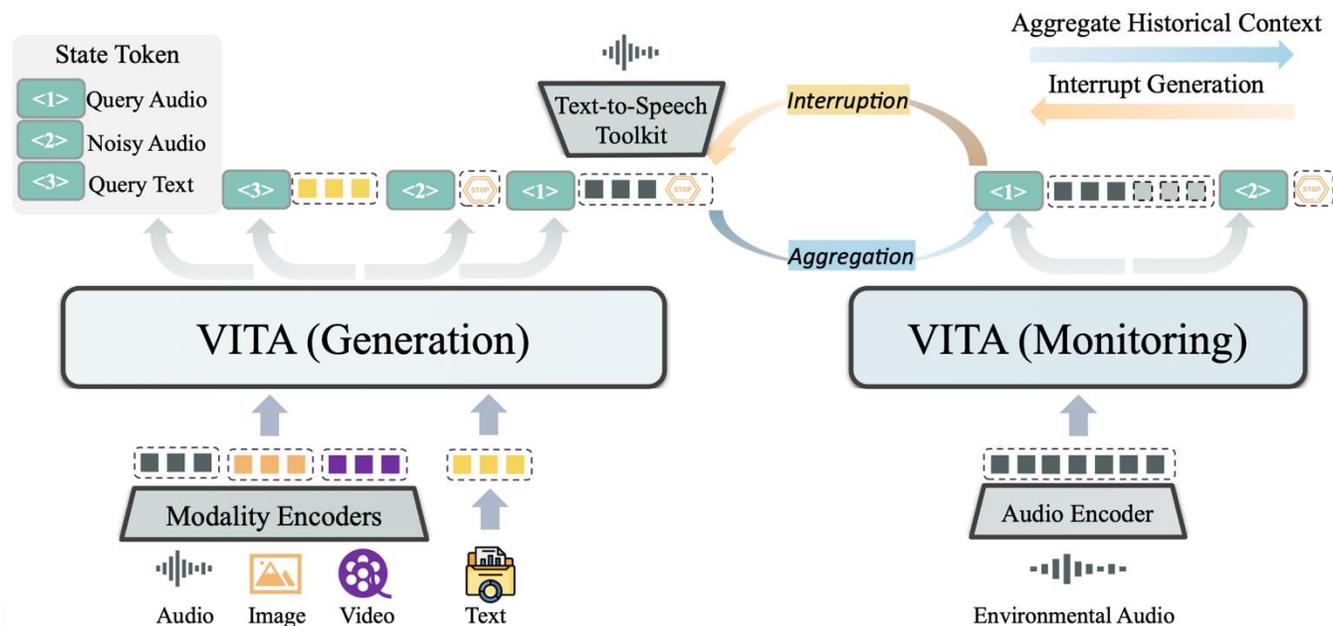
- 背景:探索提升对话交互体验的开源方案: [VITA](#)
 - 不需要专门的唤醒机制来激活语音助手, 如唤醒词 “Hi! Siri”
 - 支持打断语音输出, 插入新用户请求, 而不需要等候输出完成再输入新的用户请求
 - 屏蔽环境噪声, 提升模型对话感知的准确性



相关工作-多模态交互体验提升

• 实现: 两大核心设计

- 屏蔽背景声中的噪音: LLM词表中加入特殊的状态token, 让模型判别输入的音频是有效query还是背景噪音, 实时监听的同时又不会受噪音干扰。
- 打断输出后快速响应新query: 双工机制。同时运行两个模型—生成和监听模型, 分别用于输出回答和音频监听。当生成模型被新音频query打断时切换另一个模型响应新请求, 两个模型身份互换。





05

未来展望



未来展望

- 未来需要什么？
- 支持更长的context/有效的压缩视觉token
 - 输入的视觉token越来越多，多图、高分辨率、长视频理解...
 - 长上下文理解能力，如电影级别的视频理解
- 多模态Agent
 - 如智能手机助手，自动完成某项操作，如将手机亮度调整至夜间模式
 - 轻量化的部署
- 统一多模态生成和理解
 - 统一的训练范式，统一的推理流程
 - 直接生成任意图文交错的内容



THANKS

DataFunSummit # 2024

