

多模态大模型的实践与思考

朱优松 副研究员

中国科学院自动化研究所大模型研究中心

yousong.zhu@nlpr.ia.ac.cn

2024.08



大模型开启人工智能新一代技术革命

■ 大模型技术的突破掀起了人工智能的新一轮技术革命，产业需求呈现爆炸式增长，全新的应用场景和应用模式正在迅速涌现。

大模型+人形机器人

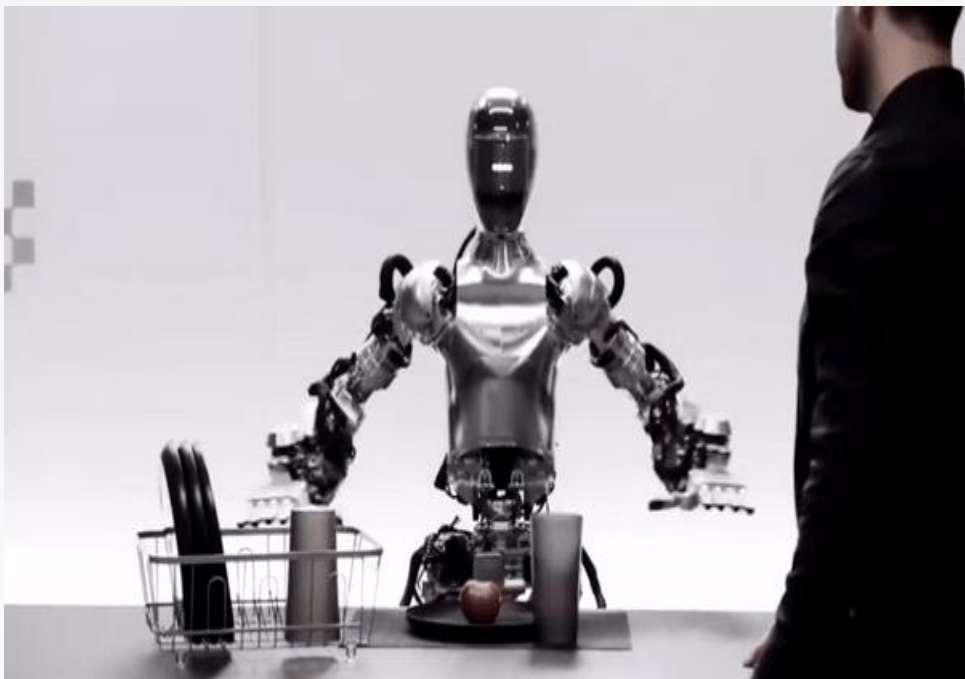
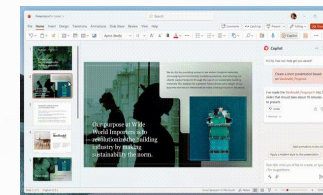


Figure 01**机器人**：通过GPT4大模型加持，机器人可以听懂人类指令和提问，能对话、能思考、能学习

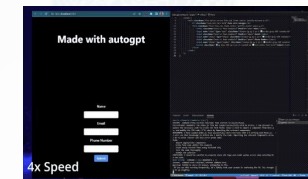
大模型+自动驾驶



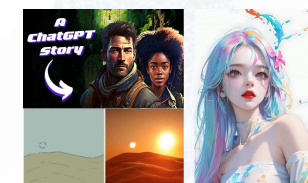
特斯拉 FSD V12：感知-规划-决策端对端，不需要为减速带、停车标志或环形路口编写程序



重塑办公模式



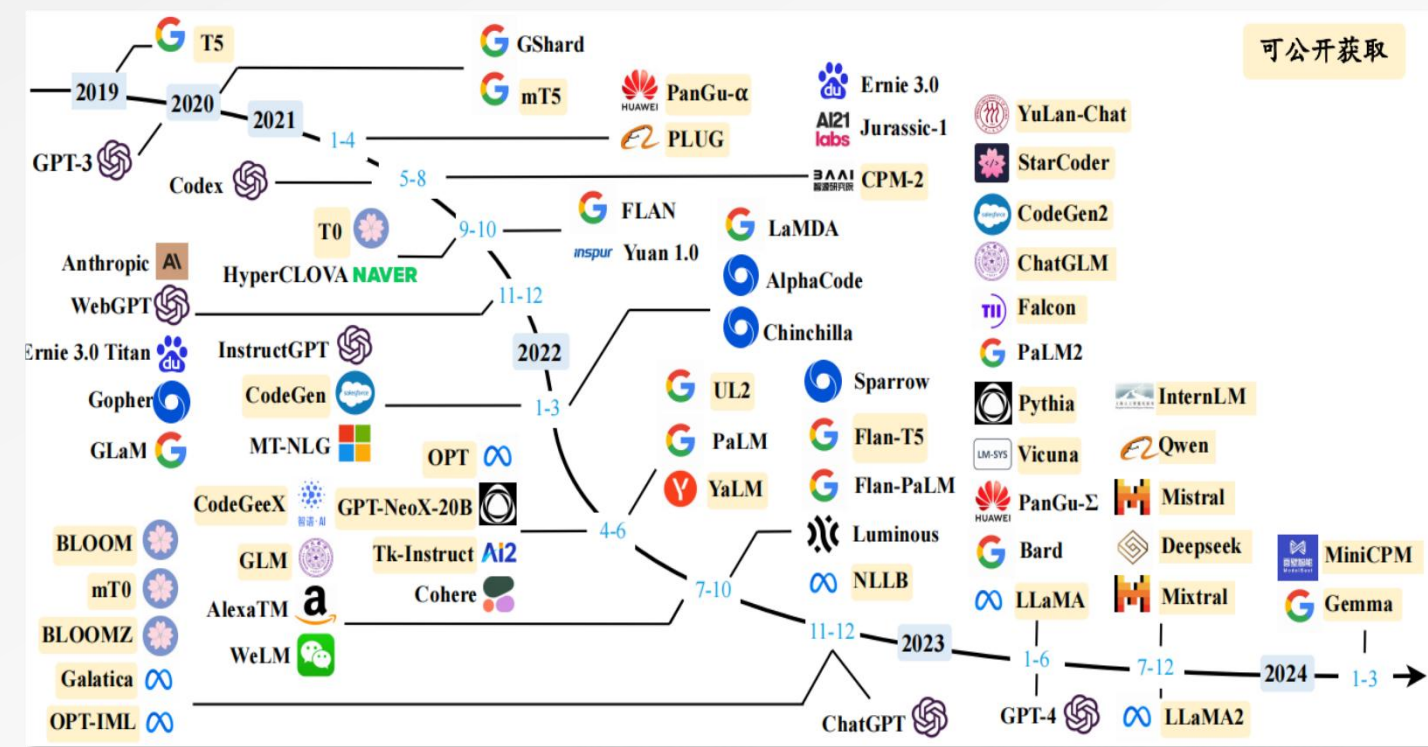
程序自动设计



AIGC

...

大语言模型快速发展，百花齐放

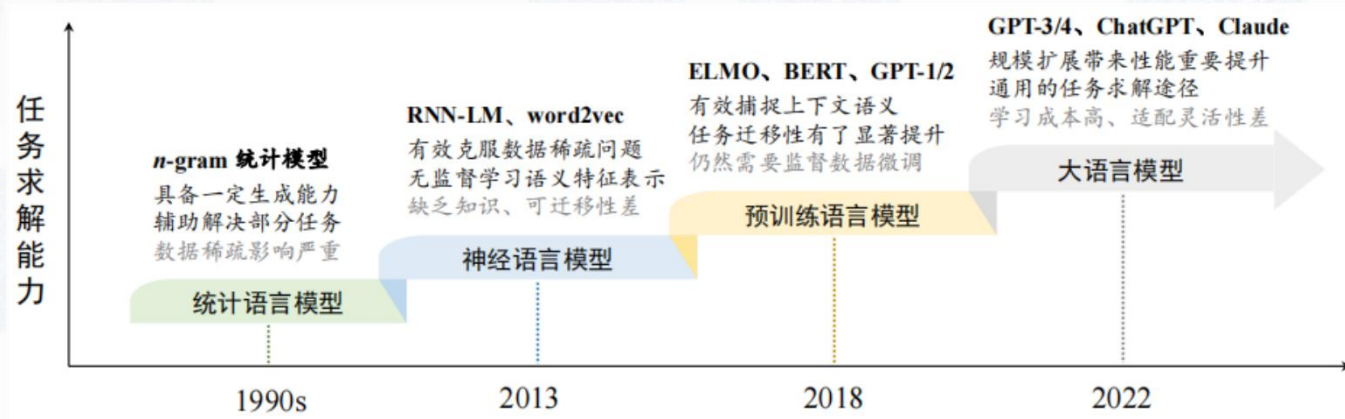


大语言模型典型特征：

- **参数量大**：从几十亿参数到数千亿参数，例如GPT-3模型的参数量高达1750亿(与人脑神经元数目一个量级)；
- **训练数据大**：数万亿Tokens (all data in internet) ；
- **训练算力大**：数千甚至数万个GPU，耗费几个月才能完成一次训练；

语言模型任务求解能力的演化过程

通用性是大语言模型的显著特征，在广泛的下游任务（语言理解、语言生成）中达到甚至超过人类的水平





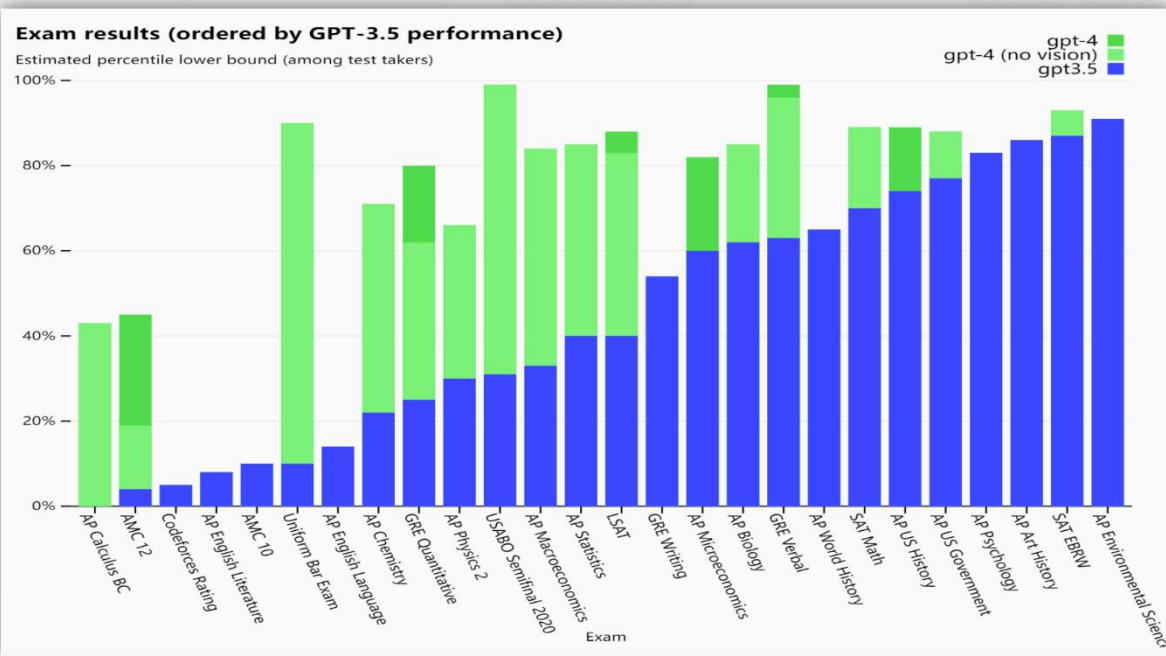
大语言模型给通用人工智能带来了无限可能

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
 Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
 Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

ChatGPT系列被称为是“通用人工智能的火花”



GPT4在一系列考试中表现优异

Prompt:

What is the integral of $\ln(\cos(x)) - x \tan(x)$?

GPT-4:

We can split the integral of $\ln(\cos(x)) - x \tan(x)$ into two parts:

$$\int \ln(\cos(x)) - x \tan(x) dx = \int \ln(\cos(x)) dx - \int x \tan(x) dx$$

To find the first part, we can use integration by parts, which states that:

$$\int u dv = uv - \int v du$$

Let $u = \ln(\cos(x))$ and $dv = dx$. Then $du = -\frac{\sin(x)}{\cos(x)} dx = -\tan(x) dx$ and $v = x$. Substituting these into the formula, we get:

$$\int \ln(\cos(x)) dx = x \ln(\cos(x)) + \int x \tan(x) dx$$

So the first part is equal to the second part plus $x \ln(\cos(x))$. This means we can simplify the original integral as:

$$\int \ln(\cos(x)) - x \tan(x) dx = x \ln(\cos(x)) + \int x \tan(x) dx - \int x \tan(x) dx = x \ln(\cos(x)) + C$$

where C is an arbitrary constant of integration. Therefore, the answer is:

$$\int \ln(\cos(x)) - x \tan(x) dx = x \ln(\cos(x)) + C$$

GPT4解高等数学题

Bubeck S, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv:2303.12712, 2023.

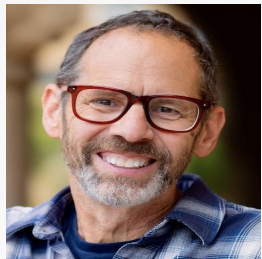


Foundation Models (基础模型)

2021年8月，斯坦福大学众多学者，包括Feifei Li, Dan Jurafsky, Christopher D. Manning (三位美国科学院院士)，联合撰写文章“On the Opportunities and Risks of Foundation Models”，将预训练大模型称之为基础模型 (Foundation models)，并指出基础模型是未来人工智能发展的重要方向。



Feifei Li



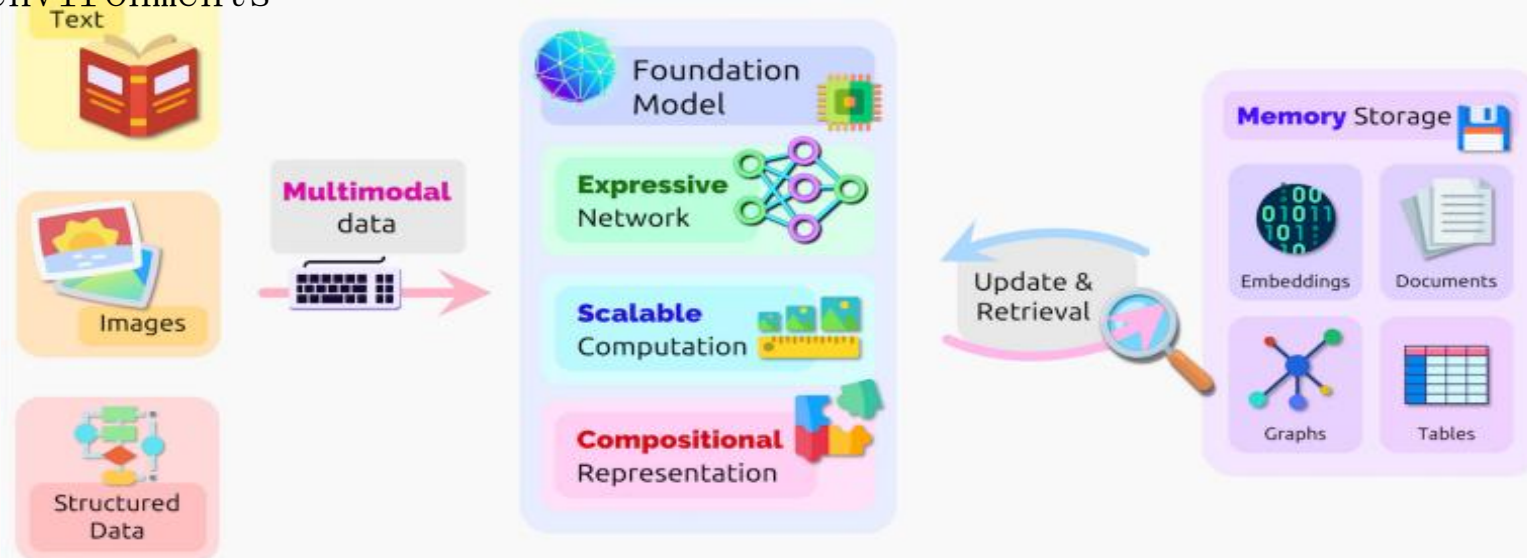
Dan Jurafsky



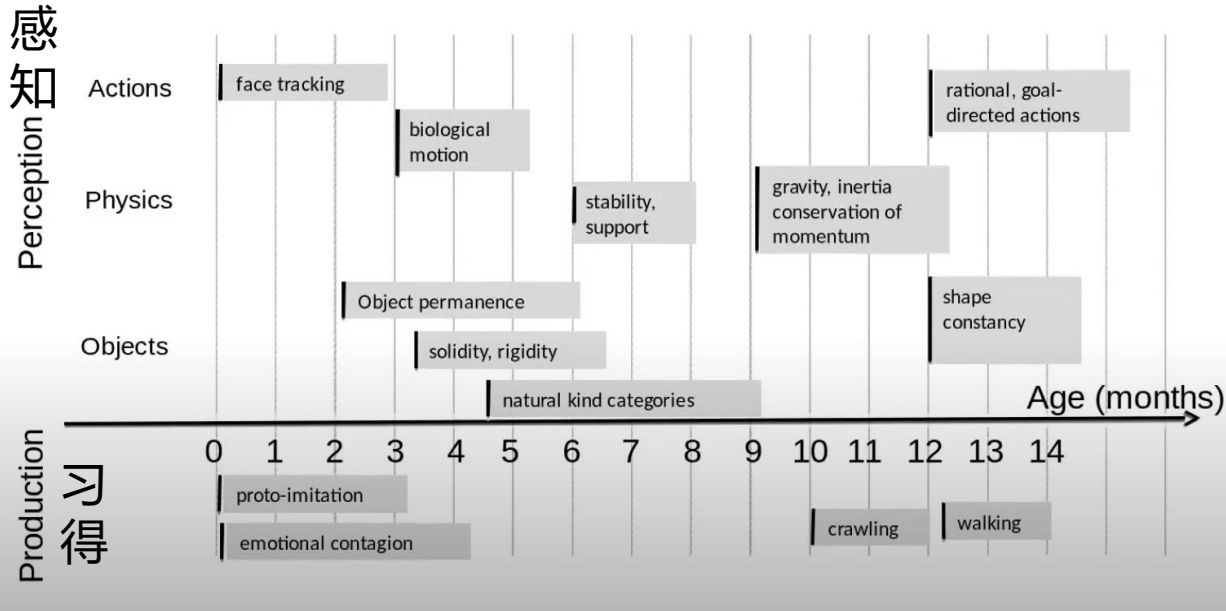
Christopher D. Manning

The five key properties of a foundation model:

- Expressivity: to flexibly capture and represent rich information
- Scalability: to efficiently consume large quantities of data
- **Multimodality: to connect together various modalities and domains**
- Memory capacity: to store the vast amount of accumulated knowledge
- Compositionality: to generalize to new contexts, tasks and environments



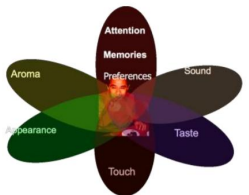
婴儿阶段的早期概念习得过程 (from Emmanuel Dupoux)



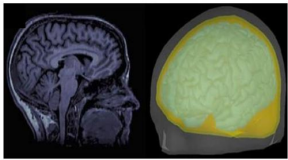
各类模态的示例:

- 自然语言 (说/写)
- 视觉 (图像/视频)
- 听觉 (声音/语音/音乐等)
- 触觉
- 嗅觉/味觉
- 生理信号 (心电图ECG, 皮肤电传导等)
- 其他 (红外图像, 深度图, 核磁共振图像)

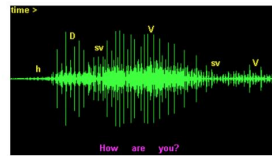
应用和行为方面:



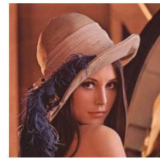
Psychology
心理学



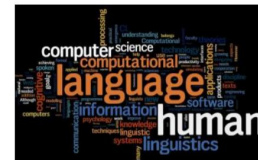
Medical
医学



Speech
声学



Vision
视觉



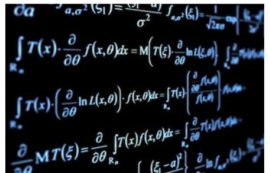
Language
语言



Multimedia
多媒体



Robotics
机器人



Learning
学习

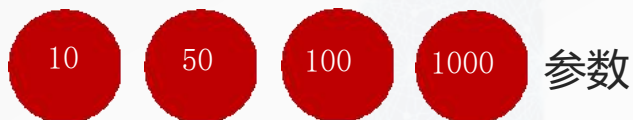


紫东太初大模型攻关历程



不同场景对应不同小模型

一个场景一个模型
模型参数量小，泛化性差



▶从0开始 ▶独立调优 ▶艰难迭代 ▶推倒重来

多模态大模型吸收海量知识

多个场景一个模型
模型参数大，泛化性强



预训练 ▶ 微调

算力中心
(北京-武汉)



海量弱关联数据



图像

在照片的前半部分，我们可以看到一个干草区，那里有一个女人站着，穿着浅蓝色牛仔裤和灰白色长袖衬衫。

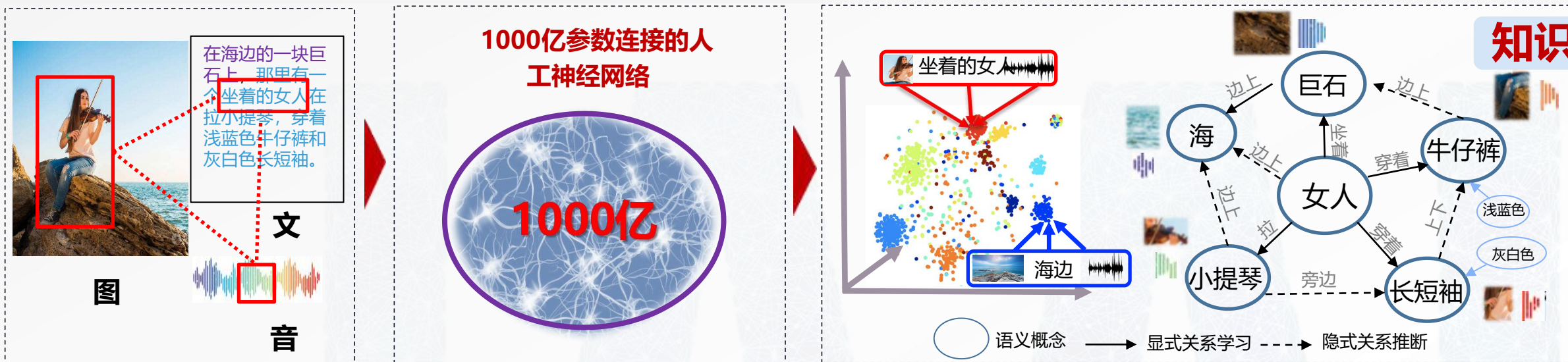


语音



紫东太初1.0：首个千亿参数图文音大模型

紫东太初突破当前AI技术局限，通过跨模态关联自监督学习，开拓性地实现了图像、文字、语音不同模态数据间的统一表示和互相生成，从“一专一能”迈向“多专多能”，形成了完整的智能表示、推理和生成能力。



9亿弱关联图音文数据

多模态关联自监督学习

自动获得100万级实体概念及关联关系



国产化智能计算中心
200P算力

模型
工具链

大模型微调
一键式微调

大模型小型化
剪枝、量化、蒸馏

推理开发套件
加速库、预处理

数据传输
数据安全及加密

模型
算子库

图像定制
算子集

视频定制
算子集

文本定制
算子集

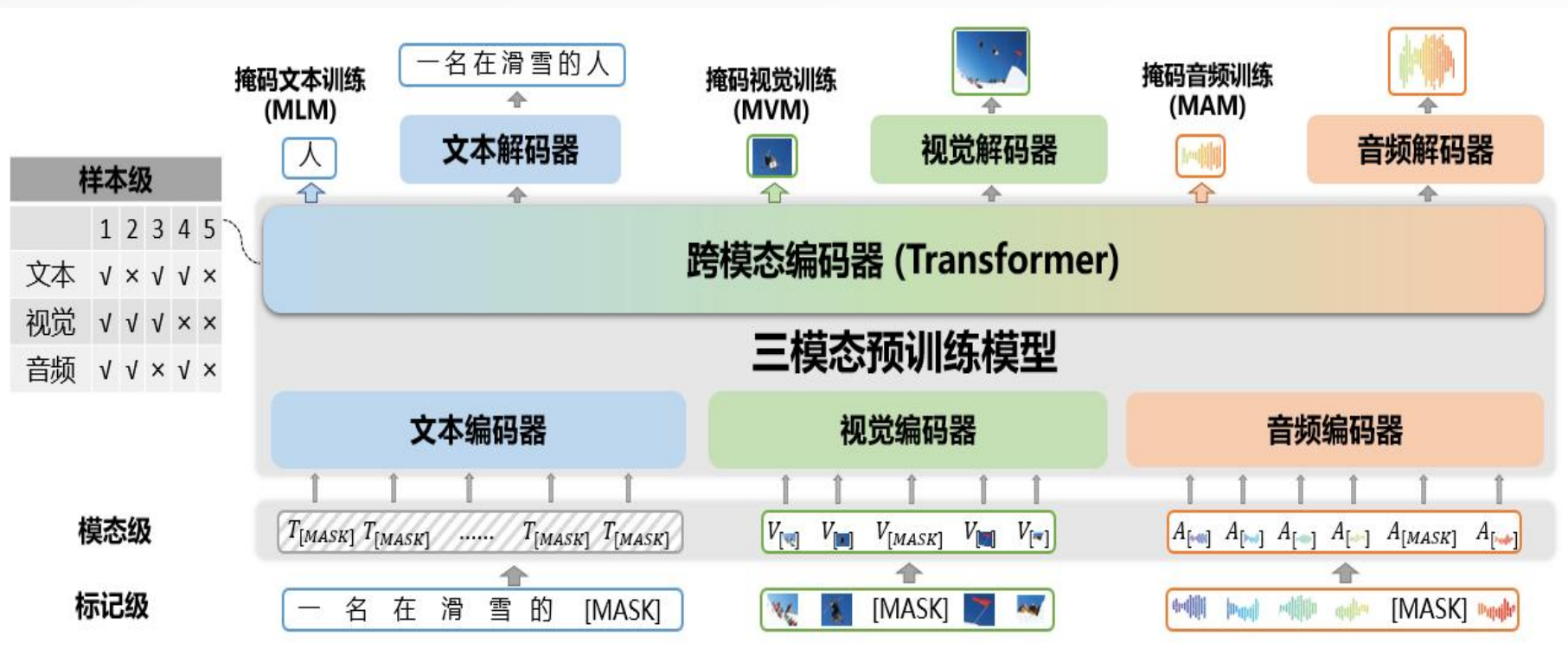
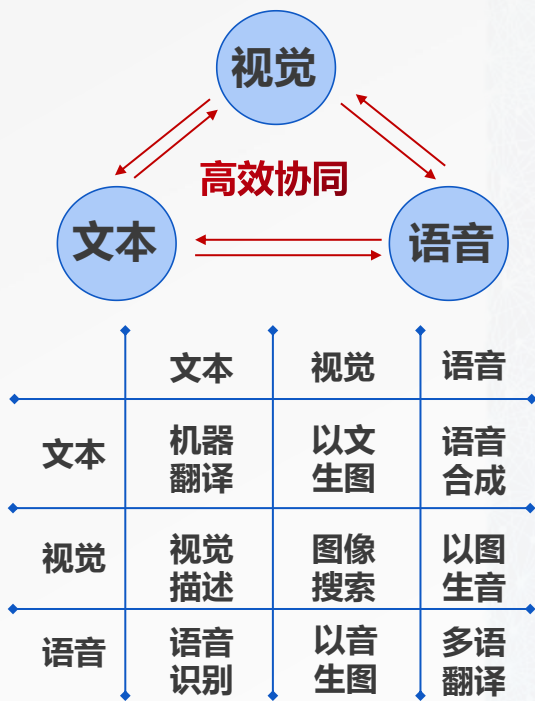
语音定制
算子集

多模态定
制算子集

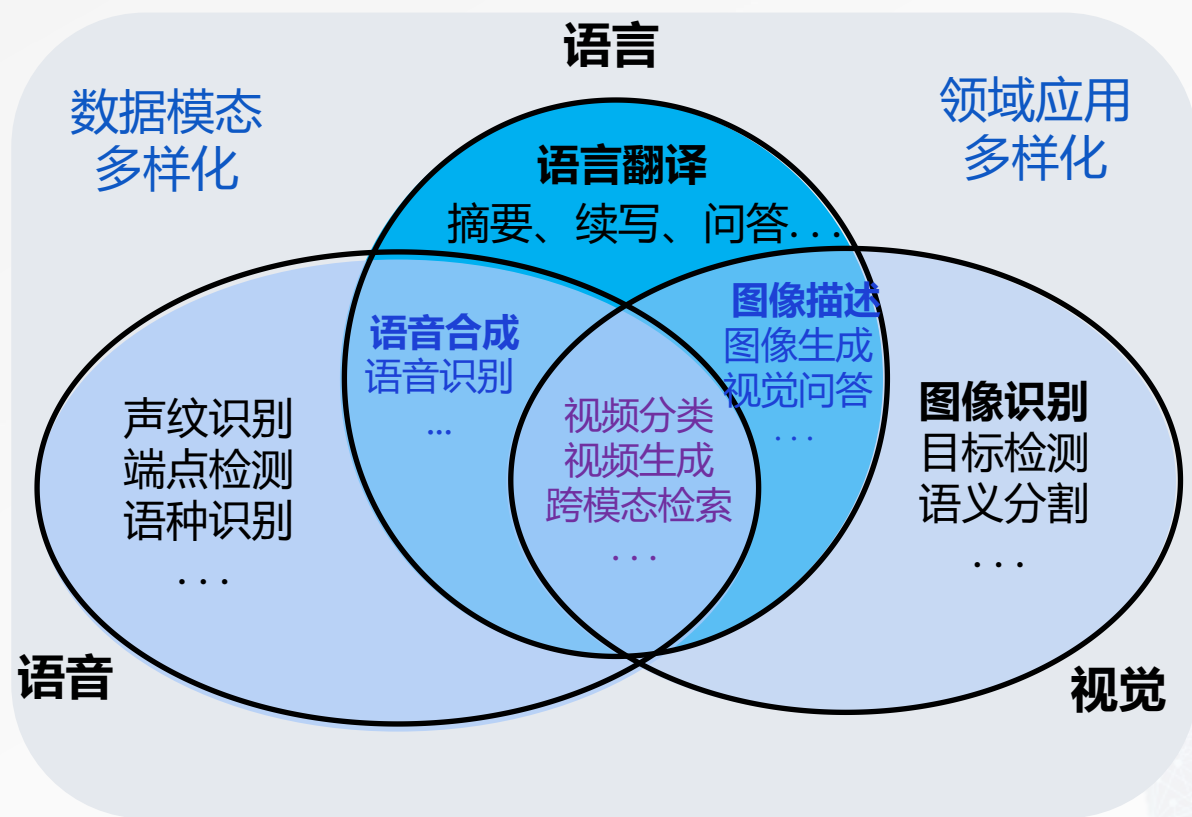


紫东太初1.0：首个千亿参数图文音大模型

- 多模态多任务自监督学习：**Token级别、模态级别与样本级别**的多任务自监督学习理论
- 弱关联多模态数据语义统一表示：支持**三种或任两种模态弱关联**数据进行训练
- **跨模态理解与生成统一建模**：跨模态检索、多模态分类、语音识别、文本生成、图像生成、音频合成等30+多领域的理解与生成任务



紫东太初1.0可以一个模型实现图像识别、语言翻译、图像描述、语音合成等50余种功能，并初步具备了人机交互、认知推理、艺术创作等创新能力。例如我们可以以音生图、可以理解图像含义。



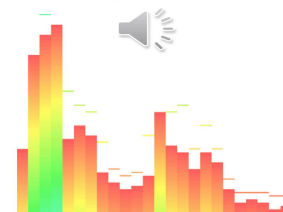
一个模型实现50余种功能

输入：桌子上的蓝色小碗里盛着水果和坚果



以文搜图

输入语音：



以音生图

输入视频：



以图生音



紫东太初2.0认知强化：全模态多任务统一生成

- 提出全模态多任务统一生成式学习框架：全模态分组对齐、关联融合和认知解码的统一框架
- 全模态低成本协同关联优化学习，融合多任务全模态能力，更具拓展性；全模态认知增强，面向物联网世界多源输入的认知，具有全模态能力涌现，认知效率更高



语言能力

➤ 主流LLM自动化评测：评测百度、科大讯飞、阿里、百川等多家的主流大语言模型的7大基础能力、35个子能力¹，紫东太初2.0在七大能力方向均获得具备竞争力的分数，在语言理解、逻辑推理方面表现尤为突出

图像能力

➤ 通用视觉分割模型FastSAM：基于COCO数据集评测AR1000指标和AUC指标²，FastSAM方法性能接近SAM方法，但参数量仅为其1/10，推理时间仅为其1/50

➤ 多模态对话模型：使用GPT4/人工进行准确性、丰富度、流利度分数指标评估，紫东太初获分847.5，高于visual-glm-6b、mplug-owl等模型

语音能力

语音能力	评价指标 ³	自研模型	SOTA ⁴
语音识别 ASR	WER ⁴ 越低越好	流式 1.71%	3.73%
		离线 1.59%	1.95%
语音鉴伪 ASV	EER 越低越好	0.63%	0.80%
语音活性检 VAD	accuracy	92.6%	—
	Recall	99%	—
	F1	95.7%	—

注：1) 七大基础能力包括文本生成、语言理解、知识问答、逻辑推理、数学、编程、角色扮演；2) AR1000指标代表在对每个图片产生1000个包围框时COCO数据集的目标召回率，AUC指标代表平均召回率，该实验用于测试方法在目标检出率上的性能；3) WER为认识词错率，EER为等错误率；4) SOTA指在该项研究任务中，目前最好/最先进的模型，其中ASR SOTA为阿里paraformer模型，在ASR leaderboard综合排名第一；ASV SOTA为End-to-End Dual-Branch Network Towards Synthetic Speech Detection 2023.3.

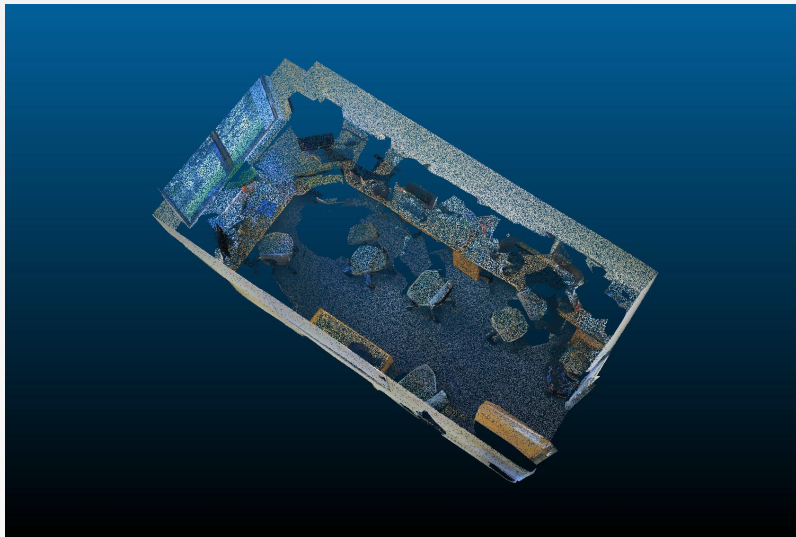


语言综合能力领先

AlignBench(CritiqueLLM 66B)

模型	总分	中文推理			中文语言						
		推理总分	数学计算	逻辑推理	语言总分	基本任务	中文理解	综合问答	文本写作	角色扮演	专业能力
gpt-4-1106-preview	7.58	7.11	7.39	6.83	8.05	7.69	7.07	8.66	8.23	8.08	8.55
gpt-4-0613	6.83	6.41	6.49	6.33	7.26	7.16	6.76	7.26	7.31	7.48	7.56
chatglm-turbo (智谱清言)	6.36	4.99	4.88	5.09	7.73	7.5	7.03	8.45	8.05	7.67	7.7
Taichu	6.20	5.04	5.10	4.98	7.37	6.40	6.69	8.42	7.84	7.69	7.18
InternLM2-Chat-20B	6.00	4.74	4.74	4.74	7.26	6.47	6.64	8.03	7.59	7.54	7.27
erniebot-3.5 (文心一言)	5.91	4.75	4.34	5.15	7.07	6.46	7.21	7.29	7.73	7.03	6.72
chatglm-pro (智谱清言)	5.73	4.49	4.55	4.43	6.96	6.47	6.81	7.26	7.25	7.29	6.7
llama_70b_sft_cp1186	5.45	4.20	3.88	4.51	6.70	5.66	5.00	8.16	7.08	7.32	6.95
gpt-3.5-turbo-0613	5.68	4.85	4.9	4.79	6.52	6.01	5.6	6.97	7.27	6.98	6.29
spark_desk_v2 (讯飞星火)	5.51	4.58	4.53	4.62	6.44	5.76	6.29	6.37	7.25	7.03	5.96
Qwen-14B-Chat	5.41	4.52	4.54	4.5	6.31	6.46	5.84	6.71	6.47	6.38	5.98
Baichuan2-13B-Chat	5.26	3.96	3.83	4.08	6.56	5.74	6.19	7.03	7.21	6.72	6.49
Baichuan2-7B-Chat	5.05	3.68	3.23	4.13	6.42	5.72	5.71	7.08	7.41	6.86	5.73
ChatGLM3-6B	5.01	3.7	3.44	3.95	6.33	6.13	5.72	6.92	7.11	6.31	5.77
InternLM-20B	4.97	3.67	3.46	3.87	6.27	5.65	5.52	6.71	6.77	6.35	6.61
Qwen-7B-Chat	4.74	3.66	3.51	3.8	5.83	6.01	5.52	5.89	6.28	6.16	5.12
ChatGLM2-6B	4.57	3.32	3.28	3.35	5.83	5.24	5.12	6.68	6.83	5.95	5.15
Chinese-LLaMA-2-7B-Chat	3.44	2.42	2.13	2.7	4.46	4.59	4.29	4.39	4.64	4.91	3.94
InternLM-Chat-7B	3.24	2.1	2.34	1.85	4.39	3.43	3.76	5.37	4.63	5.01	4.15
LLaMA-2-13B-Chinese-Chat	3.14	2.35	2.12	2.58	3.93	4.31	2.9	4.34	3.52	4.04	4.47

评测网站Url: <https://lmbench.ai/align/data>



办公室有四张桌子，五张椅子，桌子上三台电脑，墙上有一面黑板，一扇窗户，上面三盏灯，桌子下面四个柜子，一扇门，上面三盏灯，墙上一面黑板，一扇窗户。这是一个办公室。



有8个房间，包括1间客厅，1间厨房，2间卧室，2间浴室，1个阳台，1个未定义的房间，有8个门。卧室和浴室有门连通。客厅和浴室有门连通。客厅和卧室有门连通。



各种健康的水果放在桌子上



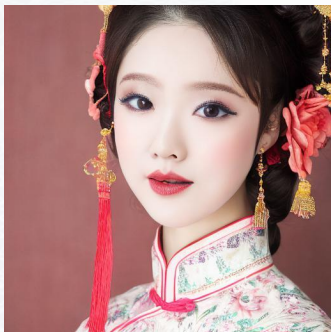
托盘上各种食品与调味品



红烧狮子头



唯美浪漫的故宫春色



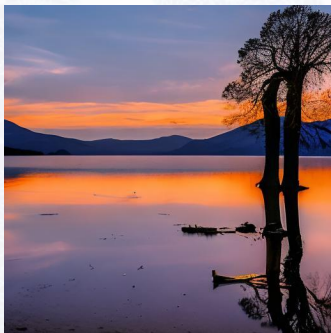
穿着旗袍的美丽中国女孩



月球上的飞船 赛博朋克



水墨画的竹子



夕阳下的树和湖



烟雨江南的山间村落




多模态大一统模型：基于Token-based的统一自回归预测

问题：现有多模态大模型通常通过Agent或任务特定的解码器实现对目标级、像素级任务的预测，缺乏对不同粒度任务的统一建模。

路径：将**定位、检测、分割**等不同粒度任务通过**自回归预测** (next token prediction) 统一到多模态大模型里边，实现一个模型既能完成VQA、图像描述等多模态任务，同时具备视觉定位、检测、分割等局部细粒度感知能力，不依赖专家模型和特定结构，减少模型碎片化，提高大模型的**可解释性**，并**降低幻觉**。

Detection & REC & Phrase Grounding



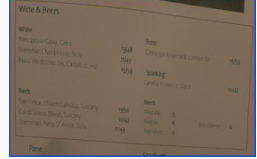
Report the coordinates of each detected object. The category set includes ...

Segmentation & RES...

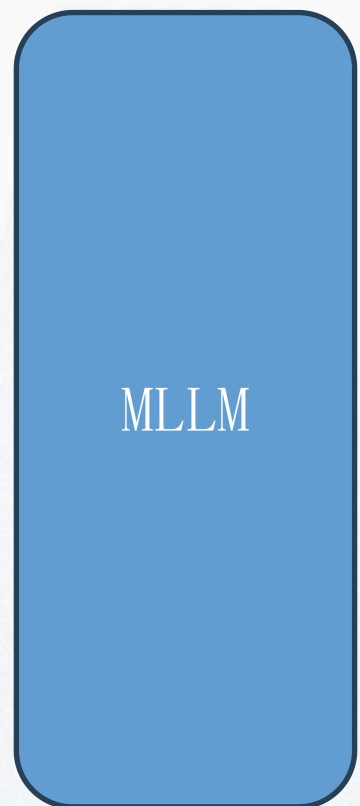


Segment all the people in the image.

OCR & VQA...




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




Output:

```
train-[0.079, 0.470, 0.820, 0.717]&traffic light-
[0.396, 0.400, 0.408, 0.431]&car-[0.416, 0.625,
0.730, 0.982]&traffic light-[0.433, 0.399, 0.450,
0.429]&bus-[0.510, 0.542, 0.644, 0.660].
```



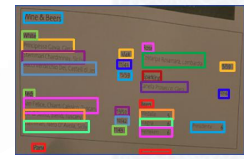
Output:

```
people-[0.015,0.485,0.081,0.521,...0.112,0.465]&
people-[0.015,0.485,0.081,0.521,...0.112,0.465]&
people-[0.015,0.485,0.081,0.521,...0.112,0.465]&
people-[0.015,0.485,0.081,0.521,...0.112,0.465].
```



Output:

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.





Griffon v1: 语言模型直接实现任意粒度目标定位

□ **背景**: 现有的多模态大模型视觉局部理解能力弱, 如何构建**统一的训练框架**, 实现LLM对图像中的细粒度**目标位置和类别理解**, 强化多模态大模型的细粒度感知能力。

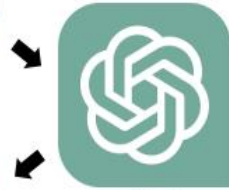
Stage 0: Data Generation

Example: "Can you point out the objects in the image possibly appear in the category..."

Instruction: Please refer to the given example and write 100 instruction templates to guide the model to detect objects of those categories in the category set and output the coordinates of each detected object.

Annotations (Optimal): horse-[0.123, 0.325, 0.498, 0.567]; house-[0.321, 0.465, 0.785, 0.892];

Image (Optimal): <image>

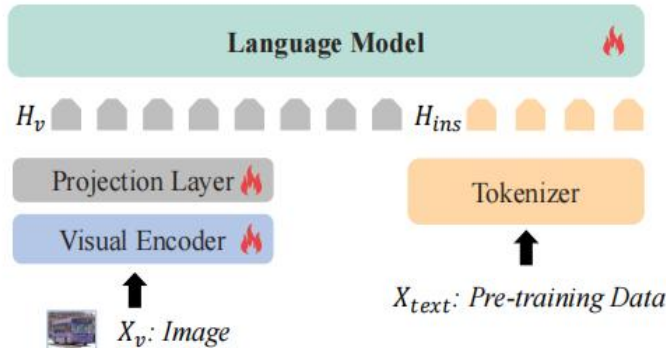


GPT-4V

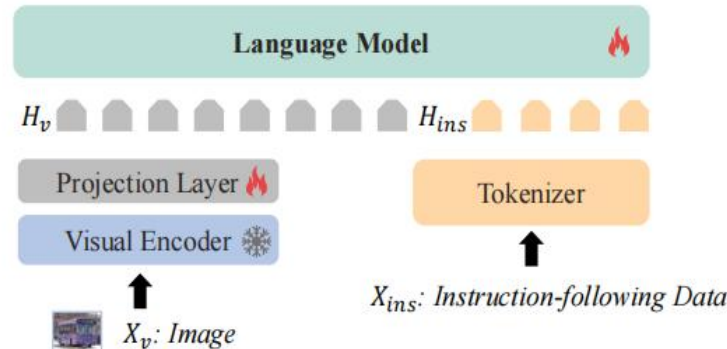
Sure! Here are 100 similar and non-repetitive instructions for commanding the model to detect objects in a category set and provide the coordinates of each detected object:

1. Identify and locate all the objects from the category set in the image. Please provide the coordinates for each detected object.
2. Detect the presence of objects belonging to the category set in the...

Stage I: Basic Scenarios Pre-training



Stage II: Full Scenarios Instruction Tuning



□ **方法**: 将坐标统一归一化[0-1]编码表示, 利用原有的语言词典, 采用LLM文本自回归训练及预测坐标:

✓ Stage1-**预训练定位基础模型**: 收集COCO、Grounding等开源数据, 使用GPT4V构建100余个prompt模板, 最终共计构建**600万**定位预训练数据;

✓ Stage2-**定位指令微调**: 强化模型对用户指令的跟随能力, 攻击构建**500K**定位指令数据。

ECCV2024-- Griffon: Spelling out All Object Locations at Any Granularity with Large Language Models.

Code: <https://github.com/jefferyZhan/Griffon>



Griffon v1: 语言模型直接实现任意粒度目标定位

- 视觉Grounding任务**超越同期Qwen-VL和苹果的FERRET-13B模型**
- 首个**在目标检测任务上进行评测的模型，并且**精度接近相同分辨率下的Faster R-CNN**
- 在更复杂的短语定位 (Phrase Grounding) 任务上也表现优异

Type	Model	Data	RefCOCO			RefCOCO+			RefCOCOg	
			val	test-A	test-B	val	test-A	test-B	val	test
Specialists	MDETR[19]	-	87.5	90.4	82.7	81.1	85.5	73.0	83.3	83.3
	TransVG[12]		81.0	82.7	78.4	64.8	70.7	56.9	68.7	67.7
	G-DINO-L[30]		90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
	UNINEXT-L[49]		91.4	93.7	88.9	83.1	87.9	76.2	86.9	87.5
Generalists	OFA-L[45]	10M	80.0	83.7	76.4	68.3	76.0	61.8	67.8	67.5
	Qwen-VL[3]	77M	89.4	92.3	85.3	83.1	88.3	77.2	85.6	85.5
	PINK[48]	5M	88.7	92.1	84.0	81.8	88.2	73.9	83.9	84.3
	FERRET-13B[52]	8.7M	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
	Shikra-13B[6]	4M	87.8	90.6	80.2	82.9	87.8	74.4	82.6	83.2
	Griffon-13B	6M	88.0	92.1	81.9	81.5	88.2	73.3	82.9	84.3
	Griffon-13B*	4M	90.1	93.4	86.1	84.8	90.5	77.8	86.1	87.2

RefCOCO系列视觉Grounding结果

Type	Model	Input Size	Epochs	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Specialists	FRCNN-R50[38]	448	1x	26.3	42.1	27.5	4.6	27.7	49.9
	Pix2Seq-R50[7]	1333	1x	43.0	61.0	45.6	25.1	46.9	59.4
	DETR-DC5[5]	1333	1x	15.5	29.4	14.5	4.3	15.1	26.9
Generalists	Griffon-13B	448	1	23.2	37.6	23.4	4.6	22.8	47.9

COCO目标检测结果

Type	Model	ANY		MERGED	
		val	test	val	test
Specialists	BAN	-	67.9	-	-
	DDPN	-	-	72.8	73.5
	VisualBert	70.4	71.3	-	-
	MDETR	82.5	83.4	82.3	83.8
Generalists	UniTAB	-	-	78.8	79.6
	FERRET-13B	-	-	81.1	84.8
	Shikra-13B	-	-	77.4	78.4
	Griffon-13B+	83.7	84.2	82.0	82.8

Flickrs30K Entities短语定位结果

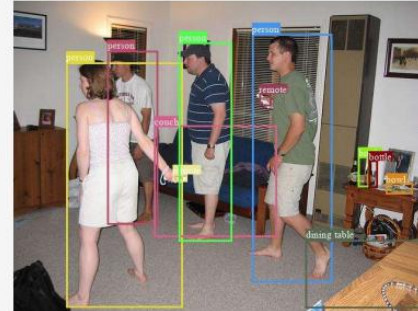


Tell me the location of the teacher in class.

[0.513, 0.266, 0.618, 0.580]



单个指代单个目标场景



Scan the image for any objects from the category set and report the coordinates of each detected object. The category set includes person, coach, cat, dog, remote, bottle, bowl, photo, dining table, boat, car... The output format for each detected object is class-name-[top-left coordinate, bottom-right coordinate] e.g. bottle-[0.001, 0.345, 0.111, 0.678]. Concatenate them with &.

person-[0.127, 0.198, 0.407, 0.987]&person-[0.226, 0.161, 0.347, 0.720]&couch-[0.339, 0.400, 0.633, 0.760]&remote-[0.383, 0.557, 0.418, 0.582]&person-[0.398, 0.135, 0.525, 0.774]&person-[0.574, 0.108, 0.766, 0.905]&remote-[0.591, 0.296, 0.620, 0.338]&dining table-[0.703, 0.769, 1.000, 0.989]&bottle-[0.829, 0.501, 0.856, 0.599]&bottle-[0.854, 0.513, 0.876, 0.600]&bowl-[0.896, 0.585, 0.960, 0.622]



多个类别多个目标场景



In this picture, identify and locate all the people in the front.

all the people-[0.268, 0.489, 0.446, 0.942]&all the people-[0.480, 0.453, 0.645, 0.920]&all the people-[0.797, 0.426, 0.951, 0.913]&all the people-[0.036, 0.185, 0.194, 0.975]&all the people-[0.384, 0.208, 0.502, 0.565]&all the people-[0.176, 0.154, 0.312, 0.915]&all the people-[0.799, 0.232, 0.875, 0.542]&all the people-[0.683, 0.214, 0.826, 0.571]&all the people-[0.496, 0.241, 0.609, 0.571] &all the people-[0.877, 0.223, 0.998, 0.817]&all the people-[0.295, 0.228, 0.406, 0.556]&all the people-[0.596, 0.223, 0.710, 0.817]&all the people-[0.018, 0.174, 0.194, 0.969]



单个类别多个目标场景



Can you help me find any car in the picture?

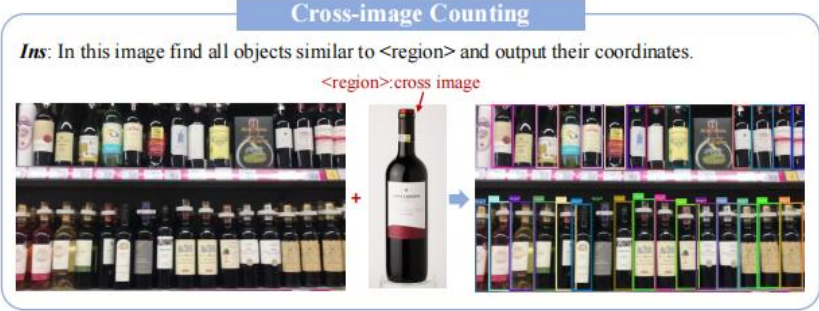
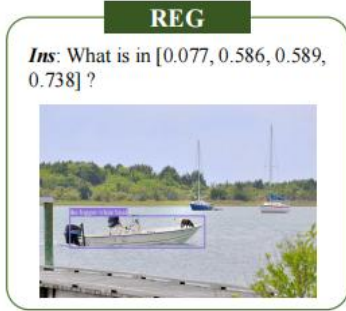
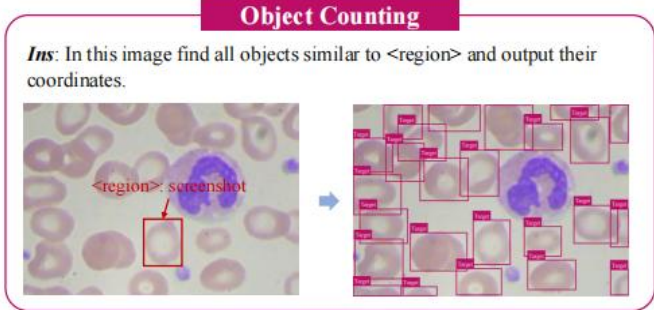
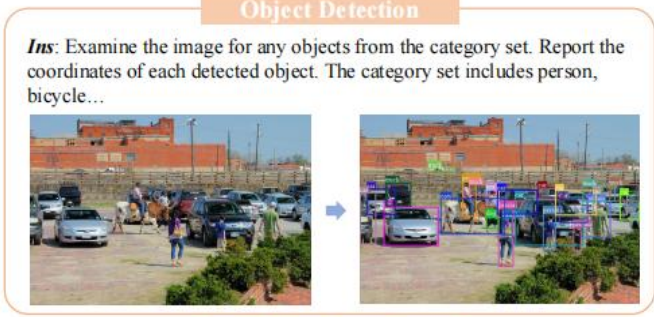
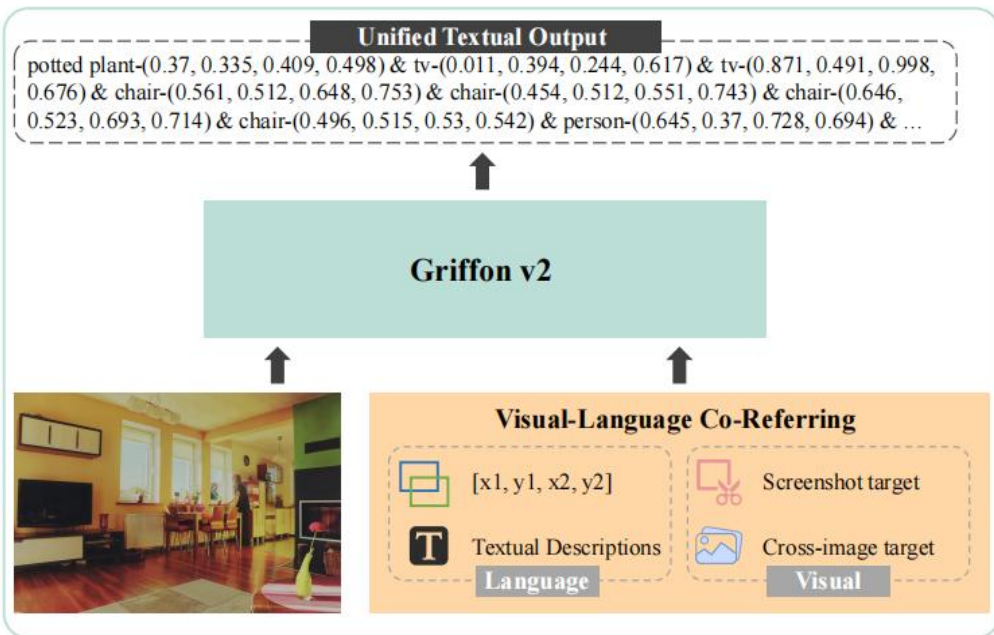
None.



目标不存在判断场景

Griffon v2: 视觉-文本双指代统一大模型

□ **目标**: 继续围绕增强多模态大模型视觉局部理解能力, 从**数据、分辨率、指代方式、模型结构**等角度入手, 构建具备定位、分割、计数、区域描述能力的可交互性统一大模型。



- **数据**: 构建千万级局部定位预训练数据, 包含box、mask、细粒度标签、区域描述和密集视觉指示数据, 并扩充指令微调数据 (900K)。
- **分辨率**: 全面支持高分辨率 (1024*1024) 输入, **非切图/切块**, 强化对细节理解。
- **视觉-文本指代**: 融入多种指代信息, 图片+文本, 支持REC、REG、detection以及Counting等视觉任务。
- **分割**: 持续推进模型对像素级的语义理解, LLM自回归**统一建模**。



Griffon v2: 视觉-文本双指代统一大模型

- 视觉Grounding任务**超越目前最优定位优化模型CogVLM-17B**
- 在短语定位 (Phrase Grounding) 及视觉区域描述任务 (REG) **性能表现优异**
- 首次**在目标检测、开放目标计数任务上精度**超越多个目标检测、目标计数专有模型**

Type	Model	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
Specialists	MDETR [17]	87.5	90.4	82.7	81.1	85.5	73.0	83.3	83.3
	G-DINO-L [29]	90.6	93.2	88.2	82.8	89.0	75.9	86.1	87.0
	UNINEXT-L [48]	91.4	93.7	88.9	83.1	87.9	76.2	86.9	87.5
	Griffon-13B† [55]	90.1	93.4	86.1	84.8	90.5	77.8	86.1	87.2
	CogVLM [44]	92.5	93.9	88.7	87.5	91.8	81.5	89.5	90.1
	Griffon v2†	92.3	94.0	89.5	88.7	92.3	82.8	90.2	90.2
Generalist	OFA-L [43]	80.0	83.7	76.4	68.3	76.0	61.8	67.8	67.5
	KOSMOS-2 [33]	52.3	57.4	47.3	45.5	50.7	42.2	60.6	61.7
	Shikra-13B [8]	87.8	90.6	80.2	82.9	87.8	74.4	82.6	83.2
	Qwen-VL [4]	89.4	92.3	85.3	83.1	88.3	77.2	85.6	85.5
	Ferret-13B [51]	89.5	92.4	84.4	82.8	88.1	75.2	85.8	86.3
	Griffon-13B [55]	88.0	92.1	81.9	81.5	88.2	73.3	82.9	84.3
	Griffon v2	89.6	91.8	86.5	81.9	85.5	76.2	85.9	86.0

RefCOCO系列视觉Grounding结果

Type	Model	CIDEr	Meteor
Spec.	SLR [53]	66.2	15.9
	ASM [45]	41.9	13.6
	Grit [46]	71.6	15.2
Gen.	KOSMOS-2 [33]	60.3	12.2
	Griffon v2	72.5	12.1

RefCOCOg 区域描述结果

Type	Model	ANY	MERGED
Spec.	DDPN [54]	-	73.5
	VisualBert [21]	71.3	-
	MDETR [17]	83.4	83.8
Gen.	UniTAB [50]	-	79.6
	Ferret-13B [51]	-	84.8
	Shikra-13B [8]	-	78.4
	Griffon-13B [55]	84.2	82.8
	Griffon v2	84.8	83.1

Flickrs30K Entities短语定位结果



Griffon v2: 视觉-文本双指代统一大模型

首次在目标检测、开放目标计数任务上精度超越多个目标检测、目标计数专有模型

Type	Model	Res.	Epochs	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Specialists	FRCNN-FPN [38]	1022	12	37.9	58.6	40.9	20.4	41.1	50.3
	FRCNN-C4 [38]	1022	12	35.6	55.7	37.8	17.0	40.6	50.3
	DAB-DETR [28]	1333	12	38.0	60.3	39.8	19.2	40.9	55.4
	Pix2Seq [9]	1333	300	43.0	61.0	45.6	25.1	46.9	59.4
	DETR [56]	1333	500	42.0	62.4	44.2	20.5	45.8	61.1
Generalist	Griffon-13B [55]	448	1	24.8	40.6	25.1	5.9	25.5	48.7
	Griffon v2	1022	1	38.5	54.3	41.2	19.4	43.2	57.6

COCO目标检测结果

Type	Model	MAE(↓)	NAE(↓)
Specialists	FamNet [37]	68.5	2.3
	FSDetView [47]	29.0	0.8
	Counting-DETR [32]	23.5	0.6
Generalist	Griffon v2	20.3	0.5

FSCD-LVIS 开放目标计数结果

Griffon v2: 视觉-文本双指代统一大模型

文本指代: 支持文本、坐标指代对话输入

视觉指代: 支持跨图、截图指代对话输入

Conversation

T Where are **the motorcycle** in the image?

Motorcycle-[0.000, 0.358, 0.272, 0.679] & Motorcycle-[0.261, 0.360, 0.521, 0.675]

What is [0.000, 0.358, 0.272, 0.679]?

Motorcycle on the far left.

Visualization



Conversation

Any objects similar to **<region>** in the image.

2: Motorcycle-[0.000, 0.358, 0.272, 0.679] & Motorcycle-[0.261, 0.360, 0.521, 0.675]

Locate and count **<region>** in the image.

2: Motorcycle-[0.000, 0.358, 0.272, 0.679] & Motorcycle-[0.261, 0.360, 0.521, 0.675]

Visualization



(a) Language Referring

(b) Visual Referring



Griffon v2: 视觉-文本双指代统一大模型

- REG: 输入坐标框位置, 生成局部细粒度描述

Referring Expression Generation



Question Region:
[0.468,0.667,0.792,0.888]

Answer: An apple laptop computer that is being used by the girl with the gray shirt



Question Region:
[0.545,0.084,0.869,0.540]

Answer: a orange container holding a type of food that is shaped like a triangle



Question Region:
[0.349,0.209,0.546,0.744]

Answer: a shirtless man with blue shorts skateboarding down a ramp

- REC: 输入描述或词组, 预测对应的坐标框 (多个)

Visual/Phrase Grounding

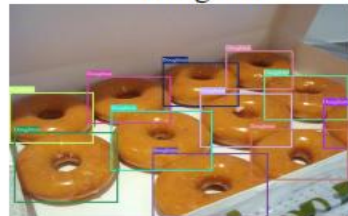
Penguin



Window



Doughnut

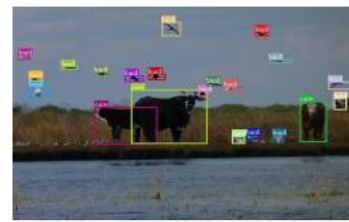


Referring Expression Comprehension

Grounding with referring expressions shown in images



Object Detection with MSCOCO Categories



Object Counting with Cross and Cropped Images



Cross-image:



Cross-image:



Cross-image:



Screenshot:



Screenshot:



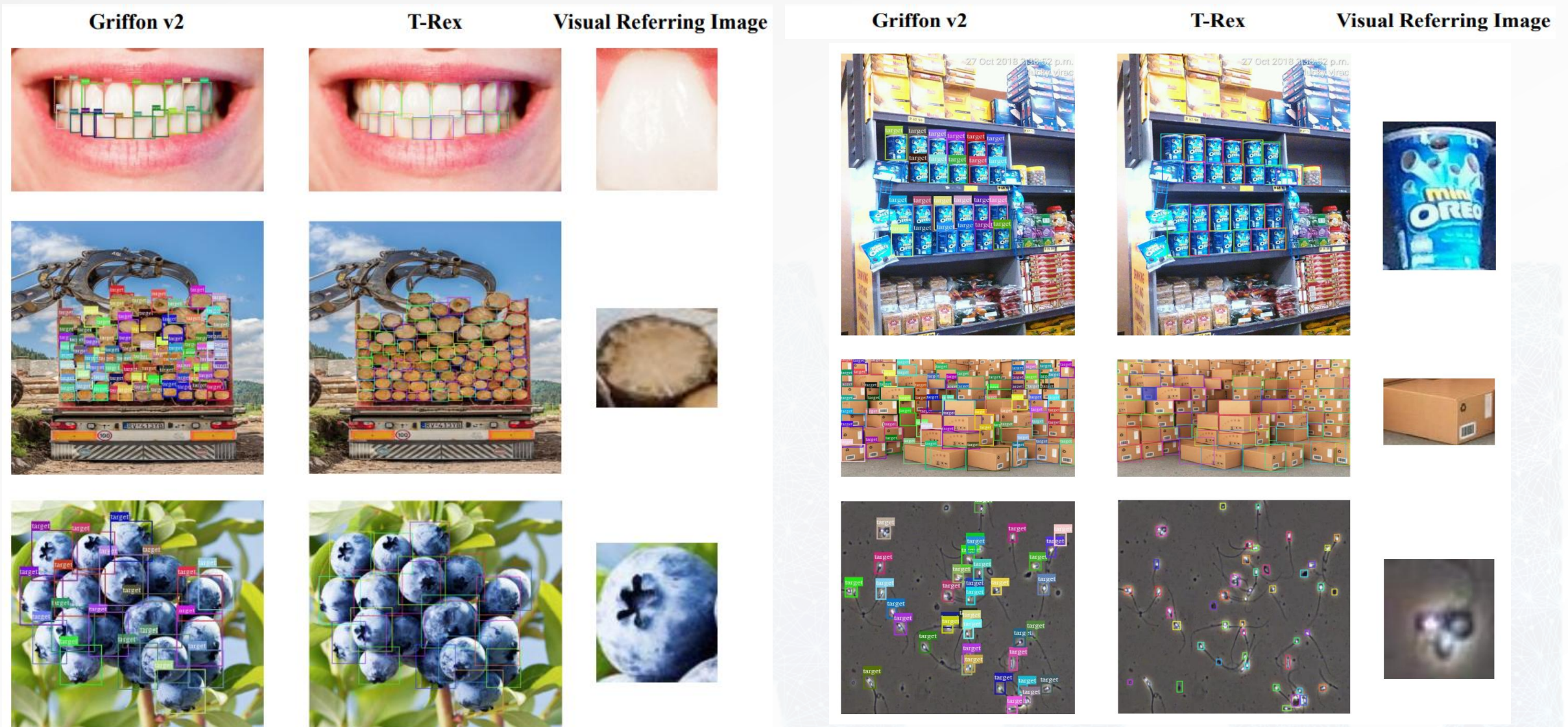
Screenshot:



- 目标计数: 输入跨图或截图目标, 定位相似目标



Griffon v2: 视觉-文本双指代统一大模型



视觉指代目标计数场景 (跨图、截图输入方式)



多模态产业联合体打造健全的AI产业生态

多模态人工智能产业联合体



紫东.太初

100家成员

理事长单位

中科院自动化所

成员单位

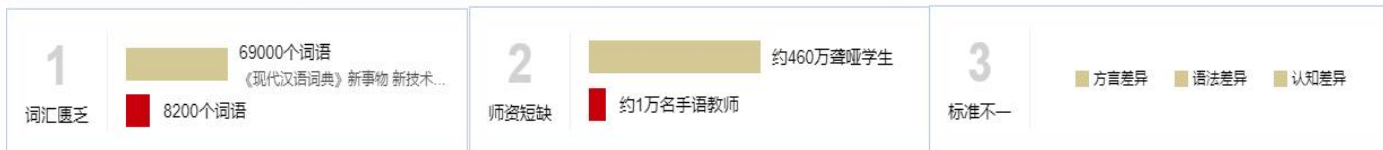
整合产学研用各方资源，打造多模态人工智能行业应用，探索通用人工智能产业化路径



应用案例：千博-多模态手语教考一体机

- ◆ 基于“紫东太初”开发手语多模态手语对话系统，已在300个电视频道的新闻播报中使用
- ◆ 多模态手语一体机已在湖南湘潭特校等数十所学校陆续上线，**开创了智能化聋哑教育新范式**

听障群体持续学习成为难题



AI赋能，让手语学习更简单、可持续

持续丰富手语词汇库 → 推动融合手语国家标准制定 → 服务听障认识产品开发与推广



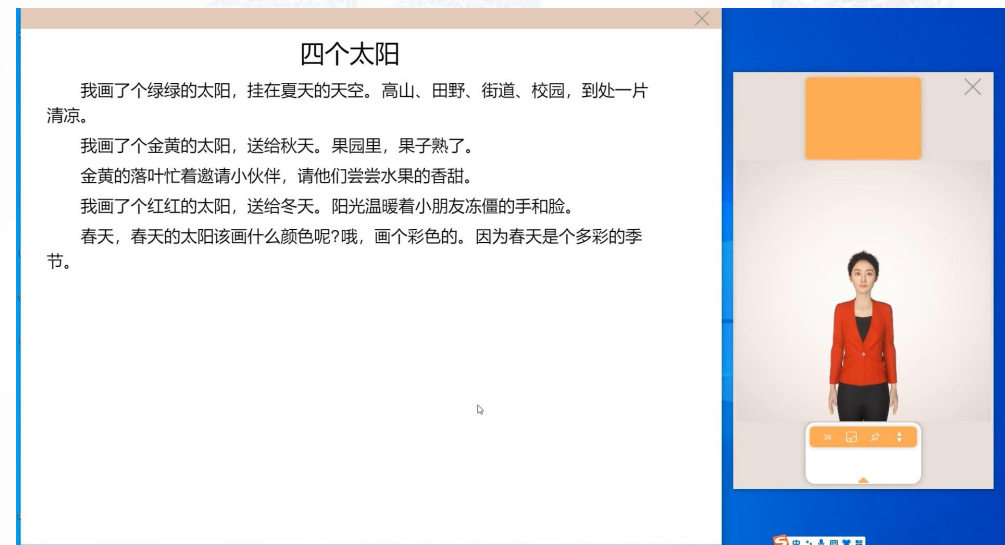
手语识别



手语生成



多模态手语教学



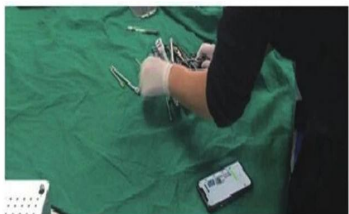


应用案例：九州通-智慧医疗器械管理

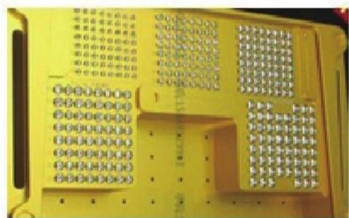
骨科类器械/耗材 管理痛点



规格复杂、数量繁多，
清点复核耗时长

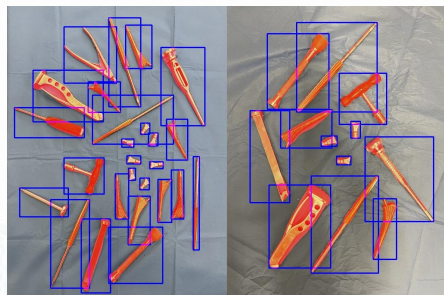


严重依赖人工清点，
出错率高

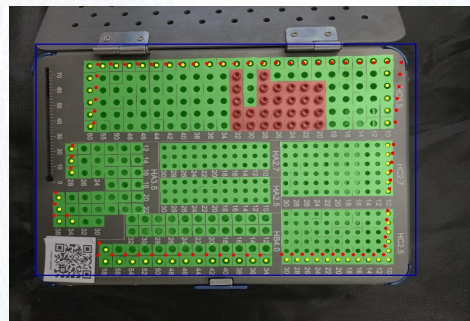


对成熟的工具管理员培养
成本高，企业负担大

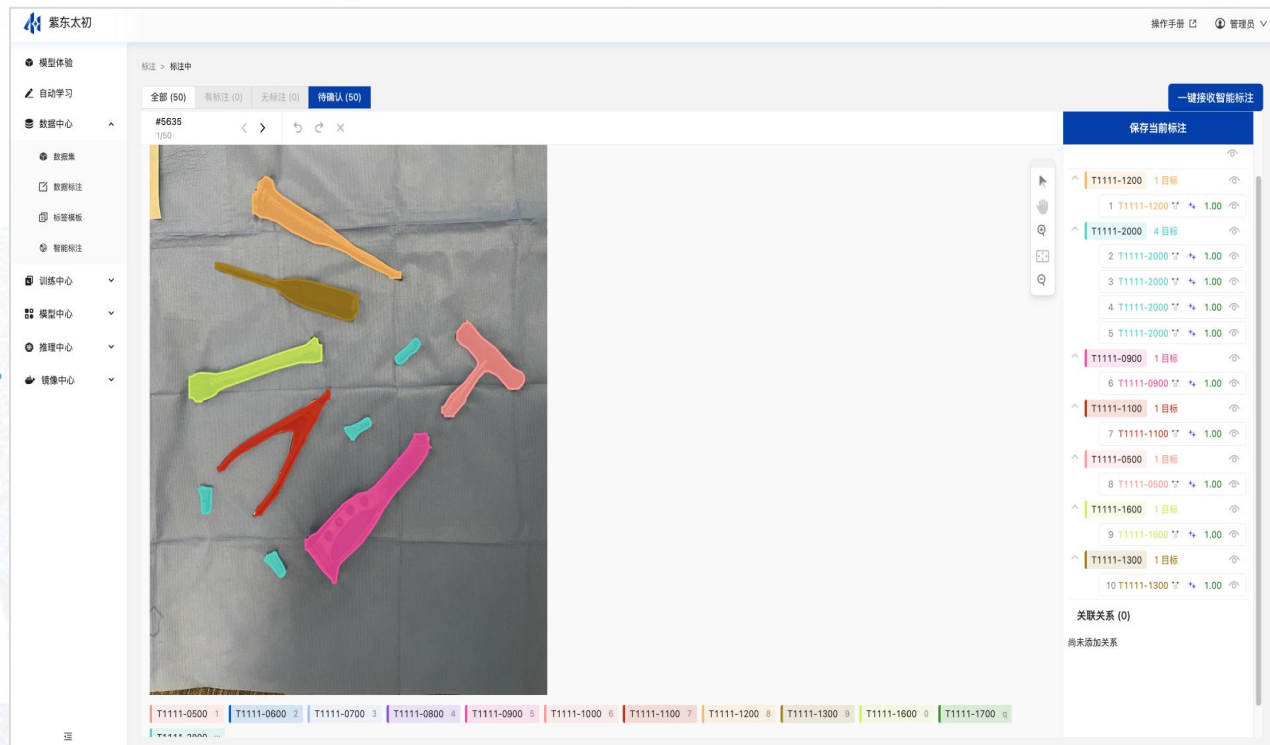
基于紫东太初大模型，实现数据智能标注、高效模型训练、模型灵活部署，实现骨科器械/耗材的自动识别和清点，实现**智能化、精细化管理**



医疗器械智能识别



耗材（钉盘）智能识别



✔ 一键微调，降低开发成本

✔ 仅需少量数据样本

5分钟/台 手术

准确率≥97%

人效提升6倍



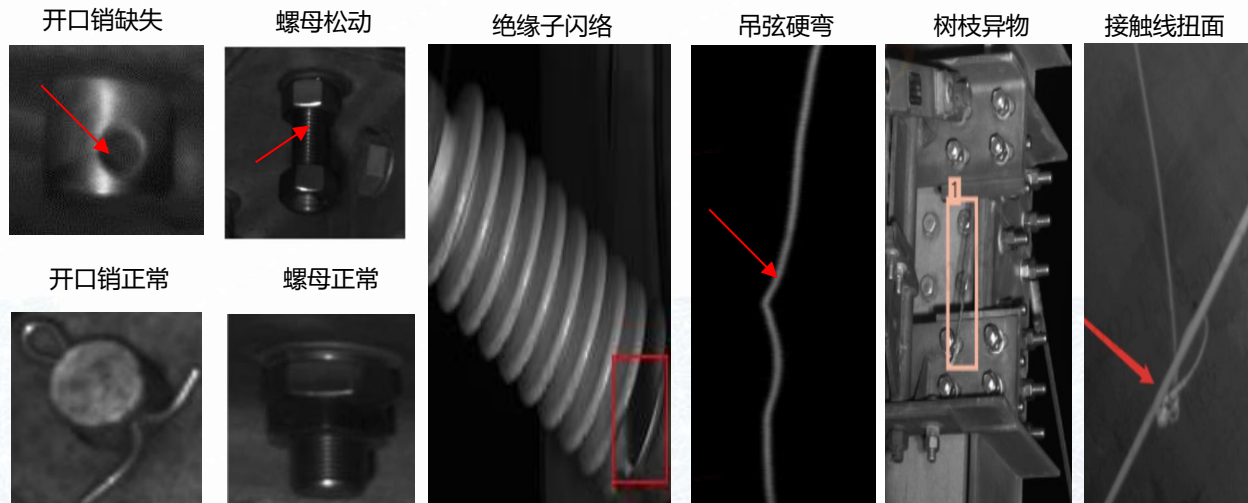
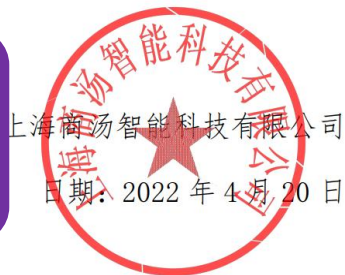
应用案例：高铁缺陷检测精度提升10%



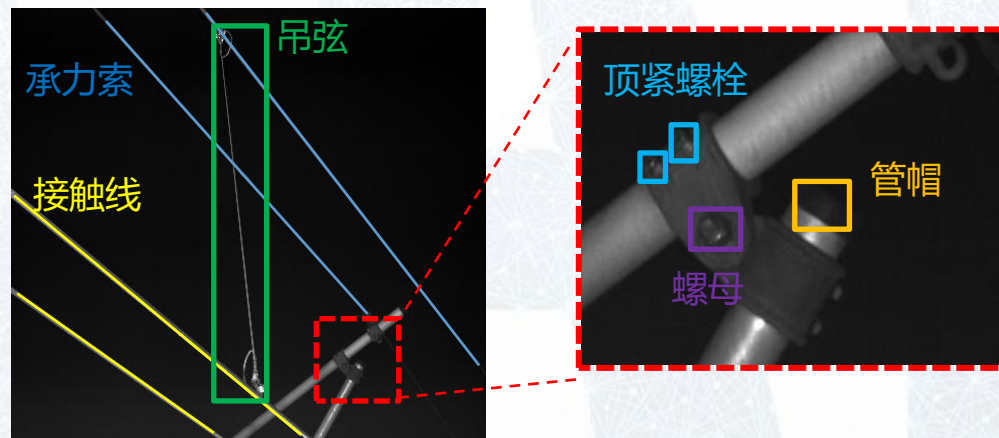
我司与中国科学院自动化研究所开展产学研合作项目，共同研发适用于高铁等工业缺陷巡检场景的通用视觉预训练模型，通过使用广泛的无标注数据，迁移到工业场景。该技术极大地提升了缺陷检测性能，有效地减少了缺陷检测标注样本的数量，降低了模型的开发成本，平均精度提升10%以上。

特此证明。

通用视觉预训练模型泛化到高铁工业缺陷场景，降低了标注样本数量，平均精度提升10%
-- 商汤科技



项点名称	商汤 (@1%)	Torchvision (@1%)	MoCoV3 (@1%)	自动化所 (@1%)
管帽缺失	13.4±1.0	32.2±0.3	38.3±0.4	38.4±0.9
管帽松脱	6.0±0.9	23.4±1.2	24.1±0.7	22.4±0.9
管帽破损	0.0±0.0	0.0±0.0	0.0±0.0	66.7±47.1
开口销掰开不足	93.55±0.00	93.5±0.0	93.5±0.0	93.5±0.0
螺栓松动	81.33±1.69	81.8±0.6	81.3±2.2	85.8±0.6
螺母松动	78.40±0.53	78.2±0.6	78.7±1.1	79.1±1.3
副螺母松动	86.77±0.65	86.0±0.8	87.1±0.0	86.0±1.5
AVG	55.40	58.67	58.62	66.07(+10.67)





应用案例：一带一路多模态对话虚拟人赋能沙特国家博物馆

- 基于紫东太初的多模态对话虚拟人通过一带一路应用在沙特国家博物馆
- 沙特王储参观了多模态对话虚拟人并高度评价



شركة هواوي

TECH. INVESTMENT SAUDI ARABIA
CO., LTD.
C.R. 1010181305

Letter No: HW-2022053001
Date: 2022/06/01

To: Institute of Automation, Chinese Academy of Sciences

Subject:
Appreciation Letter for Multi-modality Large Model Cooperation

Greetings

First of all, I would like to express our greetings and sincere appreciation.

Saudi Data and AI Authority (SDAIA) is the ministry responsible for driving and owning the national data and AI agenda and creating a data-driven and AI-supported government and economy which was established in line with the Saudi National Vision 2030.

As a global AI leader, Huawei partners with SDAIA to define the artificial intelligence strategy, innovation and development to place the Kingdom as a global leader in AI.

SDAIA, Huawei and the Institute of Automation of Chinese Academy of Sciences jointly deployed and exhibited the world's first 100-billion parameter triple-modality pretrained model "Zidongtaichu", which allowed Saudi Arabia to see the leading technology and potential wide application scenarios of large models, and more importantly it greatly improved the Saudi public's awareness of AI.

We express our high admiration for the development of the "Zidongtaichu" large-scale model by the Chinese Academy of Sciences, and our deep recognition for the world's leading AI capabilities of the Chinese Academy of Sciences. We also express our deep gratitude to the team for their great efforts in deployment, demonstration and joint innovation in Saudi Arabia.

Looking forward to the future, SDAIA, Huawei and the Chinese Academy of Sciences can carry out more in-depth and extensive cooperation on behalf of the Chinese and Saudi governments.

Yours sincerely,

Tao Geng
Enterprise Business Manager Director
Huawei Tech. Investment Saudi Arabia Co., Ltd.

شركة هواوي

تلك انفسمستت العربية السعودية
المحدودة
س.ت ١٠١٠١٨١٣٠٥

خطاب رقم: HW-2022053001
م 2022/06/01

إلى: السادة معهد الأتمتة ، الأكاديمية الصينية للعلوم

الموضوع: رسالة تقدير للتعاون حول النموذج الكبير متعدد الأبعاد

تحية طيبة وبعد

بداية أود أن أعرب عن تحياتنا وتقديرنا الصادق لكم.

الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا) هي الوزارة المسؤولة عن قيادة وإسناد البيانات الوطنية وقيادة الذكاء الاصطناعي وبناء حكومة واقتصاد مركزيين على البيانات و مدعومة بالذكاء الاصطناعي والتي تم إنشاؤها بما يتماشى مع الرؤية الوطنية السعودية 2030.

ويصفنا شركة عالمية رائدة في مجال الذكاء الاصطناعي، نتعاون هواوي مع سدايا لتحديد استراتيجية الذكاء الاصطناعي والابتكار والتطوير لوضع المملكة كرائد عالمي في مجال الذكاء الاصطناعي.

قامت لكم من سدايا و هواوي ومعهد الأتمتة التابع للأكاديمية الصينية للعلوم بتطبيق وعرض أول نموذج مسبق التدريب و 100 مليار متغير ثلاثي الأبعاد في العالم "Zidongtaichu" ، مما سمح للمملكة العربية السعودية بالإطلاع على التكنولوجيا الرائدة وسيناريوهات التطبيقات الواسعة المحتملة للنماذج الكبيرة ، والأهم من ذلك أنها رفعت بشكل كبير من وعي الجمهور السعودي بالذكاء الاصطناعي.

نعرب لكم عن إعجابنا الكبير بتطوير نموذج "Zidongtaichu" واسع النطاق من قبل الأكاديمية الصينية للعلوم ، وإعترافنا العميق بالقدرة العالمية والرائدة للأكاديمية الصينية للعلوم في مجال الذكاء الاصطناعي. كما نعرب عن امتناننا العميق لتفريق العمل على جهودهم الكبيرة في التطبيق والعرض والابتكار المشترك في المملكة العربية السعودية.

نتطلع إلى مستقبل تتكمن فيه سدايا و هواوي والأكاديمية الصينية للعلوم من إجراء تعاون أكثر عمقا ولسانغا ثابرة عن الحكومتين الصينية والسعودية.

وتقديروا فائق إحترامي...

تار جينج
الرئيس التنفيذي للطابع الأصل
شركة هواوي تلك انفسمستت العربية السعودية المحدودة



应用案例：AIMed燧石大模型，打造医药领域垂类GPT



紫东太初基座大模型

行业模型预训练

目的：通过学习特定领域的
数据，使模型可以理解
对应的内容和结构

数据集

- 医疗文献数据
- 医疗知识图谱
- 临床试验数据
- 医疗百科
- 医疗政策
- 新闻数据
- 报告数据
- 电子病例
- 医患问诊

指令微调

数据类型：开源数据+业务数据

自反馈模型

Reward模型训练

目的：通过医学反馈建立更好
的范式的经验性学习

强化反馈

目的：通过强化建立基于范式的
习惯表达

围绕业务应用（信息抽取，阅读理解，总结，问答对抽取等任务）
由数智内部专业的医学经理进行数据标注和校验

燧石（火石数智自有的数据湖仓）



清洗



去重



结构化



特征性标签

互联网/三方数据

医疗垂类大模型

未来展望

大模型（大语言模型和多模态模型）展现出了强大的多任务能力和泛化能力，已成为人工智能研究的新范式：

- **多任务大一统结构**：结构上进一步**做减法**，没有模态相关的Encoder，没有任务相关的Decoder，将理解与生成进一步统一，人工智能大一统模型？
- **可靠性**：多模态大模型面临“**幻觉**”现象，生成不真实、虚构、不一致或者无意义的内容，推理与回答**缺少可靠依据**，如何克服“幻觉”提高大模型的可靠性？
- **具身智能**：如何将多模态大模型与具身智能技术结合，开发机器人大模型，在真实环境中进行**感知、推理、规划、执行**一体化的智能代理？
- **跨学科交叉**：多模态大模型带来全新的技术创新，可以促进科学研究中跨学科的融合（融合不同学科、不同领域的知识），大模型+物理/医学/生命科学.....

谢谢大家！