上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

OpenGVLab

书生·万象 InternVL 2.0:
通过渐进式策略扩展开源多模态大模型的性能边界

王文海

2024.08

万象——理解真实世界一切事物和景象

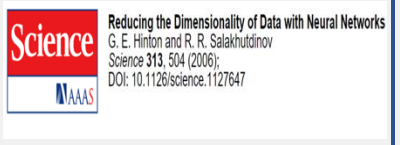
https://github.com/OpenGVLab/InternVL

目 录

研究背景：大语言模型&多模态大模型

历史：“特定任务+大数据”取得巨大成功
一个模型解决一个问题

深度学习理论突破 (2006)



深度置信网络

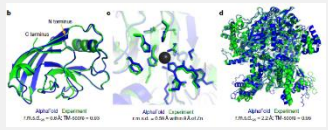
ImageNet竞赛 (2012)

1000类，100万数据

围棋比赛 (2016)

AlphaGo 4:1 李世乭

AlphaFold (2021)



蛋白质结构预测准确率新高

大规模语音识别 (2011)

Switchboard错误降低9%

人脸识别 (2014)

LFW识别率99%,超过人类

德州扑克 (2019)

首次在多人复杂对局中超越人类

未来：“通用性”
一个模型多种任务多种模态

通用人工智能将会成为我们和未来科技交互的一部分。

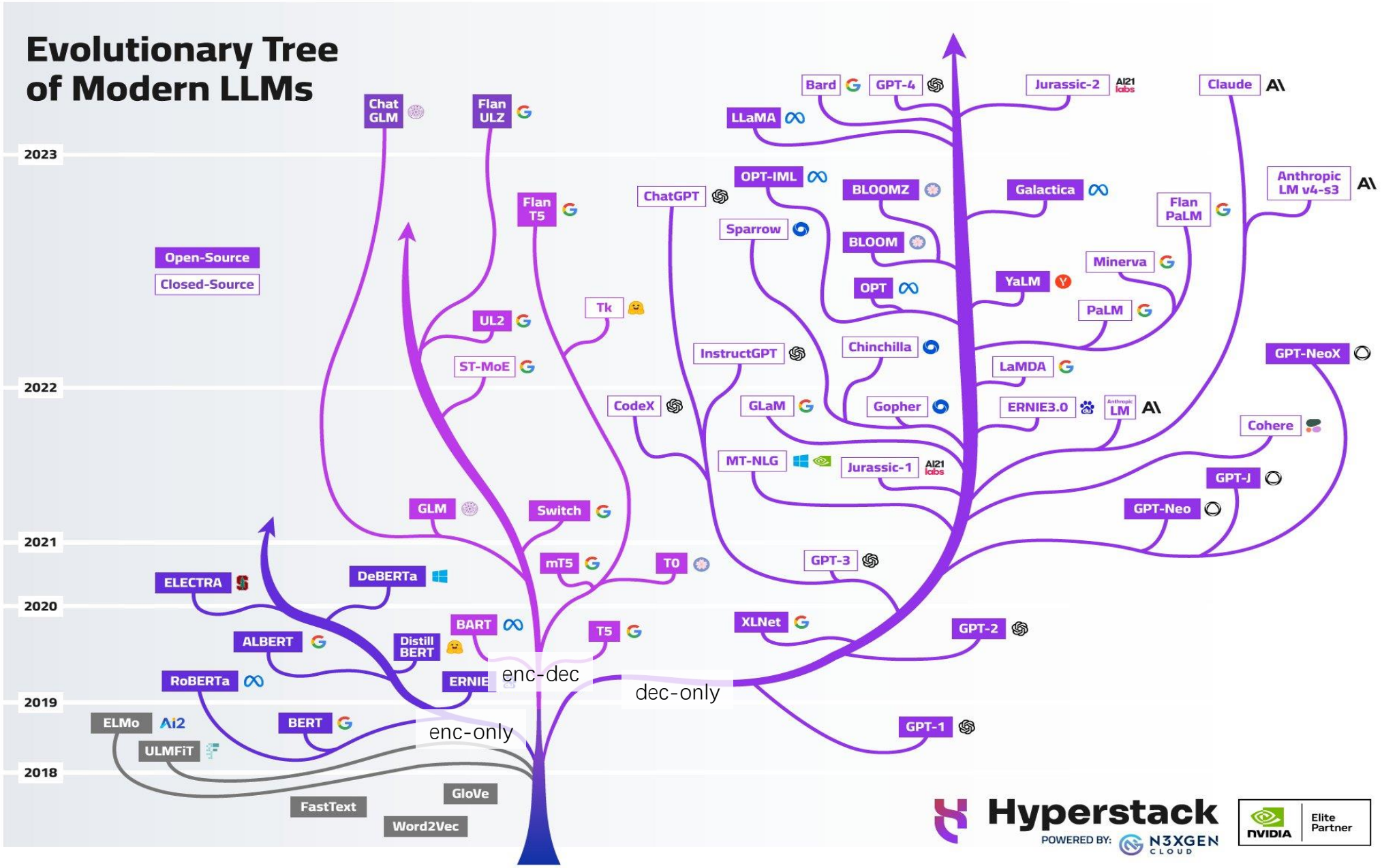
“AGI is going to be a part of how we interact with future tech.”

Yann Lecun
图灵奖获得者

以视觉为核心的多模态大模型有望在众多领域带来AI生产力革命

https://github.com/OpenGVLab/InternVL

研究背景：大语言模型&多模态大模型



Qwen

DeepSeek

InternLM

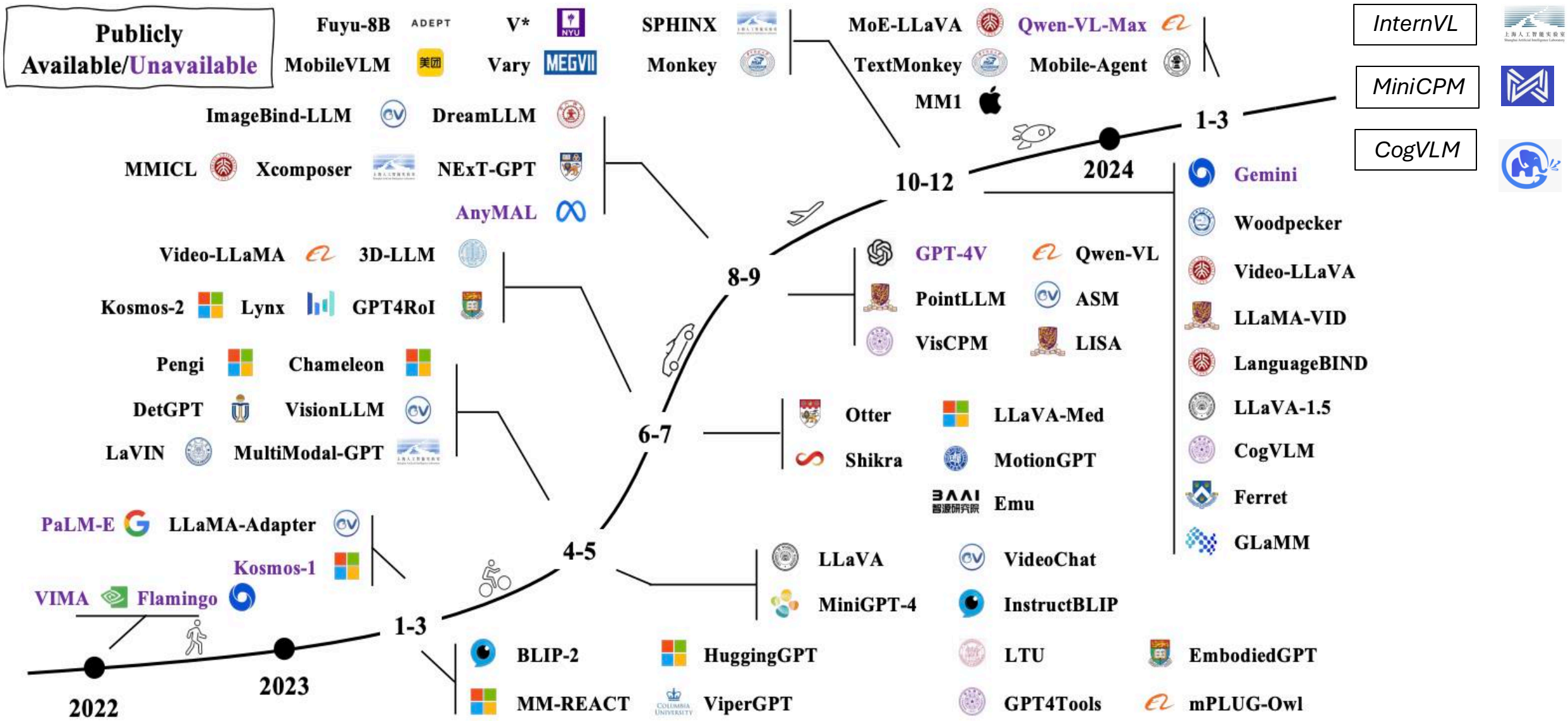
Baichuan

ChatGLM

Hyperstack POWERED BY: N3XGEN CLOUD

NVIDIA Elite Partner

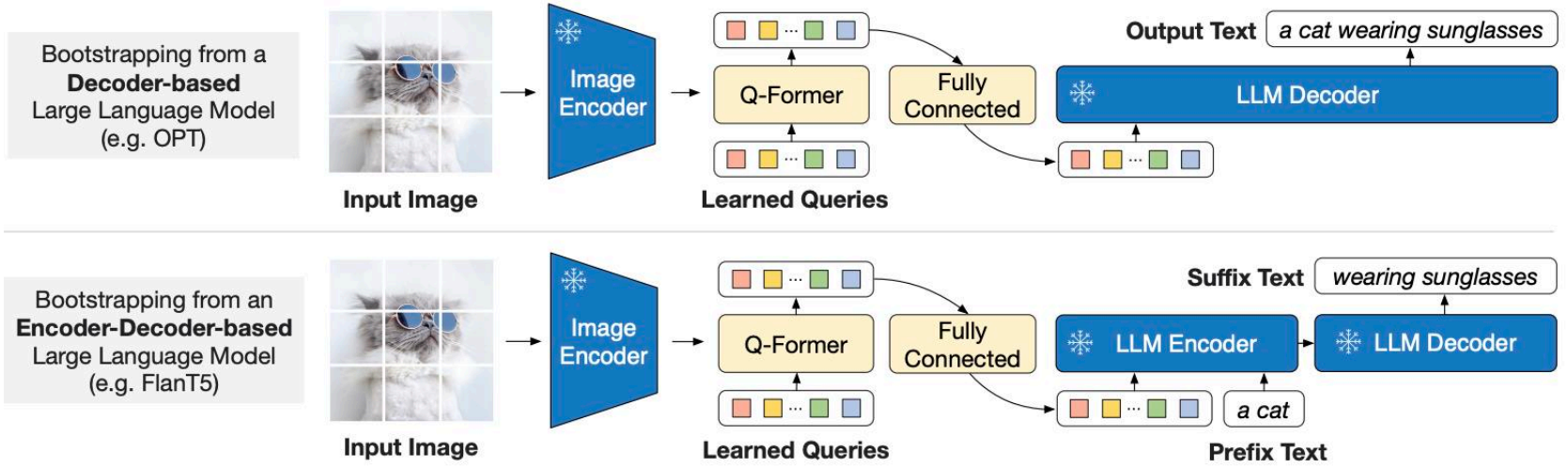
研究背景：大语言模型&多模态大模型



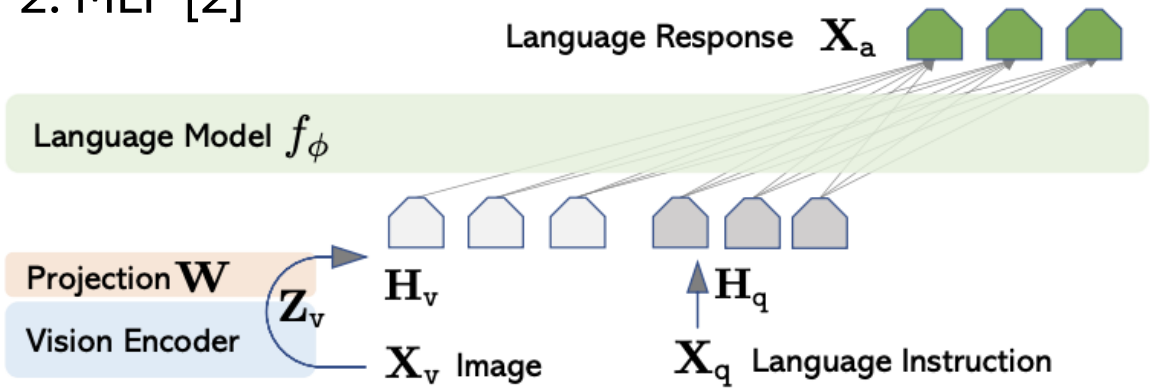
https://github.com/OpenGVLab/InternVL

研究背景：大语言模型&多模态大模型

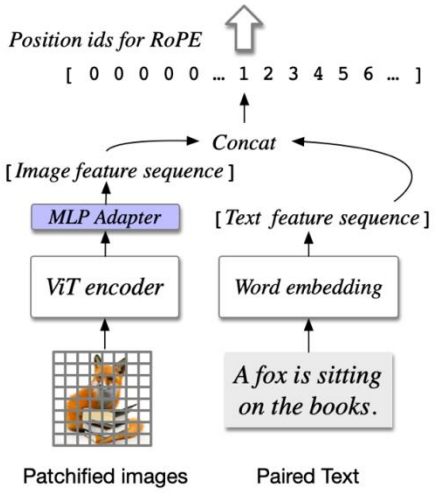
1. QFormer [1]



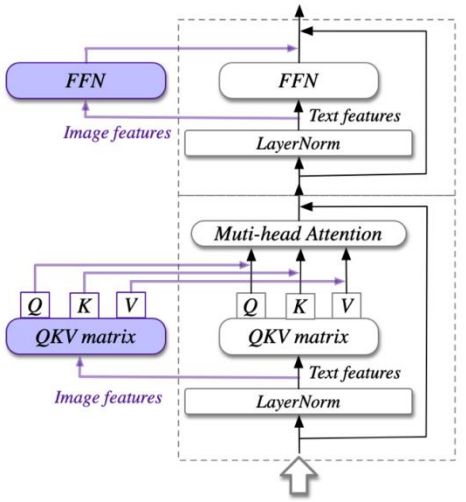
2. MLP [2]



3. MoE [3]



(a) The input of visual language model



(b) The visual expert built on the language model

[1] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.

[2] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2024, 36.

[3] Wang W, Lv Q, Yu W, et al. Cogvlm: Visual expert for pretrained language models[J]. arXiv preprint arXiv:2311.03079, 2023.

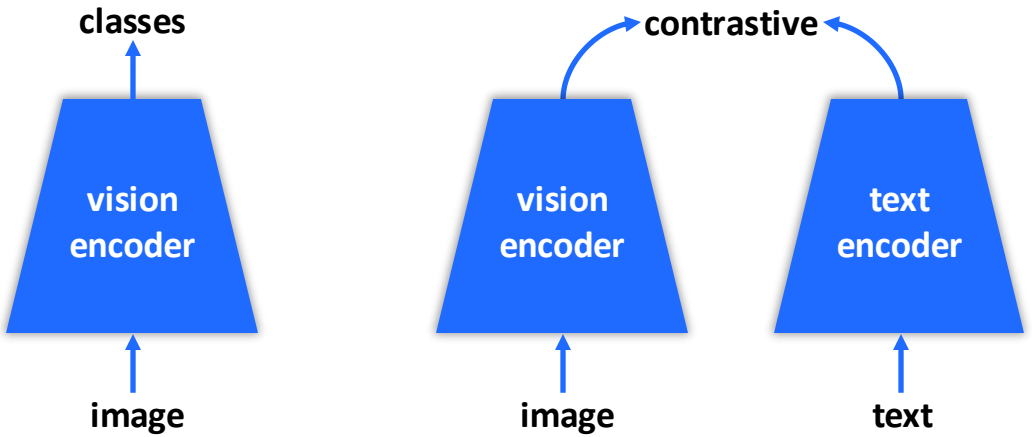
https://github.com/OpenGVLab/InternVL

目 录

InternVL：大规模视觉语言模型对齐

传统视觉/视觉-语言基础模型范式已落后于大语言模型的发展，亟需新的范式来推动其发展

2012 2021



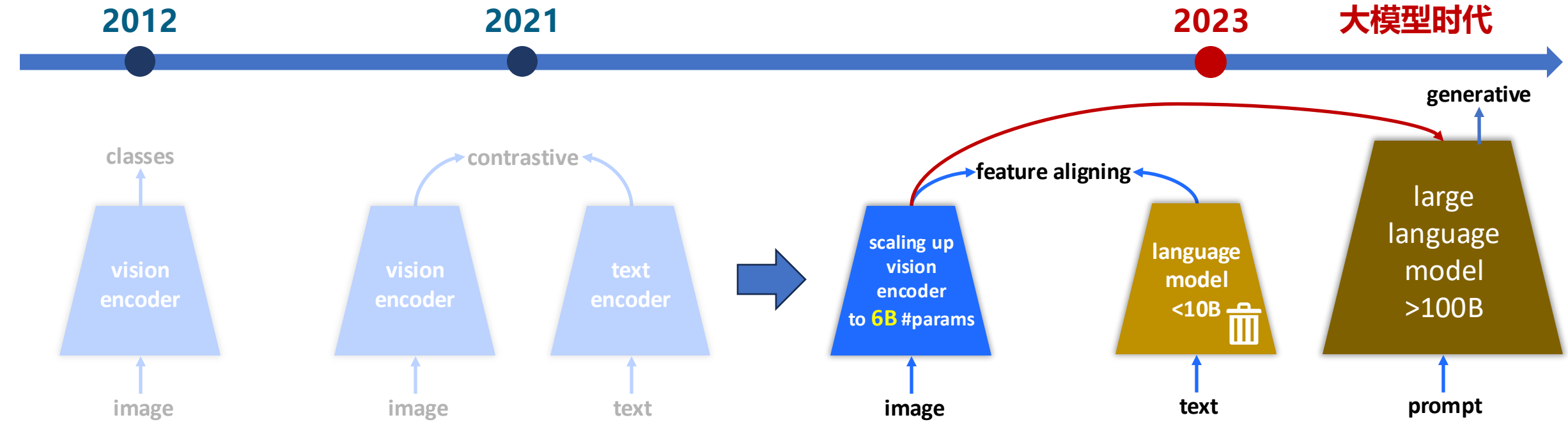
判别式预训练
AlexNet、ResNet

对比式预训练
CLIP、ALIGN

- 与LLM参数量差距过大
- 与LLM表征不一致
- 训练数据单一、数据量小

InternVL：大规模视觉语言模型对齐

传统视觉/视觉-语言基础模型范式已落后于大语言模型的发展，亟需新的范式来推动其发展



判别式预训练
AlexNet、ResNet

对比式预训练
CLIP、ALIGN

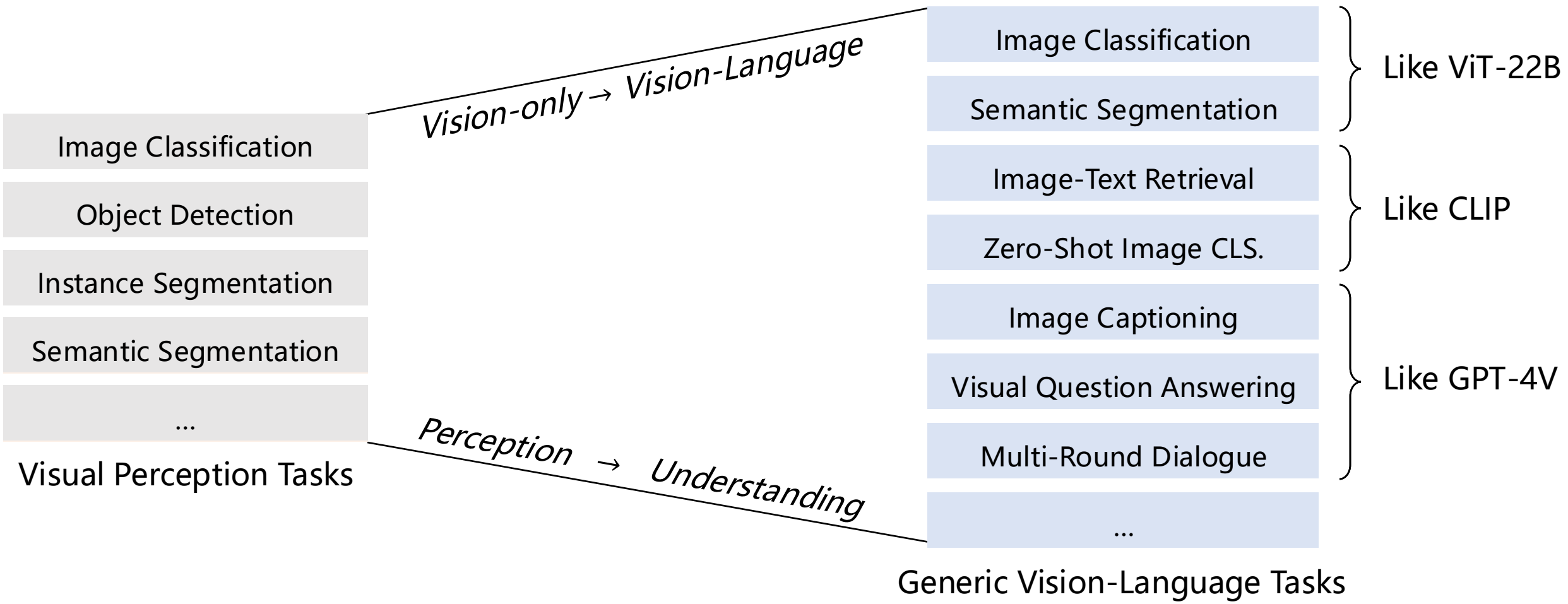
渐进式对齐训练
书生图文大模型-InternVL

- 与LLM参数量差距过大
- 与LLM表征不一致
- 训练数据单一、数据量小

- 60亿参数视觉模型+1000亿参数语言模型
- 渐进式对齐视觉基础模型和语言模型表征
- 大规模、多来源图文多模态训练数据

InternVL：大规模视觉语言模型对齐

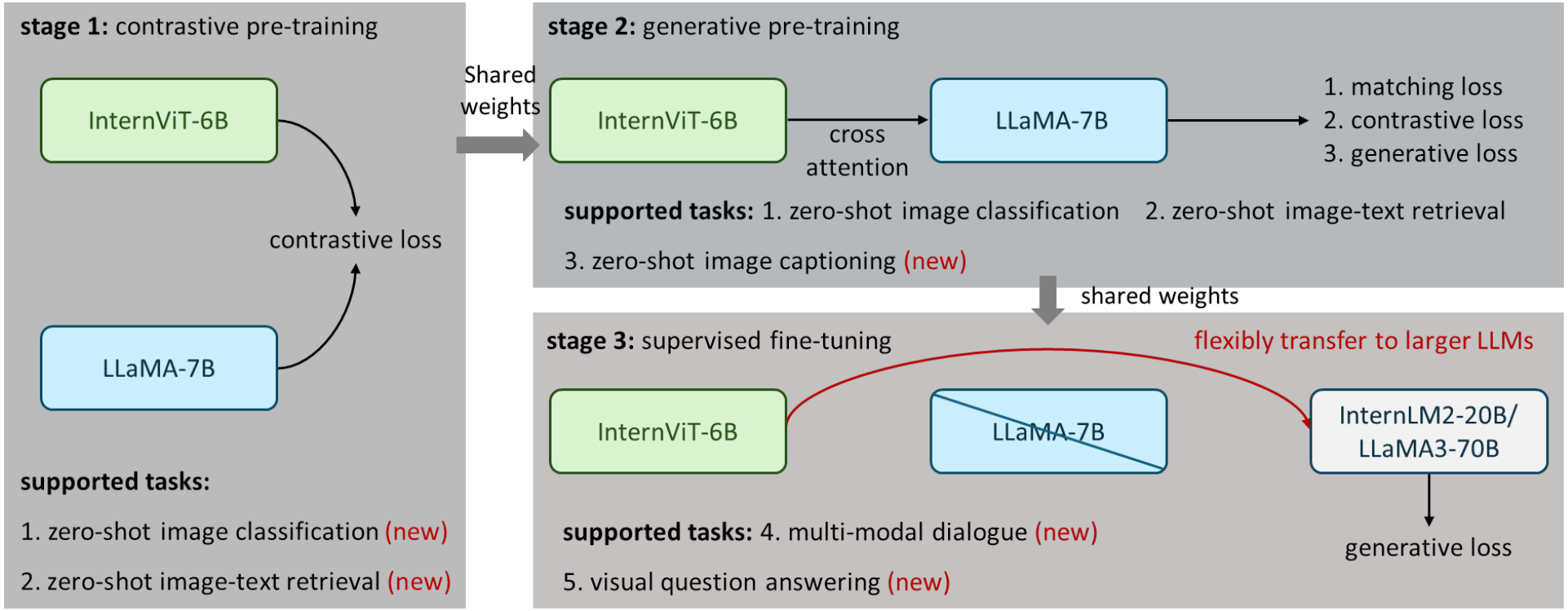
从适配视觉感知任务，到适配通用视觉语言任务，极大地扩宽了模型的适用范围



https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

核心思想：扩大视觉基础模型并为通用视觉语言任务进行对齐



设计1：扩大视觉模型至6B参数

步骤1：固定 60 亿参数，网格搜索模型宽度、深度、MLP Ratio和Attention Head维度

步骤2：使用CLIP 作为代理任务，找到在速度、准确性和稳定性之间取得平衡的模型

name	width	depth	MLP	#heads	#param (M)
ViT-G [173]	1664	48	8192	16	1843
ViT-e [23]	1792	56	15360	16	3926
EVA-02-ViT-E [130]	1792	64	15360	16	4400
ViT-6.5B [128]	4096	32	16384	32	6440
ViT-22B [37]	6144	48	24576	48	21743
InternViT-6B (ours)	3200	48	12800	25	5903

https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

设计1：扩大视觉模型至6B参数

基于原始ViT结构，通过搜索模型深度{32, 48, 64, 80}，注意力头维度{64, 128}， 以及MLP比率{4, 8}，将视觉模型扩大至6B参数，找到速度、精度、稳定性平衡的模型

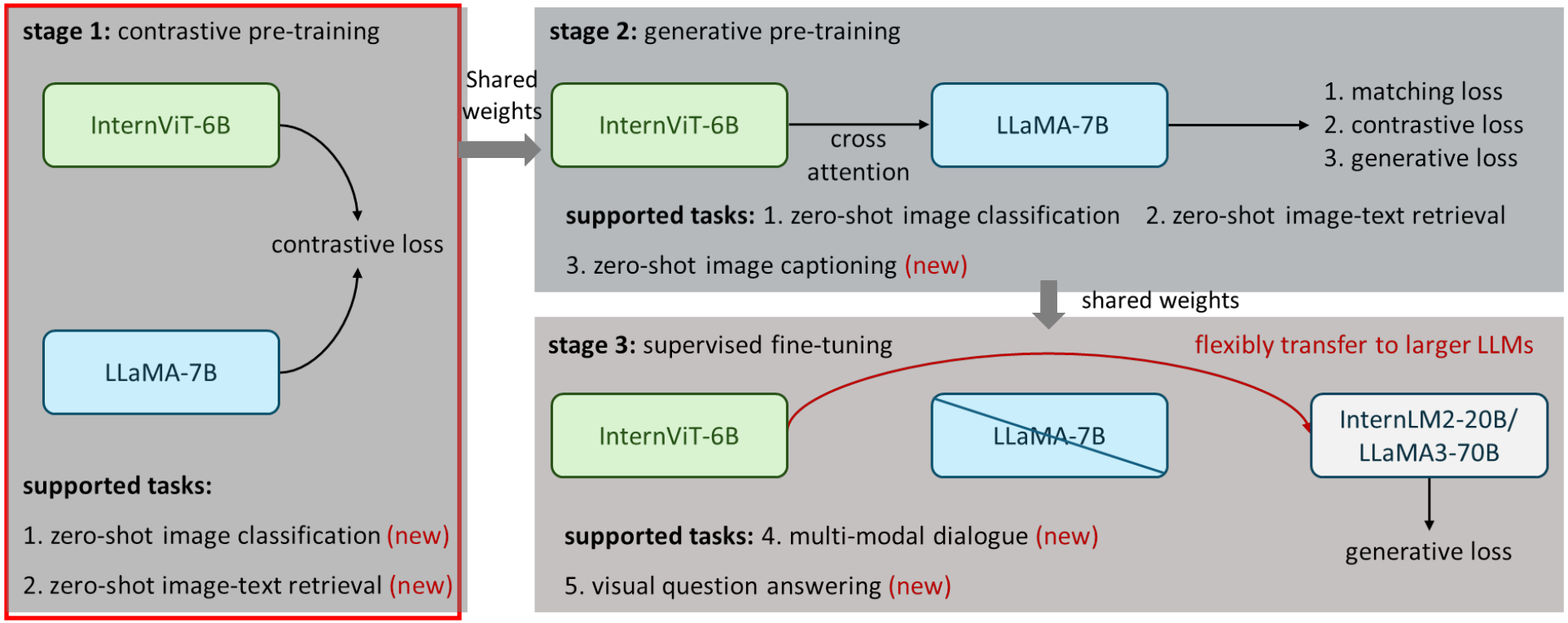
name	width	depth	MLP	#heads	#param	FLOPs	throughput	zs IN
variant 1	3968	32	15872	62	6051M	1571G	35.5 / 66.0	65.8
variant 2	3200	48	12800	50	5903M	1536G	28.1 / 64.9	66.1
variant 3	3200	48	12800	25	5903M	1536G	28.0 / 64.6	66.2
variant 4	2496	48	19968	39	5985M	1553G	28.3 / 65.3	65.9
variant 5	2816	64	11264	44	6095M	1589G	21.6 / 61.4	66.2
variant 6	2496	80	9984	39	5985M	1564G	16.9 / 60.1	66.2

Table 11. **Comparison of hyperparameters in InternViT-6B.** The throughput (img/s) and GFLOPs are measured at 224×224 input resolution, with a batch size of 1 or 128 on a single A100 GPU. Flash Attention [35] and bf16 precision are used during testing. "zs IN" denotes the zero-shot top-1 accuracy on the ImageNet-1K validation set [38]. The final selected model is marked in gray.

https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

核心思想：扩大视觉基础模型并为通用视觉语言任务进行对齐



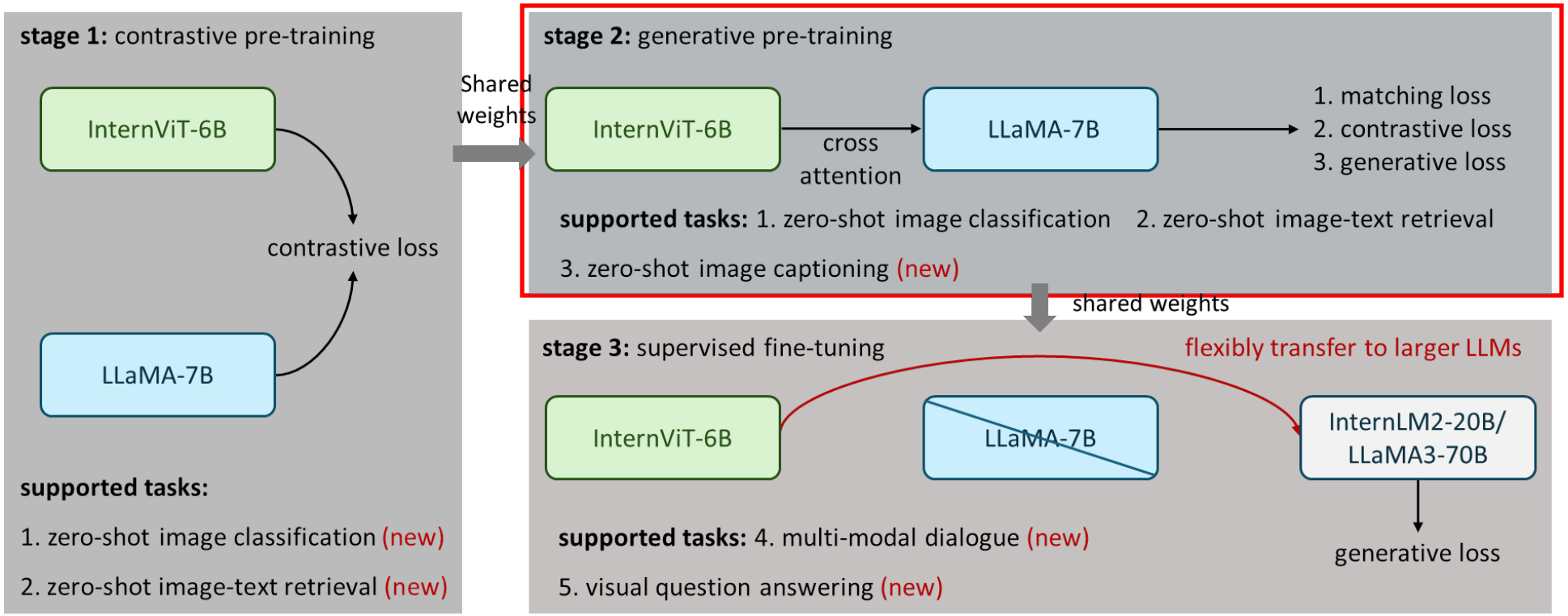
设计2：渐进式的图像-文本对齐策略

阶段1：利用海量带噪声的图文数据进行对比学习预训练（~5B图像）

https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

核心思想：扩大视觉基础模型并为通用视觉语言任务进行对齐



设计2：渐进式的图像-文本对齐策略

阶段1：利用海量带噪声的图文数据进行对比学习预训练（~5B图像）

阶段2：利用过滤后的高质量图文数据进行对比学习和生成式联合训练（~1B图像）

https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

设计2：渐进式的图像-文本对齐策略

阶段1：利用海量带噪声的图文数据进行对比学习预训练（~5B图像）

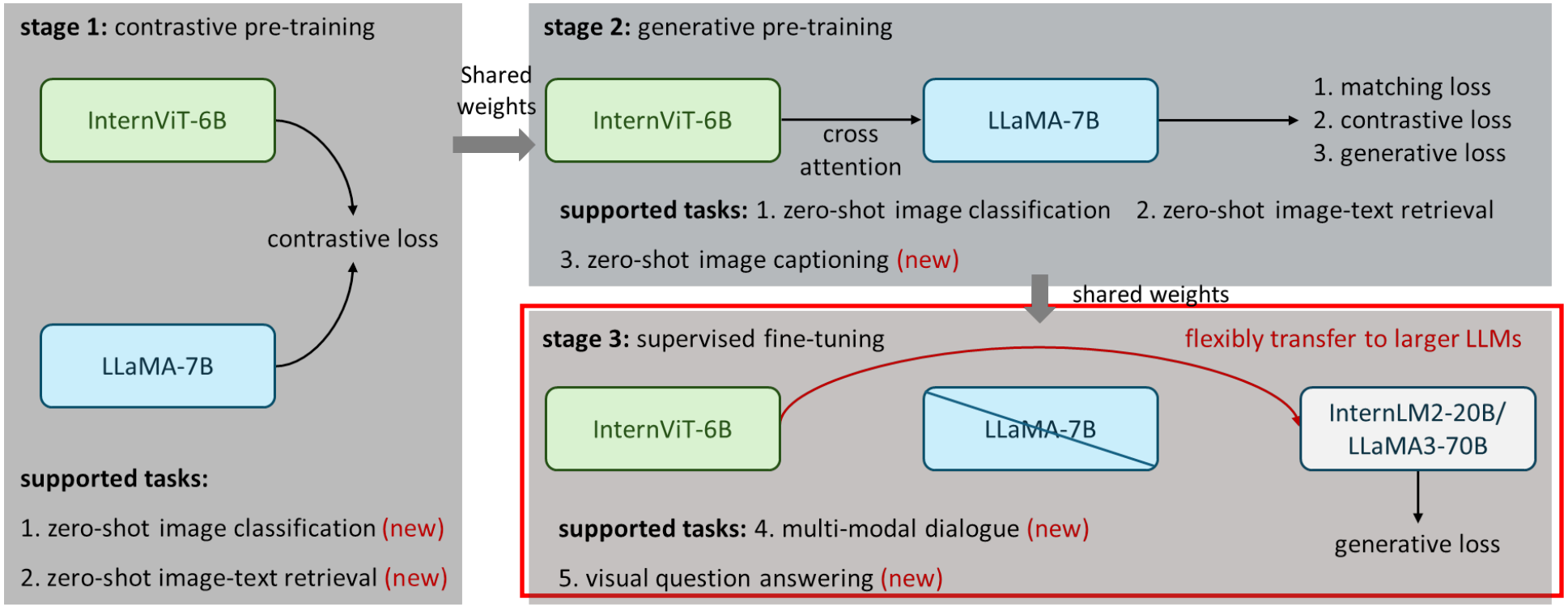
阶段2：利用过滤后的高质量图文数据进行对比学习和生成式联合训练（~1B图像）

dataset	characteristics		stage 1		stage 2	
	language	original	cleaned	remain	cleaned	remain
LAION-en [120]	English	2.3B	1.94B	84.3%	91M	4.0%
LAION-COCO [121]		663M	550M	83.0%	550M	83.0%
COYO [14]		747M	535M	71.6%	200M	26.8%
CC12M [20]		12.4M	11.1M	89.5%	11.1M	89.5%
CC3M [124]		3.0M	2.6M	86.7%	2.6M	86.7%
SBU [112]		1.0M	1.0M	100%	1.0M	100%
Wukong [55]	Chinese	100M	69.4M	69.4%	69.4M	69.4%
LAION-multi [120]	Multi	2.2B	1.87B	85.0%	100M	4.5%
Total	Multi	6.03B	4.98B	82.6%	1.03B	17.0%

筛选指标：CLIP相似度, 水印概率, unsafe概率, 美学指标, 图片分辨率, caption长度等

InternVL：大规模视觉语言模型对齐

核心思想：扩大视觉基础模型并为通用视觉语言任务进行对齐



设计2：渐进式的图像-文本对齐策略

阶段1：利用海量带噪声的图文数据进行对比学习预训练（~5B图像）

阶段2：利用过滤后的高质量图文数据进行对比学习和生成式联合训练（~1B图像）

阶段3：利用高质量Caption/VQA/多轮对话数据进行SFT训练（~4M图像）

https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

多模态对话数据收集

包含图像描述、物体检测、OCR、科学、图表、数学、常识、文档、多轮对话、文本对话...

task	ratio	dataset
Captioning	53.9%	Laion-EN (en) [93], Laion-ZH (zh) [93], COYO (zh) [10], GRIT (zh) [90], COCO (en) [17], TextCaps (en) [99]
Detection	5.2%	Objects365 (en&zh) [97], GRIT (en&zh) [90], All-Seeing (en&zh) [119]
OCR (large)	32.0%	Wukong-OCR (zh) [29], LaionCOCO-OCR (en) [94], Common Crawl PDF (en&zh)
OCR (small)	8.9%	MMC-Inst (en) [61], LSVT (zh) [105], ST-VQA (en) [9] RCTW-17 (zh) [98], ReCTs (zh) [137], ArT (en&zh) [19], SynthDoG (en&zh) [41], COCO-Text (en) [114], ChartQA (en) [81], CTW (zh) [134], DocVQA (en) [82], TextOCR (en) [101], PlotQA (en) [85], InfoVQA (en) [83]

(a) Datasets used in the pre-training stage.

task	dataset
Captioning	TextCaps (en) [99], ShareGPT4V (en&zh) [16]
General QA	VQAv2 (en) [28], GQA (en) [34], OKVQA (en) [80], VSR (en) [59], VisualDialog (en) [22]
Science	AI2D (en) [39], ScienceQA (en) [73], TQA (en) [40]
Chart	ChartQA (en) [81], MMC-Inst (en) [61], DVQA (en) [38], PlotQA (en) [85], LRV-Instruction (en) [60]
Mathematics	GeoQA+ (en) [12], TabMWP (en) [74], MathQA (en) [132], CLEVR-Math/Super (en) [54, 58], Geometry3K (en) [72]
Knowledge	KVQA (en) [96], A-OKVQA (en) [95], ViQuAE (en) [45], Wikipedia (en&zh) [31]
OCR	OCRVQA (en) [86], InfoVQA (en) [83], TextVQA (en) [100], ArT (en&zh) [19], COCO-Text (en) [114], CTW (zh) [134], LSVT (zh) [105], RCTW-17 (zh) [98], ReCTs (zh) [137], SynthDoG (en&zh) [41], ST-VQA (en) [9]
Document	DocVQA (en) [20], Common Crawl PDF (en&zh)
Grounding	RefCOCO/+/g (en) [79, 131], Visual Genome (en) [42]
Conversation	LLaVA-150K (en&zh) [63], LVIS-Instruct4V (en) [115], ALLaVA (en&zh) [14], Laion-GPT4V (en) [44], TextOCR-GPT4V (en) [37], SVIT (en&zh) [140]
Text-only	OpenHermes2.5 (en) [109], Alpaca-GPT4 (en) [106], ShareGPT (en&zh) [141], COIG-CQIA (zh) [6]

(b) Datasets used in the fine-tuning stage.

https://github.com/OpenGVLab/InternVL

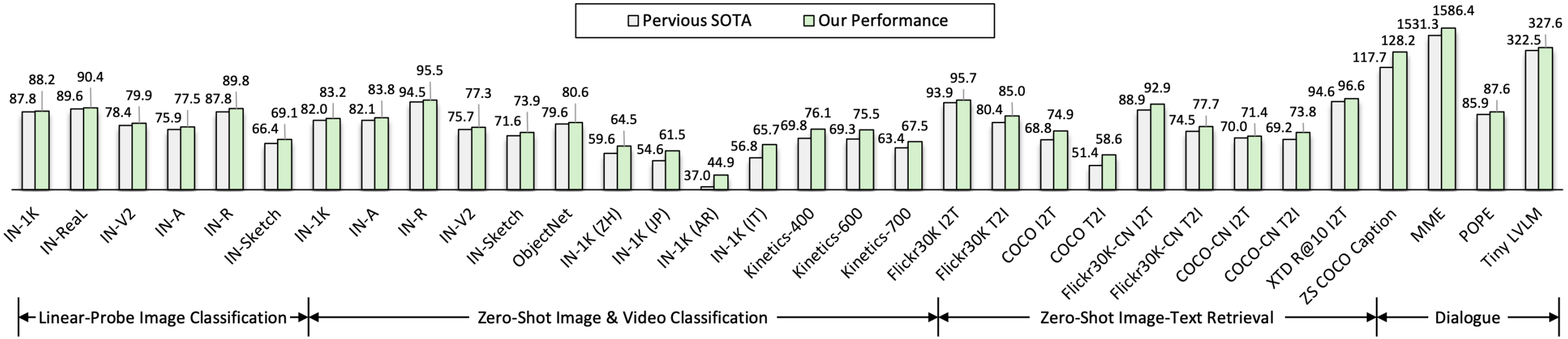
InternVL：大规模视觉语言模型对齐

在多种通用视觉语言任务上的取得了最好的性能，包括：

1）视觉任务：图像/视频分类，语义分割；

2）视觉-语言任务：图像/视频-文本检索，零样本图像分类；

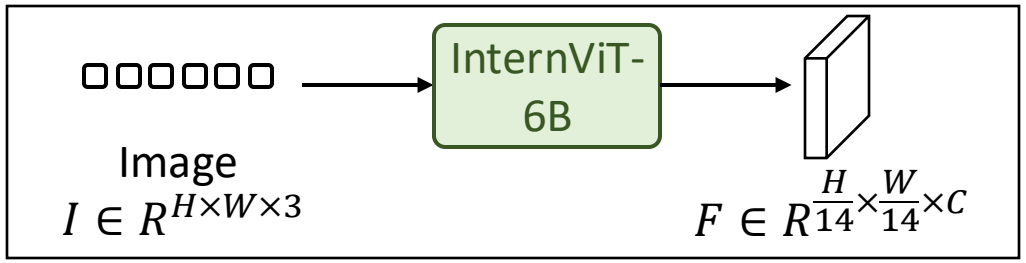
3）通用视觉问答：图像描述，视觉问答，多轮对话



https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

对于视觉任务， InternVL的视觉编码器，即InternViT-6B，可以直接用作视觉主干网络



method	#param	IN-1K	IN-ReaL	IN-V2	IN-A	IN-R	IN-Ske	avg.
OpenCLIP-H [67]	0.6B	84.4	88.4	75.5	–	–	–	–
OpenCLIP-G [67]	1.8B	86.2	89.4	77.2	63.8	87.8	66.4	78.5
DINOv2-g [111]	1.1B	86.5	89.6	78.4	75.9	78.8	62.5	78.6
EVA-01-CLIP-g [46]	1.1B	86.5	89.3	77.4	70.5	87.7	63.1	79.1
MAWS-ViT-6.5B [128]	6.5B	87.8	–	–	–	–	–	–
ViT-22B* [37]	21.7B	89.5	90.9	83.2	83.8	87.4	–	–
InternViT-6B (ours)	5.9B	**88.2**	**90.4**	**79.9**	**77.5**	**89.8**	**69.1**	**82.5**

Table 4. **Linear evaluation on image classification.** We report the top-1 accuracy on ImageNet-1K [38] and its variants [10, 60, 61, 119, 141]. *ViT-22B [37] uses the private JFT-3B dataset [173].

Image-Level Tasks

method	#param	crop size	1/16	1/8	1/4	1/2	1
ViT-L [137]	0.3B	504^2	36.1	41.3	45.6	48.4	51.9
ViT-G [173]	1.8B	504^2	42.4	47.0	50.2	52.4	55.6
ViT-22B [37]	21.7B	504^2	44.7	47.2	50.6	52.5	54.9
InternViT-6B (ours)	5.9B	504^2	**46.5**	**50.0**	**53.3**	**55.8**	**57.2**

(a) Few-shot semantic segmentation with limited training data. Following ViT-22B [37], we fine-tune the InternViT-6B with a linear classifier.

method	decoder	#param (train/total)	crop size	mIoU
OpenCLIP-G$_{frozen}$ [67]	Linear	0.3M / 1.8B	512^2	39.3
ViT-22B$_{frozen}$ [37]	Linear	0.9M / 21.7B	504^2	34.6
InternViT-6B$_{frozen}$ (ours)	Linear	0.5M / 5.9B	504^2	**47.2**
ViT-22B$_{frozen}$ [37]	UperNet	0.8B / 22.5B	504^2	52.7
InternViT-6B$_{frozen}$ (ours)	UperNet	0.4B / 6.3B	504^2	**54.9**
ViT-22B [37]	UperNet	22.5B / 22.5B	504^2	55.3
InternViT-6B (ours)	UperNet	6.3B / 6.3B	504^2	**58.9**

(b) Semantic segmentation performance in three different settings, from top to bottom: linear probing, head tuning, and full-parameter tuning.

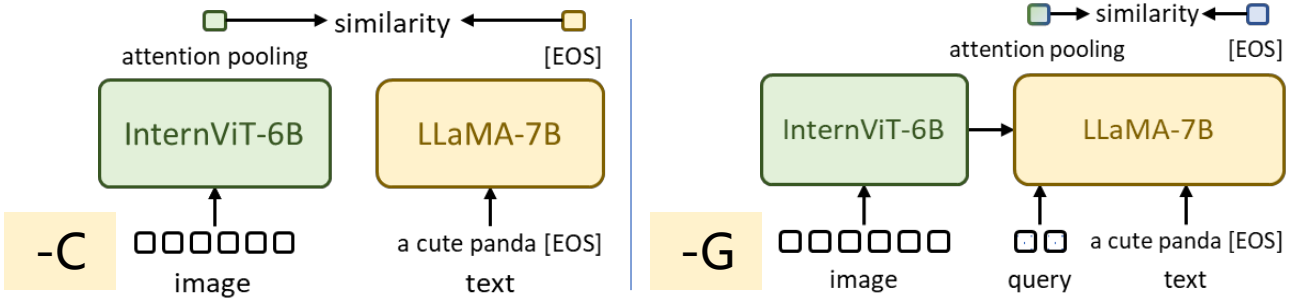
Pixel-Level Tasks

仅用不到不到三分之一参数量，实现了与 ViT-22B 相当的性能

https://github.com/OpenGVLab/InternVL

InternVL：大规模视觉语言模型对齐

对于视觉语言任务，有两种变体：
InternVL-C and InternVL-G



method	multi-lingual	Flickr30K (English, 1K test set) [88] Image → Text R@1	R@5	R@10	Text → Image R@1	R@5	R@10	COCO (English, 5K test set) [20] Image → Text R@1	R@5	R@10	Text → Image R@1	R@5	R@10	avg.
Florence [133]	×	90.9	99.1	–	76.7	93.6	–	64.7	85.9	–	47.2	71.4	–	–
ONE-PEACE [110]	×	90.9	98.8	99.8	77.2	93.5	96.2	64.7	86.0	91.9	48.0	71.5	79.6	83.2
OpenCLIP-g [51]	×	91.4	99.2	99.6	77.7	94.1	96.9	66.4	86.0	91.8	48.8	73.3	81.5	83.9
EVA-01-CLIP-g+ [99]	×	91.6	99.3	99.8	78.9	94.5	96.9	68.2	87.5	92.5	50.3	74.0	82.1	84.6
CoCa [131]	×	92.5	99.5	99.9	80.4	95.7	97.7	66.3	86.2	91.8	51.2	74.2	82.0	84.8
OpenCLIP-G [51]	×	92.9	99.3	99.8	79.5	95.0	97.1	67.3	86.9	92.6	51.4	74.9	83.0	85.0
EVA-02-CLIP-E+ [99]	×	93.9	99.4	99.8	78.8	94.2	96.8	68.8	87.8	92.8	51.1	75.0	82.7	85.1
BLIP-2† [61]	×	97.6	100.0	100.0	89.7	98.1	98.9	–	–	–	–	–	–	–
InternVL-C (ours)	✓	94.7	99.6	99.9	81.7	96.0	98.2	70.6	89.0	93.5	54.1	77.3	84.6	86.6
InternVL-G (ours)	✓	**95.7**	**99.7**	**99.9**	**85.0**	**97.0**	**98.6**	**74.9**	**91.3**	**95.2**	**58.6**	**81.3**	**88.0**	**88.8**

method		Flickr30K-CN (Chinese, 1K test set) [58]						COCO-CN (Chinese, 1K test set) [63]						avg.
WuKong-ViT-L [41]	×	76.1	94.8	97.5	51.7	78.9	86.3	55.2	81.0	90.6	53.4	80.2	90.1	78.0
R2D2-ViT-L [123]	×	77.6	96.7	98.9	60.9	86.8	92.7	63.3	89.3	95.7	56.4	85.0	93.1	83.0
Taiyi-CLIP-ViT-H [137]	×	–	–	–	–	–	–	–	–	–	60.0	84.0	93.3	–
AltCLIP-ViT-H [24]	✓	88.9	98.5	99.5	74.5	92.0	95.5	–	–	–	–	–	–	–
CN-CLIP-ViT-H [126]	×	81.6	97.5	98.8	71.2	91.4	95.5	63.0	86.6	92.9	69.2	89.9	96.1	86.1
OpenCLIP-XLM-R-H [51]	✓	86.1	97.5	99.2	71.0	90.5	94.9	70.0	91.5	97.0	66.1	90.8	96.0	87.6
InternVL-C (ours)	✓	90.3	98.8	99.7	75.1	92.9	96.4	68.8	92.0	96.7	68.9	91.9	96.5	89.0
InternVL-G (ours)	✓	**92.9**	**99.4**	**99.8**	**77.7**	**94.8**	**97.3**	**71.4**	**93.9**	**97.7**	**73.8**	**94.4**	**98.1**	**90.9**

检索性能优于CLIP、OpenCLIP、CoCa等模型

InternVL：大规模视觉语言模型对齐

零样本图像分类能力评测

method	IN-1K	IN-A	IN-R	IN-V2	IN-Sketch	ObjectNet	Δ↓	avg.
OpenCLIP-H [67]	78.0	59.3	89.3	70.9	66.6	69.7	5.7	72.3
OpenCLIP-g [67]	78.5	60.8	90.2	71.7	67.5	69.2	5.5	73.0
OpenAI CLIP-L+ [117]	76.6	77.5	89.0	70.9	61.0	72.0	2.1	74.5
EVA-01-CLIP-g [130]	78.5	73.6	92.5	71.5	67.3	72.3	2.5	76.0
OpenCLIP-G [67]	80.1	69.3	92.1	73.6	68.9	73.0	3.9	76.2
EVA-01-CLIP-g+ [130]	79.3	74.1	92.5	72.1	68.1	75.3	2.4	76.9
MAWS-ViT-2B [128]	81.9	–	–	–	–	–	–	–
EVA-02-CLIP-E+ [130]	82.0	82.1	94.5	75.7	71.6	79.6	1.1	80.9
CoCa* [169]	86.3	90.2	96.5	80.7	77.6	82.7	0.6	85.7
LiT-22B* [37, 174]	85.9	90.1	96.0	80.9	–	87.6	–	–
InternVL-C (ours)	**83.2**	**83.8**	**95.5**	**77.3**	**73.9**	**80.6**	**0.8**	**82.4**

(a) ImageNet variants [38, 60, 61, 119, 141] and ObjectNet [8].

method	EN	ZH	JP	AR	IT	avg.
M-CLIP [16]	–	–	–	–	20.2	–
CLIP-Italian [11]	–	–	–	–	22.1	–
Japanese-CLIP-ViT-B [102]	–	–	54.6	–	–	–
Taiyi-CLIP-ViT-H [176]	–	54.4	–	–	–	–
WuKong-ViT-L-G [55]	–	57.5	–	–	–	–
CN-CLIP-ViT-H [162]	–	59.6	–	–	–	–
AltCLIP-ViT-L [26]	74.5	59.6	–	–	–	–
EVA-02-CLIP-E+ [130]	82.0	3.6	5.0	0.2	41.2	–
OpenCLIP-XLM-R-B [67]	62.3	42.7	37.9	26.5	43.7	42.6
OpenCLIP-XLM-R-H [67]	77.0	55.7	53.1	37.0	56.8	55.9
InternVL-C (ours)	**83.2**	**64.5**	**61.5**	**44.9**	**65.7**	**64.0**

(b) Multilingual ImageNet-1K [38, 76].

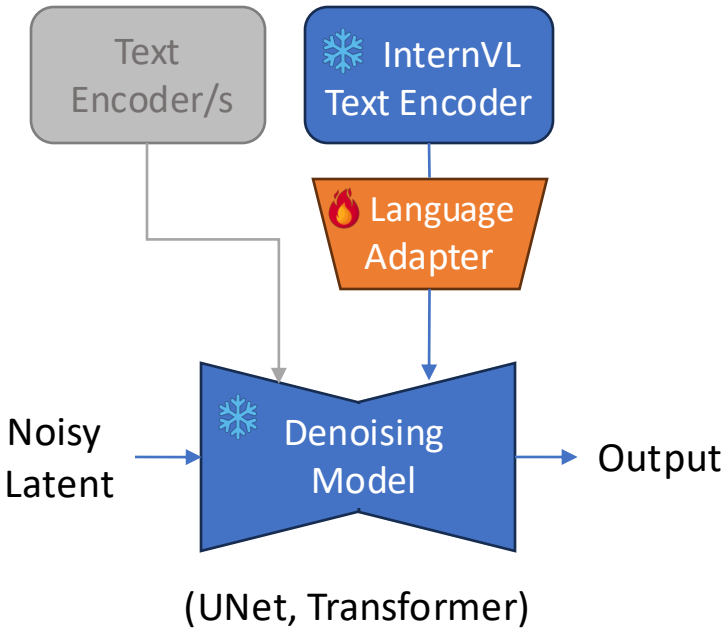
零样本视频分类能力评测

method	#F	K400 [15]		K600 [16]		K700 [17]	
		top-1	avg.	top-1	avg.	top-1	avg.
OpenCLIP-g [51]	1	–	63.9	–	64.1	–	56.9
OpenCLIP-G [51]	1	–	65.9	–	66.1	–	59.2
EVA-01-CLIP-g+ [99]	1	–	66.7	–	67.0	–	60.9
EVA-02-CLIP-E+ [99]	1	–	69.8	–	69.3	–	63.4
InternVL-C (ours)	1	–	**71.0**	–	**71.3**	–	**65.7**
ViCLIP [117]	8	64.8	75.7	62.2	73.5	54.3	66.4
InternVL-C (ours)	8	**69.1**	**79.4**	**68.9**	**78.8**	**60.6**	**71.5**

强零样本图像、视频分类能力

https://github.com/OpenGVLab/InternVL

InternVL + Language Adapter -> Zeroshot 多语言内容生成



(1) Overall Architecture

```
# pip install mulankit
from diffusers import StableDiffusionPipeline
+ import mulankit

pipe = StableDiffusionPipeline.from_pretrained('Lykon/dreamshaper-8')
+ pipe = mulankit.transform(pipe)
image = pipe('一只蓝色的🐶 in the 바다').images[0]
```

- 即插即用的为现有扩散模型增加多语言能力
- 只需要英文数据训练，即可泛化到其他语言
- 支持多种语言的混合输入，甚至是 emoji
- 无需额外训练，即可兼容社区模型，如 ControlNet，LCM，LoRA 等

https://github.com/mulanai/MuLan

https://github.com/OpenGVLab/InternVL

InternVL + Language Adapter -> Zeroshot 多语言内容生成

只需要英文数据，即可支持超多语言

https://github.com/mulanai/MuLan


繁体中文


希腊语


Emoji


简体中文


阿塞拜疆语


中英混合


斯洛伐克语


阿尔巴尼亚语


土耳其语


匈牙利语


波斯语


日语


德语


印尼语


加泰罗尼亚语


越南语


乌克兰语


荷兰语


英文


阿拉伯语


韩语


法语


捷克语

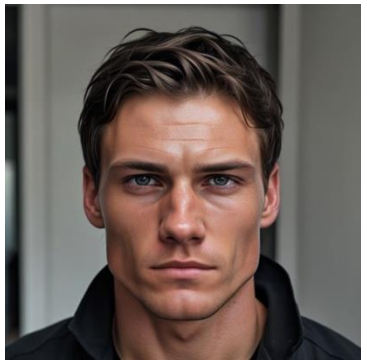

俄语

https://github.com/OpenGVLab/InternVL

InternVL + Language Adapter -> Zeroshot 多语言内容生成

即插即用，无需对Diffusion Model做额外训练


Dreamshaper

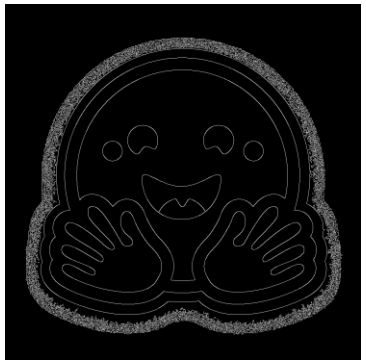

Realistic Vision


Cartoonmix


3D Animation


LoRA (Lego)


MVDream


ControlNet


LCM


SDXL Turbo


SDXL Lightning


AnimateDiff

https://github.com/mulanai/MuLan

https://github.com/OpenGVLab/InternVL

目录

InternVL 1.5：接近商用性能的开源多模态大模型

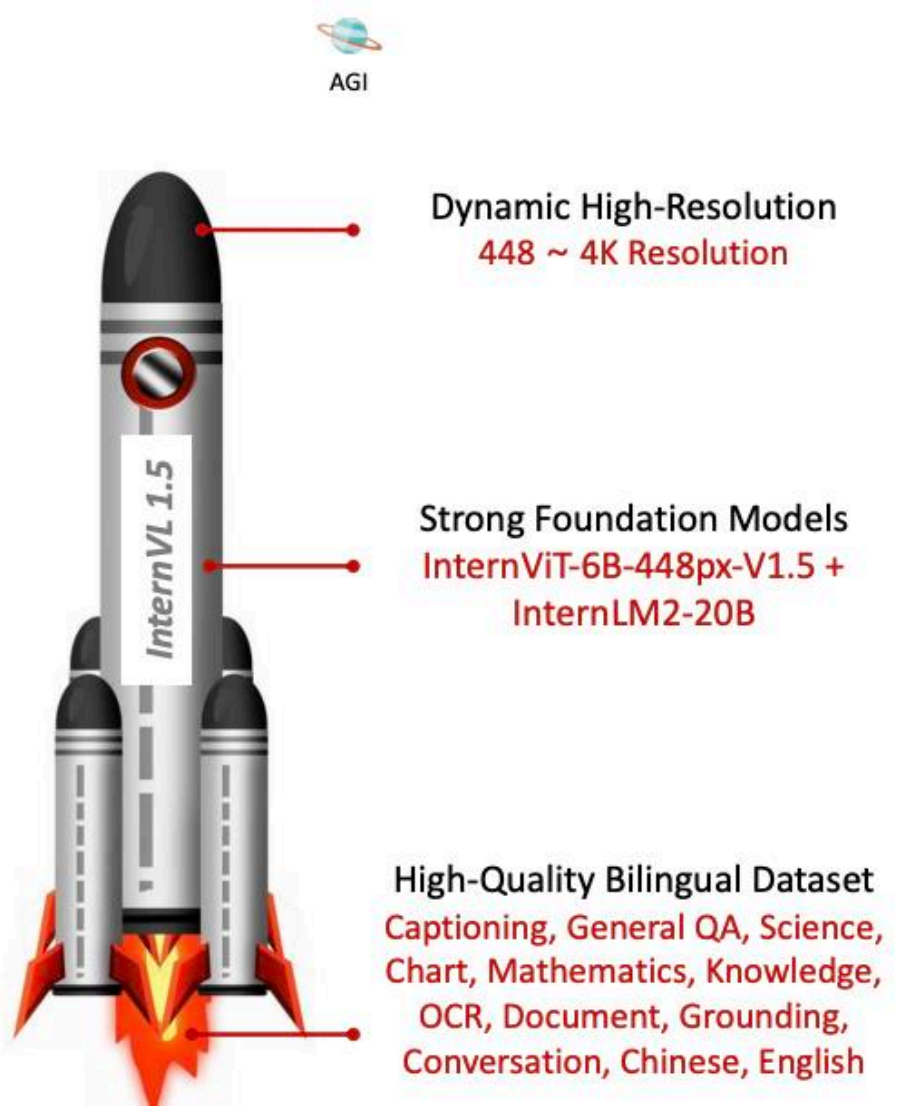
增强图文多模态对话能力

3个关键点

主体（强基础模型）：更大的视觉模型可以包含更广的视觉domain，抽取更强的视觉表征，更强的语言模型有更强的语言能力、世界知识和推理能力

动态分辨率（火箭头）：模型需要根据任务调整不同的分辨率。对于一些图像细节的理解任务，如：文档理解，高分辨率非常重要。但是对于一些常见的问答任务又不需要大分辨率。

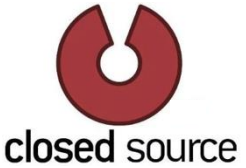
燃料（高质量数据集）：多语言、多来源、精细标注



https://github.com/OpenGVLab/InternVL

InternVL 1.5：接近GPT-4V的开源多模态对话模型

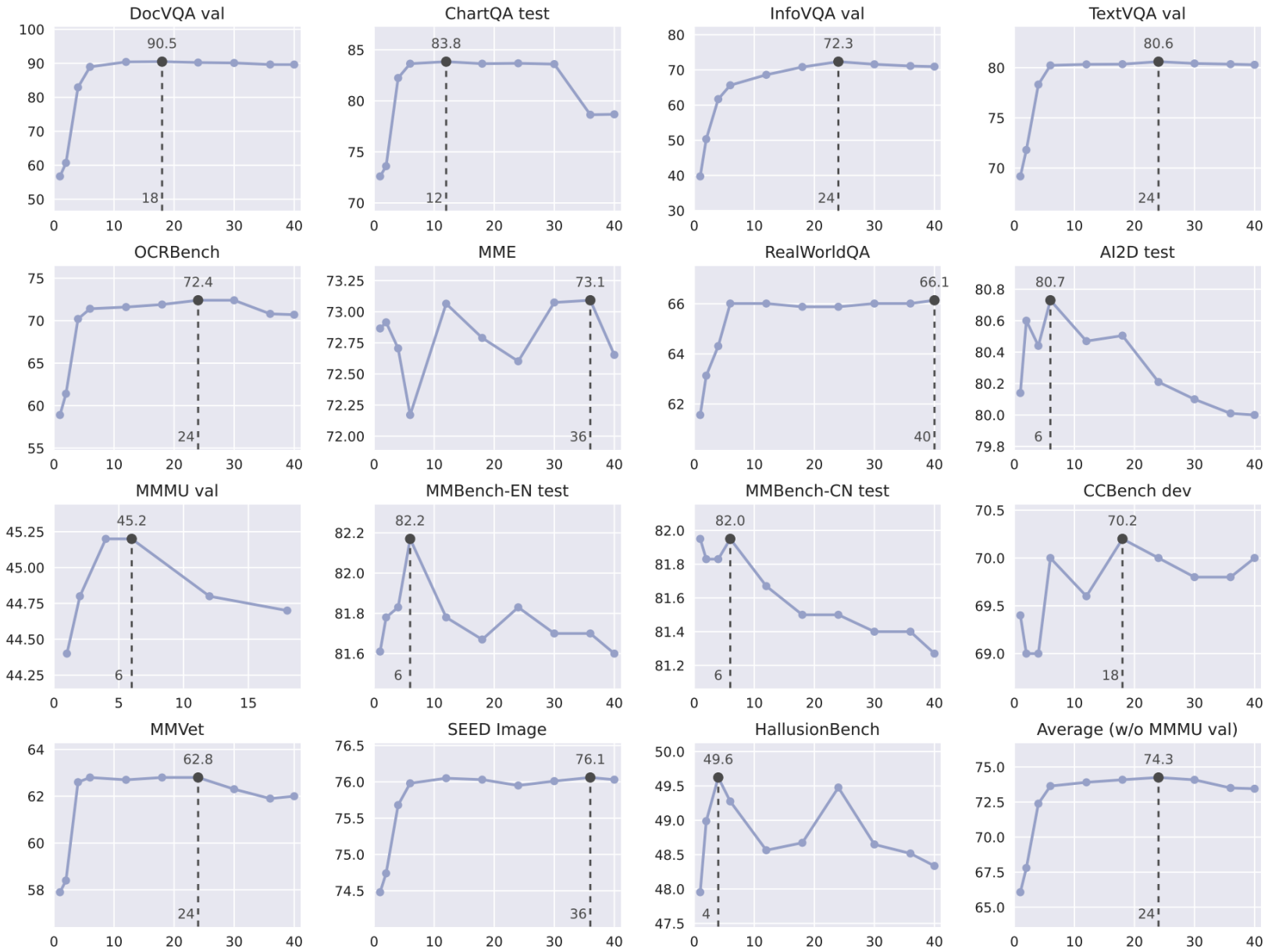
和头部商用模型对比



Benchmark	InternVL 1.5	Grok-1.5V	GPT-4V	Claude-3 Opus	Gemini Pro 1.5
MMMU Multi-discipline	45.2%	53.6%	56.8%	**59.4%**	58.5%
MathVista Math	**53.5%**	52.8%	49.9%	50.5%	52.1%
AI2D Diagrams	80.7%	88.3%	78.2%	**88.1%**	80.3%
TextVQA Text reading	**80.6%**	78.1%	78.0%	-	73.5%
ChartQA Charts	**83.8%**	76.1%	78.5%	80.8%	81.3%
DocVQA Documents	**90.9%**	85.6%	88.4%	89.3%	86.5%
RealWorldQA Real-world understanding	66.0%	**68.7%**	61.4%	49.8%	67.5%

https://github.com/OpenGVLab/InternVL

InternVL 1.5：接近GPT-4V的开源多模态对话模型



← 分辨率对性能的影响

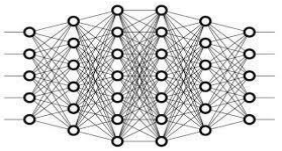
https://github.com/OpenGVLab/InternVL

书生·万象 InternVL 2.0：全方面提升

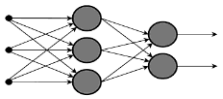
渐进式对齐训练，通过模型"从小到大"、数据"从粗到精"的渐进式的训练策略，以较低的成本完成了大模型的训练，在有限资源下展现出卓越的性能表现

海量带噪数据

筛选数据



千亿参数模型

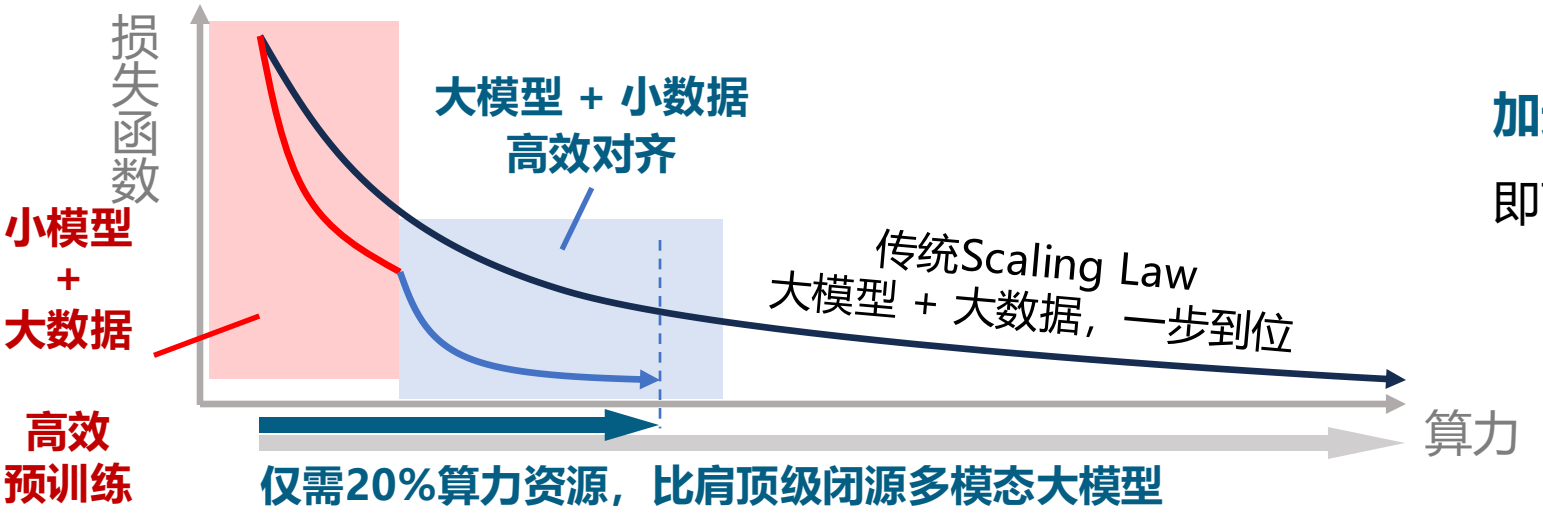


百亿参数模型

扩大模型

高质量精选数据

在MMMU，MMBench等评测上**比肩GPT-4o和Gemini Pro 1.5**



加速Scaling Law曲线，仅需原有的**1/5的算力**即可取得同等的效果

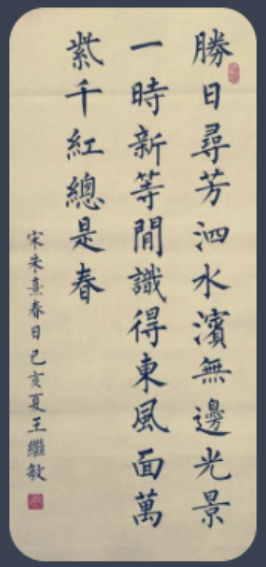
https://github.com/OpenGVLab/InternVL

书生·万象 InternVL 2.0：全方面提升

和头部商用模型对比

Benchmark	Mini-InternVL 4B	InternVL 1.5 26B	书生·万象 InternVL2-8B	书生·万象 InternVL2 Pro	Other Open-Source MLLMs	GPT-4o 20240513	Claude-3 Opus	Gemini Pro 1.5
MMMU 多学科问答	43.3%	45.2%	49.3 (4.1↑)	62.0% (16.8↑)	53.4%	**69.1%**	59.4%	58.5%
MathVista 数学	53.7%	53.5%	58.3 (4.8↑)	**66.3%** (12.8↑)	59.5%	63.8%	50.5%	52.1%
AI2D 科学图表	76.9%	80.7%	83.8 (3.1↑)	87.3% (6.6↑)	81.2%	**94.2%**	88.1%	80.3%
ChartQA 通用图表	81.0%	83.8%	84.9 (1.1↑)	**87.1%** (3.3↑)	81.4%	85.7%	80.8%	81.3%
DocVQA 文档	87.7%	90.9%	92.9 (2.0↑)	**94.3%** (3.4↑)	85.6%	92.8%	89.3%	86.5%
InfographicVQA 信息图表	64.9%	72.4%	74.8 (2.4↑)	**82.0%** (9.6↑)	-	-	-	72.7%
OCRBench OCR	638	724	794 (70↑)	**837** (113↑)	776	736	694	754
MMBench v1.1 通用视觉问答	69.7%	79.7%	81.5 (1.8↑)	**86.4%** (6.7↑)	77.8%	82.2%	59.1%	73.9%

更强的OCR能力：毛笔字+竖排+繁体



这是啥

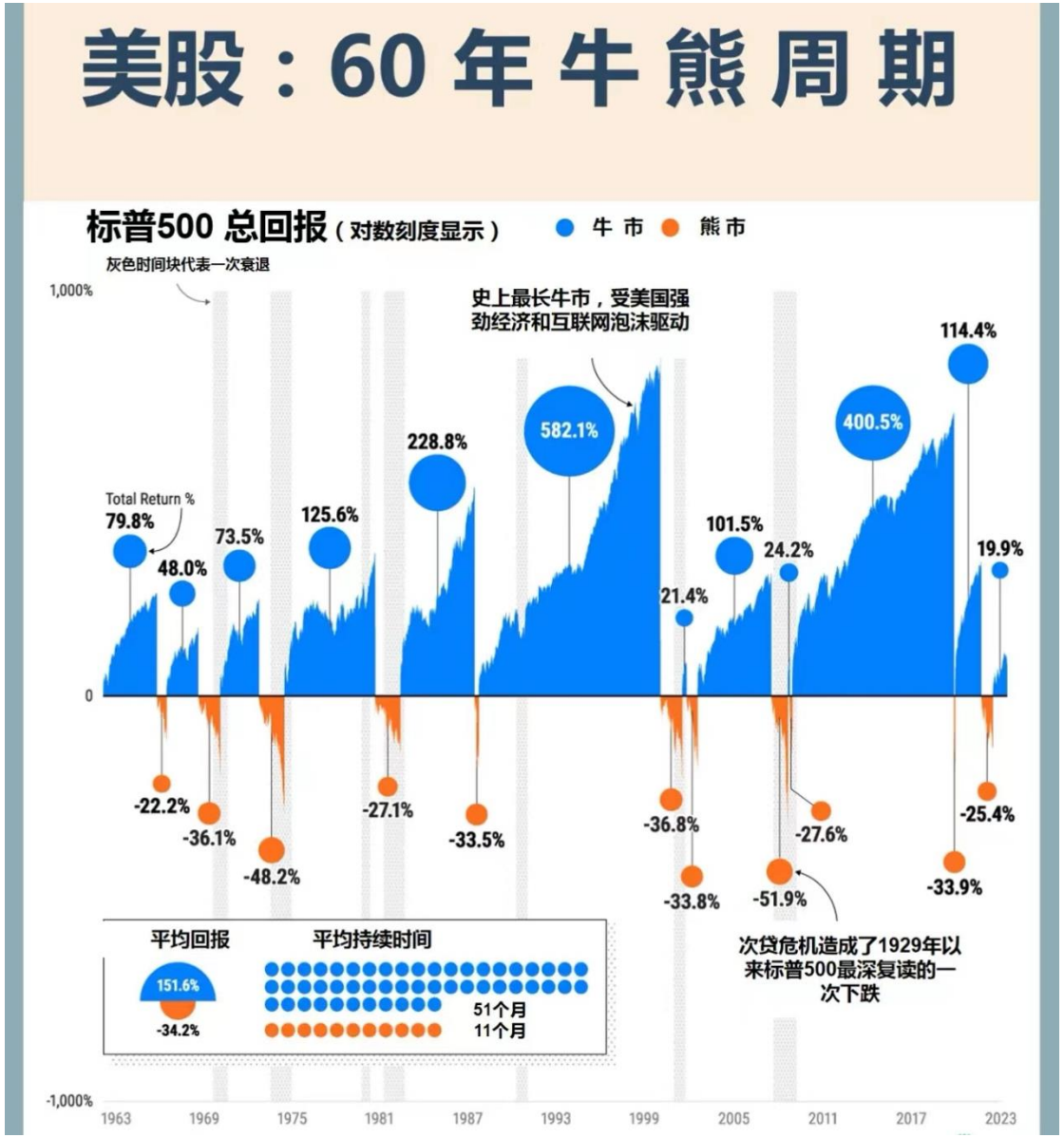
这是一幅书法作品，书写的是宋朝朱熹的诗句。诗句内容如下：

胜日寻芳泗水滨，无边光景一时新。等闲识得东风面，万紫千红总是春。

这幅书法作品展示了中国传统书法艺术的精髓，字迹流畅，布局合理，给人一种美的享受。

https://github.com/OpenGVLab/InternVL

更强的图表理解能力



这张图描述了什么

这张图展示了美股60年牛熊周期的标普500总回报情况。图中用蓝色和橙色分别表示牛市和熊市，并用对数刻度显示总回报百分比。每个牛市和熊市周期都有对应的总回报百分比标注。

图中关键信息包括：

- 牛市的平均回报为151.6%。
- 熊市的平均回报为-34.2%。
- 牛市的平均持续时间为51个月。
- 熊市的平均持续时间为11个月。

此外，图中还特别标注了历史上最长的牛市，该牛市受到美国强劲经济和互联网泡沫的驱动，总回报达到了582.1%。次贷危机造成了1929年以来标普500最深的跌幅，下跌了51.9%。

https://github.com/OpenGVLab/InternVL

细节文字理解+文字深层含义

请问杯子里是什么?

杯子里装着绿色的液体，看起来像是某种饮料。

https://github.com/OpenGVLab/InternVL

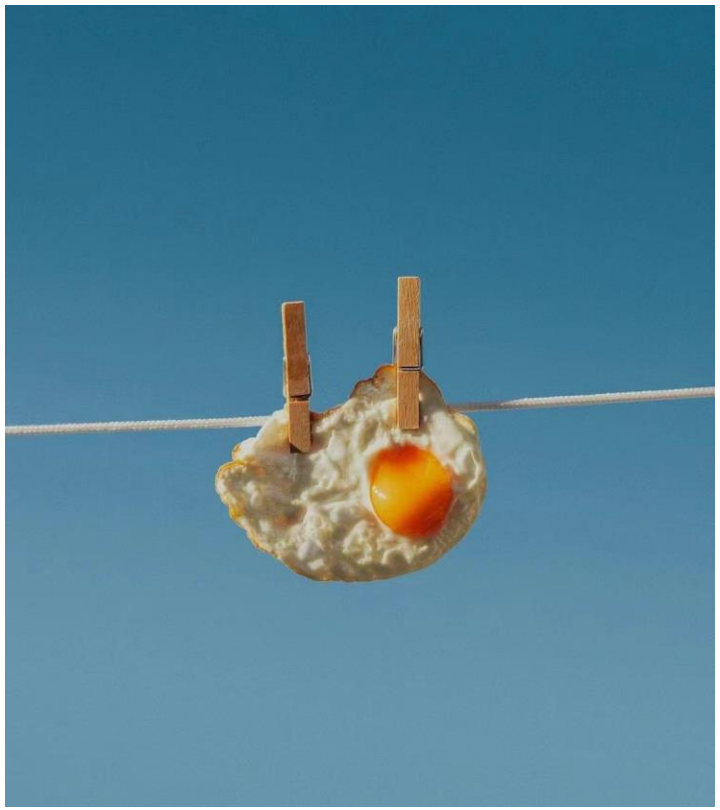
细节文字理解+文字深层含义

瓶子里是什么

瓶子里的是费尔利（Fairy）洗洁精。图片中，瓶子旁边有几杯绿色液体，但这些液体并不是饮料，而是洗洁精。

https://github.com/OpenGVLab/InternVL

结合专业知识问答



这张图表现了什么，可能出现在哪位导演的镜头中?

这张图片展示了一个煎蛋被两个木制夹子夹在一条绳子上，背景是蓝天。这种场景表现了一种幽默和超现实的风格，可能出现在一些以创意和非传统视角著称的导演的镜头中，例如大卫·林奇（David Lynch）或简·坎皮恩（Jane Campion）。这些导演以其独特的视觉风格和对日常物品的不寻常运用而闻名。

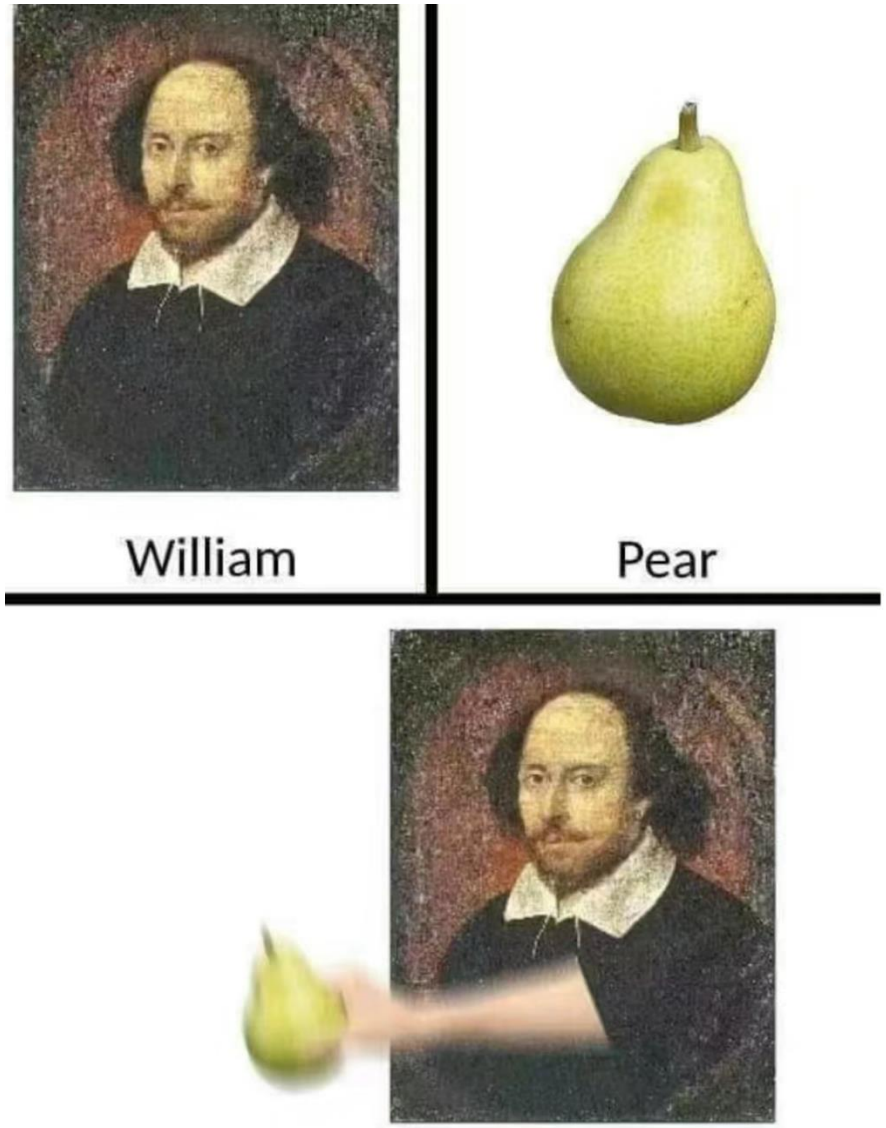
https://github.com/OpenGVLab/InternVL

理解人工布置的巧妙之处

这是什么东西?

这是一些大白菜，上面装饰了辣椒和大蒜，看起来像是人脸的造型。这些装饰可能用于吸引顾客的注意力，使蔬菜更加有趣和吸引人。

图-文-动作结合回答



第三张图是人摇晃一个梨，它和前两张图的对应关系是什么

在第三张图中，一个人摇晃着一个梨，这与前两张图的对应关系如下：

1. 第一张图：威廉·莎士比亚的画像，标注为“William”。
2. 第二张图：一个梨的图片，标注为“Pear”。
3. 第三张图：一个人摇晃着一个梨，这个动作可以理解为“Shake a Pear”。

将这三张图结合起来，我们可以得到“William Shake a Pear”，这是一个文字游戏，巧妙地将“William Shakespeare”这个名字与“Shake a Pear”这个动作结合在一起，制造了幽默效果。

图像细节理解+精确定位

画面最远处是什么?

画面最远处是一个著名的旅游景点——圣米歇尔山（Mont Saint-Michel）。圣米歇尔山位于法国诺曼底和布列塔尼之间的海湾中，是一座小岛，以其宏伟的修道院和独特的地理位置而闻名。

你怎么知道呢

我是通过图像识别技术来确定画面内容的。这张图片中，远处的小岛和修道院的建筑风格与圣米歇尔山非常相似，因此可以判断出这是圣米歇尔山。此外，我也根据图片中的地理特征和标志性建筑进行识别和确认。

开源模型权重

可在3090运行

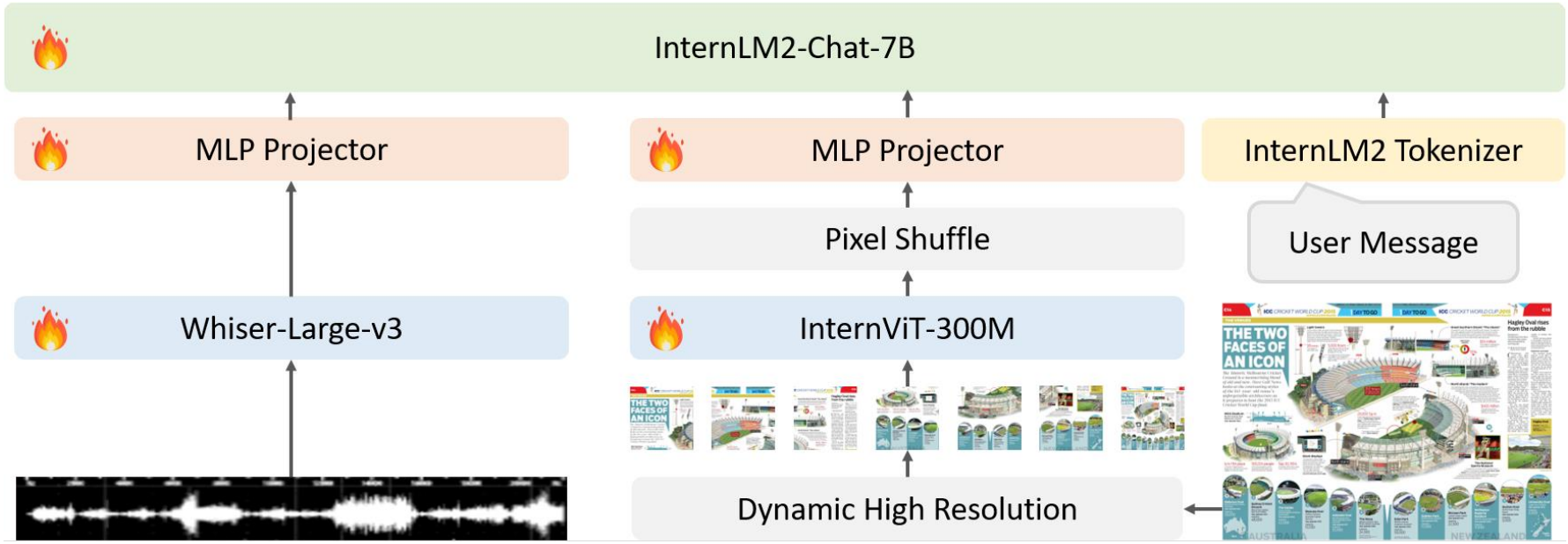
Model Name	Vision Part	Language Part	HF Link	MS Link	Document
InternVL2-1B	InternViT-300M-448px	Qwen2-0.5B-Instruct	link	link	doc
InternVL2-2B	InternViT-300M-448px	internlm2-chat-1-8b	link	link	doc
InternVL2-4B	InternViT-300M-448px	Phi-3-mini-128k-instruct	link	link	doc
InternVL2-8B	InternViT-300M-448px	internlm2_5-7b-chat	link	link	doc
InternVL2-26B	InternViT-6B-448px-V1-5	internlm2-chat-20b	link	link	doc
InternVL2-40B	InternViT-6B-448px-V1-5	Nous-Hermes-2-Yi-34B	link	link	doc
InternVL2-Llama3-76B	InternViT-6B-448px-V1-5	Hermes-2-Theta-Llama-3-70B	link	link	doc

InternVL2-Pro API

We welcome everyone to use our API for research. For better management, please submit (application form) / (申请表) to obtain free API access.

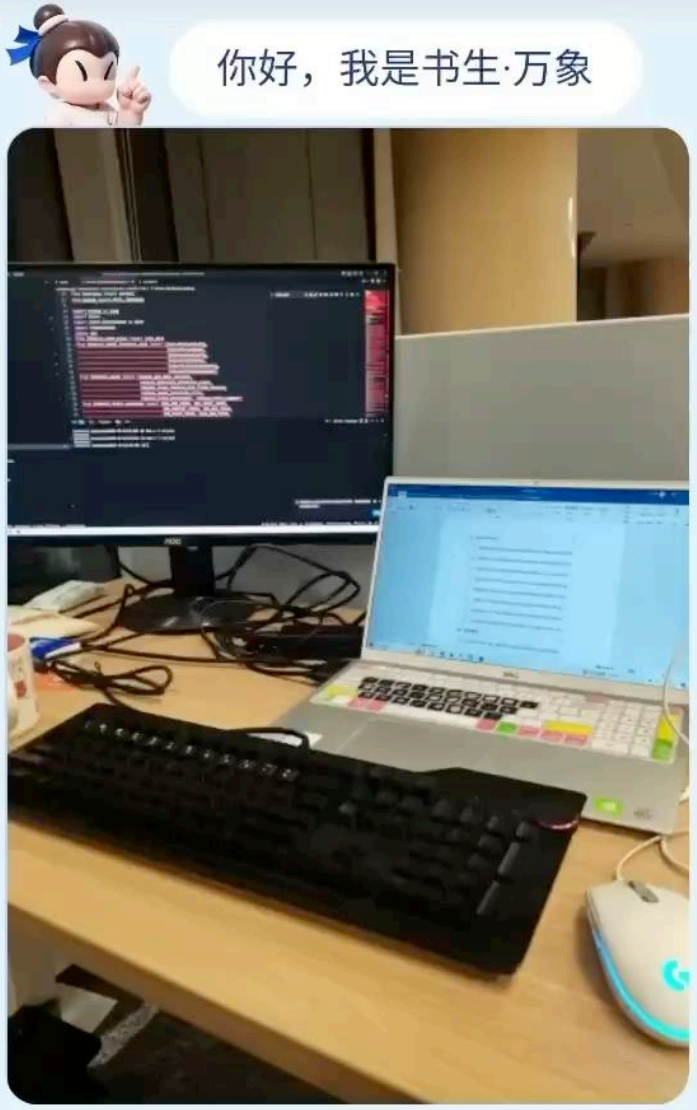
https://github.com/OpenGVLab/InternVL

InternOmni: Extending InternVL with Audio Modality



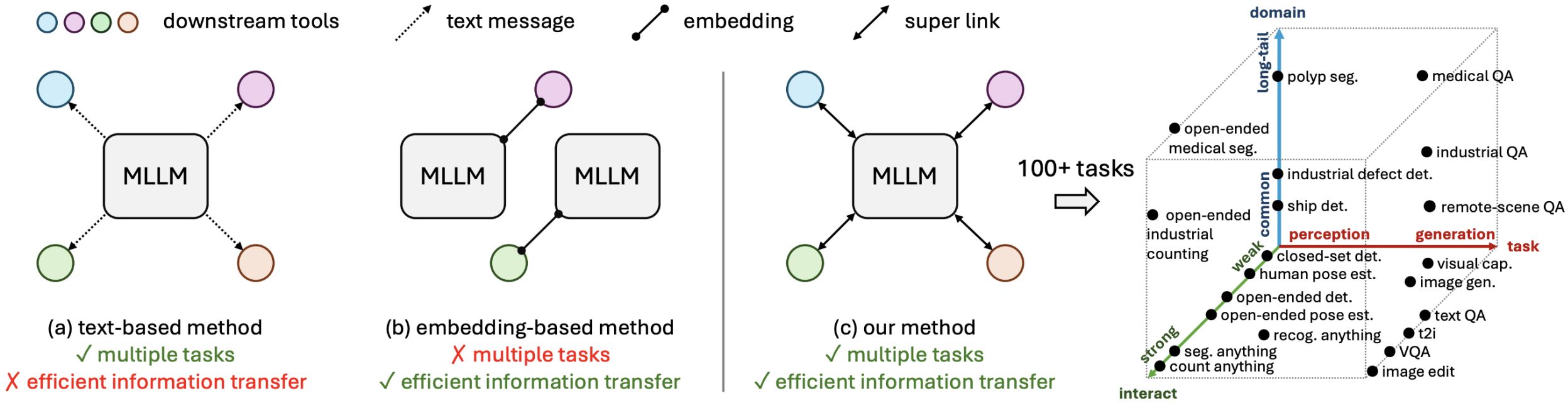
更多详情看blog

https://github.com/OpenGVLab/InternVL



目 录

不止于语言输出：通专融合



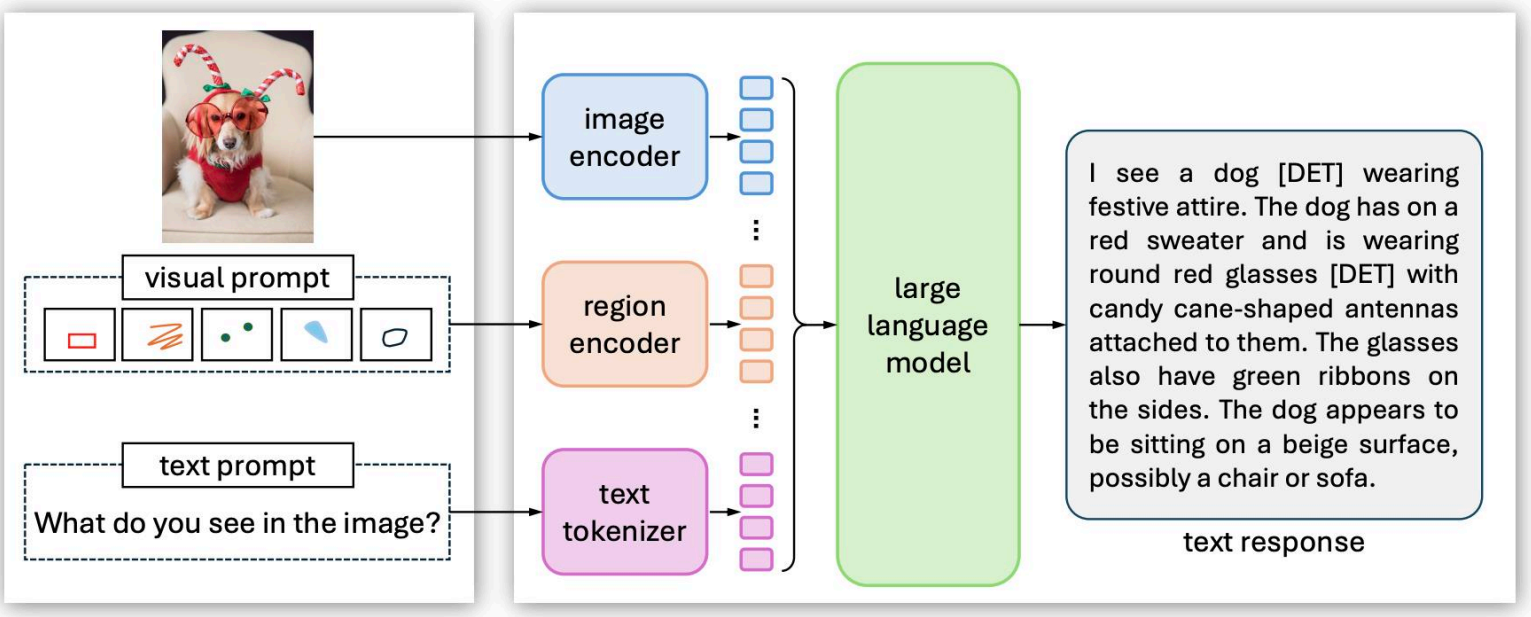
(a) text-based method
✓ multiple tasks
✗ efficient information transfer

(b) embedding-based method
✗ multiple tasks
✓ efficient information transfer

(c) our method
✓ multiple tasks
✓ efficient information transfer

https://github.com/OpenGVLab/InternVL

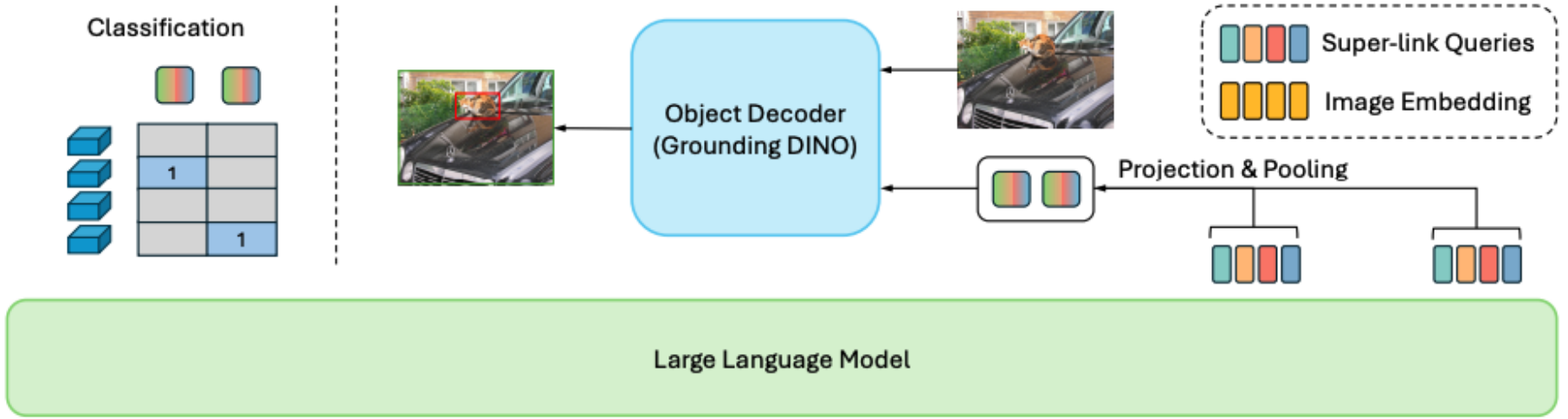
不止于语言输出：通专融合



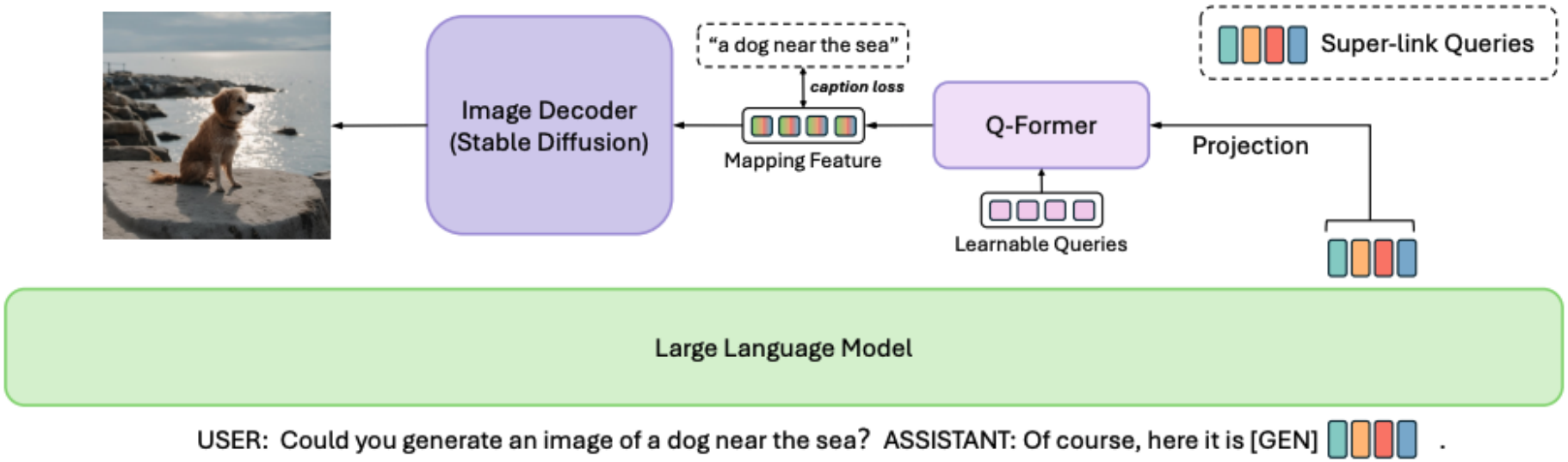
Example: Text Prompt + Visual Prompt for Interactive Segmentation.
USER: <image> Could you please segment all the corresponding objects according to the visual prompts as region1 <region>, region2 <region>?
ASSISTANT: Sure, these objects are region1 [SEG], region2 [SEG].

https://github.com/OpenGVLab/InternVL

不止于语言输出：通专融合



(a) Connecting with object decoder for visual perception.



https://github.com/OpenGVLab/InternVL

不止于语言输出：通专融合

method	query/token number	inst seg. AP_b	inst seg. AP_m	ground. P@.5	pose AP	interact seg. mIoU	interact seg. cIoU
super-link queries	1	50.4	39.6	85.8	43.0	43.2	60.0
	4	52.0	41.0	85.7	71.0	44.8	60.4
	8	52.1	40.7	86.4	71.6	45.9	61.9

Table 5: **Ablation on the super-link queries number.** We evaluate the results on the four crucial visual perception tasks: instance segmentation (COCO), visual grounding (RefCOCO), pose estimation (COCO), and interactive segmentation (COCO using scribble). Our default setting is marked in gray.

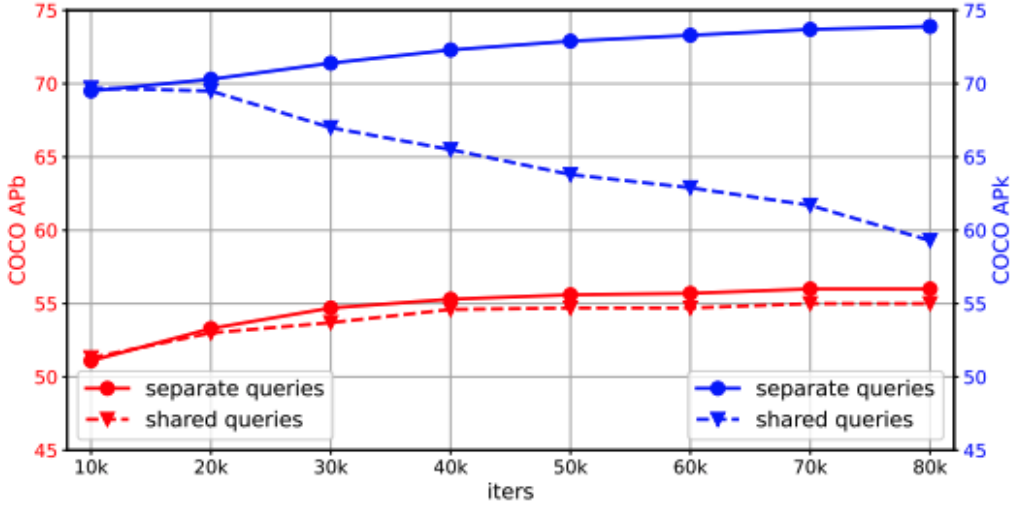


Figure 4: **Shared *vs.* unshared super-link queries for different decoders.** We report the box/keypoint AP on COCO.

1）query不同任务不能共享；2）感知任务8个query就够了；3）图像生成要64个query

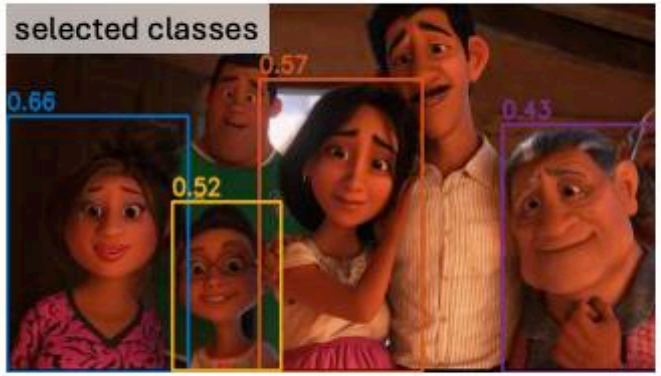
https://github.com/OpenGVLab/InternVL

不止于语言输出：通专融合

Please conduct object detection to any [List of COCO classes] that may be present.

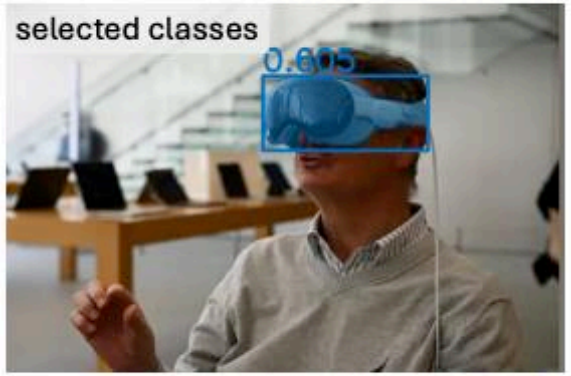
Please conduct object detection to any [List of COCO classes] that may be present.

← 开放检测 & 分割



Can you carry out object detection on this image and identify the women it contains?

I'm trying to detect bottles and forks in the image. Can you help me?



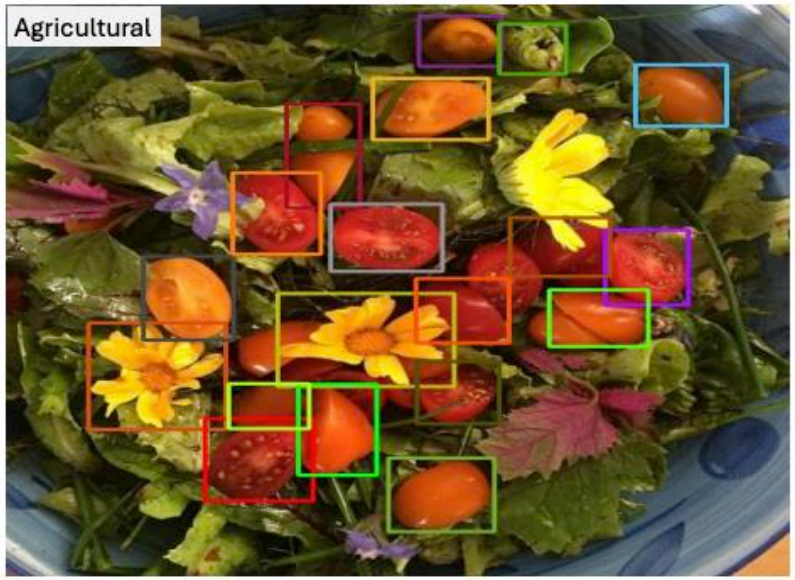
Are you capable of identifying Apple Vision Pro within this image?

https://github.com/OpenGVLab/InternVL

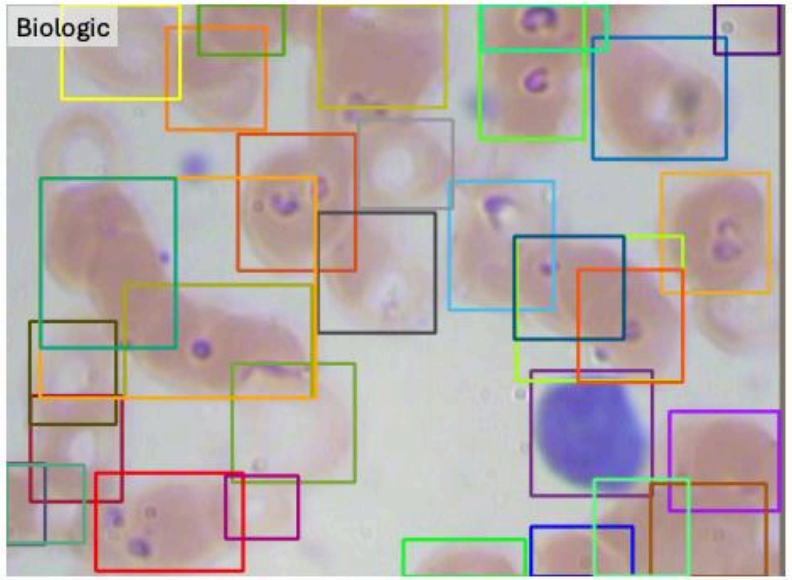
不止于语言输出：通专融合

Please assist me in identifying the gas cylinders within the image.

Please assist me in identifying the workers within the image.



Please assist me in identifying the vegetables within the image.

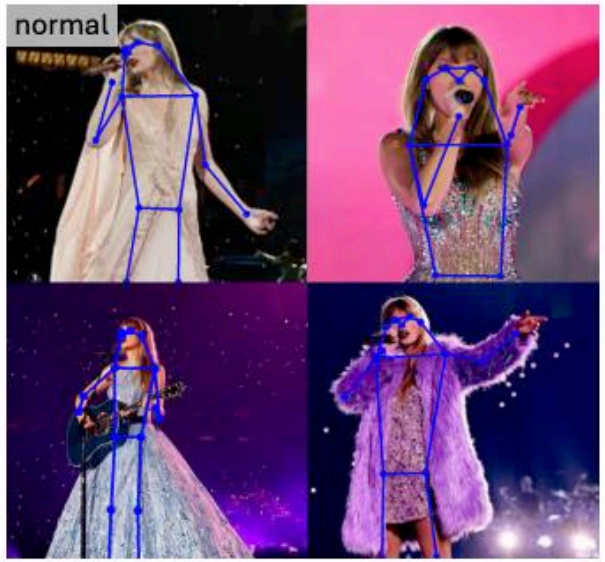


Please assist me in identifying the red blood cells within the image.

不同domain的开放检测 & 分割

https://github.com/OpenGVLab/InternVL

不止于语言输出：通专融合



I need your expertise to locate any **person** *in this image. Can you pinpoint the keypoint locations of* **[List of 17 COCO keypoints]** *?*

I need your expertise to locate any **person** *in this image. Please analyze this image and find the keypoint of* **the right elbow***?*

⬅ **不同domain的开放姿态估计**

Detect any **person** *in the given image. Can you pinpoint the keypoint locations of the* **nose, left_eye, right_eye, left_ear, right_ear** *?*

Please perform object detection on this image for identifying **elephants***. Detect the keypoint positions of the* **List of 17 AP-10K keypoints]**.

Can you detect **horses** *within the image? Can you pinpoint the keypoint locations of* **[List of 17 AP-10K keypoints]** *?*

https://github.com/OpenGVLab/InternVL

不止于语言输出：通专融合

An astronaut riding a horse X, where X ∈ {"", "in the style of van gogh", "in the style of ink painting", "in the style of black and white sketch"}

Panda mad scientist mixing sparkling chemicals, art station

Fox with wine cup

Dog with sunglasses

Watercolor painting of plant

Watercolor of sunset

Spectacular Tiny World in the Transparent Jar On the Table, interior of the Great Hall

A person standing on a mountaintop, looking out over a vast and rugged landscape

A red car near the sea

There is a boat on the foggy lake

A tiger in a lab coat with a 1980s Miami vibe, digital art

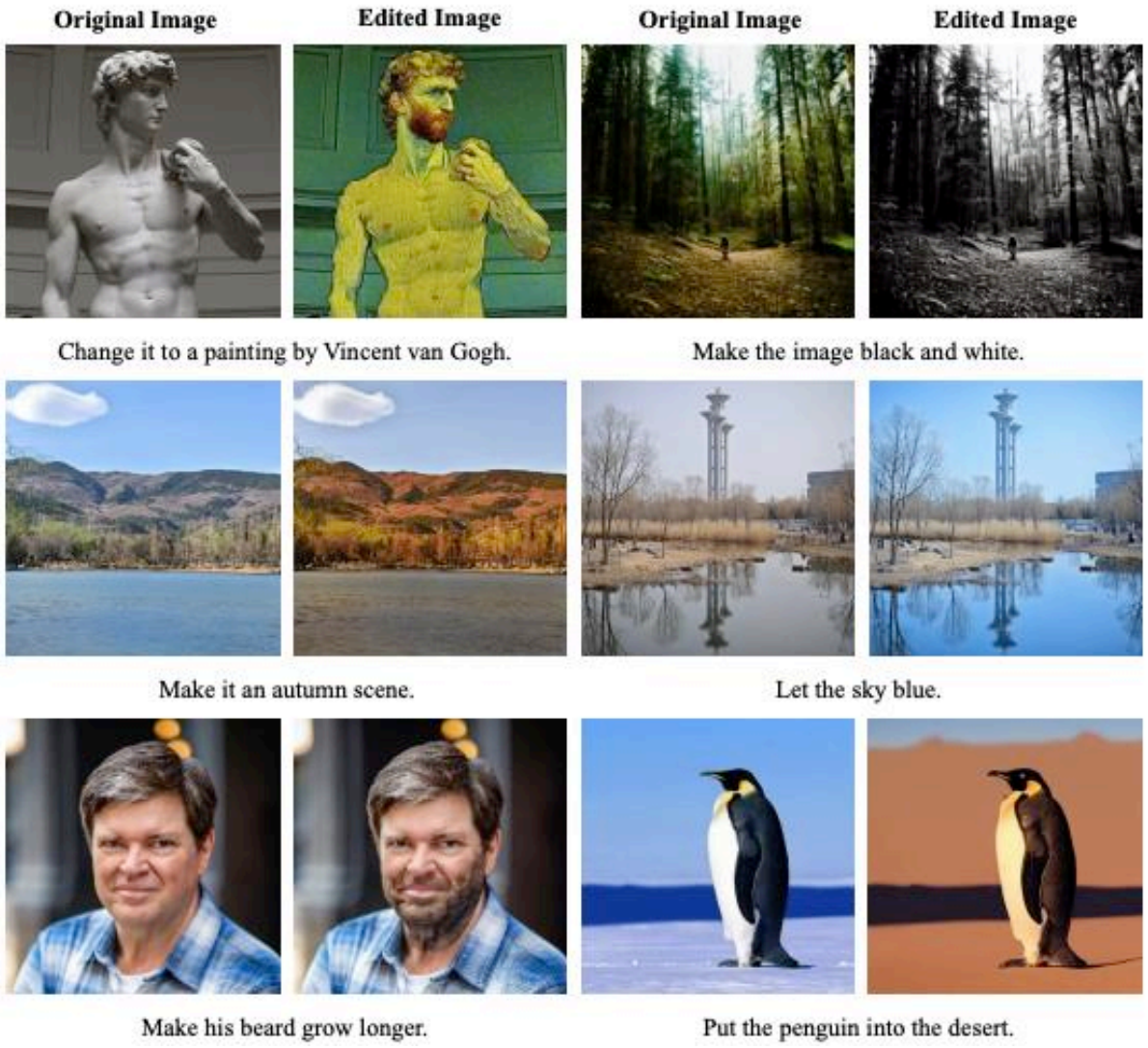
Guizhou Huangguoshu Waterfall

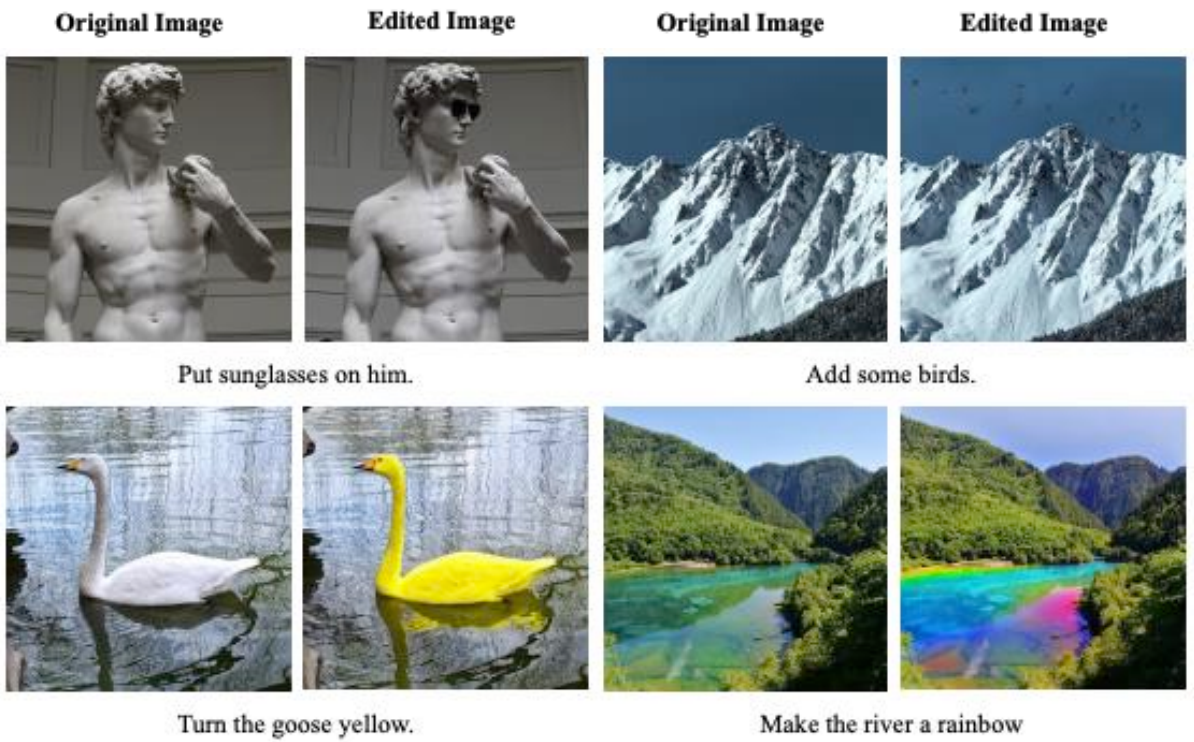
A farmyard surrounded by beautiful flowers

文生图

https://github.com/OpenGVLab/InternVL

不止于语言输出：通专融合





图像编辑

https://github.com/OpenGVLab/InternVL

感谢观看

通用视觉组交流群小助手

InternVL 2.0 在线试玩

https://github.com/OpenGVLab/InternVL

青稞社区

青年AI研究员Idea加油站，转行码农的新能源充电桩。

Talk主页：qingkelab.github.io/talks/talks

H5主页：https://appodzjvyp51702.h5.xiaoeknow.com

🔥青稞Talk： 青年AI研究员idea加油站

2024/04/10	19:00:00	青稞Talk		慕尼黑工业大学博士生陈振宇	SceneTex: 高质量三维室内场景纹理图生成
2024/04/15	19:00:00	青稞Talk		清华大学自然语言处理实验室（THUNLP）博士后钱忱	ChatDev—大语言模型驱动的多智能体协作与演化
2024/04/	09:00:00	青稞Talk		加州大学洛杉矶分校在读博士洪逸宁	从 3D-LLM 到 MultiPLY，3D 具身基础模型的构建
2024/04/24	19:00:00	青稞Talk		香港中文大学MMLAB在读博士李彦玮	Mini-Gemini：挖掘多模态视觉语言大模型的潜力
2024/05/10	19:00:00	青稞Talk		上海交通大学本科生甄昊宇	3D-VLA：构建生成式三维具身世界模型
2024/05/29	19:00:00	青稞Talk		南洋理工大学在读博士姜瑜铭	VideoBooth：文本和图像提示共同驱动的视频生成
2024/05/24	19:00:00	青稞Talk		NUS博士后倪瑾杰	MixEval：混合评测数据集来拟合大语言模型的人类评估
2024/05/21	19:00:00	青稞Talk		华南理工大学在读博士梁智源	实时渲染 3DGS 中的反走样及逆渲染应用
2024/06/06	19:00:00	青稞Talk		上海人工智能实验室青年研究员、OpenDriveLab具身智能方向负责人曾嘉博士	具身多模态大模型的视觉表征预训练研究
2024/06/13	19:00:00	青稞Talk		北京大学在读博士孟繁续	PiSSA：收敛快、误差小的大模型参数高效微调方法
2024/06/17	19:00:00	青稞Talk		香港大学在读博士吴成岳	LLaMA Pro：扩展Transformer块优化的大型语言模型继续预训练
2024/06/27	19:00:00	青稞Talk		浙江大学硕士董玉博	VillagerAgent：减少幻觉、提高任务分解效率的多智能协作体框架
2024/07/08	19:00:00	青稞Talk		北航博士郑耀威	LLaMA Factory：从预训练到RLHF，大模型高效训练框架
2024/07/11	19:00:00	青稞Talk		通义实验室高级算法专家徐海洋	Mobile-Agent：基于多模态Agent架构的手机智能体
2024/07/15	19:00:00	青稞Talk		清华大学在读博士余天予	MiniCPM-V：端侧可用的 GPT-4V 级多模态大模型
2024/07/23	19:00:00	青稞Talk		华中科技大学在读博士程天恒	YOLO-World：基于视觉语言模型的实时开放词汇物体检测
2024/07/30	19:00:00	青稞Talk		香港科技大学在读博士杨帅	SEED-story：生成长篇图文故事的多模态大型语言模型
2024/08/06	19:00:00	青稞Talk		香港中文大学博士后王文海	InternVL 2.0：通过渐进式策略扩展开源多模态大模型的性能边界
2024/08/14	19:00:00	青稞Talk		麻省理工学院博士生唐嘉铭	AWQ大模型低比特权重量化
2024/09/01	11:00:00	青稞Talk		斯坦福在读博士盛颖	SGLang

欢迎飞书订阅：直播日历

添加小助手入群

关注青稞社区公众号

Update: 2024.07