



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



# 我们离GPT-4V还有多远? 通过开源套件缩小与商用多模态模型之间的差距

王文海、陈喆

2024.05.14

# 背景：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

Science

Reducing the Dimensionality of Data with Neural Networks  
G. E. Hinton and R. R. Salakhutdinov  
Science 313, 504 (2006);  
DOI: 10.1126/science.1127647

深度置信网络

2006

ImageNet竞赛



1000类, 100万数据

2012

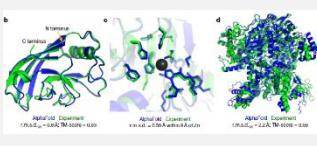
围棋比赛



AlphaGo 4:1 李世乭

2016

AlphaFold



蛋白质结构预测准确率新高

2021

大规模语音识别



Switchboard错误降低9%

人脸识别



LFW识别率99%,超过人类

2011

2014

2019

德州扑克



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



Yann LeCun  
图灵奖获得者

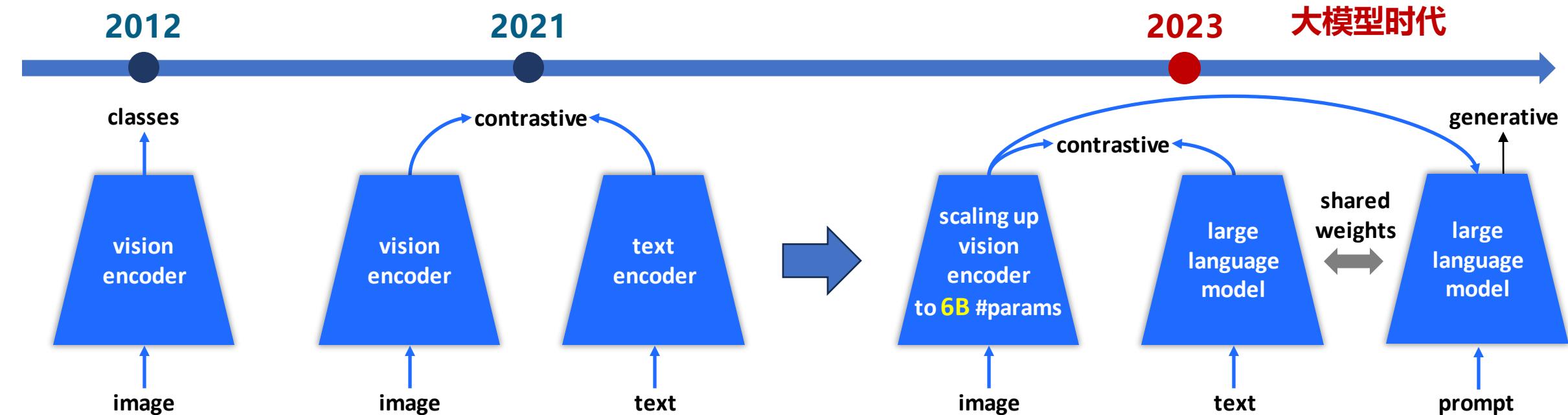
通用人工智能将会成为我们和未来科技交互的一部分。

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

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

# InternVL: 大规模视觉语言基础模型

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

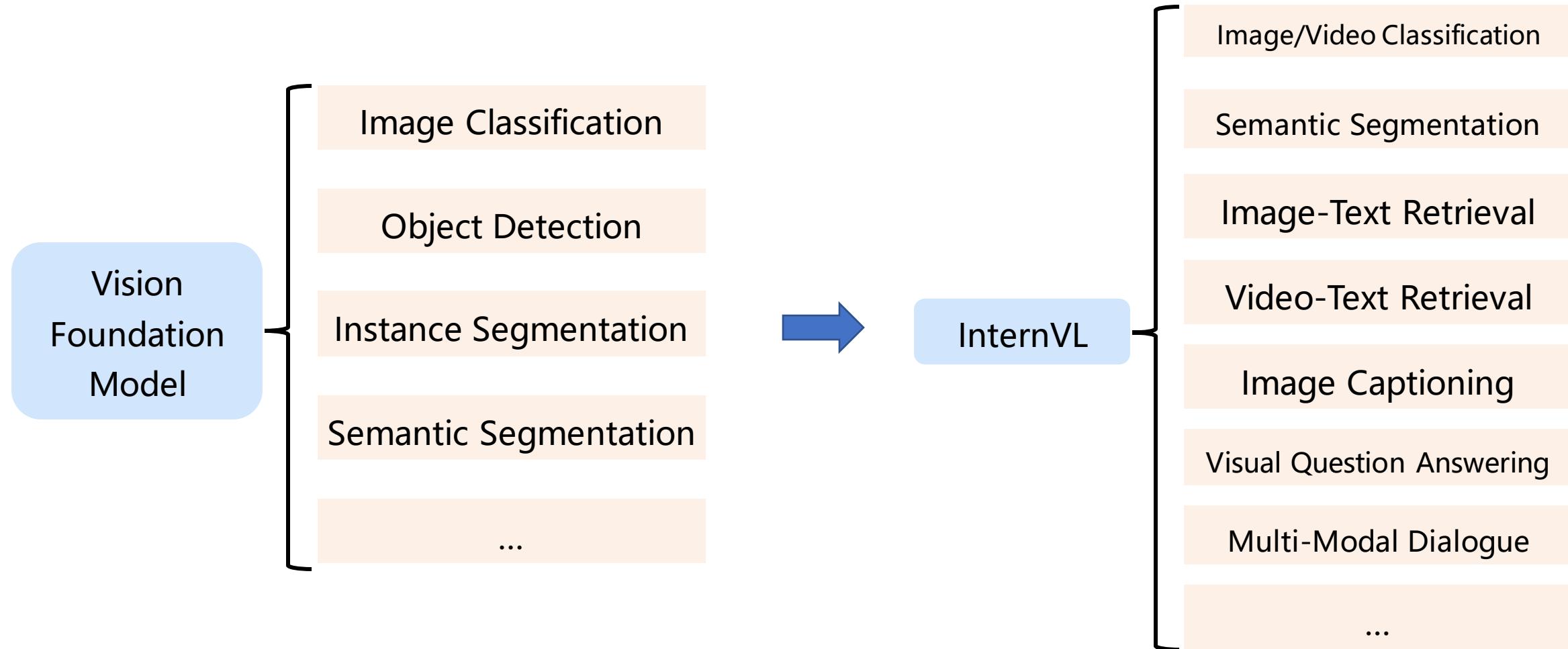
对比式预训练  
CLIP、ALIGN

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

- 60亿参数视觉基础模型
- 渐进式对齐视觉基础模型和LLM表征
- 大规模、多来源图文多模态训练数据

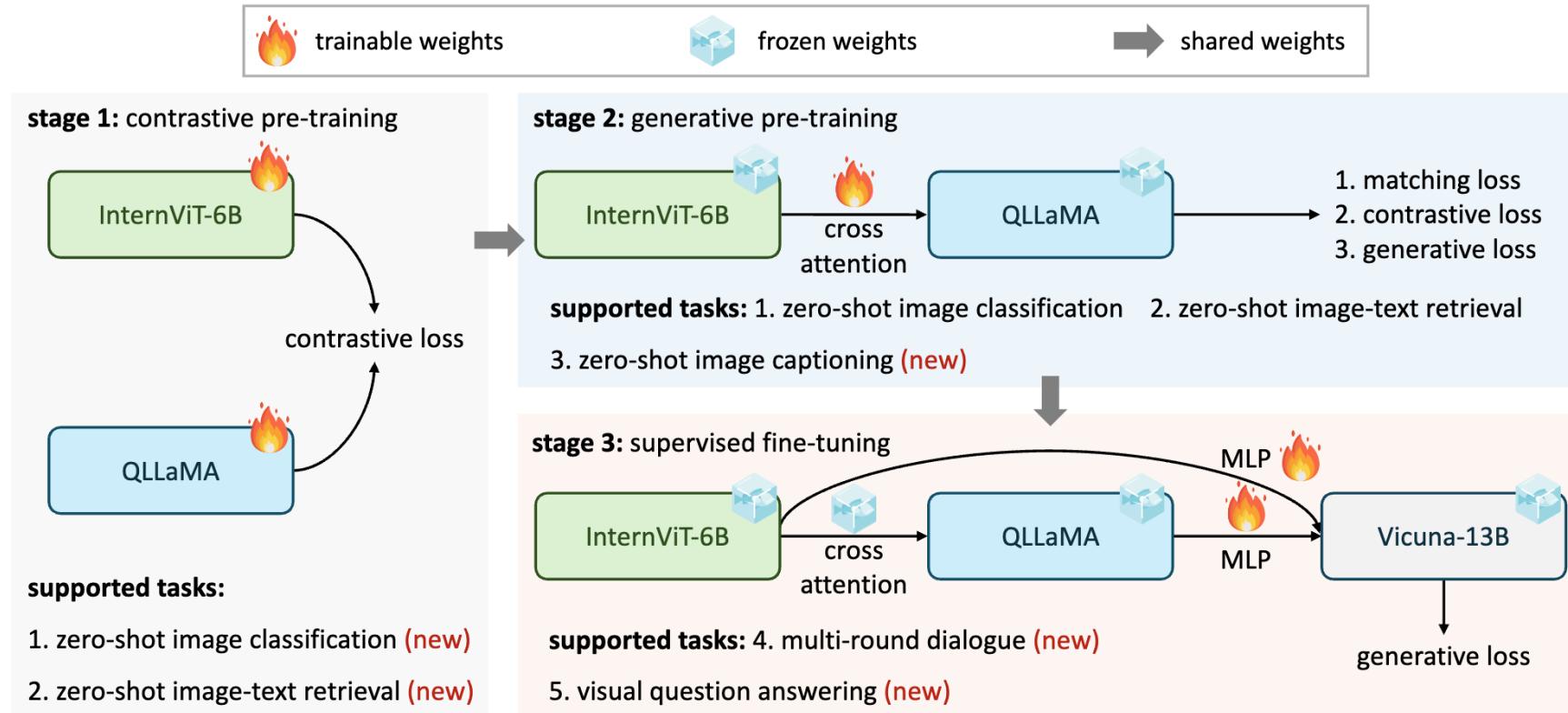
从适配视觉感知任务，到适配通用视觉语言任务，极大地拓宽了模型的适用范围



视觉感知任务：分类、检测、分割 ...

视觉语言任务：感知、图文、对话 ...

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

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

## 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

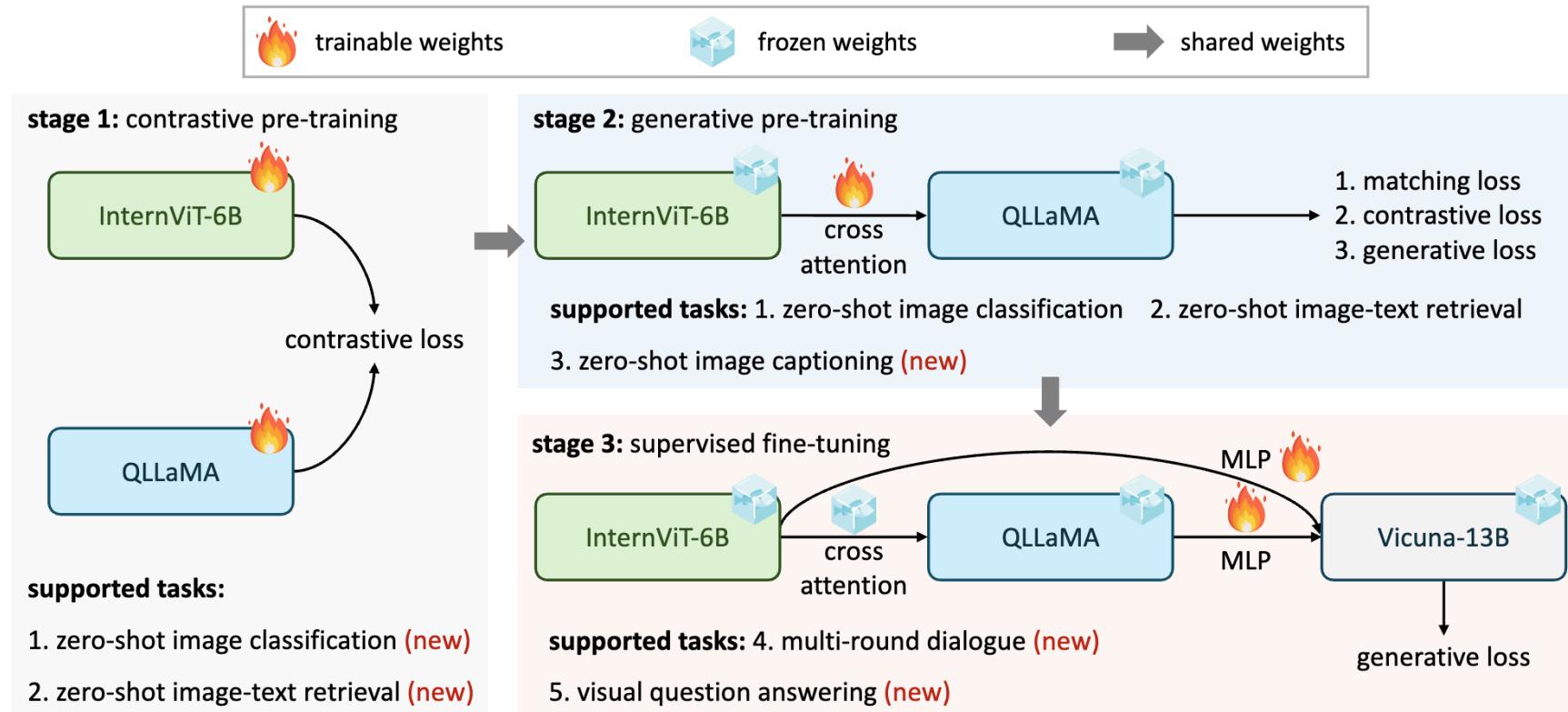
name	width	depth	MLP	#heads	#param	FLOPs	throughput	zs IN
variant 1	3968	32	15872	62	6051M	1571G	35.5 / 66.0	65.8
variant 2	3200	48	12800	50	5903M	1536G	28.1 / 64.9	66.1
variant 3	3200	48	12800	25	5903M	1536G	28.0 / 64.6	66.2
variant 4	2496	48	19968	39	5985M	1553G	28.3 / 65.3	65.9
variant 5	2816	64	11264	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  $224 \times 224$  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.

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

Table 1. Architecture details of the InternViT-6B model.

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

阶段一：利用海量带噪声的图文数据（5B）进行对比学习预训练

阶段二：利用过滤后的高质量图文数据（1B）进行生成式预训练

## 设计2: 渐进式的图像-文本对齐策略

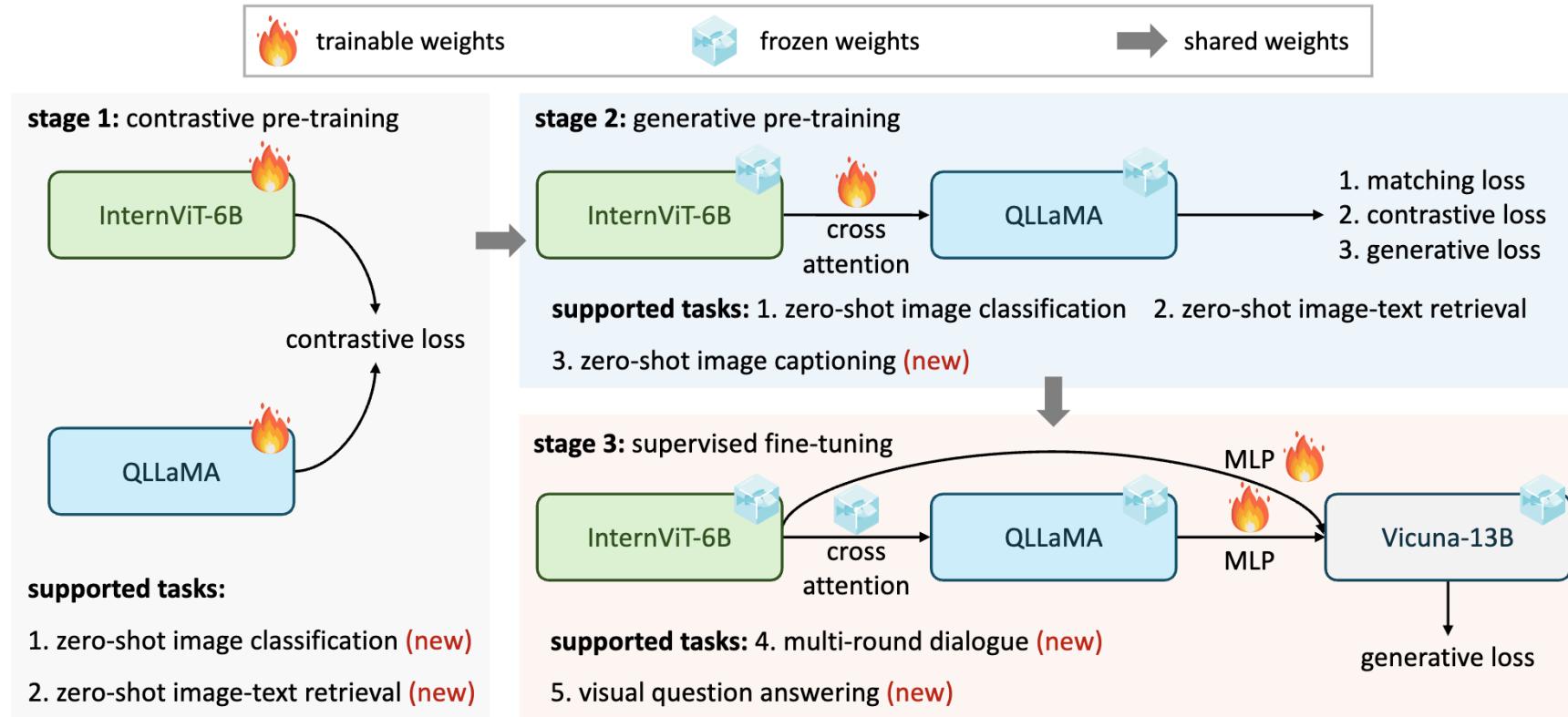
阶段一：利用海量带噪声的图文数据（5B）进行对比学习预训练

阶段二：利用过滤后的高质量图文数据（1B）进行生成式预训练

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%

主要考虑：CLIP相似度, 水印概率, unsafe概率, 美学指标, 图片分辨率, caption长度等

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

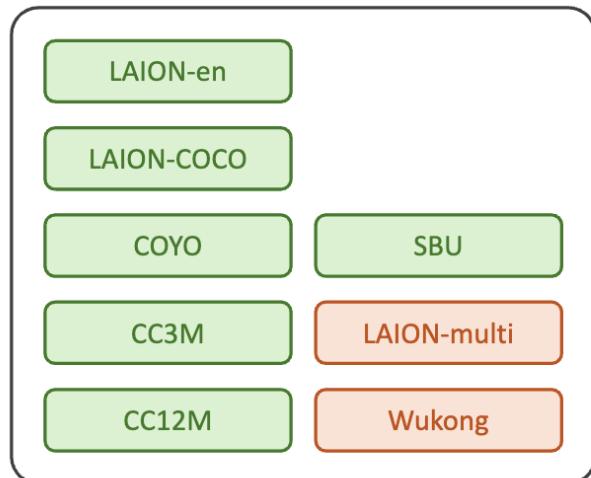


## 设计2：渐进式的图像-文本对齐策略

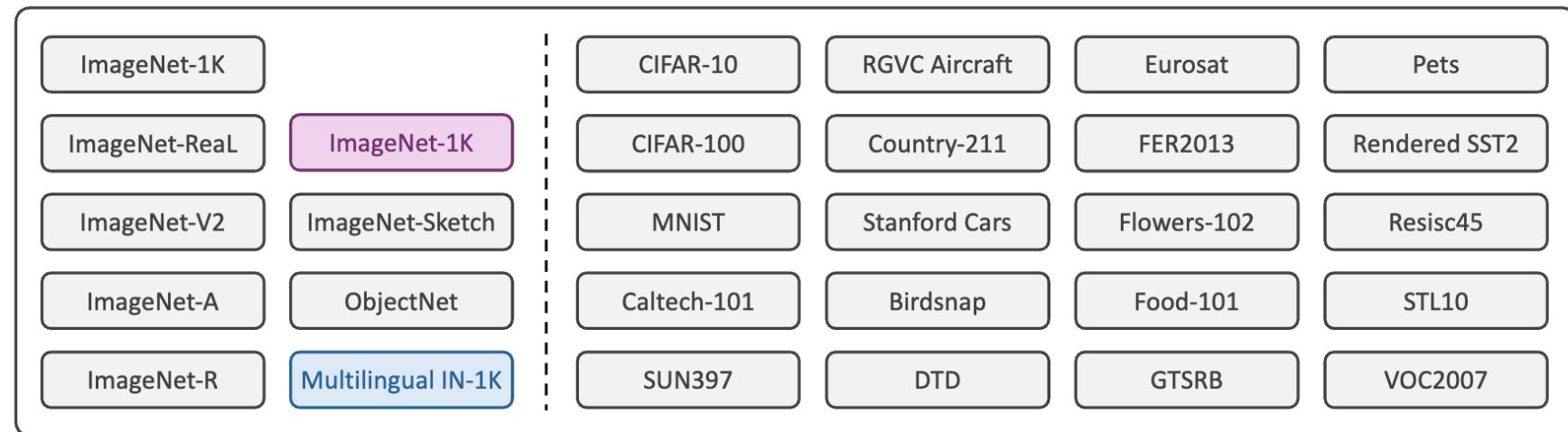
阶段三：利用高质量Caption/VQA/多轮对话数据（4M）进行SFT训练

## 预训练阶段1 & 阶段2 的训练/评测数据集全景图

Training Sets (English)  
  Training Sets (Multilingual)  
  Zero-Shot Test Sets (English)  
  Zero-Shot Test Sets (Multilingual)  
  Datasets for Transfer Learning



(a) Training Data for Stage 1 & 2



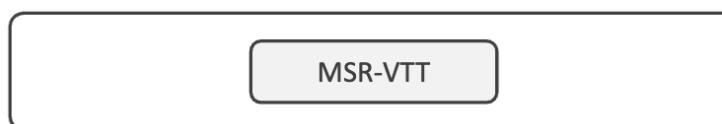
(b) Testing Datasets for Image Classification



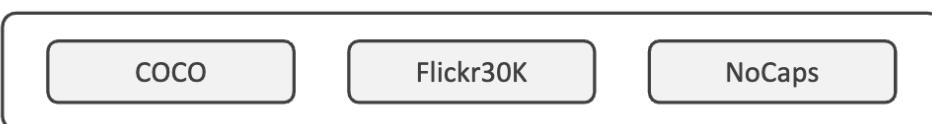
(c) Testing Datasets for Video Classification



(d) Testing Datasets for Image-Text Retrieval



(e) Testing Dataset for Video-Text Retrieval



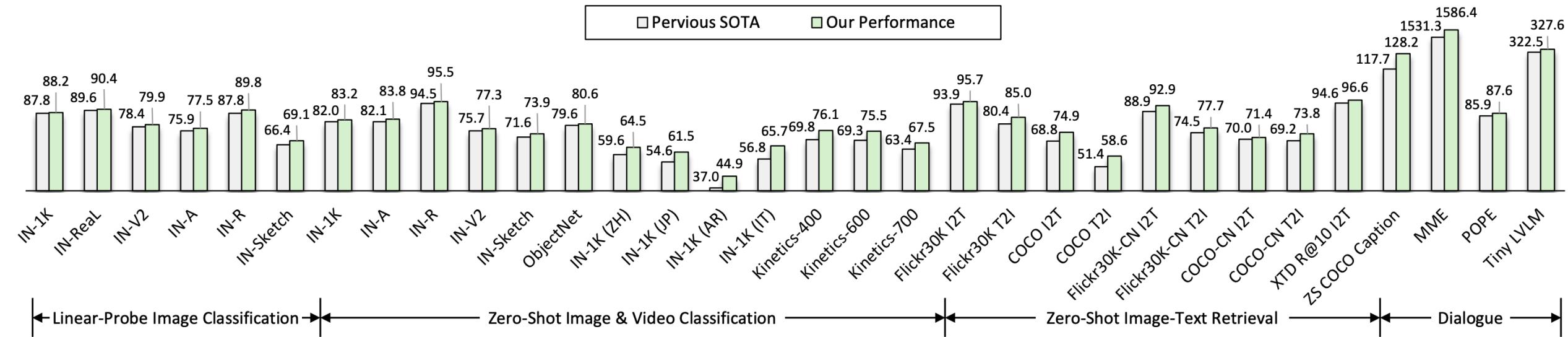
(f) Testing Datasets for Image Captioning



(g) Testing Dataset for Segmentation

搜集各种来源的图像、图文数据

同时支持图像分类、视频分类、图文检索、视频文本检索、图像描述、视觉问答、多模态对话等通用视觉语言任务，并达到最好性能



## InternViT-6B能力评测

从InternVL中单独抽出InternViT-6B，  
对其视觉特征提取能力进行评测：

- **Image-level:** 图像分类
- **Pixel-level:** 语义分割

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	<b>88.2</b>	<b>90.4</b>	<b>79.9</b>	<b>77.5</b>	<b>89.8</b>	<b>69.1</b>	<b>82.5</b>

图像分类线性探针实验

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

(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 <sub>frozen</sub> [67]	Linear	0.3M / 1.8B	512 <sup>2</sup>	39.3
ViT-22B <sub>frozen</sub> [37]	Linear	0.9M / 21.7B	504 <sup>2</sup>	34.6
InternViT-6B <sub>frozen</sub> (ours)	Linear	0.5M / 5.9B	504 <sup>2</sup>	<b>47.2</b>
ViT-22B <sub>frozen</sub> [37]	UperNet	0.8B / 22.5B	504 <sup>2</sup>	52.7
InternViT-6B <sub>frozen</sub> (ours)	UperNet	0.4B / 6.3B	504 <sup>2</sup>	<b>54.9</b>
ViT-22B [37]	UperNet	22.5B / 22.5B	504 <sup>2</sup>	55.3
InternViT-6B (ours)	UperNet	6.3B / 6.3B	504 <sup>2</sup>	<b>58.9</b>

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

语义分割能力评测

## InternVL能力评测

### 零样本图像分类能力评测

method	IN-1K	IN-A	IN-R	IN-V2	IN-Sketch	ObjectNet	$\Delta \downarrow$	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)	<b>83.2</b>	<b>83.8</b>	<b>95.5</b>	<b>77.3</b>	<b>73.9</b>	<b>80.6</b>	<b>0.8</b>	<b>82.4</b>

(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)	<b>83.2</b>	<b>64.5</b>	<b>61.5</b>	<b>44.9</b>	<b>65.7</b>	<b>64.0</b>

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

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	—	<b>71.0</b>	—	<b>71.3</b>	—	<b>65.7</b>
ViCLIP [117]	8	64.8	75.7	62.2	73.5	54.3	66.4
InternVL-C (ours)	8	<b>69.1</b>	<b>79.4</b>	<b>68.9</b>	<b>78.8</b>	<b>60.6</b>	<b>71.5</b>

### 零样本视频分类能力评测

## InternVL能力评测

InternVL具备很强的多语言的零样本图文检索能力

method	multi-lingual	Flickr30K (English, 1K test set) [88]						COCO (English, 5K test set) [20]						avg.	
		Image → Text			Text → Image			Image → Text			Text → Image				
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10		
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 <sup>†</sup> [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)	✓	<b>95.7</b>	<b>99.7</b>	<b>99.9</b>	<b>85.0</b>	<b>97.0</b>	<b>98.6</b>	<b>74.9</b>	<b>91.3</b>	<b>95.2</b>	<b>58.6</b>	<b>81.3</b>	<b>88.0</b>	<b>88.8</b>	

method		Flickr30K-CN (Chinese, 1K test set) [58]						COCO-CN (Chinese, 1K test set) [63]						avg.
		Image → Text			Text → Image			Image → Text			Text → Image			
R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
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)	✓	<b>92.9</b>	<b>99.4</b>	<b>99.8</b>	<b>77.7</b>	<b>94.8</b>	<b>97.3</b>	<b>71.4</b>	<b>93.9</b>	<b>97.7</b>	<b>73.8</b>	<b>94.4</b>	<b>98.1</b>	<b>90.9</b>

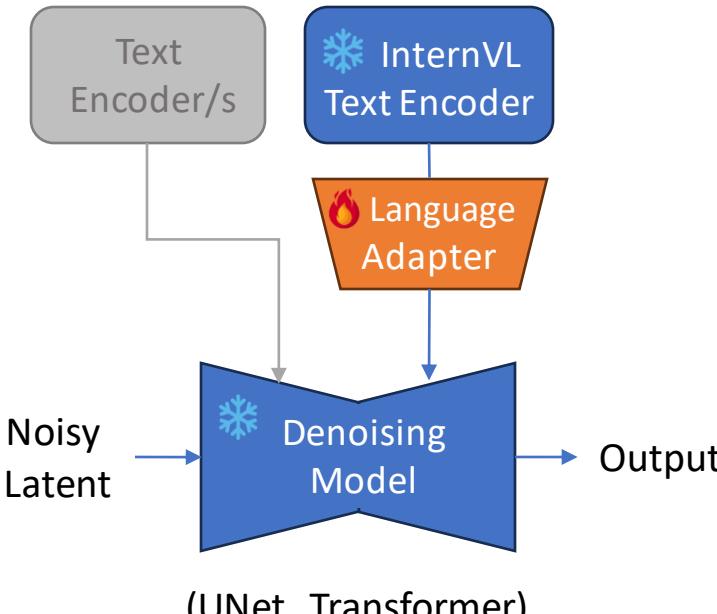
多语言的零样本图文检索评测

# InternVL: 大规模视觉语言基础模型

最强开源视觉基础模型，不到1/3的参数，能力比肩谷歌闭源ViT-22B

能力类型	模型	OpenCLIP-G (LAION)	InternViT-6B (Ours)	ViT-22B (谷歌)
开源		是	是	否
视觉能力	ImageNet (Linear probing)	86.2	88.2	<b>89.5</b>
	ADE20K (Linear probing)	-	<b>47.2</b>	34.6
图文对比能力	MS COCO 以图搜文	67.3	<b>74.9</b>	-
	Flickr30K 以文搜图	79.5	<b>85.0</b>	-
图文生成能力	NoCaps 以图生文	-	126.2 (持平)	126.3 (+PALI-X)

# 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

# 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



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

即插即用，无需对Diffusion Model做额外训练



Dreamshaper



Realistic Vision



Cartoonmix



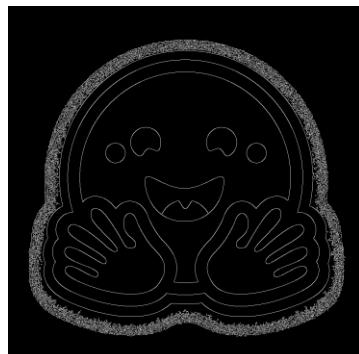
3D Animation



LoRA (Lego)



MVDream



ControlNet



LCM



SDXL Turbo

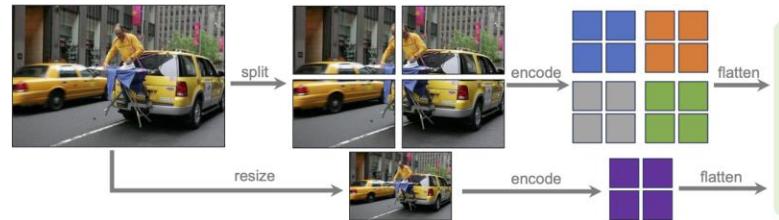


SDXL Lightning

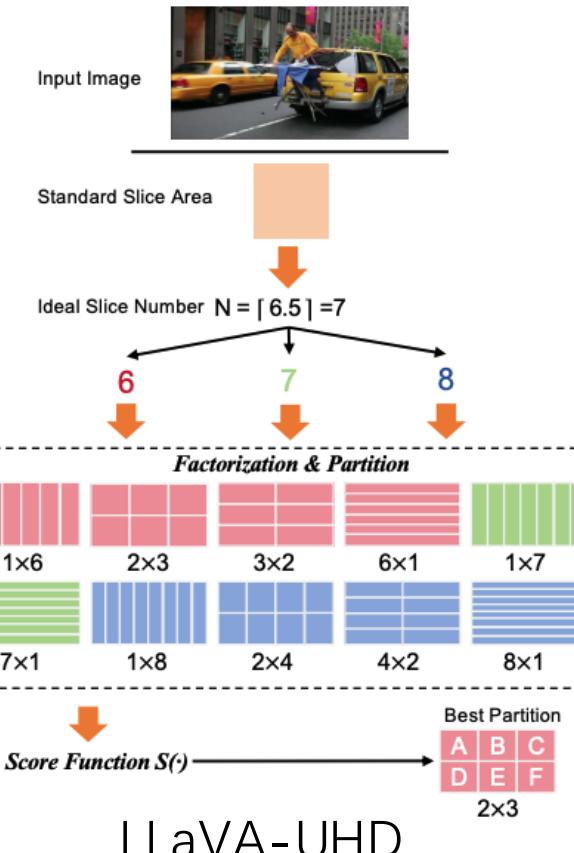


AnimateDiff

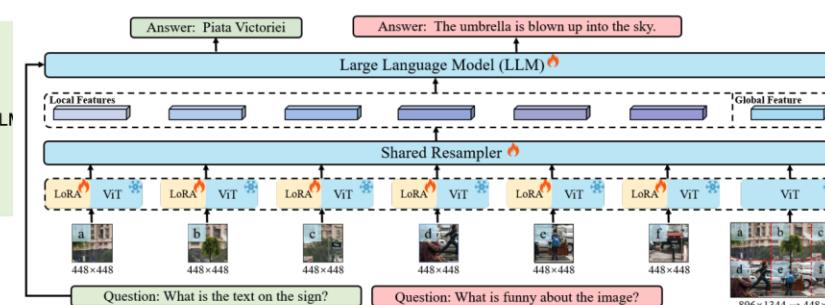
# 来自开源社区的启发：高分辨率很重要



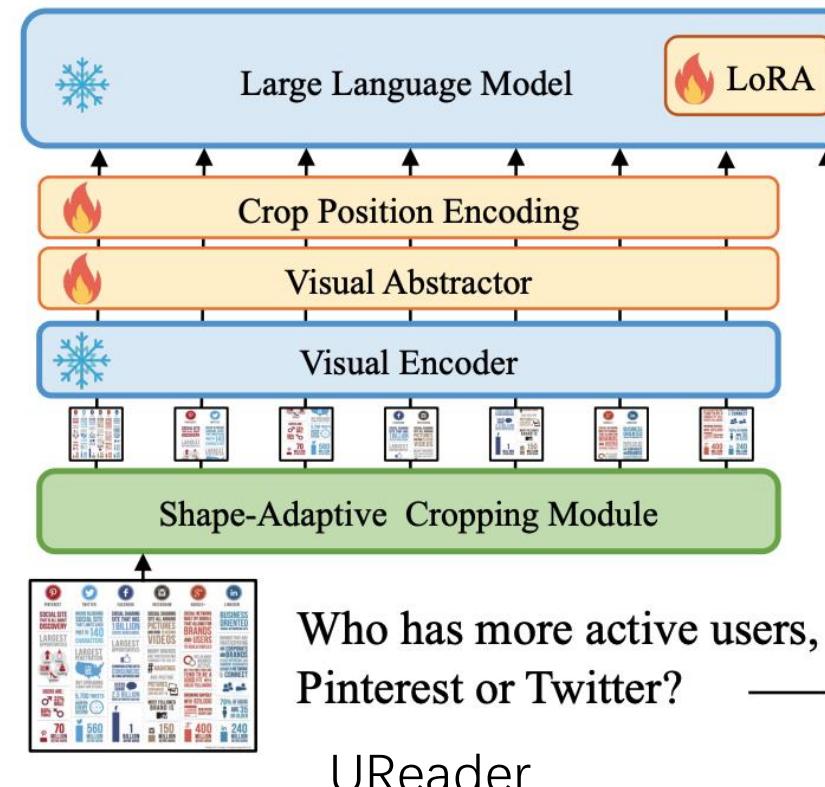
LLaVA-NeXT



LLaVA-UHD

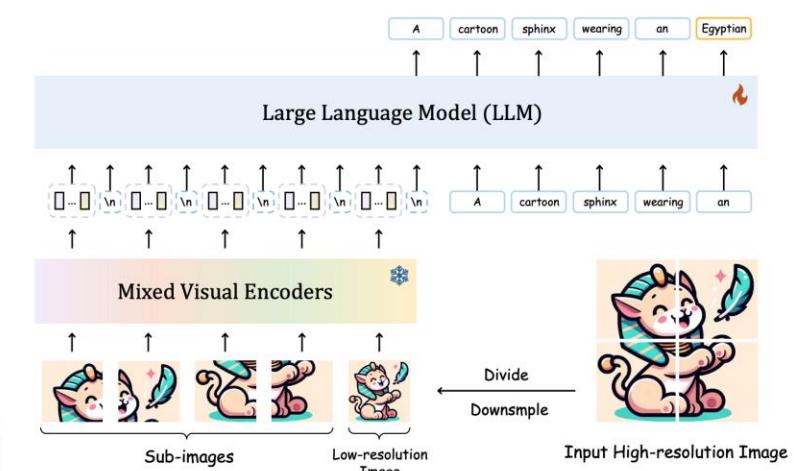


Monkey

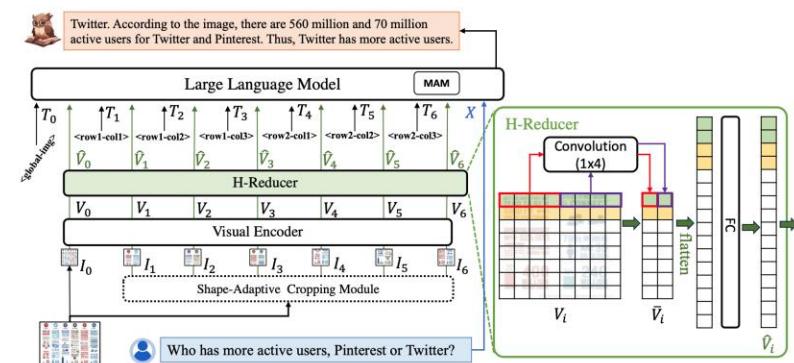


URader

<https://github.com/OpenGVLab/InternVL>



SPHINX



mPLUG-DocOwl 1.5

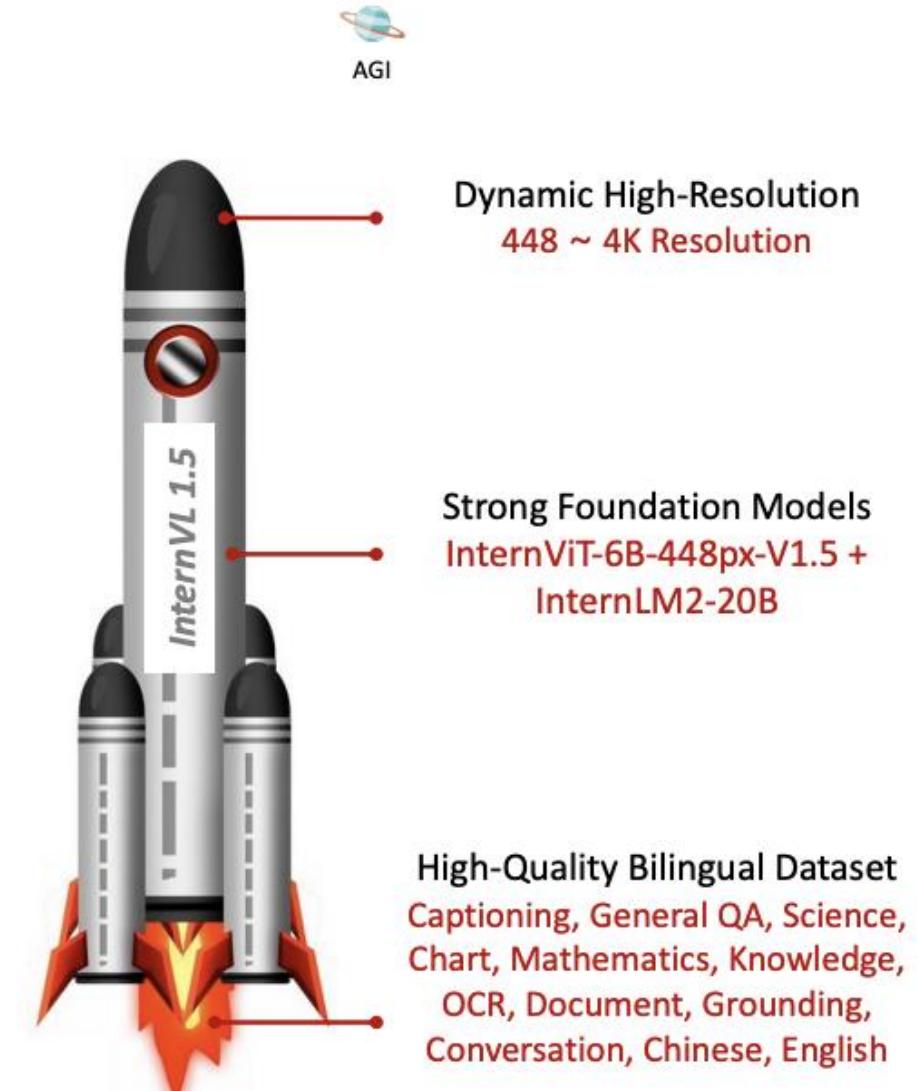
## 增强图文多模态对话能力

### 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：**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。

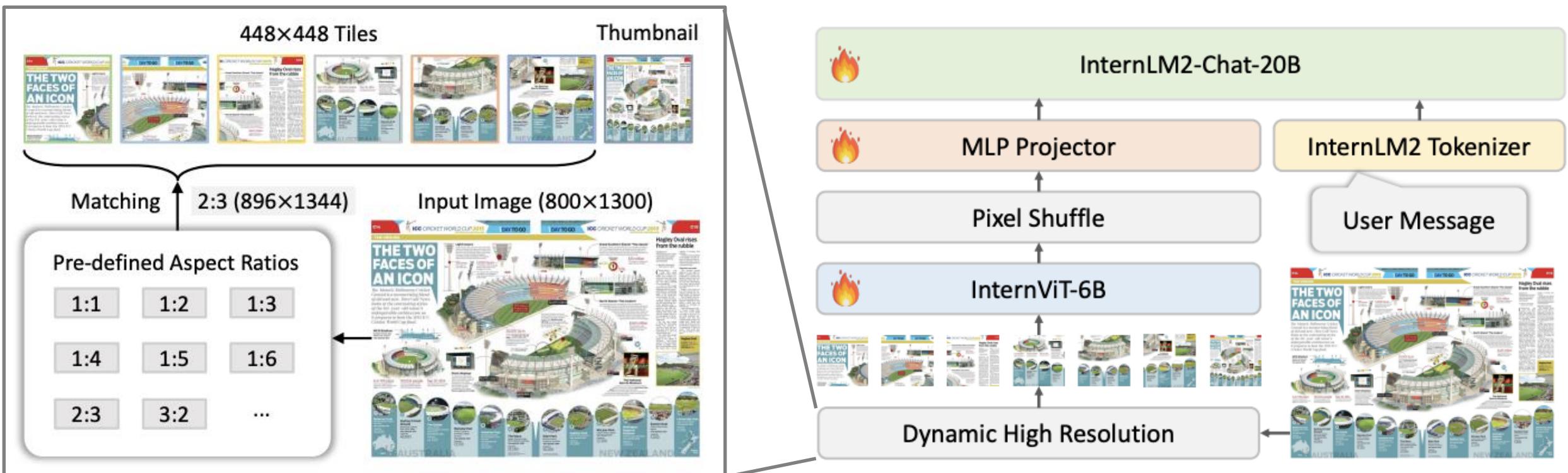
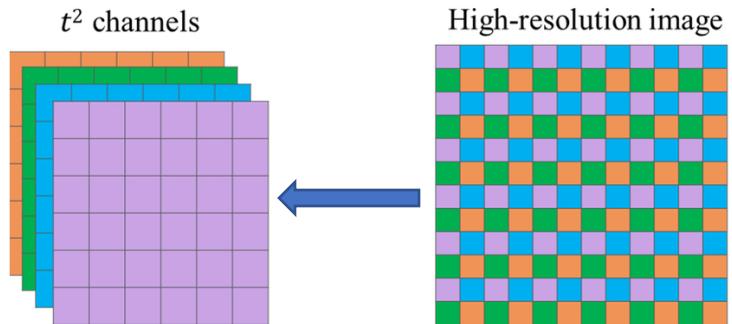
**燃料（高质量数据集）：**多语言、多来源、精细标注



# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

模型结构: 动态分辨率 + ViT-6B + PS + MLP + InternLM2-20B

每个 $448 \times 448$ 的图像块被表示为256个视觉token



## 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.

# 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



VS



4 : 3

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%

# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

## OCR能力强，InfoVQA和MMMU仍有差距

model	open-source	#param	OCR-related Benchmarks					General Multimodal Benchmarks									Math MathVista
			DocVQA	ChartQA	InfoVQA	TextVQA	OCRBench	MME	RWQA	AI2D	MMMU	MMB-EN/CN	CCB	MMVet	SEED	HallB	
GPT-4V [1]	✗	—	88.4	78.5	—	78.0	645	1926.6	61.4	78.2	56.8	77.0 / 74.4	46.5	<b>67.6</b>	71.6	46.5	49.9
Gemini Ultra 1.0 [107]	✗	—	90.9	80.8	<b>80.3</b>	<b>82.3</b>	—	—	—	79.5	<b>59.4</b>	— / —	—	—	—	—	53.0
Gemini Pro 1.0 [107]	✗	—	88.1	74.1	75.2	74.6	659	1933.4	—	73.9	47.9	73.6 / 74.3	52.5	64.3	70.7	45.2	45.2
Gemini Pro 1.5 [92]	✗	—	86.5	81.3	72.7	73.5	—	—	67.5	<b>80.3</b>	58.5	— / —	—	—	—	—	52.1
Qwen-VL-Max [5]	✗	—	<b>93.1</b>	79.8	73.4	—	723	<b>2433.6</b>	—	79.3	51.3	77.6 / 75.7	63.5	66.6	—	41.2	51.0
Qwen-VL-Plus [5]	✗	—	91.4	78.1	—	—	694	2183.4	—	75.9	45.2	67.0 / 70.7	55.1	61.1	72.7	40.6	43.3
Claude-3 Opus [3]	✗	—	89.3	80.8	—	—	694	1586.8	49.8	88.1	<b>59.4</b>	63.3 / 59.2	26.3	58.1	—	37.8	50.5
Claude-3 Sonnet [3]	✗	—	89.5	81.1	—	—	646	1625.9	51.9	<b>88.7</b>	53.1	67.8 / 64.2	27.8	—	—	41.3	47.9
Claude-3 Haiku [3]	✗	—	88.8	81.7	—	—	658	1453.2	—	<b>86.7</b>	50.2	60.7 / 57.2	24.5	—	—	39.2	46.4
HPT Pro [35]	✗	—	—	—	—	—	—	—	—	—	52.0	77.5 / 76.7	—	—	73.1	—	—
MM1 [84]	✗	30B	—	—	—	73.5	—	2069.0	—	—	44.7	75.1 / —	—	48.7	72.1	—	39.4
Step-1V [102]	✗	100B	—	—	—	—	625	2206.4	—	79.2	49.9	80.7 / 79.9	<b>71.2</b>	63.3	70.3	48.4	44.8
Grok-1.5V [125]	✗	—	85.6	76.1	—	78.1	—	—	<b>68.7</b>	88.3	—	— / —	—	—	—	—	52.8
Text-Monkey [68]	✓	10B	66.7	59.9	28.6	64.3	561	—	—	—	—	— / —	—	—	—	—	—
DocOwl-1.5 [33]	✓	8B	82.2	70.2	50.7	<b>68.6</b>	<b>599</b>	—	—	—	—	— / —	—	—	—	—	—
Mini-Gemini [53]	✓	35B	—	—	—	74.1*	—	2141.0	—	—	48.0	80.6 / —	—	59.3	—	—	43.3
LLaVA-NeXT [64]	✓	35B	84.0	68.7	51.5	69.5*	574	2028.0	—	74.9	51.1	81.1 / 79.0	49.2	57.4	75.9	34.8	46.5
InternVL 1.2 (ours)	✓	40B	57.7	68.0	<b>39.5</b>	72.5*	569	2175.4	67.5	79.0	<b>51.6</b>	82.2 / 81.2	59.2	48.9	75.6	47.6	47.7
InternVL 1.5 (ours)	✓	26B	90.9	<b>83.8</b>	72.5	80.6	<b>724</b>	2187.8	66.0	80.7	45.2	<b>82.2 / 82.0</b>	69.8	62.8	<b>76.0</b>	<b>49.3</b>	<b>53.5</b>

# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

## 感知能力强，但多轮对话和商用模型仍有较大差距

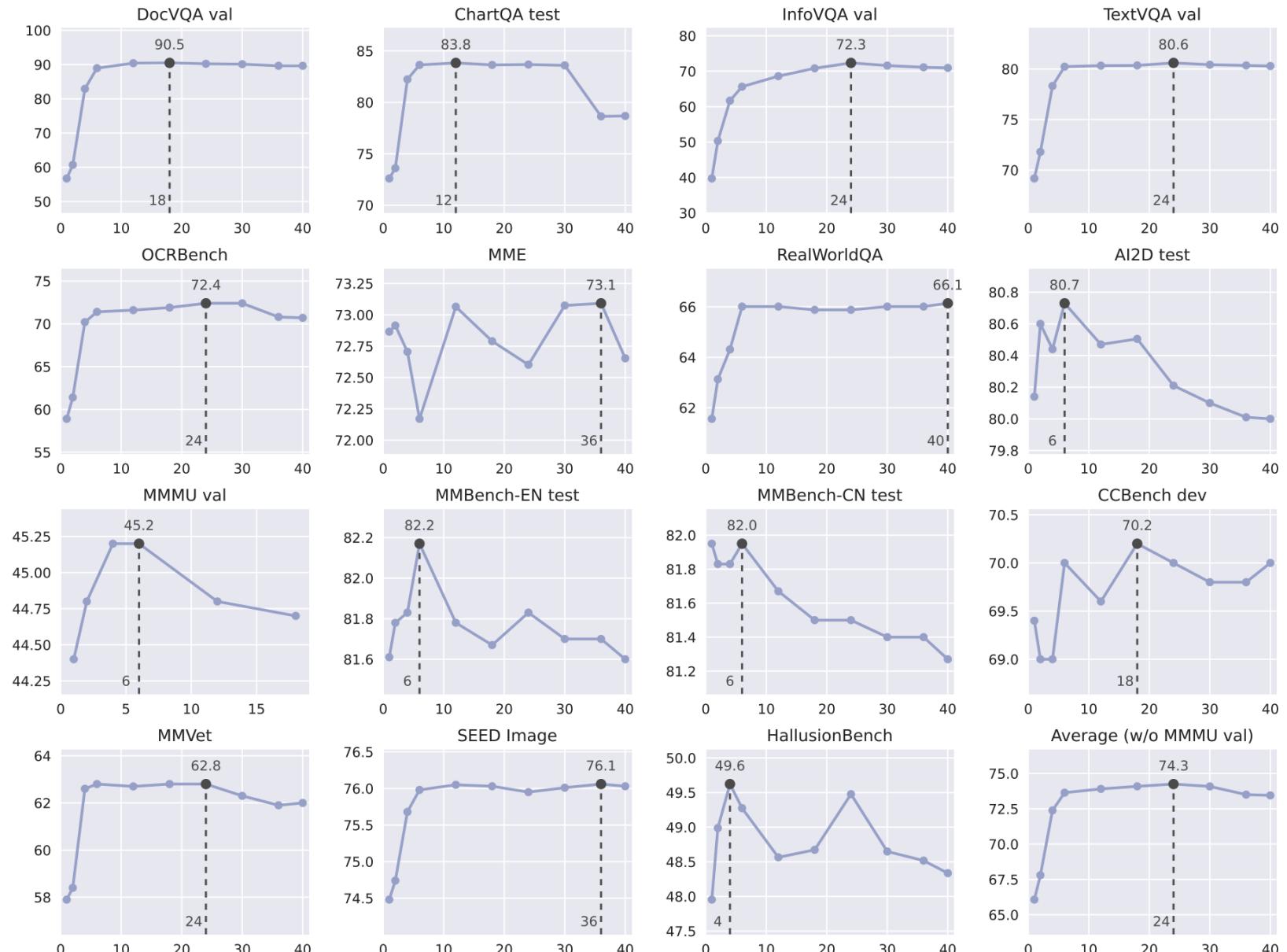
model	open-source	#param	ConvBench (Pairwise Grading)						ConvBench (Direct Grading)					
			$R_1$	$R_2$	$S_1$	$S_2$	$S_3$	$S_O$	$R_1$	$R_2$	$S_1$	$S_2$	$S_3$	$S_O$
GPT-4V [1]	✗	—	<b>39.51</b>	<b>38.47</b>	38.47	<b>39.34</b>	<b>37.61</b>	<b>40.55</b>	<b>7.09</b>	<b>7.30</b>	<b>7.30</b>	<b>7.48</b>	<b>7.12</b>	<b>6.88</b>
Claude-3 Opus [3]	✗	—	36.60	37.49	<b>38.99</b>	39.17	34.32	35.70	6.54	6.75	6.53	7.04	6.68	6.32
Reka Flash [89]	✗	—	25.60	24.67	25.13	27.56	21.32	26.52	6.78	6.86	6.93	7.25	6.41	6.70
Gemini Pro 1.0 [107]	✗	—	8.44	8.55	9.01	9.36	7.28	8.32	4.42	4.60	5.18	4.95	3.66	4.24
ShareGPT4V-13B [16]	✓	13B	17.56	17.45	17.85	18.72	15.77	17.68	4.85	5.03	5.16	5.06	4.86	4.67
LLaVA-1.5-13B [62]	✓	13B	16.93	18.08	20.45	18.02	15.77	15.77	4.94	5.14	5.03	5.41	4.99	4.74
XComposer2 [23]	✓	8B	15.83	16.41	17.16	19.06	13.00	15.25	5.82	5.98	5.98	6.17	5.78	5.66
mPLUG-Owl2 [128]	✓	8B	14.93	15.83	17.50	17.16	12.82	14.04	5.04	5.17	4.98	5.38	5.14	4.91
Qwen-VL-Chat [5]	✓	10B	14.33	14.62	16.29	18.37	9.19	14.04	5.54	5.65	5.96	5.78	5.22	5.43
MiniGPT-4 [142]	✓	8B	10.95	10.80	11.61	11.27	9.53	11.09	3.85	4.04	3.99	4.40	3.73	3.66
LLaMA-A-V2 [27]	✓	7B	9.04	9.59	8.84	10.92	9.01	8.49	4.77	4.91	4.77	5.47	4.48	4.64
InternVL 1.2 (ours)	✓	40B	21.17	22.41	24.96	21.31	20.97	19.93	5.49	5.69	5.80	5.88	5.39	5.29
InternVL 1.5 (ours)	✓	26B	17.65	20.22	26.00	17.33	17.33	15.08	5.60	5.76	6.11	5.93	5.25	5.43

model	open-source	#param	MMT-Bench	
			Overall	Overall*
GPT-4V [1]	✗	—	62.0	55.5
Qwen-VL-Plus [4]	✗	—	62.3	56.6
Gemini Pro 1.0 [107]	✗	—	61.6	55.1
Claude-3 Haiku [3]	✗	—	52.2	46.4
LLaVA-NeXT [64]	✓	35B	60.8	56.3
XComposer2 [23]	✓	8B	55.7	50.0
BLIP-2-XXL [50]	✓	12B	54.8	49.1
Yi-VL-34B [130]	✓	35B	54.2	48.6
Monkey-Chat [107]	✓	10B	53.4	46.0
DeepSeek-VL [71]	✓	7B	53.2	46.5
CogVLM-Chat [117]	✓	17B	51.6	44.2
InternVL 1.2 (ours)	✓	40B	<b>63.4</b>	<b>58.2</b>
InternVL 1.5 (ours)	✓	26B	59.0	56.2

Table 3. Comparison with Sota models on ConvBench and MMT-Bench. ConvBench [65] is a multi-turn conversation evaluation benchmark designed for MLLMs. The table presents win rates against humans, where  $S_1$ ,  $S_2$ , and  $S_3$  represent the scores for perception, reasoning, and creation, respectively.  $R_2$  is calculated as  $(S_1 + S_2 + S_3)/3$ , reflecting the average performance across three turns.  $R_1$  is derived from  $(R_2 + S_0)/2$ , indicating the model’s overall score. MMT-Bench [129] is a comprehensive benchmark designed to assess MLLMs across massive multimodal tasks requiring expert knowledge and deliberate visual recognition, localization, reasoning, and planning. The overall score is computed across 162 subtasks, excluding visual recognition as denoted by \*.

# 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:



# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

## 复杂公式的理解能力

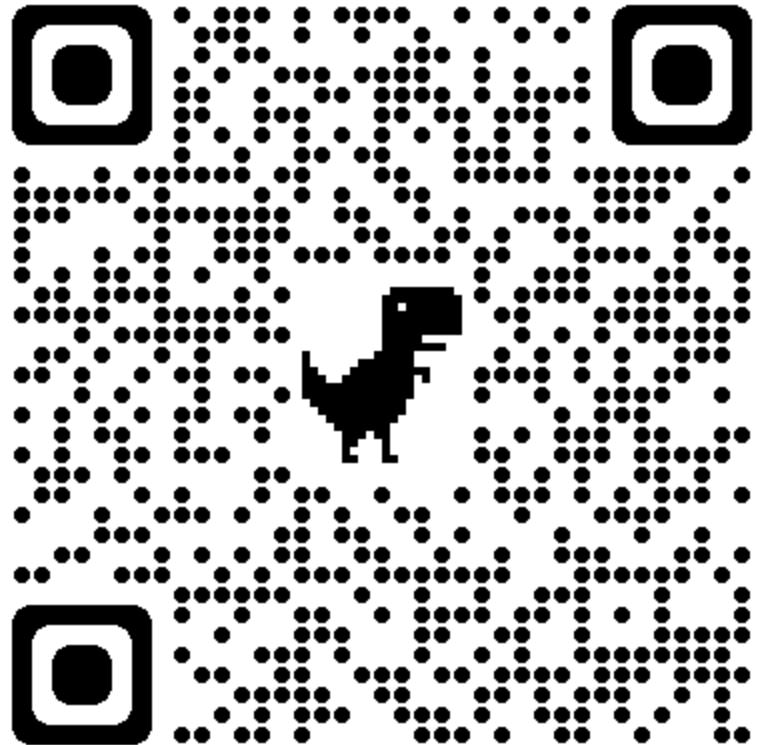
InternVL-Chat

$$B(M) = \frac{3}{L_{\text{host}}(M)} \int_{M_{\min}}^M \frac{dn}{dm} dm \int_0^1 dx_{\text{sub}} \times [1 + B_{\text{sub}}(m)] L_{\text{sub}}(m, x_{\text{sub}}) x_{\text{sub}}^2$$

convert it to latex

Certainly, I'll convert it to LaTeX for you.

```
\begin{eqnarray*}
B(M)&=&\frac{3}{L_{\text{host}}(M)}\int_{M_{\min}}^M\frac{dn}{dm}dm\int_0^1dx_{\text{sub}}\times\nonumber\\
&&[1+B_{\text{sub}}(m)]L_{\text{sub}}(m,x_{\text{sub}})x_{\text{sub}}^2\\
\end{eqnarray*}
```



# InternVL 1.5：接近GPT-4V的开源多模态对话模型

## 复杂公式的理解能力

InternVL-Chat

每一天有改变吗?

$$(1+0.01)^{365} = 37.78343433289 > 1$$
$$1^{365} = 1 = 1$$
$$(1-0.01)^{365} = 0.02551796445229 < 1$$

将图中的公式转换成latex

每一天有改变吗?

```
\begin{aligned}& (1+0.01)^{365} = 37.78343433289 > 1 \\& 1^{365} = 1 = 1 \\& (1-0.01)^{365} = 0.02551796445229 < 1\end{aligned}
```

这些公式蕴含着什么人生哲理?

这些公式可能暗示着即使每天只改变一点点，随着时间的积累，也能产生显著的影响。这可以应用于许多领域，包括个人成长、财务积累、健康习惯的养成等。

# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

极强的OCR能力



详细描述图片

这张图片显示的是一块挂在墙上的招牌，上面用中文写着“河大西门第一家”，表明这是河大西门地区的第一家店。招牌上还写着“正宗 老字号 五香味”，这可能是在强调这家店的传统和口味。

招牌上还提到“曹记烟烤鸡腿”，这可能是这家店的主打产品。招牌上列出了营业时间：“16:30-22:30”，说明这家店在下午4:30到晚上10:30之间营业。

招牌上还写着“满播萬”，这可能是一个品牌或者口号。

招牌上还提到了“7点之后在门口路灯下出夜市”，这表明这家店在晚上7点之后会提供夜市服务，顾客可以在门口的路灯下找到夜市的位置。

此外，招牌上还有一些其他的文字和图案，但由于图片分辨率的限制，这些细节可能不太容易辨认。整体上，这块招牌提供了很多关于这家店的信息，包括它的历史、主打产品、营业时间和夜市服务。



# InternVL 1.5：接近GPT-4V的开源多模态对话模型

## GUI理解能力

InternVL-Chat



详细描述图片

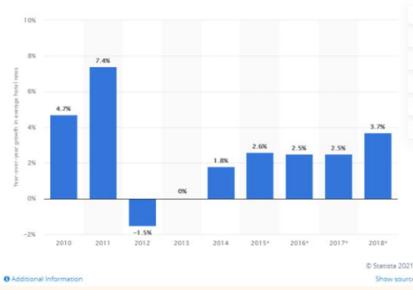
这张图片是流行视频游戏Minecraft的官方网站的欢迎页面。图片背景是一片郁郁葱葱的绿色景观，远处有蓝色的山脉和晴朗的天空。前景中有两个Minecraft玩家角色，他们站在一个绿草覆盖的小山上。其中一个角色似乎在照顾一只小狗。

欢迎页面上的文字写着：“欢迎来到Minecraft的官方网站。与新游戏、新更新和新玩法一起加入最大的游戏社区，今天开始制作吧！”这句话暗示网站提供了游戏的新版本、更新内容以及与其他玩家互动的机会。

欢迎页面还包括一个绿色的“获取Minecraft”按钮，邀请访问者下载或购买游戏。整个图片的风格和色彩方案与Minecraft游戏的标志性像素化美学相一致。

# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

## 复杂图表理解&转化



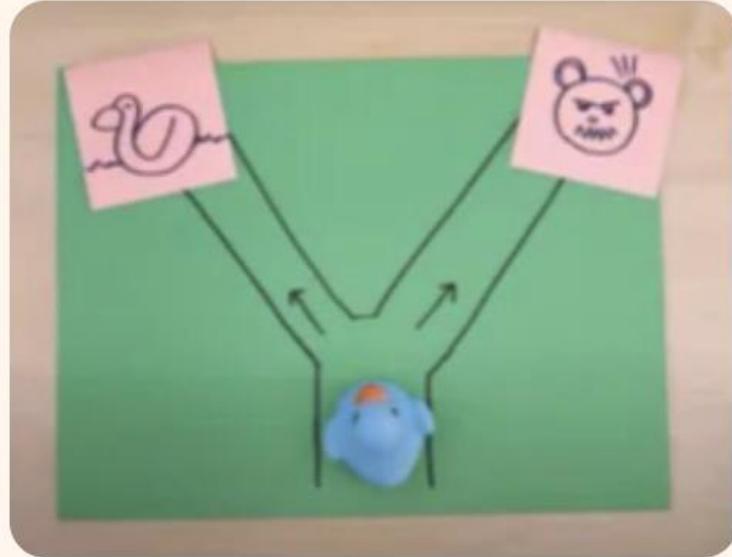
describe this table in detail

The column chart shows the year-over-year growth in average hotel rates from 2010 to 2018. In 2010, the growth was 4.7%, followed by a significant increase to 7.4% in 2011. However, in 2012, there was a decrease of -1.5%. The growth remained at 0% in 2013, followed by a slight increase to 1.8% in 2014. In 2015, the growth was 2.6%, followed by a slight increase to 2.5% in 2016 and 2017. In 2018, the growth was 3.7%.

把表格转换成markdown格式

---	---
2018*   3.7%	
2017*   2.5%	
2016*   2.5%	
2015*   2.6%	
2014   1.8%	
2013   0%	
2012   -1.5%	
2011   7.4%	
2010   4.7%	````

## 视觉逻辑推理



小鸭子应该往哪边走，请详细分析一下

根据图片中的信息，小鸭子应该往左边走。因为左边的路是通向小鸭子的，而右边的路通向一个凶恶的熊，所以小鸭子应该避免走右边的路。

## 视觉内容关联能力

InternVL-Chat



Tell me a person who had made a similar posture to the cute duck

The posture of the duck, with its hands on its hips and a confident stance, is reminiscent of the martial artist Bruce Lee. Bruce Lee was known for his dynamic fighting style and his charismatic presence, often captured in iconic photos with his hands on his hips, exuding confidence and power.

# InternVL 1.5: 接近GPT-4V的开源多模态对话模型

## 指令跟随能力

InternVL-Chat



The image shows a US Permanent Resident Card (Green Card). The card is light blue with a portrait of a woman on the left. Text on the card includes:

- UNITED STATES OF AMERICA PERMANENT RESIDENT
- 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

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}

```
{  
    "Surname": "STEVEN",  
    "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"  
}
```

# InternVL 1.5：接近GPT-4V的开源多模态对话模型

## 更多例子

InternVL-Chat



图片里男子右手指的是啥

图片中的男子右手手指指向的是一只白鲸的面部。白鲸是一种非常大的海洋哺乳动物，以其巨大的体型和温和的性格而闻名。它们主要分布在北极和亚北极地区，并且是鲸类中非常聪明的一种。

图片里男子左手呢

图片中的男子左手似乎是放在孩子的背后，以稳定和支持孩子。他正把孩子抱在怀里，让孩子可以近距离地观察和接触白鲸。这种互动方式在海洋馆和动物园中很常见，以帮助游客特别是儿童与动物建立联系，并增强他们对野生动物的认知和兴趣。

# 开源模型权重

## Vision Large Language Model

Model	Date	Download	Note
InternVL-Chat-V1.5-Int8	2024.04.28	 <a href="#">HF link</a>	The INT8 version of InternVL-Chat-V1-5
InternVL-Chat-V1.5	2024.04.18	 <a href="#">HF link</a>	support 4K image; super strong OCR; Approaching the performance of GPT-4V and Gemini Pro on various benchmarks like MMMU, DocVQA, ChartQA, MathVista, etc. (🔥new)
InternVL-Chat-V1.2-Plus	2024.02.21	 <a href="#">HF link</a>	more SFT data and stronger
InternVL-Chat-V1.2	2024.02.11	 <a href="#">HF link</a>	scaling up LLM to 34B
InternVL-Chat-V1.1	2024.01.24	 <a href="#">HF link</a>	support Chinese and stronger OCR
InternVL-Chat-19B-448px	2024.02.03	 <a href="#">HF link</a>	448 resolution
InternVL-Chat-19B	2023.12.25	 <a href="#">HF link</a>	English multimodal dialogue
InternVL-Chat-13B	2023.12.25	 <a href="#">HF link</a>	English multimodal dialogue

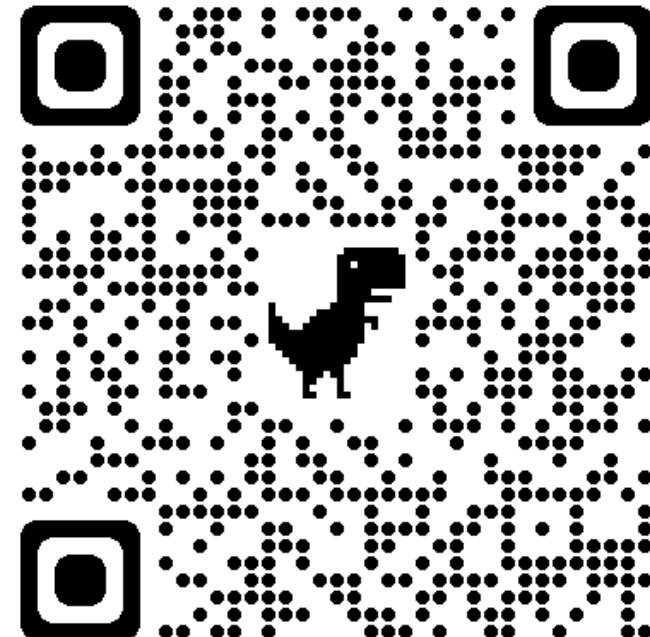
## Vision-Language Foundation Model

Model	Date	Download	Note
InternViT-6B-448px-V1.5	2024.04.20	 <a href="#">HF link</a>	support dynamic resolution, super strong OCR (🔥new)
InternViT-6B-448px-V1.2	2024.02.11	 <a href="#">HF link</a>	448 resolution
InternViT-6B-448px-V1.0	2024.01.30	 <a href="#">HF link</a>	448 resolution
InternViT-6B-224px	2023.12.22	 <a href="#">HF link</a>	vision foundation model
InternVL-14B-224px	2023.12.22	 <a href="#">HF link</a>	vision-language foundation model

# 感谢观看



通用视觉组交流群小助手



InternVL 1.5 在线试玩