

多模态通用感知大模型的技术演进与应用探索

王玮赞

2025.06

目录

- 1** | 多模态大模型研究背景
- 2 | 多模态大模型模型结构设计
- 3 | 多模态大模型训练算法设计
- 4 | 不止于语言输出：通专融合

研究背景：大语言模型&多模态大模型

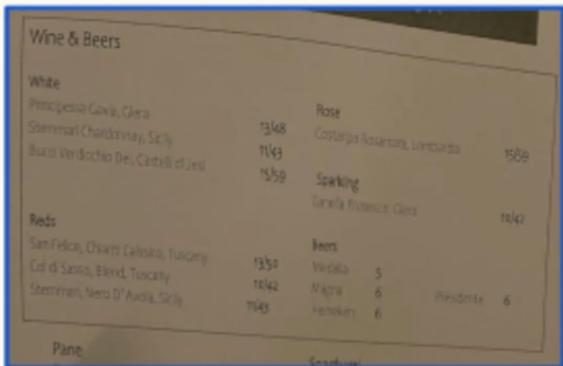
视觉问答

Prompt:

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



Here is a photo of the menu.



GPT-4V:

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

名人识别

Prompt:

Describe the image



GPT-4V:

The image shows a collage of 8 prominent personalities from different fields. On the top row, from left to right, there is singer Justin Bieber, footballer Lionel Messi, footballer Cristiano Ronaldo, and entrepreneur Elon Musk. On the bottom row, from left to right, there is model and reality TV star Kylie Jenner, singer Taylor Swift, actor Robert Downey Jr., and actor Scarlett Johansson. All individuals are shown in different poses and outfits, with various backgrounds.

场景描述

Prompt:

Describe the name of the dish.



GPT-4V:

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

Describe the name of the dish.



GPT-4V:

Unagi Donburi is the name of the dish that is originated from Japan. In Japanese cuisine, it is a type of rice bowl dish typically served in a deep bowl. The word "Unagi" means freshwater eel in Japanese and "donburi" refers to a large bowl served with rice and other toppings. Unagi Donburi is a dish made with grilled eel and a type of sweet and savory sauce called "unagi sauce" over white rice. It is often served with pickled vegetables, sesame seeds, green onions, and seaweed.

地标识别

Prompt:

Describe the landmark in the image.



GPT-4V:

The landmark in the image is the Space Needle, an observation tower located in Seattle, Washington. It stands 605 feet tall and offers panoramic views of the city and surrounding areas. The tower was built for the 1962 World's Fair and has since become a symbol of Seattle and a popular tourist attraction. The tower has a futuristic design with a large circular observation deck at the top and a restaurant that rotates 360 degrees. The base of the tower is surrounded by a park with trees and green spaces.



GPT-4V

OpenAI 研发的GPT-4v



Google研发的Gemini

文字识别

Prompt:

Count the number of apples in the image.



GPT-4V:

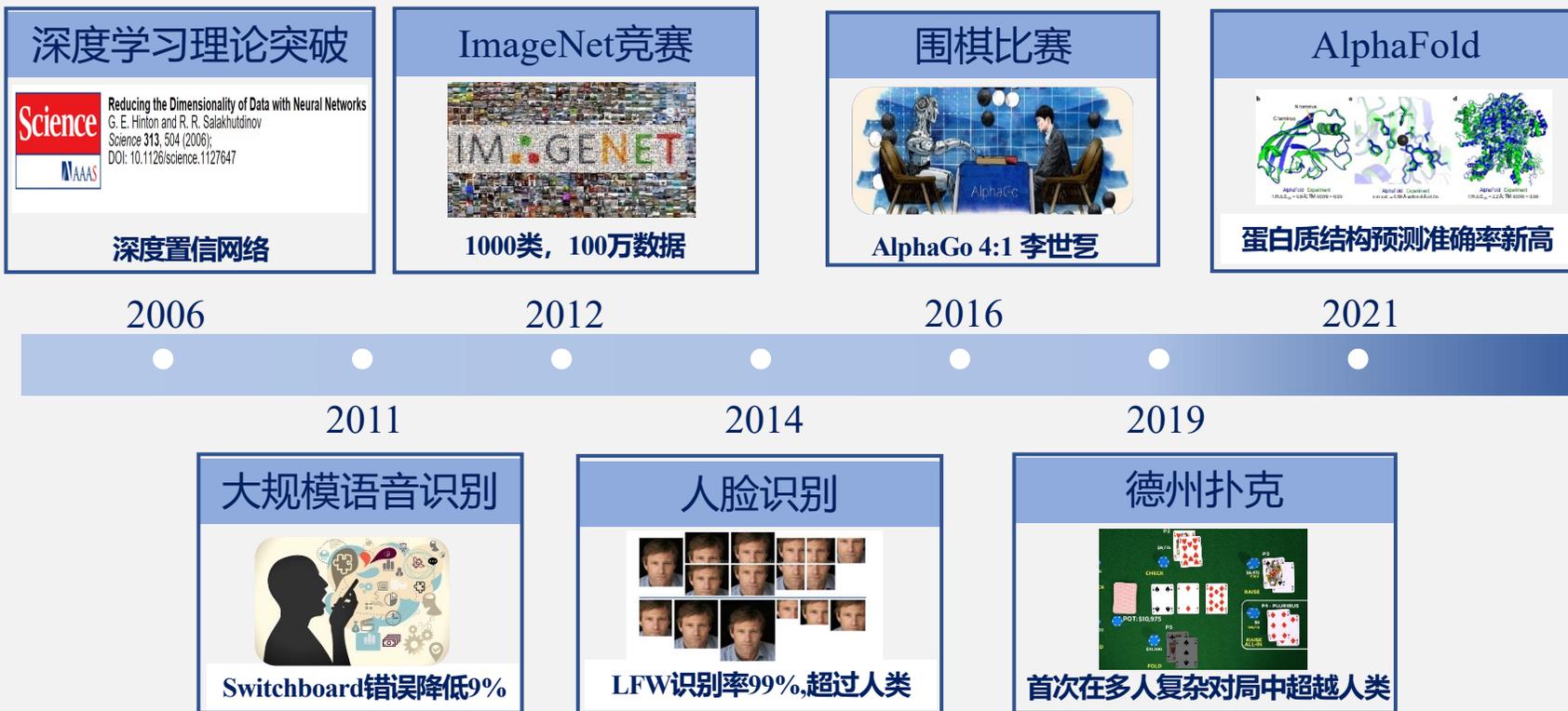
There are 12 apples in the image.

物体计数

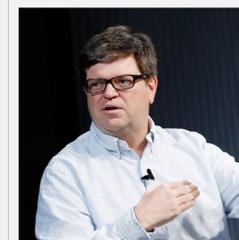
多模态感知是实现大模型与现实世界交互的必要手段

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



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

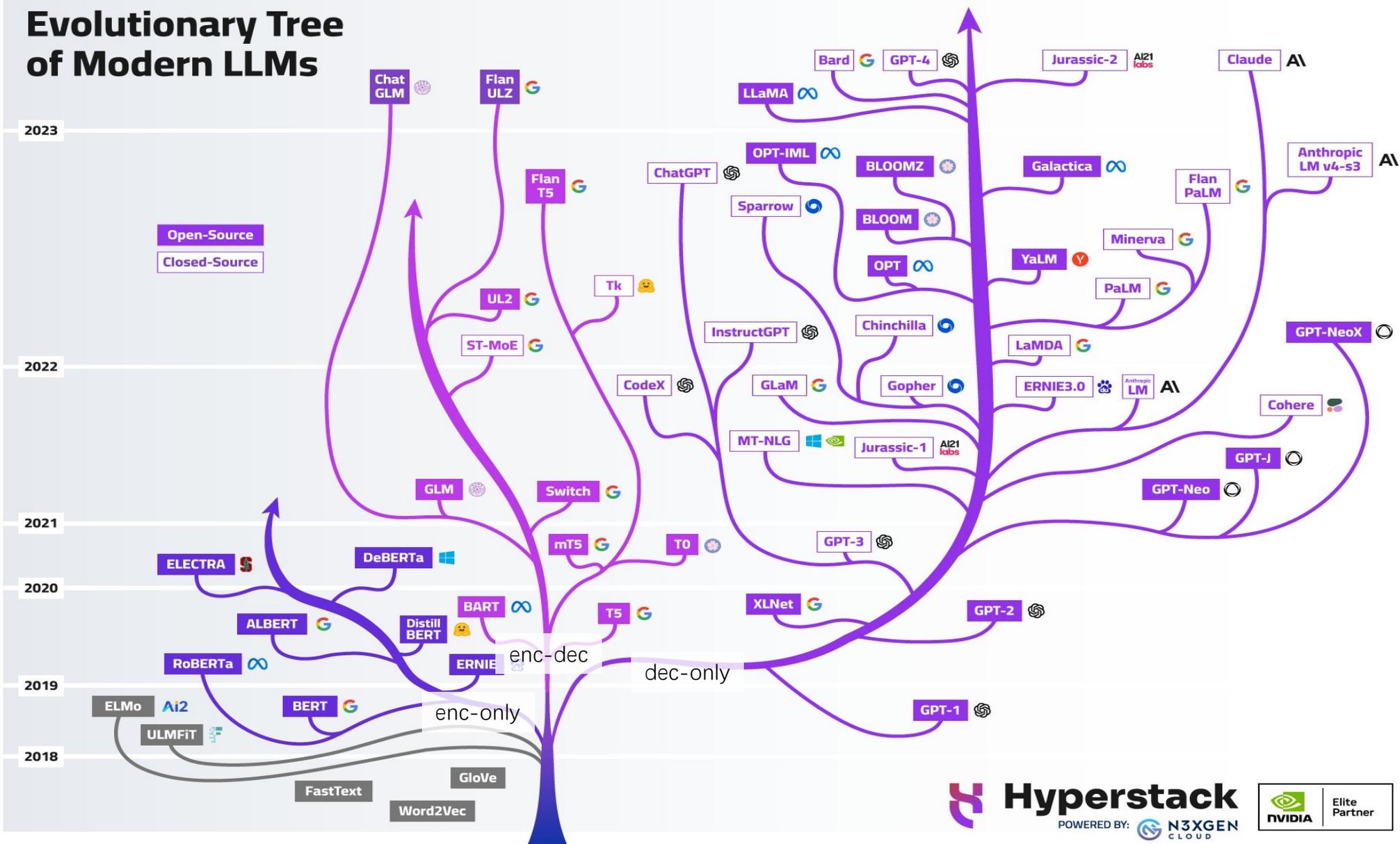
通用人工智能将会成为我们和未来科技交互的一部分。

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

多模态大模型有望在众多领域带来AI生产力革命

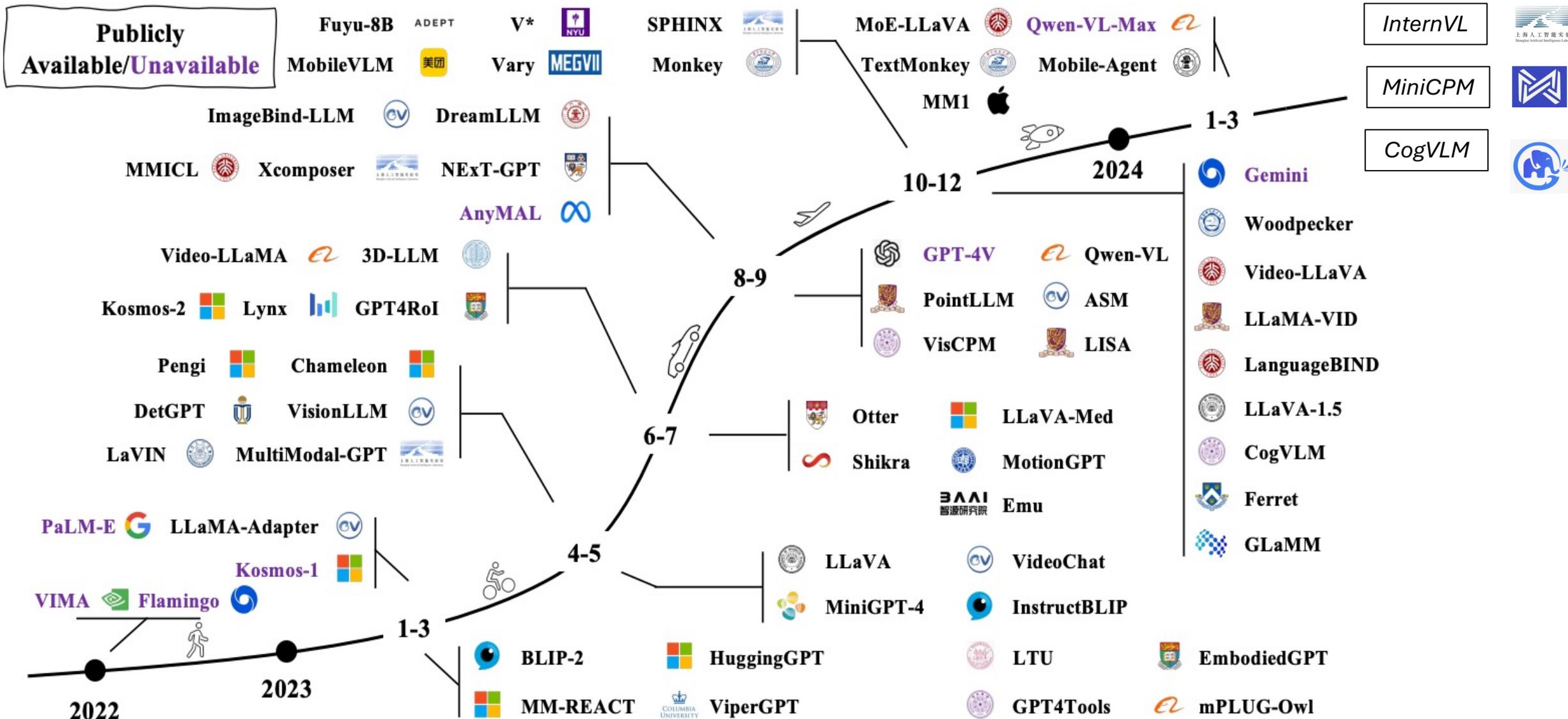
研究背景：大语言模型&多模态大模型

Evolutionary Tree of Modern LLMs



- Qwen
- DeepSeek
- InternLM
- Baichuan
- ChatGLM

研究背景：大语言模型&多模态大模型



和头部商用模型对比

- 推理及感知性能均位列开源第一!

Multi-modal Academic Leaderboard

Real-time Ranking ▾

All ▾

1	SenseNova-V6-Pro	SenseTime
2	Gemini-2.5-Pro	Google
3	InternVL3-78B	Shanghai AI Laboratory & Tsinghua University
4	InternVL3-38B	Shanghai AI Laboratory & Tsinghua University
5	Step-1o	StepFun

Multimodal Model Reasoning Leaderboard

Real-time List

All ▾

1	Doubao-1.5-Pro	ByteDance	61.6
2	Gemini-2.0-Pro	Google	56.6
3	ChatGPT-4o-latest	OpenAI	54.8
4	GPT-4.1-20250414	OpenAI	54.0
5	InternVL3-78B	Shanghai AI Laboratory & Tsinghua University	51.0

目录

1

多模态大模型研究背景

2

多模态大模型模型结构设计

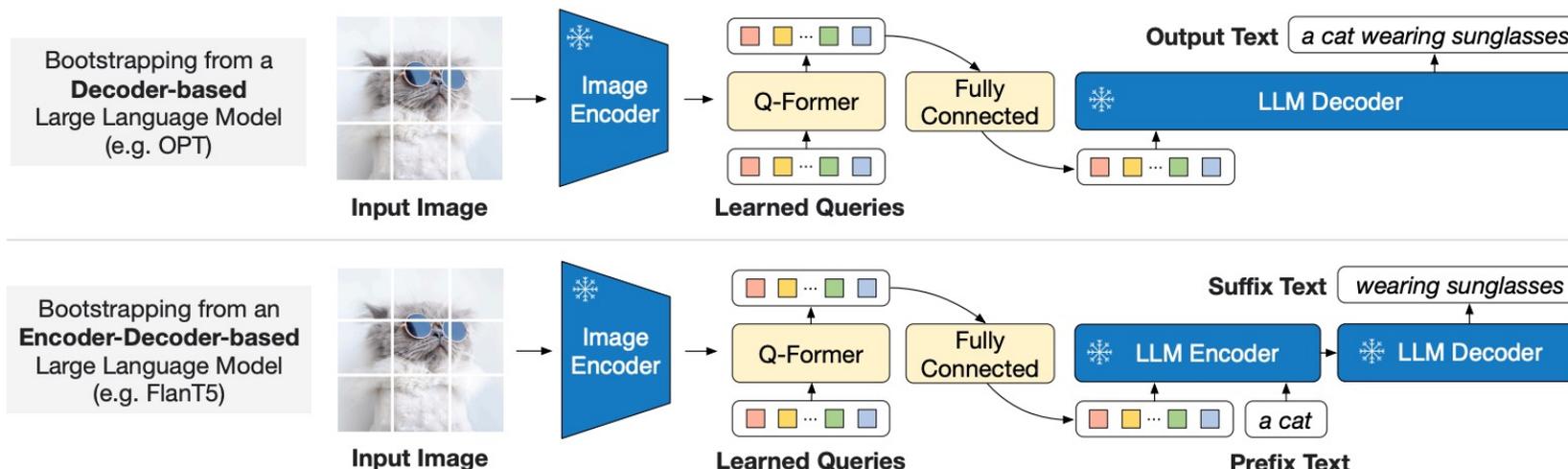
3

多模态大模型训练算法设计

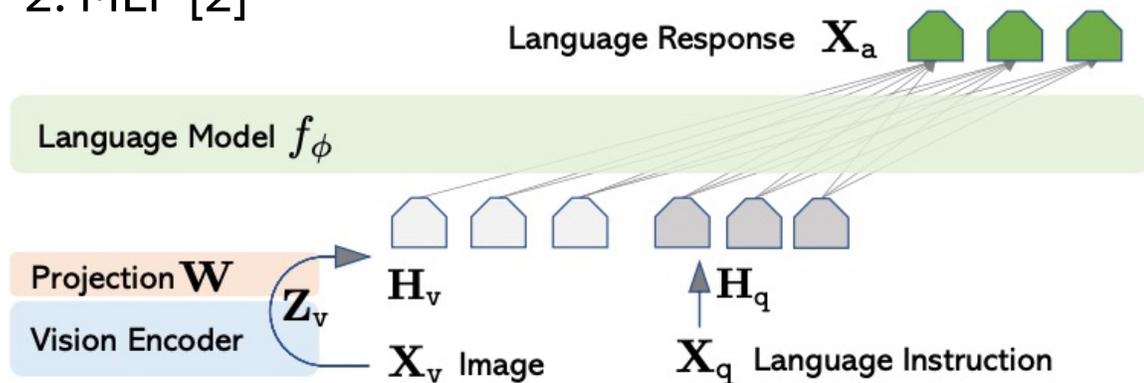
4

不止于语言输出：通专融合

1. QFormer [1]



2. MLP [2]

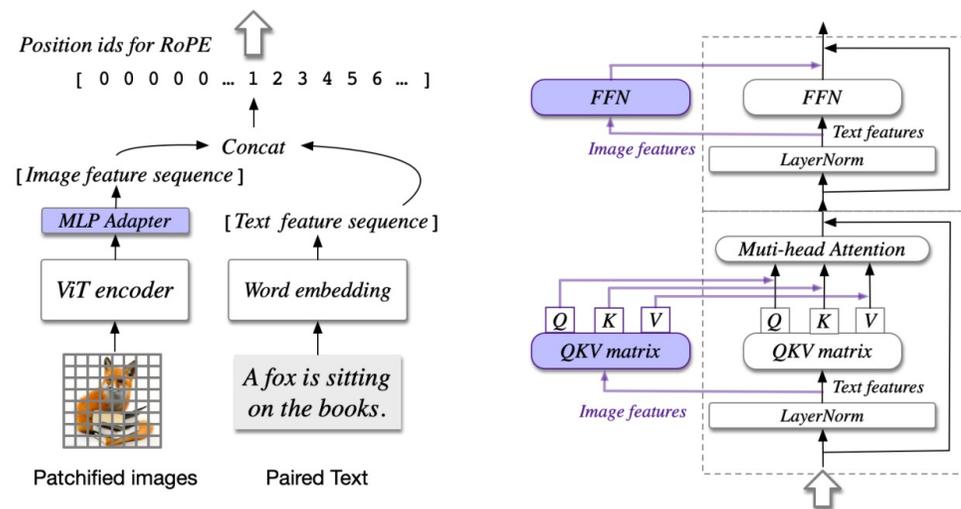


[1] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.

[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.

3. MoE [3]

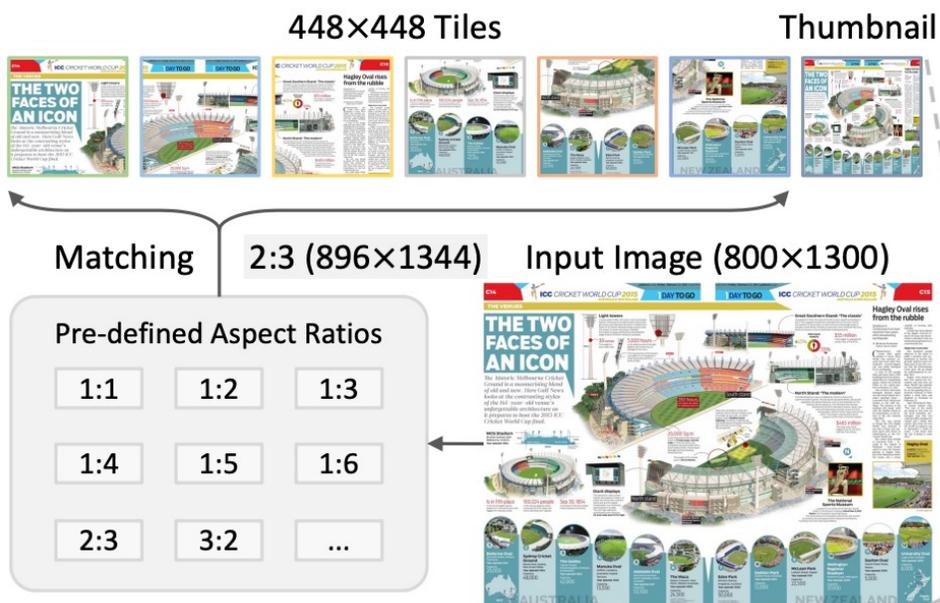


(a) The input of visual language model

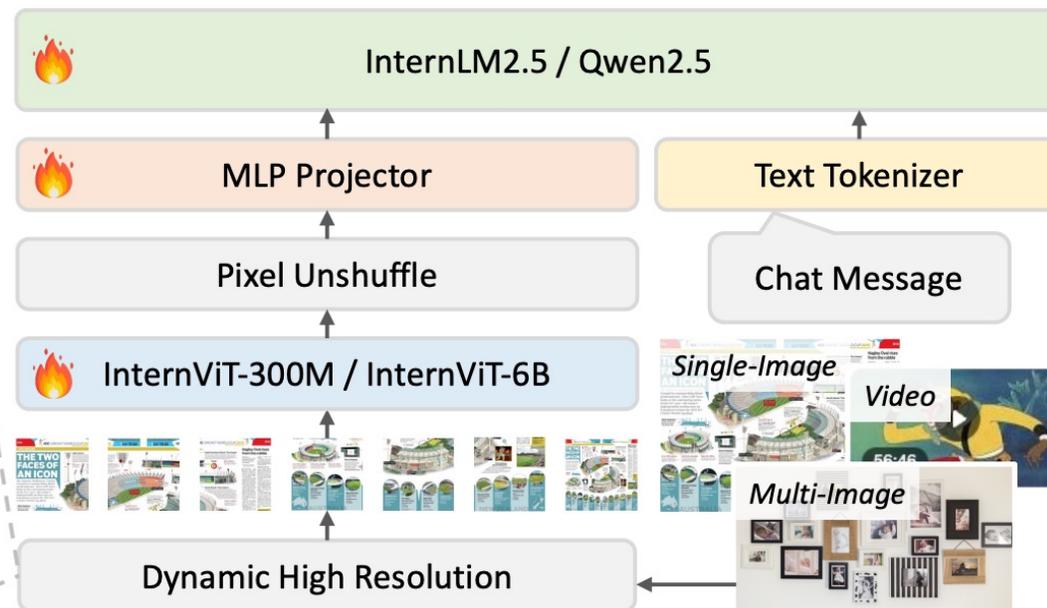
(b) The visual expert built on the language model

1. 分体式结构 (ViT-MLP-LLM) (以语言模型为核心)

- 允许单模态预训练
- 对齐阶段训练成本低, 数据复用率高 (渐进式对齐)
 - LLaVA1.5预训练数据558K, SFT数据665K
- 引入了额外的负载压力



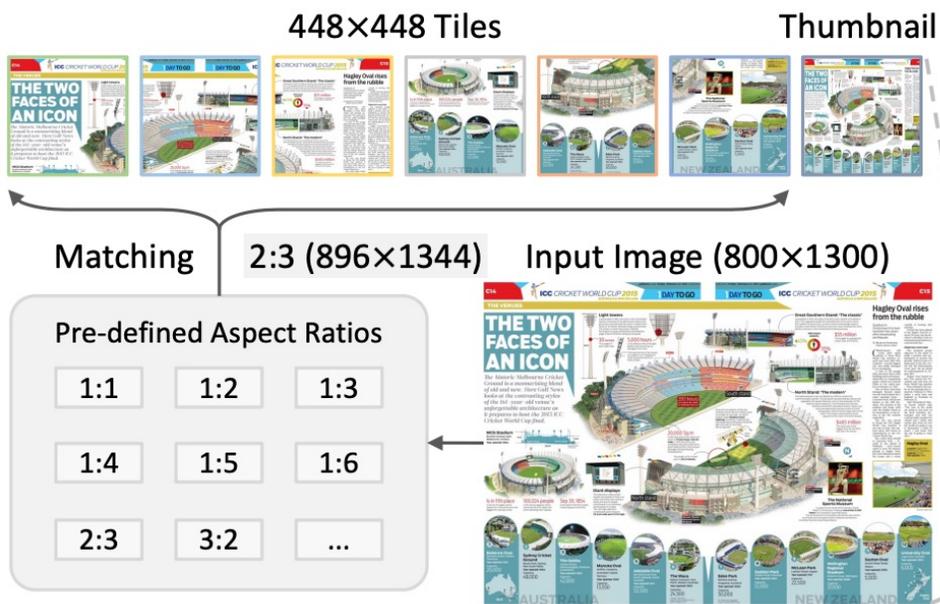
(a) Data Preprocessing



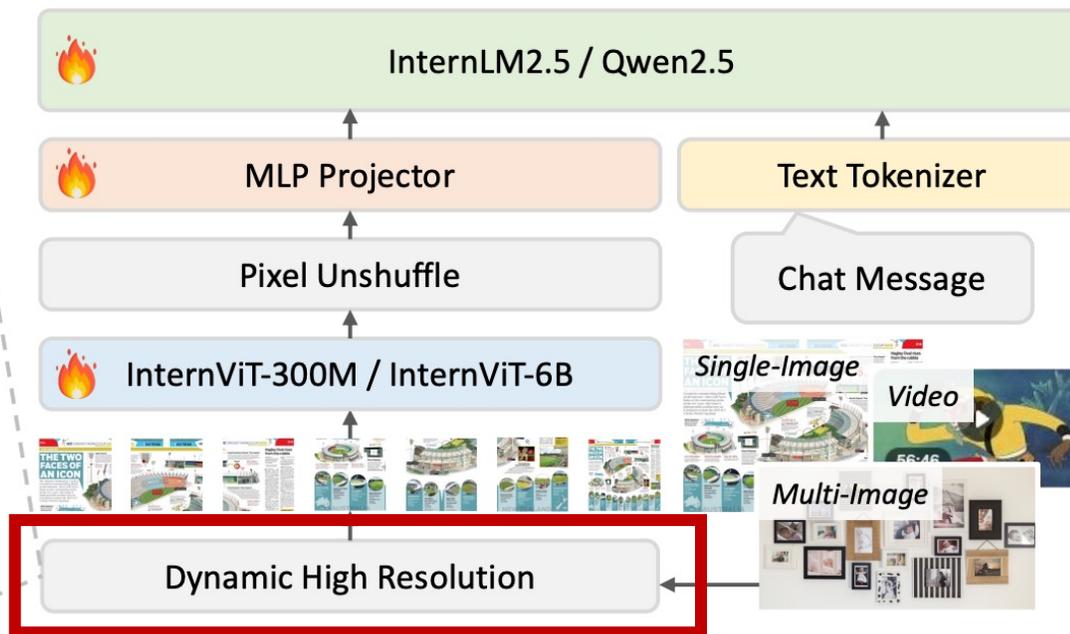
(b) Model Architecture

1. 分体式结构 (ViT-MLP-LLM)

- 动态分辨率：将一张图像切分为多个patch
- 针对不同分辨率的图像，使用不同数量的视觉token进行表征
- **在文档/OCR任务上，引入动态分辨率可以极大提升模型性能**

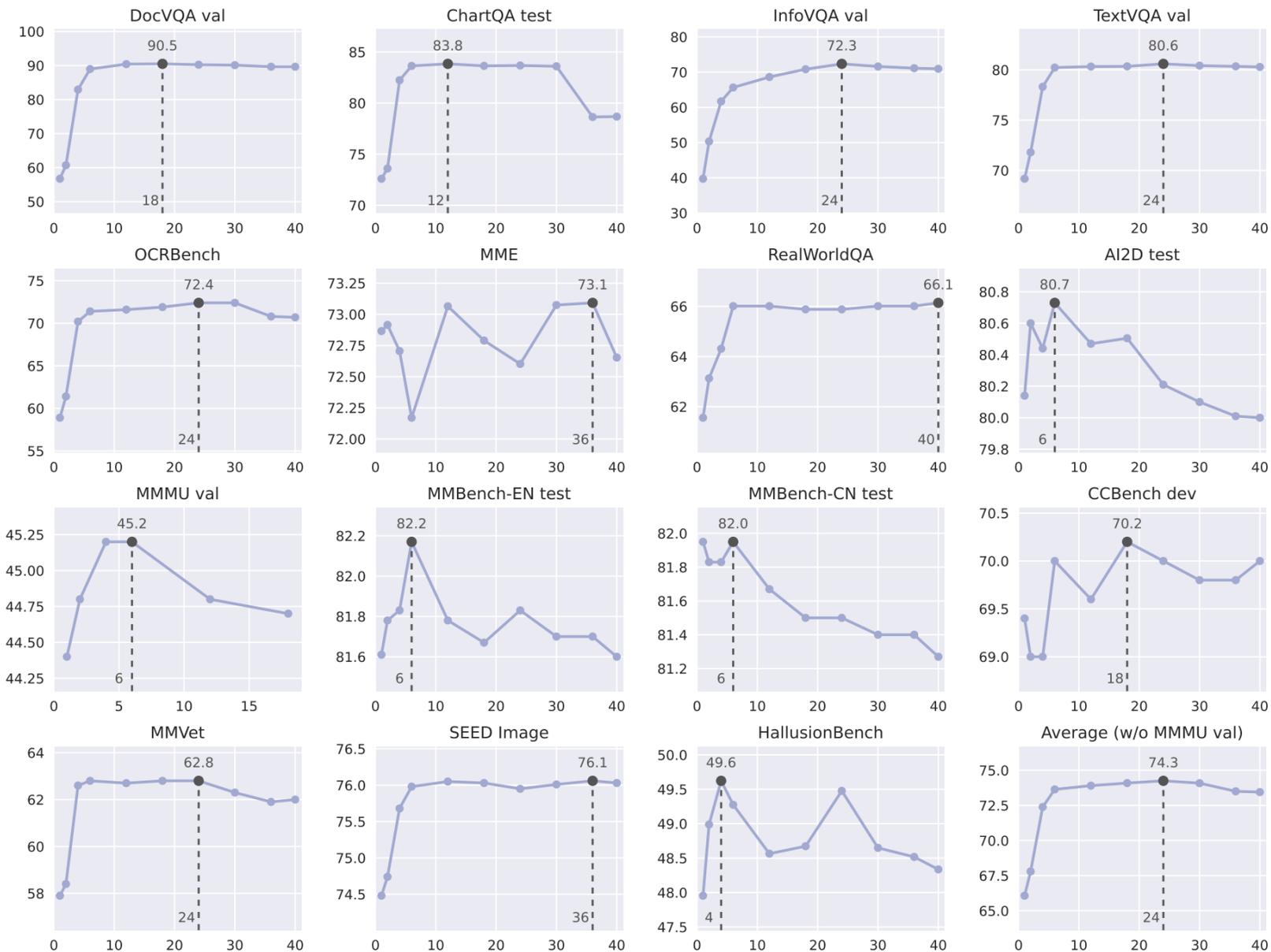


(a) Data Preprocessing



(b) Model Architecture

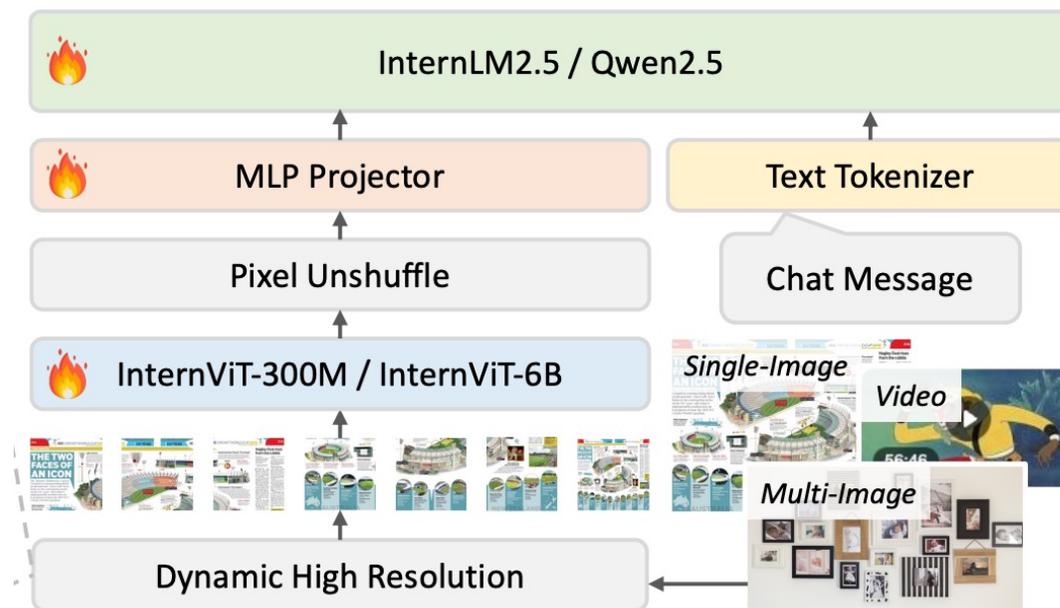
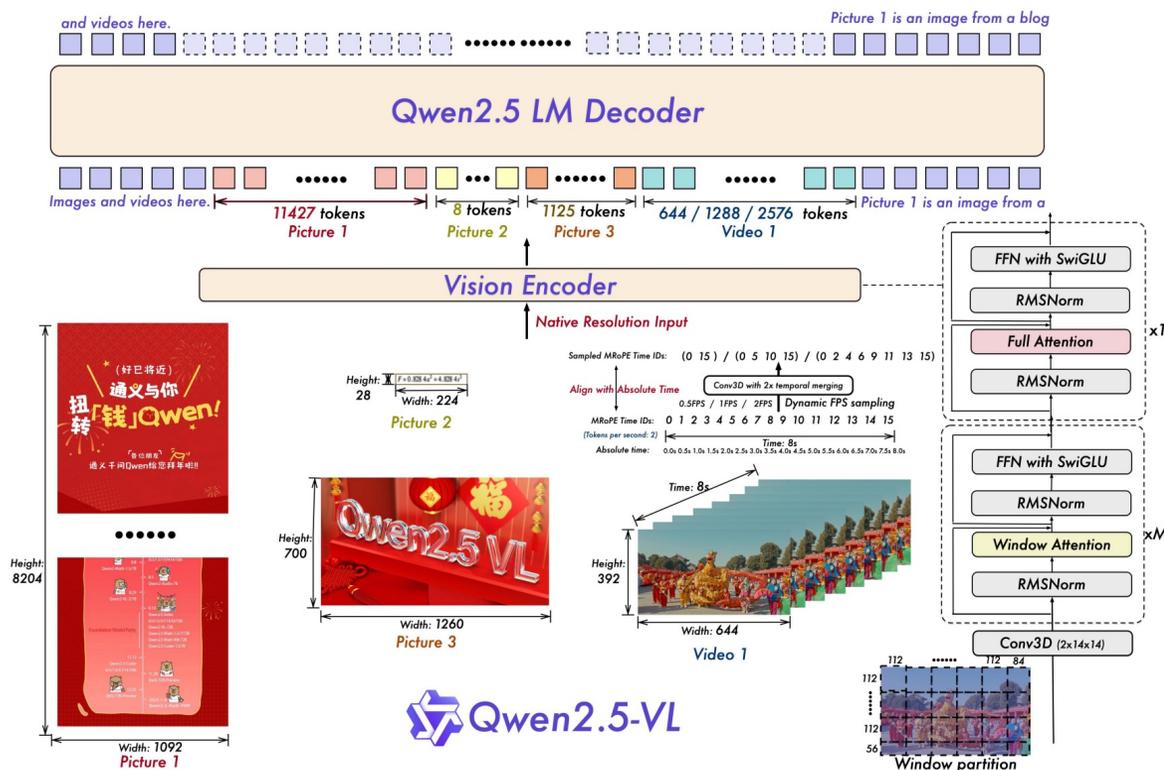
多模态大模型模型结构设计



← 分辨率对性能的影响

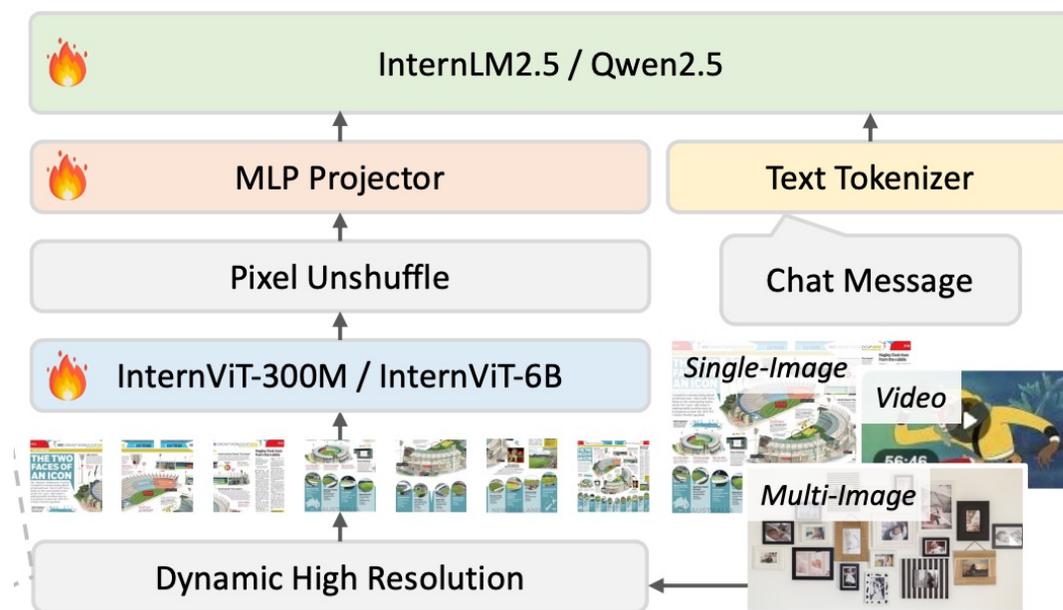
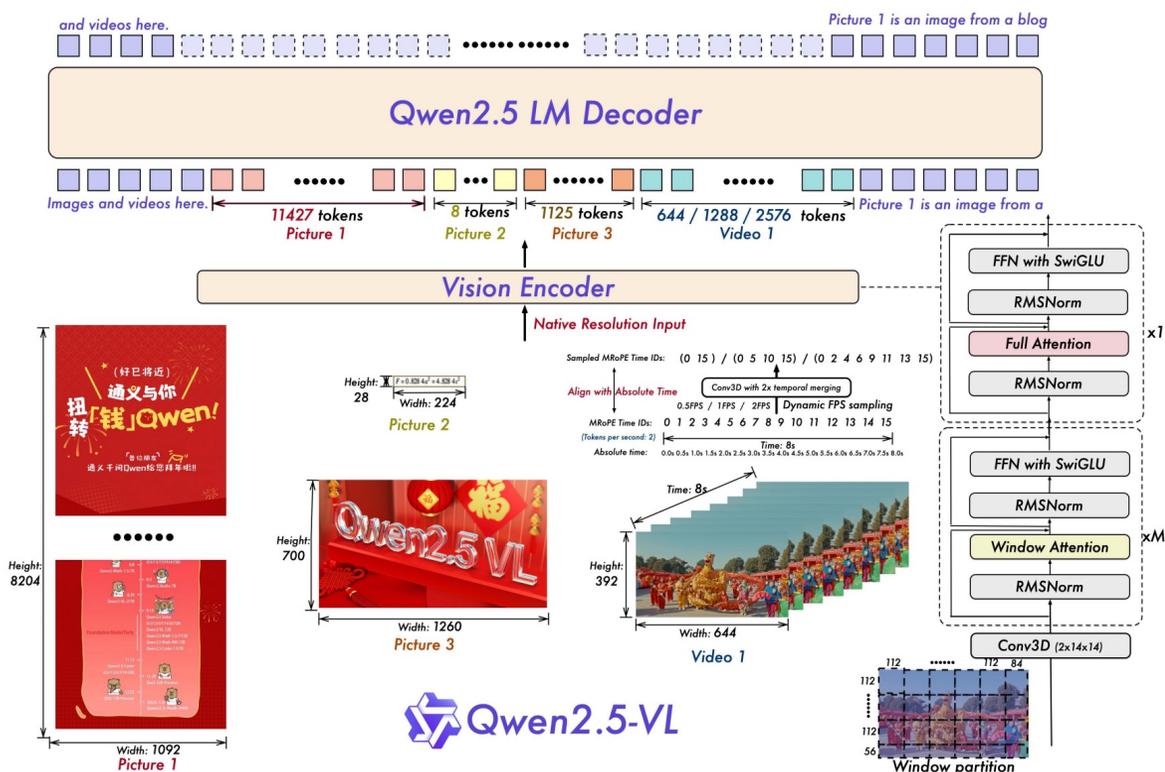
动态分辨率策略对比

- Qwen2.5-VL保持原始分辨率，维持图像尺度信息，InternVL2.5 / 3.0 仅保持图像长宽比
- Qwen的每个patch 112x112，InternVL的每个patch 448x448
- Qwen2.5-VL视觉编码引入全局注意力，InternVL2.5 / 3.0 全部为窗口注意力



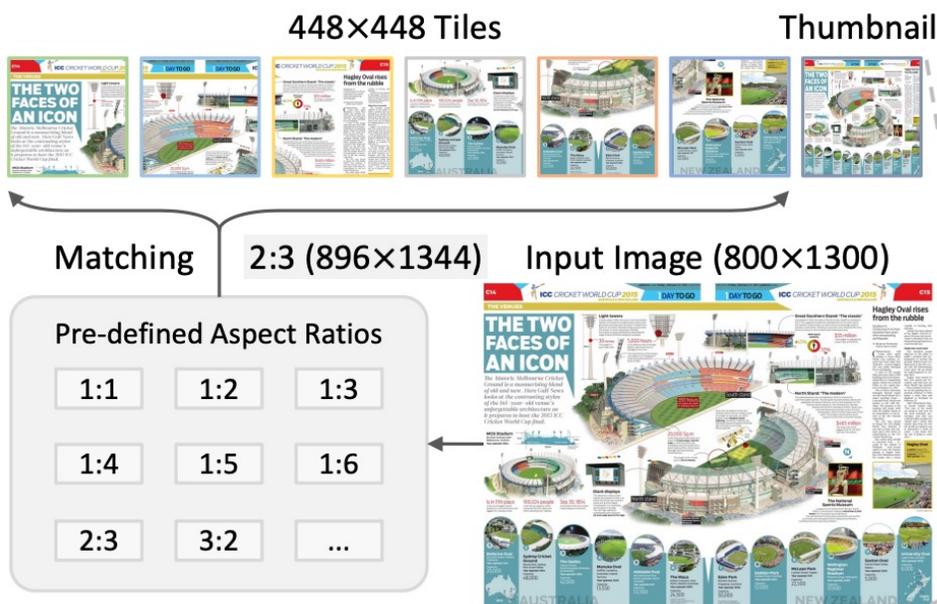
动态分辨率策略对比

- 两种策略没有本质区别，且在通用领域性能相当
- Qwen2.5-VL 与 InternVL2.5 / 3.0 的性能差异主要来自于训练数据

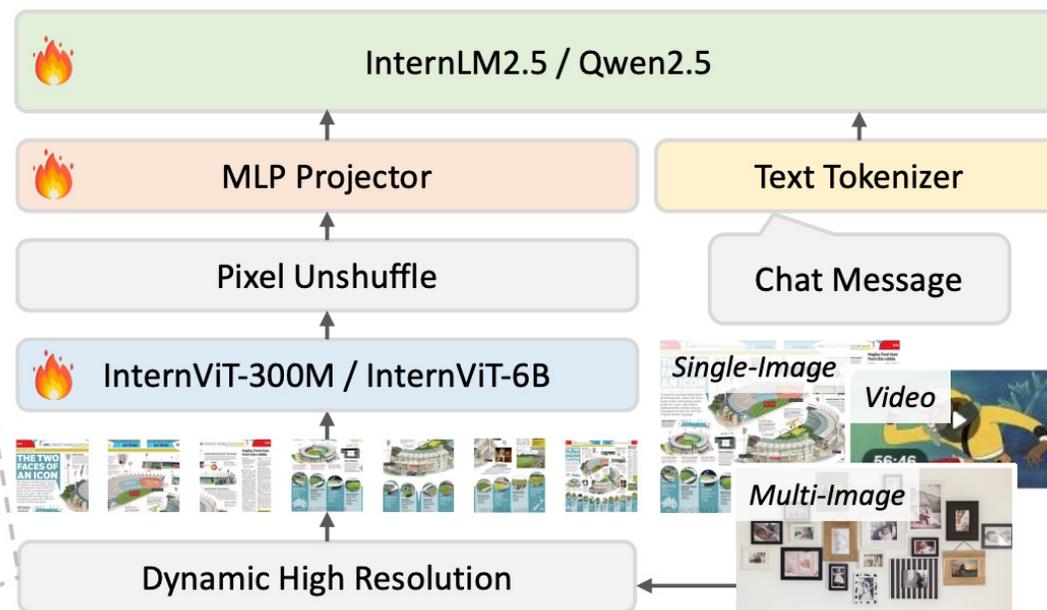


1. 分体式结构 (ViT-MLP-LLM)

- 针对ViT引入的额外的负载压力, 可以使用VL分离策略进行部署
- 推理阶段, 将ViT和LLM分别部署在不同的节点/集群
- 图像切块后, 在patch-level进行负载均衡 (不能在ViT阶段引入全局注意力)



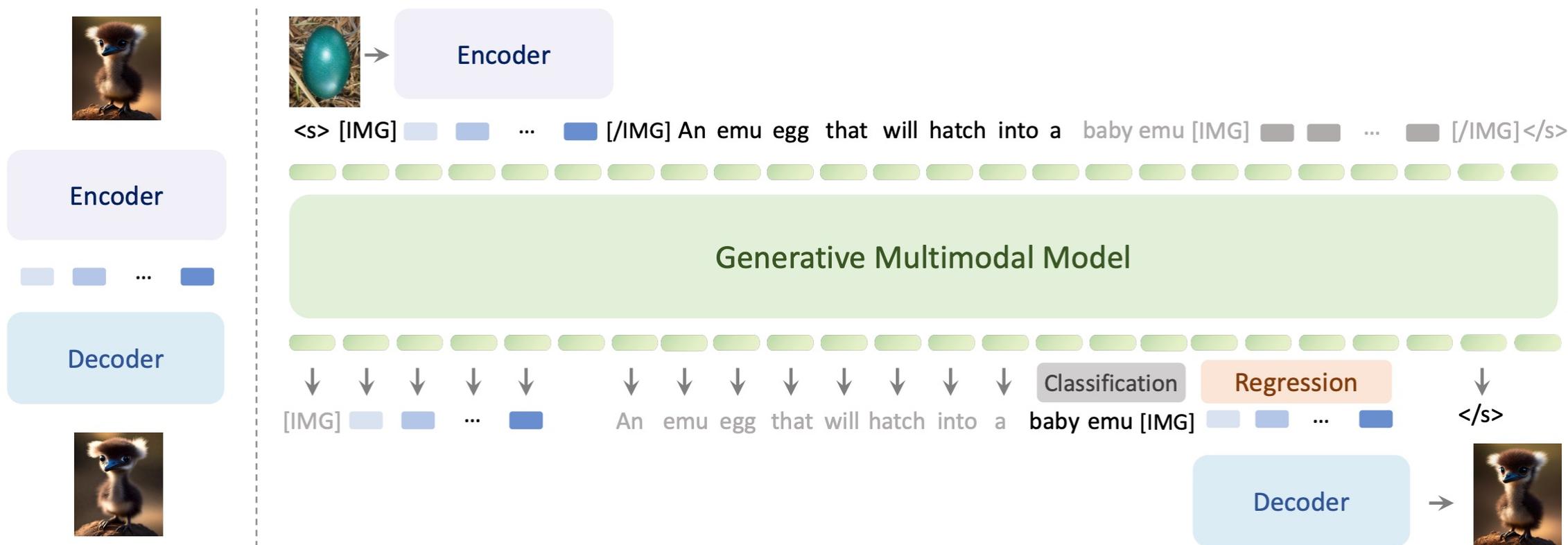
(a) Data Preprocessing



(b) Model Architecture

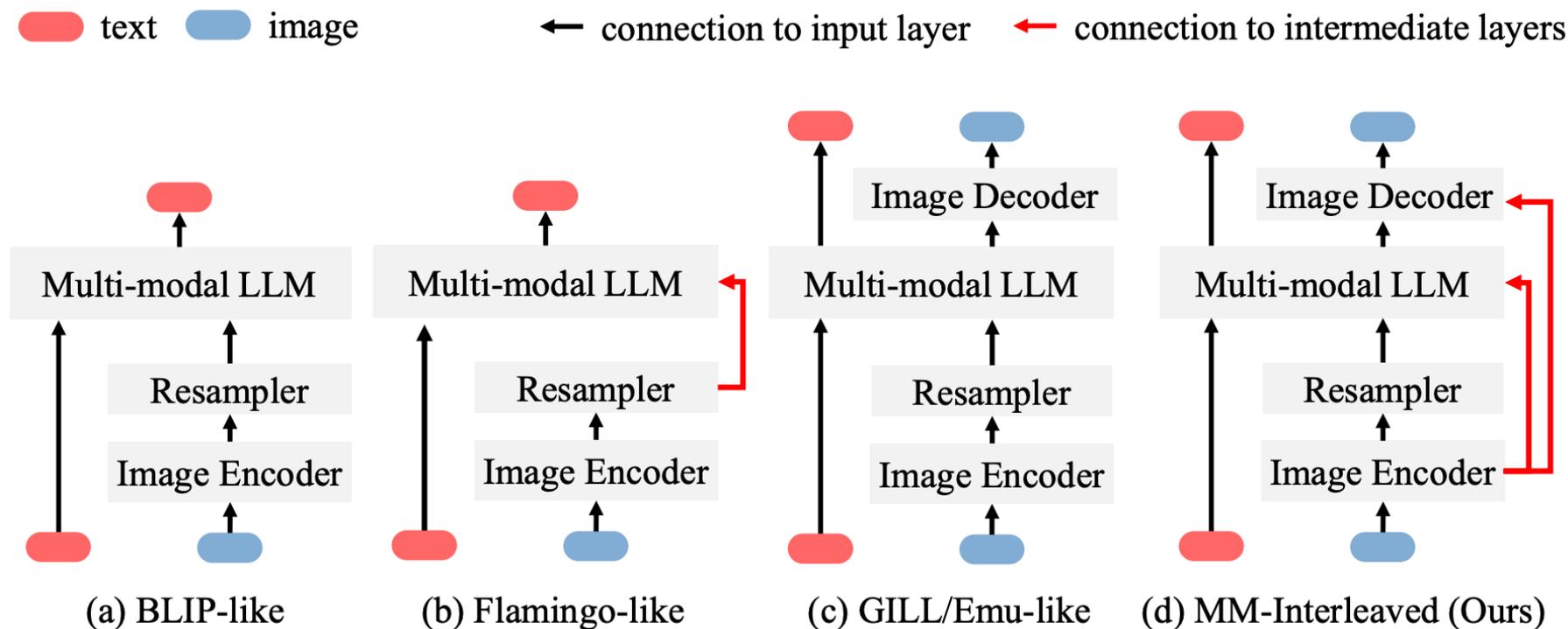
1. 分体式结构 (ViT-MLP-LLM)

- 在LLM前侧接入ViT, 可以支持图像理解
- 在LLM后侧接入Diffusion Model, 可以支持图像生成
- 都是将图像建模为一种特殊的语言, ViT/DM作为Tokenizer的Encoder/Decoder



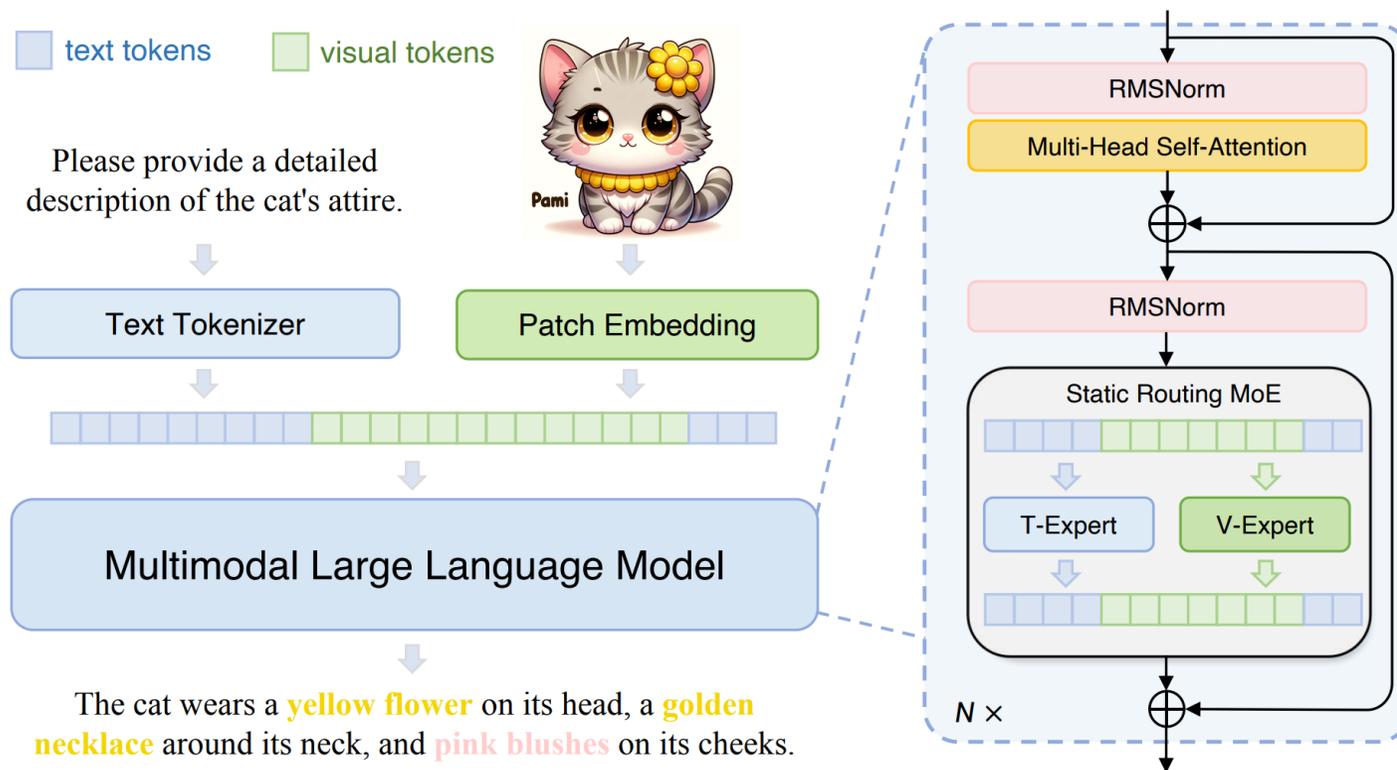
1. 分体式结构 (ViT-MLP-LLM + Cross Attn)

- 代表性工作: Flamingo、LLaMA-V
- Cross Attn部分的参数量极大 (收敛更慢, 但是性能上限没有提升)
- 参数量应该集中在ViT/LLM, 不要让Connector喧宾夺主



2. 一体式结构

- 引入基于静态路由的MoE结构，在FFN层不同模态使用不同MLP进行编码
- 目前无法引入预训练的视觉编码器，因此数据复用率低



2. 一体式结构

- 基于InternVL2的数据进行融合训练，分体式结构的性能优于一体式结构

Model	#A-Param	MMB	MMVet	MMMU	MME	MathVista	SEED-I	OCRBench	HallB	CCB	Avg _{MM}	TextVQA	SQA-I	GQA	DocVQA	AI2D	ChartQA	InfoVQA	Avg _{QA}
▼ Modular MLLMs:																			
MobileVLM-V2-3B [20]	3.0B	63.2	-	-	-	-	-	-	-	-	-	57.5	70.0	66.1	-	-	-	-	-
Mini-Gemini-2B [52]	3.5B	59.8	31.1	31.7	1653	29.4	-	-	-	-	-	56.2	-	-	34.2	-	-	-	-
MM1-3B-MoE-Chat [74]	3.5B	70.8	42.2	38.6	1772	32.6	69.4	-	-	-	-	72.9	76.1	-	-	-	-	-	-
DeepSeek-VL-1.3B [65]	2.0B	64.6	34.8	32.2	1532	31.1	66.7	409	27.6	37.6	43.4	57.8	-	-	-	51.5	-	-	-
PaliGemma-3B [7]	2.9B	71.0	33.1	34.9	1686	28.7	69.6	614	32.2	29.6	46.7	68.1	-	-	-	68.3	-	-	-
MiniCPM-V-2 [107]	2.8B	69.1	41.0	38.2	1809	38.7	67.1	605	36.1	45.3	51.2	74.1	-	-	71.9	62.9	-	-	-
†InternVL-1.5-2B [18]	2.2B	70.9	39.3	34.6	1902	41.1	69.8	654	37.5	63.5	54.4	70.5	84.9	61.6	85.0	69.8	74.8	55.4	71.7
Qwen2VL-2B [101]	2.1B	74.9	49.5	41.1	1872	43.0	-	809	41.7	-	-	79.7	-	-	90.1	74.7	73.5	65.5	-
▼ Monolithic MLLMs:																			
Fuyu-8B (HD) [6]	8B	10.7	21.4	-	-	-	-	-	-	-	-	-	-	-	-	64.5	-	-	-
SOLO [16]	7B	-	-	-	1260	34.4	64.4	-	-	-	-	-	73.3	-	-	61.4	-	-	-
Chameleon-7B ¹ [12]	7B	31.1	8.3	25.4	170	22.3	30.6	7	17.1	3.5	16.1	4.8	47.2	-	1.5	46.0	2.9	5.0	17.9
EVE-7B [26]	7B	49.5	25.6	32.3	1483	25.2	61.3	327	21.1	12.4	34.8	51.9	63.0	60.8	22.0	48.5	19.5	20.0	40.8
EVE-7B (HD) [26]	7B	52.3	25.7	32.6	1628	34.2	64.6	398	26.4	16.3	38.9	56.8	64.9	62.6	53.0	61.0	59.1	25.0	54.6
Emu3 [104]	8B	58.5	37.2	31.6	-	-	68.2	687	-	-	-	64.7	89.2	60.3	76.3	70.0	68.6	43.8	67.6
Mono-InternVL-2B	1.8B	65.5	40.1	33.7	1875	45.7	67.4	767	34.8	66.3	55.2	72.6	93.6	59.5	80.0	68.6	73.7	43.0	70.1

name	MMMU (val)	MathVista (testmini)	AI2D (test)	ChartQA (test)	DocVQA (test)	InfoVQA (test)	OCRBench	MMB-EN (test)	MMB-CN (test)	OpenCompass (avg score)
InternVL2-1B	35.4 / 36.7	37.7	64.1	72.9	81.7	50.9	754	65.4	60.7	48.3
InternVL2-2B	34.3 / 36.3	46.3	74.1	76.2	86.9	58.9	784	73.2	70.9	54.0
InternVL2-4B	47.0 / 48.3	58.6	78.9	81.5	89.2	67.0	788	78.6	73.9	60.6
InternVL2-8B	49.3 / 51.2	58.3	83.8	83.3	91.6	74.8	794	81.7	81.2	64.1
InternVL2-26B	48.3 / 50.7	59.4	84.5	84.9	92.9	75.9	825	83.4	82.0	66.4
InternVL2-40B	53.9 / 55.2	63.7	87.1	86.2	93.9	78.7	837	86.8	86.5	69.7
InternVL2-Llama3-76B	55.2 / 58.2	65.5	87.6	88.4	94.1	82.0	839	86.5	86.3	71.0
InternVL2-Pro	58.9 / 62.0	66.3	87.3 / 96.0	87.1	95.1	83.3	837	87.8	87.2	71.8

1. 分体式结构 (BLIP、LLaVA、InternVL、QwenVL)
 - 允许单模态预训练
 - 对齐阶段训练成本低, 数据复用率高, 允许快速迭代
 - 引入了额外的负载压力
 2. 一体式结构 (BEiT-3、CogVLM、EVE、MonoInternVL)
 - 结构更优雅, (理论上) 更少的负载压力
 - 数据复用率低
- 目前分体式结构在最终性能上有一定优势

目录

1

多模态大模型研究背景

2

多模态大模型模型结构设计

3

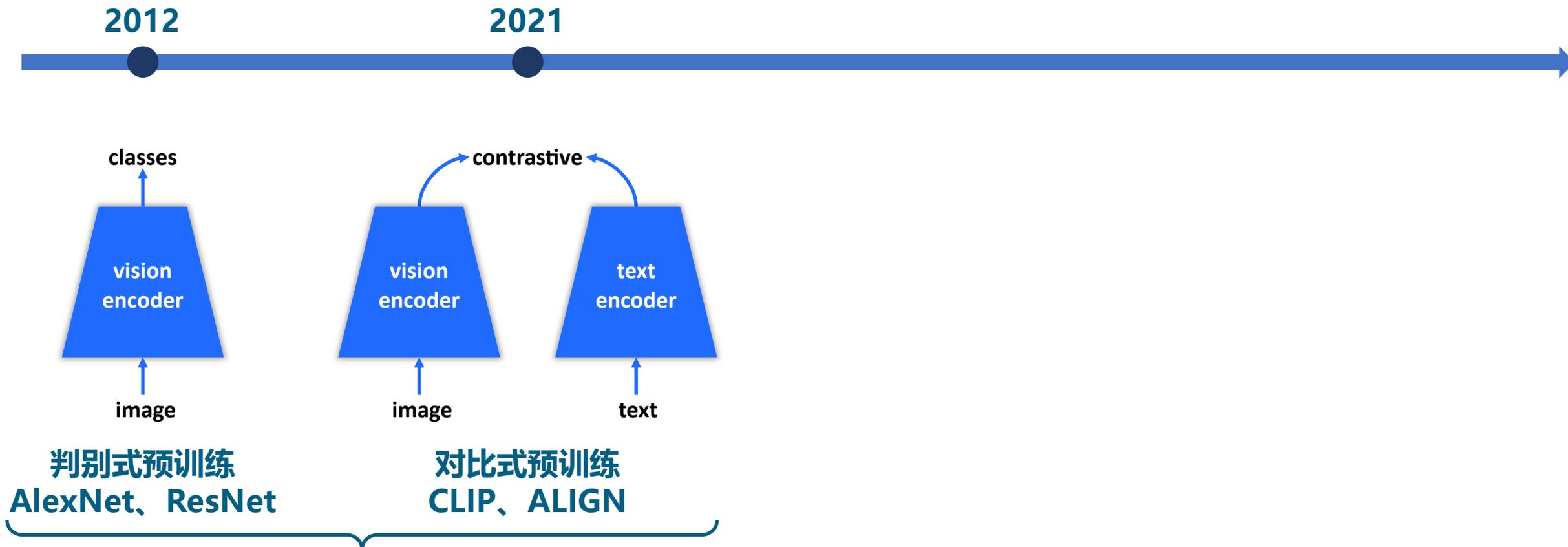
多模态大模型训练算法设计

4

不止于语言输出：通专融合

1. (大语言模型) 预训练/指令微调/后训练
 - LLaMA、Qwen、InternLM
2. (视觉编码器) 预训练
 - CLIP、InternViT、SigLIP、DINO
3. (多模态大模型) **冷启动 (optional)**
4. (多模态大模型) **融合预训练**
 - InternVL2.5
5. (多模态大模型) **后训练 (SFT+MPO)**
 - InternVL2.5-MPO
6. (多模态大模型) **测试阶段增强**
 - VisualPRM

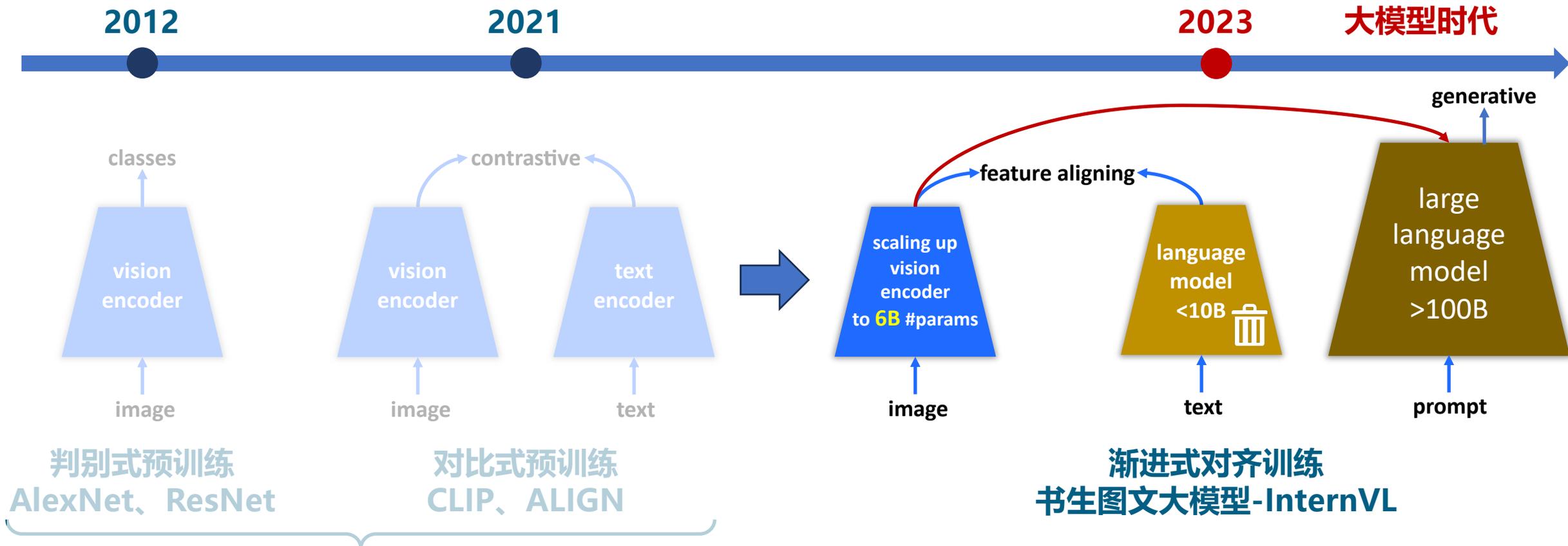
传统视觉/视觉-语言基础模型范式已落后于大语言模型的发展，亟需新的范式来推动其发展



- 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

多模态大模型训练算法设计 (视觉编码器预训练)

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

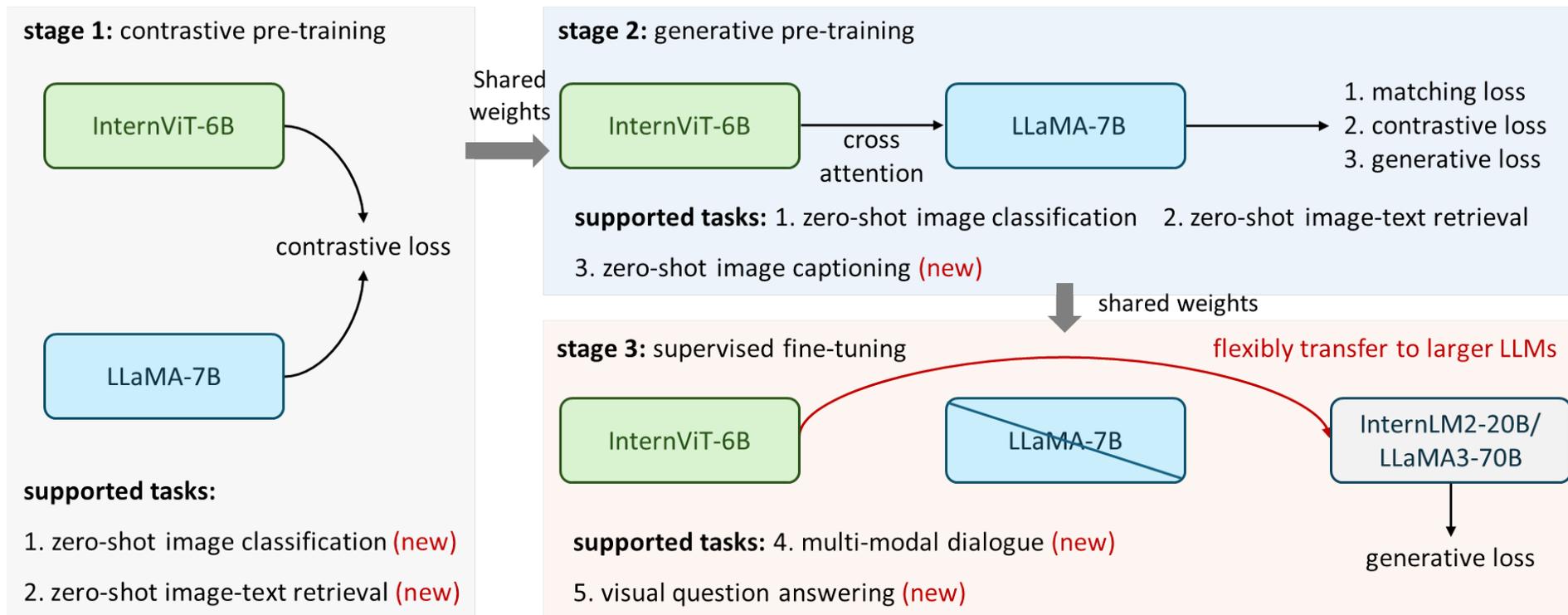


- 与LLM参数量差距过大
- 与LLM表征不一致
- 训练数据单一、数据量小

- 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

多模态大模型训练算法设计（视觉编码器预训练）

核心思想： 训练一个可以抽取包含（隐式）文本信息的视觉编码器



设计1：扩大视觉模型至6B参数

步骤1：对比学习（CLIP-like）

步骤2：对比学习与生成式训练（BLIP-like）

步骤3：生成式训练（LLaVA-like）

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

多模态大模型训练算法设计 (冷启动)

核心思想： 使得视觉编码器抽取的视觉表征可以被大语言模型理解

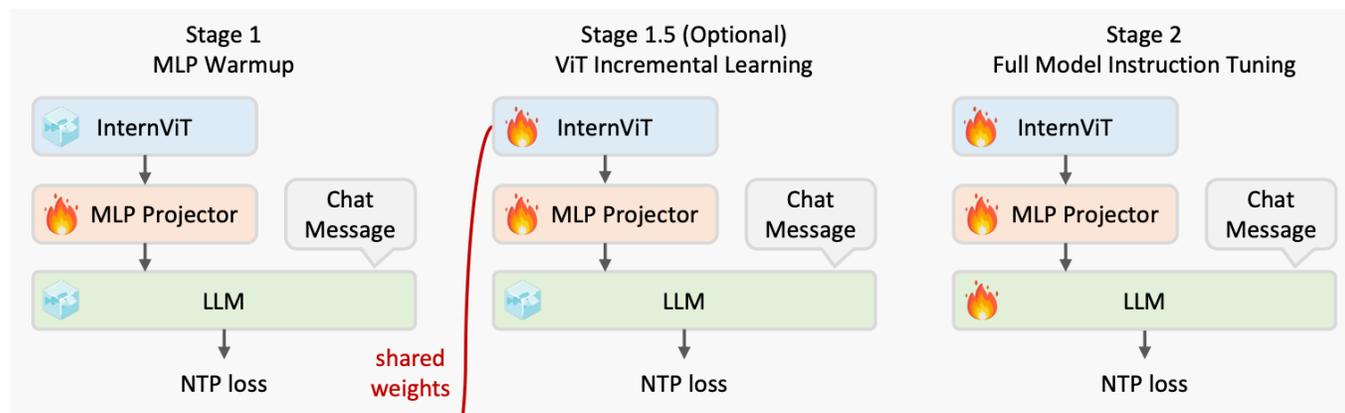
设计2：渐进式的图像-文本对齐策略

阶段1：训练MLP

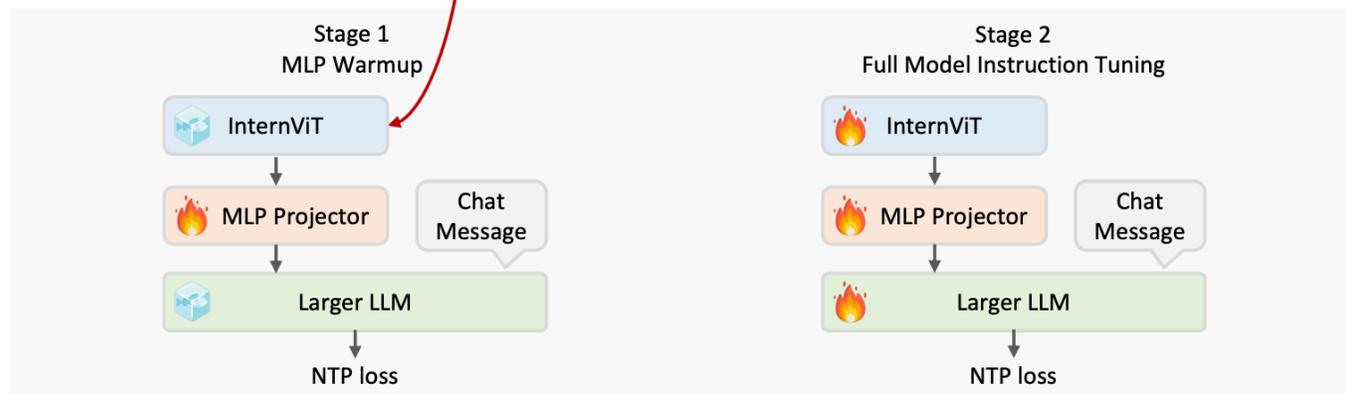
阶段1.5：训练MLP + ViT

阶段2：训练MLP + ViT + LLM

注：跳过阶段1会导致模型收敛显著变差



(a) Single Model Training Pipeline



(b) Progressive Scaling Strategy

核心思想： 使得视觉编码器抽取的视觉表征可以被大语言模型理解

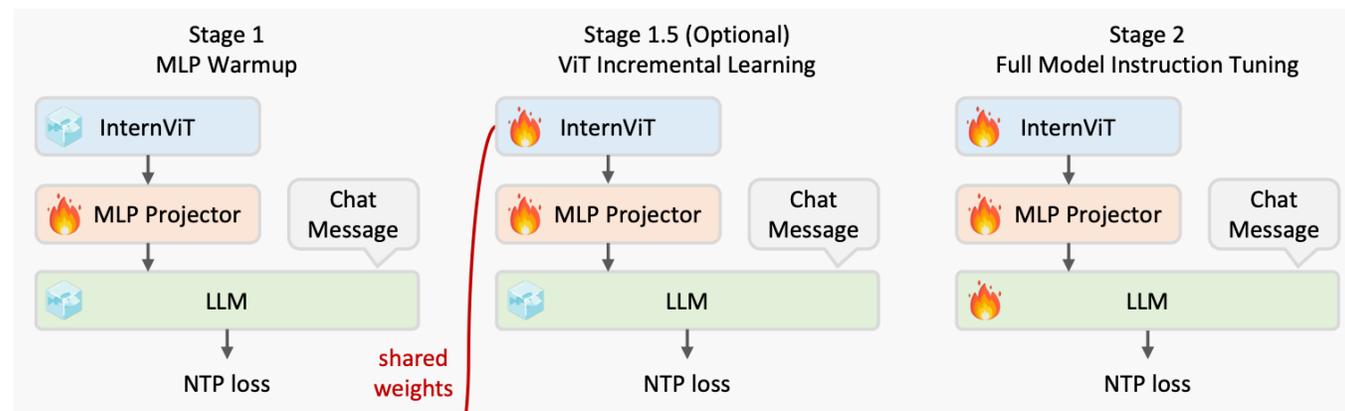
设计2：渐进式的图像-文本对齐策略

阶段1：训练MLP

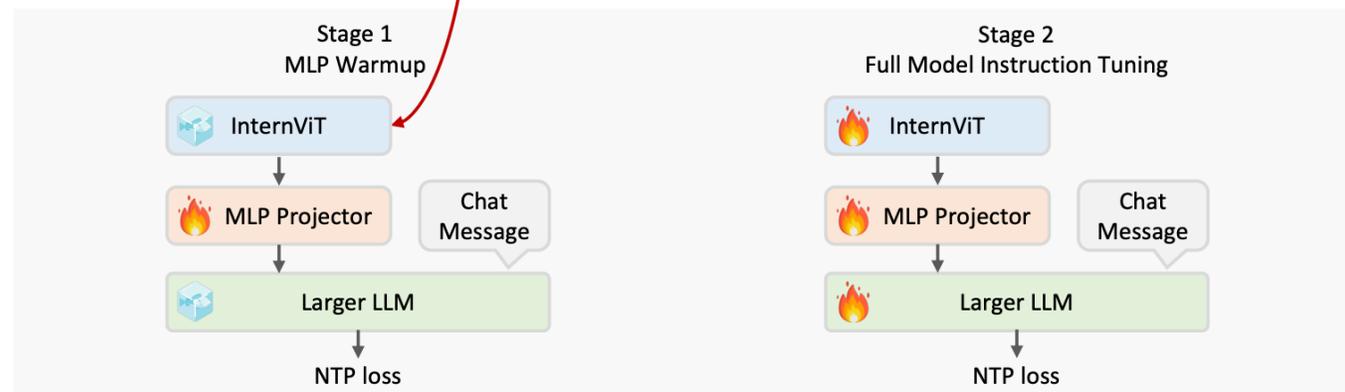
阶段1.5：训练MLP + ViT

阶段2：训练MLP + ViT + LLM

- 阶段1.5训练得到的 ViT 可以接给其他 MLP + LLM，只进行阶段1和3的训练



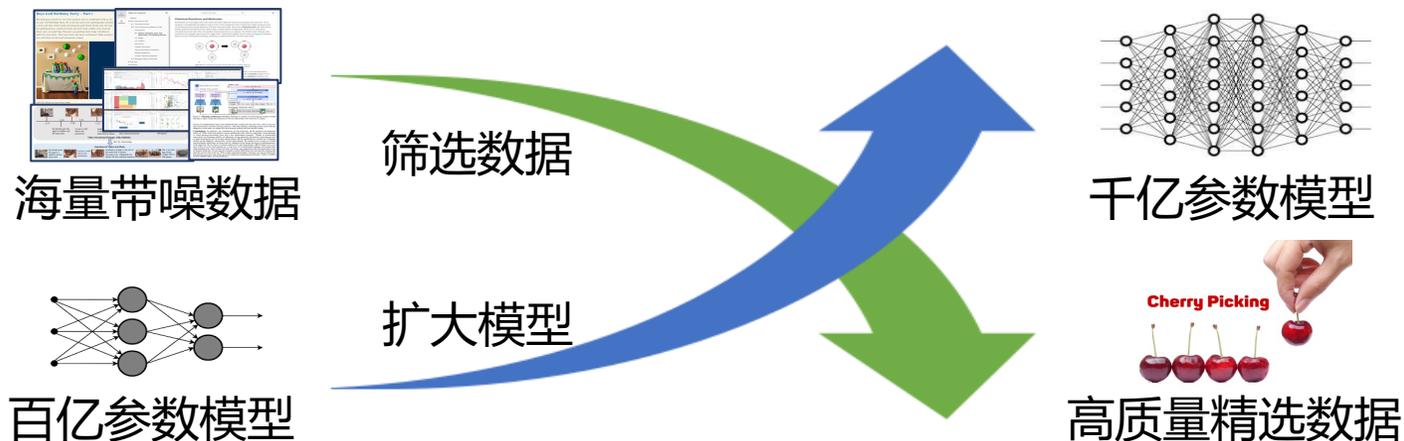
(a) Single Model Training Pipeline



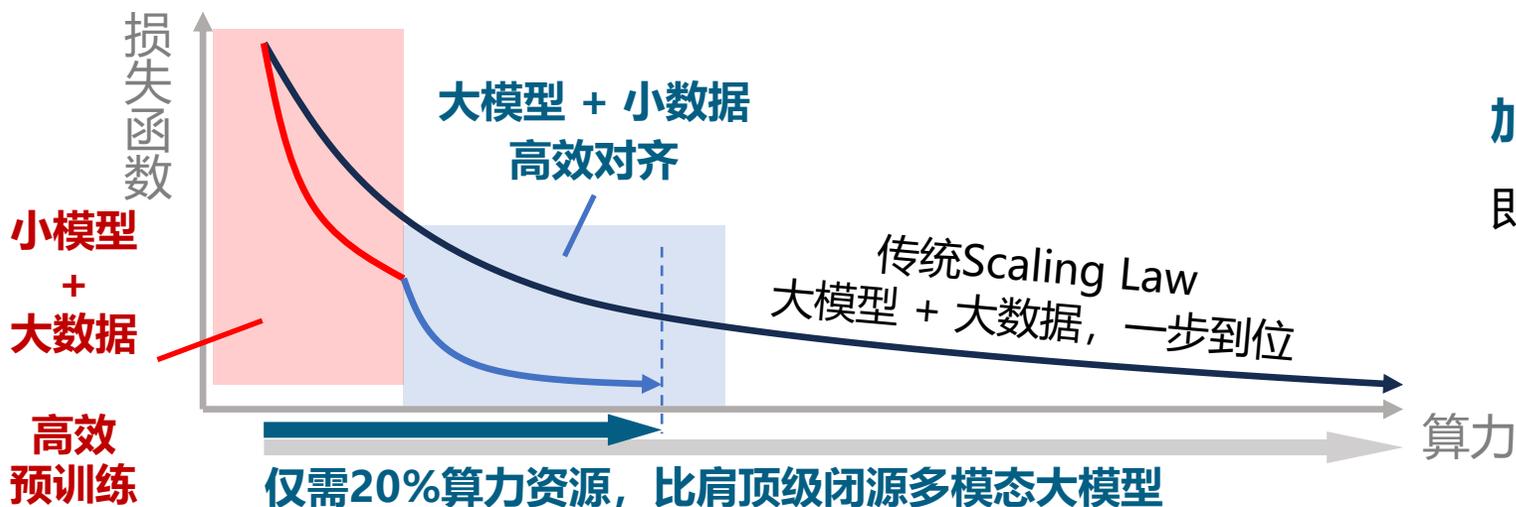
(b) Progressive Scaling Strategy

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



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



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

核心思想：使得视觉编码器抽取的视觉表征可以被大语言模型理解

设计2：渐进式的图像-文本对齐策略（InternVL2.5）

阶段1：训练MLP

（含噪声的图文预训练数据数据）

阶段1.5：训练MLP + ViT

（含噪声的图文预训练数据数据）

阶段2：训练MLP + ViT + LLM

（高质量图文对话数据）

设计2：渐进式的图像-文本对齐策略（InternVL3）

阶段1：训练LLM

（含噪声的纯文本预训练数据数据）

阶段2：训练MLP + ViT + LLM

（含噪声的图文预训练数据数据 + 纯文本数据）

阶段3：训练MLP + ViT + LLM

（高质量图文对话数据 + 纯文本对话数据）

设计2: 渐进式的图像-文本对齐策略

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, 大语言模型直接打分等

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.

设计3: Random JPEG Compression (InternVL2.5 Section 3.4)

设计4: Loss Reweighting (InternVL2.5 Section 3.4)

$$\mathcal{L} = \frac{w_i}{\sum_j w_j} \cdot \mathcal{L}_i, \quad w_i = \begin{cases} \frac{1}{x^0}, & \text{for token averaging} \\ \frac{1}{x^1}, & \text{for sample averaging,} \end{cases}$$

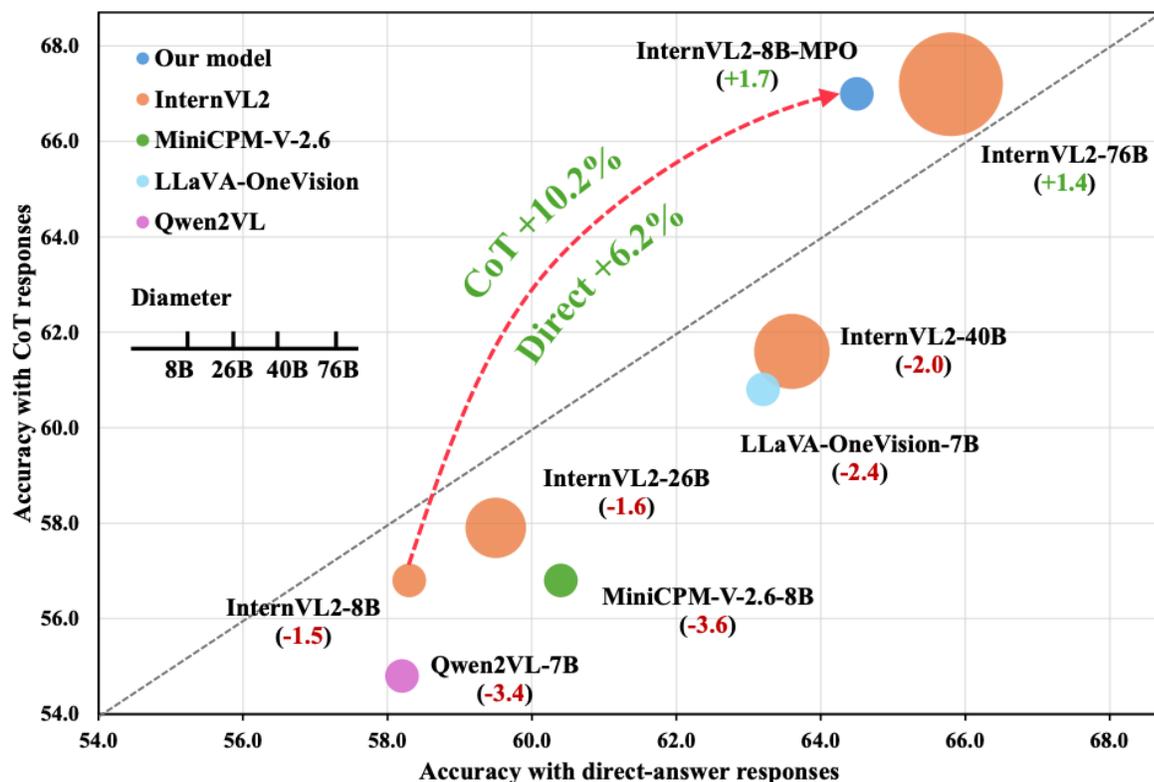
设计5: Packed Dataset (InternVL2.5 Section 3.4)

- 极大减少padding token数量, 训练速度翻倍



核心思想：融合训练中只有人类生成的正样本，需要在训练过程中引入模型自身的负样本

- 训练阶段：基于人类生成的 context / prefix 生成下一个 token
- 推理阶段：基于模型生成的 context / prefix 生成下一个 token
- **分布偏移导致模型在长序列输出场景下的性能受损**



核心思想：融合训练中只有人类生成的正样本，需要在训练过程中引入模型自身的负样本

- 通过模型自身输出构建正负样本对
 - Verifiable: 回答正确为正样本，错误为负样本
 - Unverifiable: 模型输出为正样本，（不输入图像）基于正样本前半段续写得到的为负样本

Instruction with clear ground truths



Question:

Your task is to answer the question below. Give step by step reasoning before you answer, and when you're ready to answer, please use the format "Final answer: ..."

Question:

What could be the reason for a person to have a brown teddy bear tied with a ribbon around its neck, a pack of colored pencils and a colorful patterned blanket?

- A. The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art
- B. The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art
- C. The person is a child who loves to play with teddy bears and draw with color pencils
- D. The person is a school teacher who uses these items to decorate their classroom for their students

Chosen Response:

To determine the answer, let's analyze each option:

- A. The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art. This option assumes that the person collects multiple items as part of a fan interest in art. The items are not just limited to teddy bears, colored pencils, and blankets.
- B. The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art. This option suggests that the teddy bear was received as a gift and that the person uses colored pencils as they are a fan of art. This explanation is plausible.
- C. The person is a child who loves to play with teddy bears and draw with color pencils. This option simply states that the person is a child, but it does not provide a clear connection between the other items like the vibrant blanket.
- D. The person is a school teacher who uses these items to decorate their classroom for their students. This option seems less likely because the items (teddy bear, colored pencils, and blanket) might not be common classroom decoration items.

Final answer: B

Rejected Response:

To determine the reason for a person to have a brown teddy bear tied with a ribbon around its neck, a pack of colored pencils, and a colorful patterned blanket, we need to consider the context and meaning behind these items.

Option A: "The person likes to collect teddy bears, color pencils, and blankets as they are a fan of art." This option suggests that the person has a specific interest in collecting items, which may not necessarily imply a connection to art.

Option B: "The person received the brown teddy bear as a gift from someone and owns the colored pencils because they are a fan of art." This option suggests a personal connection to a gift while considering an individual interest in art.

Option C: "The person is a child who loves to play with teddy bears and draw with color pencils." **This option assumes that the person is a child, which could be a plausible explanation for the items.**

Option D: "The person is a school teacher who uses these items to decorate their classroom for their students." This option suggests a professional context rather than personal interest or collecting items.

Given the context of the items in the image and the most plausible explanation based on the presence of children's items and a child-friendly material like a teddy bear with a ribbon, the most likely explanation is:

Final answer: C

Instruction without clear ground truths



Question:

Where is the woman in the picture located?

Chosen Response:

The woman in the picture is located in a public park. Here are the observations that support this conclusion:

1. Surroundings: The image shows a well-maintained area with grass, trees, and benches, which is typical of public parks. The presence of other people walking in the background also suggests a public space.
2. Lighting and Ambience: The bright daylight and open space are characteristic of outdoor environments like parks, rather than indoor locations like libraries.
3. Bench Design: The style of the bench, which appears to be made of wood or metal with a decorative back, is common in public parks. Such benches are often placed along walking paths for public use.
4. Vegetation: The variety and arrangement of trees and shrubs look like planned landscaping, which is typical of parks rather than the natural, untamed growth found in a virgin forest. These elements collectively indicate that the setting is a public park.

Rejected Response:

The woman in the picture is located in a public park. Here are the observations that support this conclusion:

1. Surroundings: The image shows a well-maintained area with grass, trees, and benches, which is typical of public parks. The presence of other people walking in the background also suggests a public space.
2. Lighting and Ambience: The bright daylight and open space are characteristic of outdoor environments like parks, rather than indoor locations like libraries.
3. Bench: If she were in a private garden or library, it would likely be more private and less likely to have communal spaces such as a bench.
4. Background Activity: **The presence of other people in the background suggests a public space where people gather, such as a park.**

Given these observations, the picture most likely shows a woman in a public park.

实验现象：基于DPO训练的模型容易出现塌缩现象（复读、胡言乱语）

- 问题：DPO算法在长期训练时，存在“塌缩”现象（正样本生成概率高于负样本，但是正负样本的生成概率同时下降）
- 解决方案：
 - 引入**生成Loss**，保证正样本生成概率提高（然而仍然会出现，间距拉大，但是负样本概率也被拉高的情况）
 - 进一步引入**绝对质量Loss**，保证负样本生成概率降低

核心思想： 模型应同时学习自身响应的相对质量、绝对质量、以及被偏好响应的生成方式

- 通过MPO算法让模型学习针对自身输出的正负反馈信息
- MPO的优化目标包括相对质量Loss、绝对质量Loss以及生成Loss

$$\mathcal{L} = w_p \cdot \mathcal{L}_p + w_q \cdot \mathcal{L}_q + w_g \cdot \mathcal{L}_g,$$

- 实践中，上述Loss分别采用了DPO Loss、BPO Loss以及LM Loss

实验结果

Model	MPO	MMMU	MathVista	MathVision	MathVerse	DynaMath	WeMath	LogicVista	Overall
InternVL3-1B	✗	43.4	47.2	13.8	18.1	4.2	14.7	31.1	24.6
	✓	43.4	45.8	18.8	18.7	5.8	13.4	29.8	25.1 (+0.5)
InternVL3-2B	✗	49.1	59.0	22.0	23.2	13.4	18.1	30.0	30.7
	✓	48.6	57.0	21.7	25.3	14.6	22.4	36.9	32.4 (+1.7)
InternVL3-8B	✗	61.9	67.4	24.7	36.9	22.8	32.7	43.2	41.4
	✓	62.7	71.6	29.3	39.8	25.5	37.1	44.1	44.3 (+2.9)
InternVL3-9B	✗	59.0	68.8	28.9	32.2	23.0	32.5	46.5	41.6
	✓	57.7	71.5	27.6	35.3	26.7	33.8	49.2	43.1 (+1.5)
InternVL3-14B	✗	67.1	70.5	31.2	38.8	27.9	38.1	49.9	46.2
	✓	67.1	75.1	37.2	44.4	31.3	43.0	51.2	49.9 (+3.7)
InternVL3-38B	✗	69.3	71.2	34.2	45.1	22.2	41.7	54.4	48.3
	✓	70.1	75.1	34.2	48.2	35.3	48.6	58.4	52.8 (+4.5)
InternVL3-78B	✗	72.2	74.0	35.2	44.2	31.7	42.5	53.5	50.5
	✓	72.2	79.0	43.1	51.0	35.1	46.1	55.9	54.6 (+4.1)

Table 13: Comparison of reasoning abilities before and after Mixed Preference Optimization (MPO).

核心Insights

- MPO后的模型推理能力显著优于MPO前的模型
- 不包含Generation Loss会导致模型塌缩，复读、胡言乱语的概率显著提升（16.4% v.s. 0.3%）
 - 引入Generation Loss，令正样本生成概率提升
 - 引入Quality Loss，令负样本生成概率下降，防止随正样本概率一起上升
- **实践中发现，模型A作为policy model构建的训练数据也可以用于模型B的训练**
 - InternVL3的MPO训练数据主要基于InternVL2.5和InternVL2构建
 - **数据复用率!**

和头部商用模型对比

- 推理及感知性能均位列开源第一!

Multi-modal Academic Leaderboard

Real-time Ranking

All

1	SenseNova-V6-Pro	SenseTime
2	Gemini-2.5-Pro	Google
3	InternVL3-78B	Shanghai AI Laboratory & Tsinghua University
4	InternVL3-38B	Shanghai AI Laboratory & Tsinghua University
5	Step-1o	StepFun

Multimodal Model Reasoning Leaderboard

Real-time List

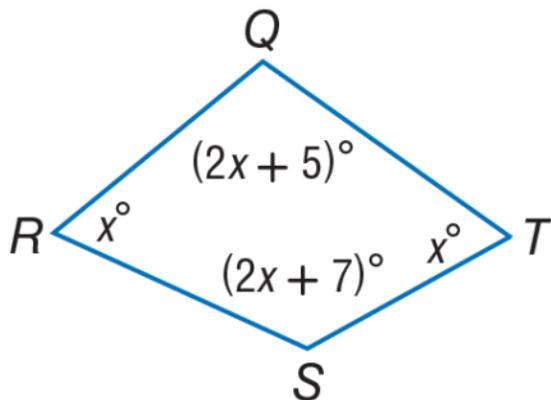
All

1	Doubao-1.5-Pro	ByteDance	61.6
2	Gemini-2.0-Pro	Google	56.6
3	ChatGPT-4o-latest	OpenAI	54.8
4	GPT-4.1-20250414	OpenAI	54.0
5	InternVL3-78B	Shanghai AI Laboratory & Tsinghua University	51.0

核心思想：基于每个步骤的预期准确率构建过程监督

- 按照双换行符 (\n \n) 切分不同步骤
- 基于当前步骤及其prefix, 采样16次, 估计预期准确率 (Monte Carlo得分)
- 实践中为了节省成本, 最多切分为12步, 超过12步, 则均匀合并

Data Example from VisualPRM400K



Question:

Find $m \angle S$.

Solution:

Step-0 ($mc=0.75$)

To find $(m \angle S)$ in the given quadrilateral, we'll use the property that the sum of the interior angles of any quadrilateral is (360°) .

Step-1 ($mc=0.5$)

Step 1: Write the equation for the sum of the angles: $[(2x + 5)^\circ + x^\circ + (2x + 7)^\circ + x^\circ = 360^\circ]$

Step-2 ($mc=0.5625$)

Step 2: Simplify the equation:

$$[2x + 5 + x + 2x + 7 + x = 360]$$

$$[6x + 12 = 360]$$

Step-3 ($mc=0.5$)

Step 3: Solve for (x) :

$$[6x + 12 = 360]$$

$$[6x = 360 - 12]$$

$$[6x = 348]$$

$$[x = 58]$$

Step-4 ($mc=0.0$)

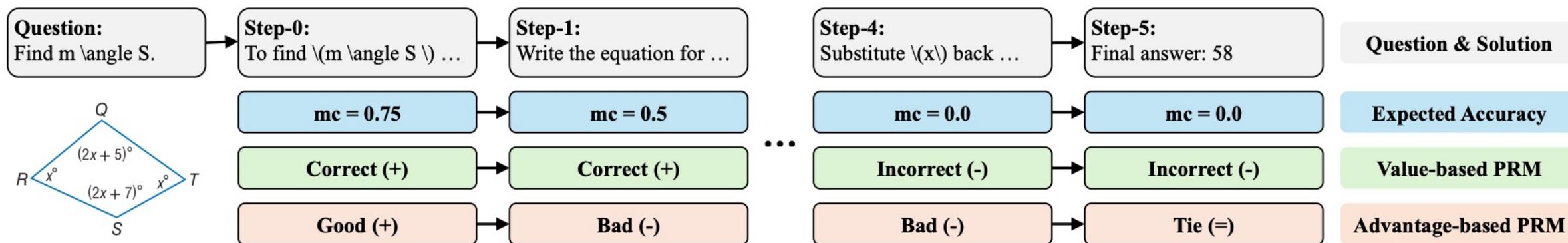
Step 4: Substitute (x) back into $(m \angle S)$ which is given as (x) : $[m \angle S = x = 58^\circ]$

Step-5 ($mc=0.0$)

Final answer: 58

核心思想： 基于每个步骤的预期准确率构建过程监督

- Value-based PRM: 预测每个步骤的对错 (mc是否大于0)
- Advantage-based PRM: 预测每个步骤相比上一个步骤的好坏 (当前mc是否高于上一步mc)
- 实践发现, Value-based PRM效果更好



多模态大模型训练算法设计 (测试阶段增强)

Model	MMMU	MathVista	MathVision	MathVerse-VO	DynaMath	WeMath	LogicVista	Overall
<i>Proprietary Models</i>								
GPT-4o [58]	70.7	60.0	31.2	40.6	34.5	45.8	52.8	47.9
Gemini-2.0-Flash [61]	69.9	70.4	43.6	47.8	42.1	47.4	52.3	53.4
Claude-3.5-Sonnet [4]	66.4	65.3	35.6	46.3	35.7	44.0	60.4	50.5
<i>Open-source Models</i>								
MiniCPM-V2.6-8B [89]	49.8	60.8	23.4	18.9	9.8	16.4	27.5	29.5
+VisualPRM	56.8	65.7	24.7	35.8	11.2	31.0	37.4	37.5
	+7.0	+4.9	+1.3	+16.9	+1.4	+14.6	+9.8	+8.0
Qwen2.5-VL-7B [7]	55.0	67.8	25.4	41.1	21.0	35.2	44.1	41.4
+VisualPRM	58.6	70.3	31.3	44.3	23.0	39.8	48.3	45.1
	+3.6	+2.5	+5.9	+3.2	+2.0	+4.6	+4.2	+3.7
InternVL2.5-8B [15]	56.2	64.5	17.0	22.8	9.4	23.5	36.0	32.8
+VisualPRM	60.2	68.5	25.7	35.8	18.0	36.5	43.8	41.2
	+4.0	+4.0	+8.7	+13.0	+8.6	+13.0	+7.8	+8.4
InternVL2.5-26B [15]	60.7	68.2	23.4	24.0	11.4	30.9	39.6	36.9
+VisualPRM	63.9	73.1	29.6	39.1	23.2	40.8	51.0	45.8
	+3.2	+4.9	+6.2	+15.1	+11.8	+9.9	+11.4	+8.9
InternVL2.5-38B [15]	63.9	71.9	32.2	36.9	20.0	38.3	47.9	44.4
+VisualPRM	69.0	73.9	35.2	46.7	30.5	46.2	53.7	50.7
	+5.1	+2.0	+3.0	+9.8	+10.5	+7.9	+5.8	+6.3
InternVL2.5-78B [15]	70.0	72.3	32.2	39.2	19.2	39.8	49.0	46.0
+VisualPRM	70.7	75.1	35.9	47.1	31.3	49.1	53.9	51.9
	+0.7	+2.8	+3.7	+7.9	+12.1	+9.3	+4.9	+5.9

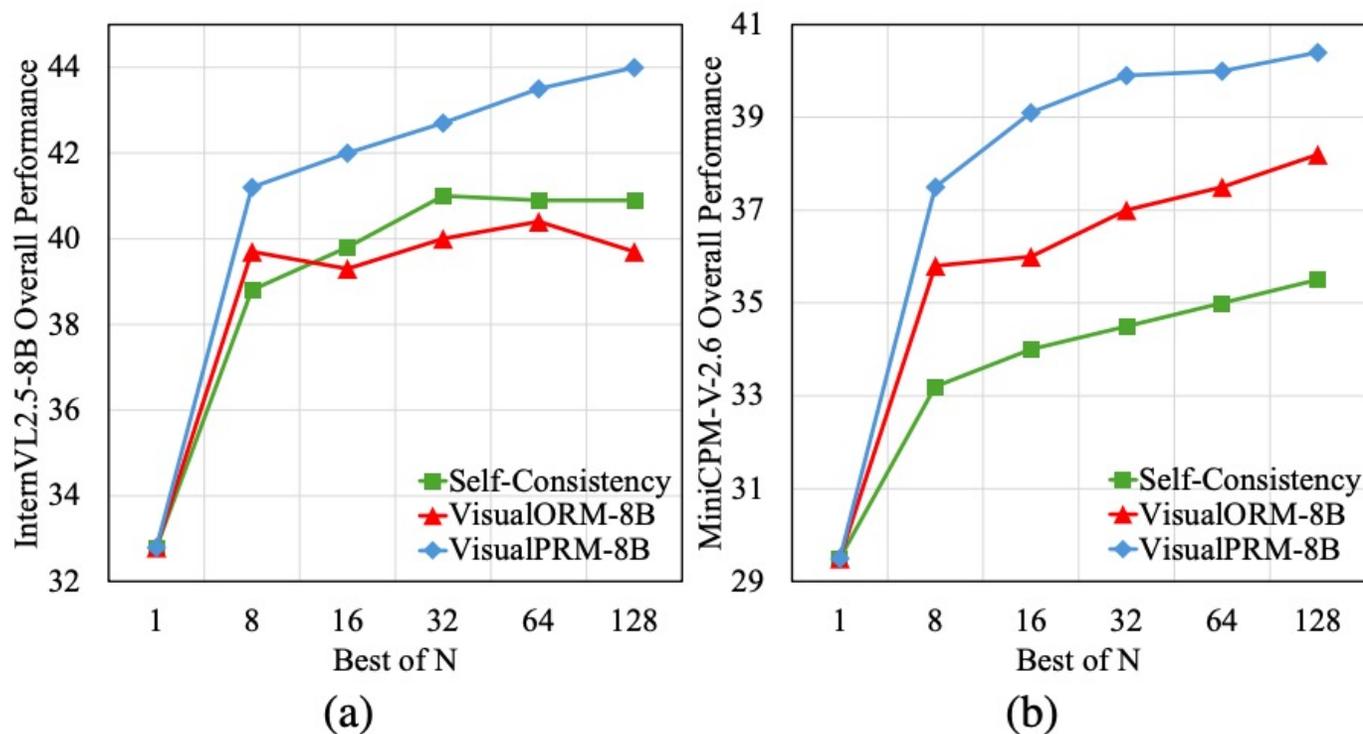
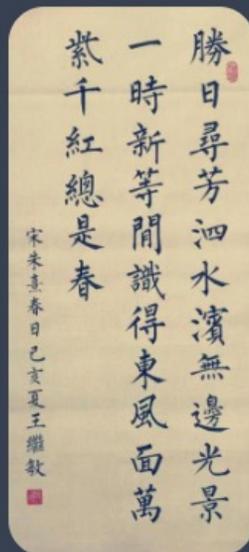


Figure 4. **Overall Best-of-N results across seven multimodal reasoning benchmarks with different policy and critic models.** VisualPRM consistently enhances reasoning performance of both InternVL2.5-8B and MiniCPM-V as N increases and outperforms the improvement introduced by Self-Consistency and ORM, highlighting the effectiveness of PRM in Test-Time Scaling.



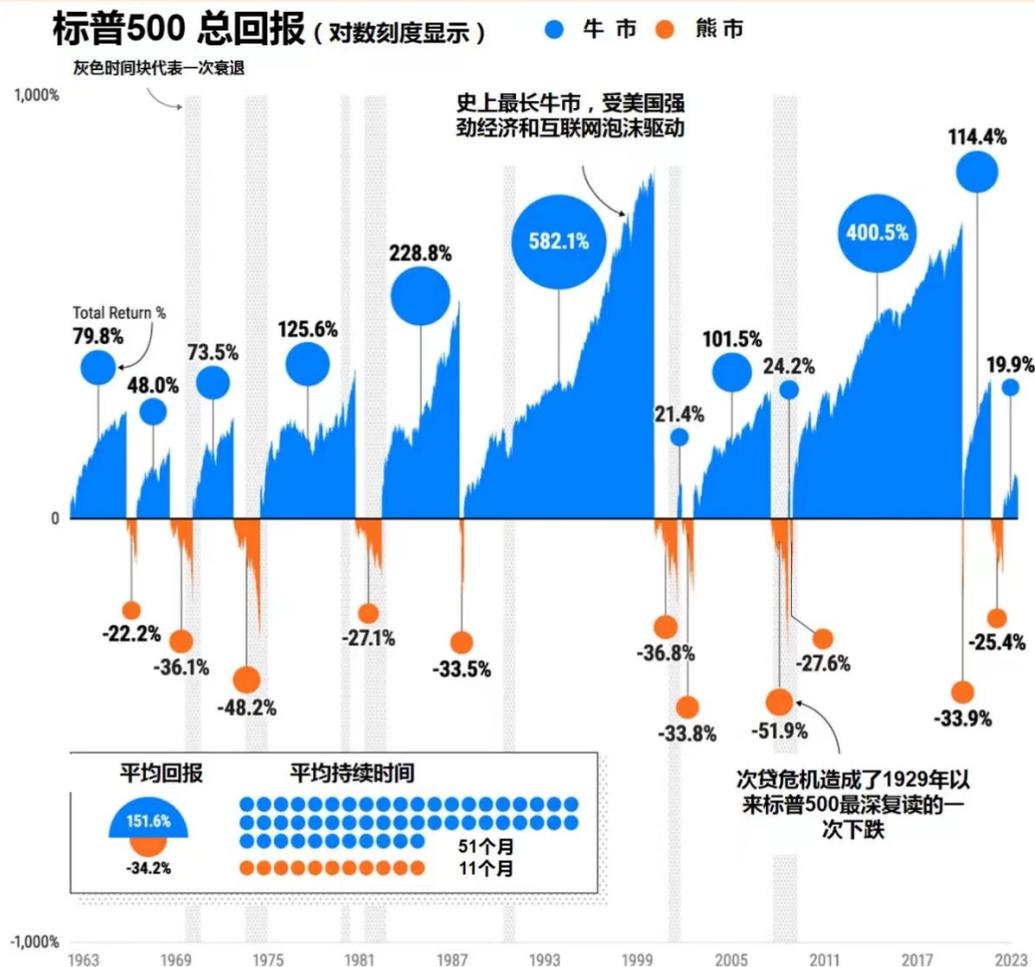
这是啥

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。

美股：60年牛熊周期



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%。

细节文字理解+文字深层含义



请问杯子里是什么?

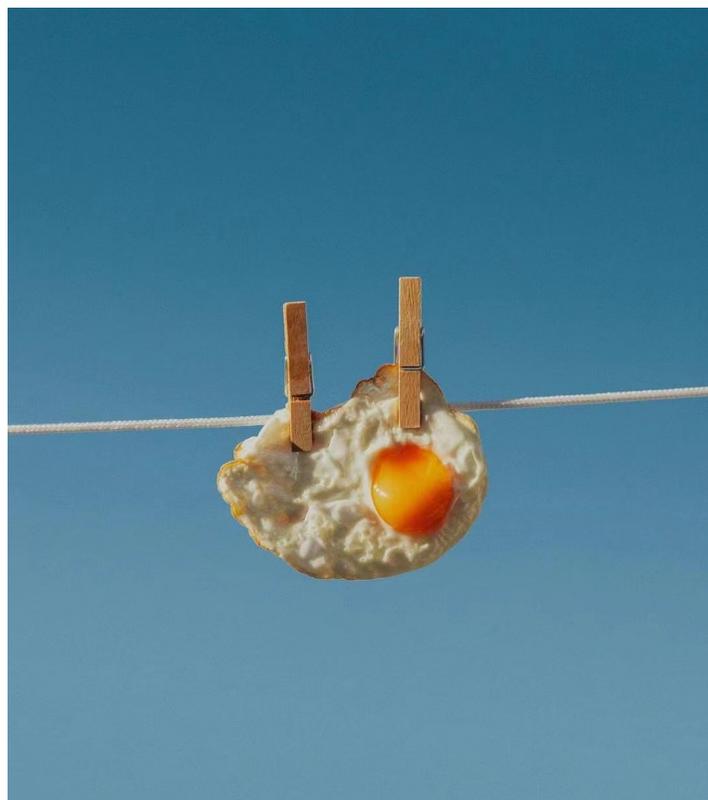
杯子里装着绿色的液体，看起来像是某种饮料。

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。



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。

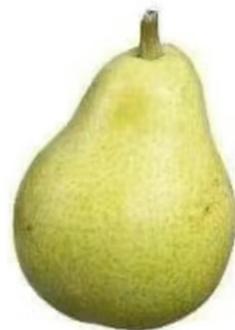


这是什么东西?

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。



William



Pear

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：

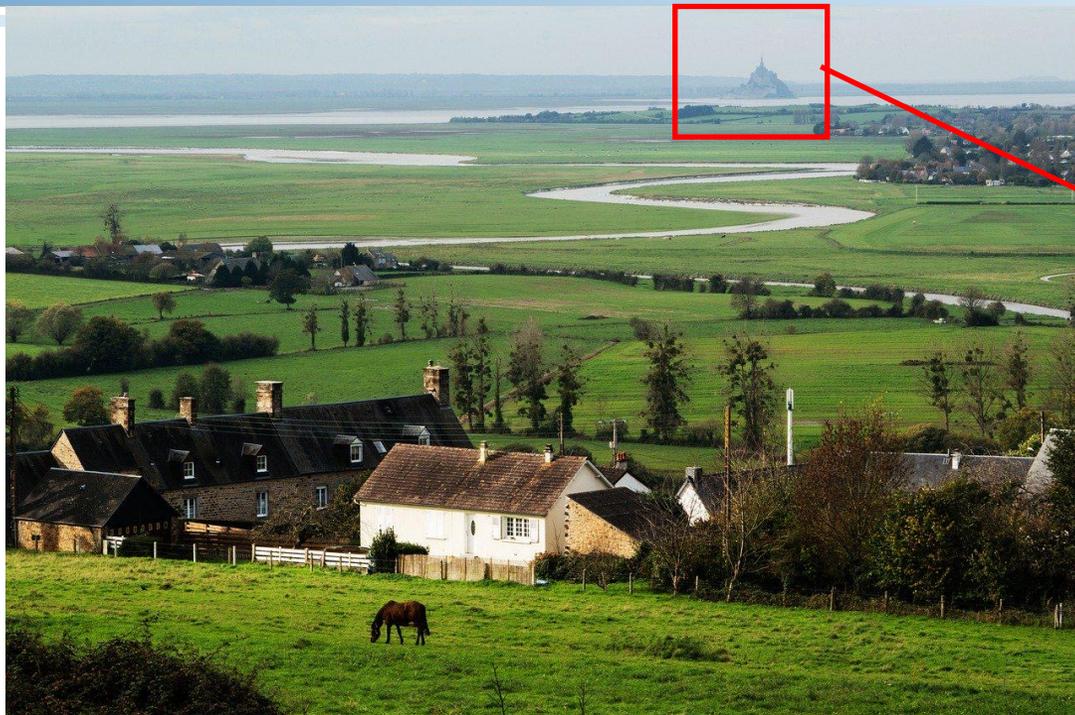
1. 第一张图：威廉·莎士比亚的画像，标注为“William”。
2. 第二张图：一个梨的图片，标注为“Pear”。
3. 第三张图：一个人摇晃着一个梨，这个动作可以理解为“Shake a Pear”。

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。



William Shakespeare

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。

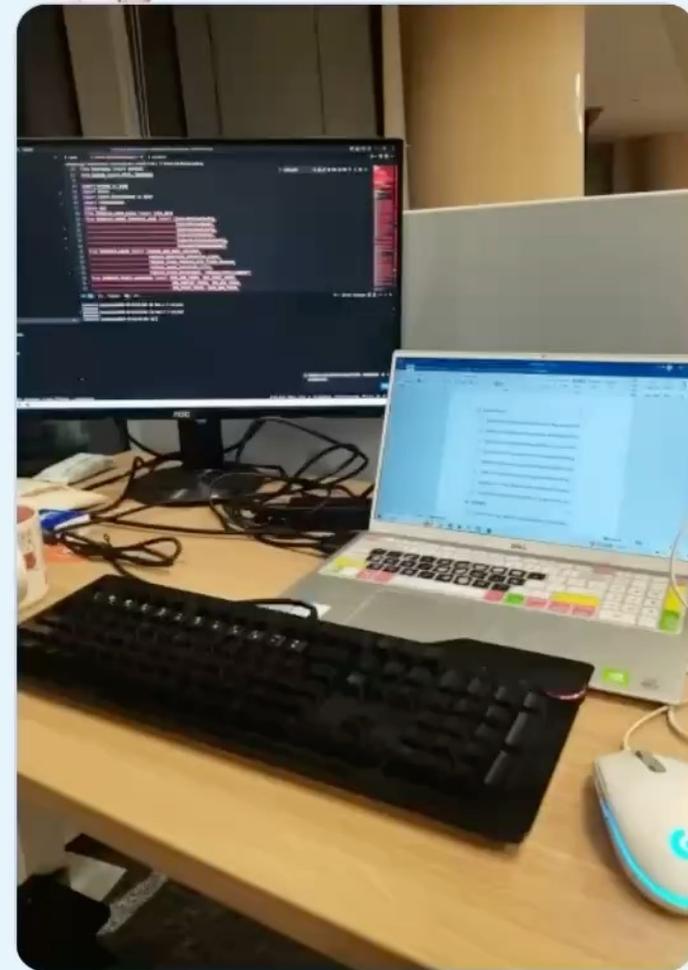
你怎么知道呢

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。

InternOmni: Extending InternVL with Audio Modality

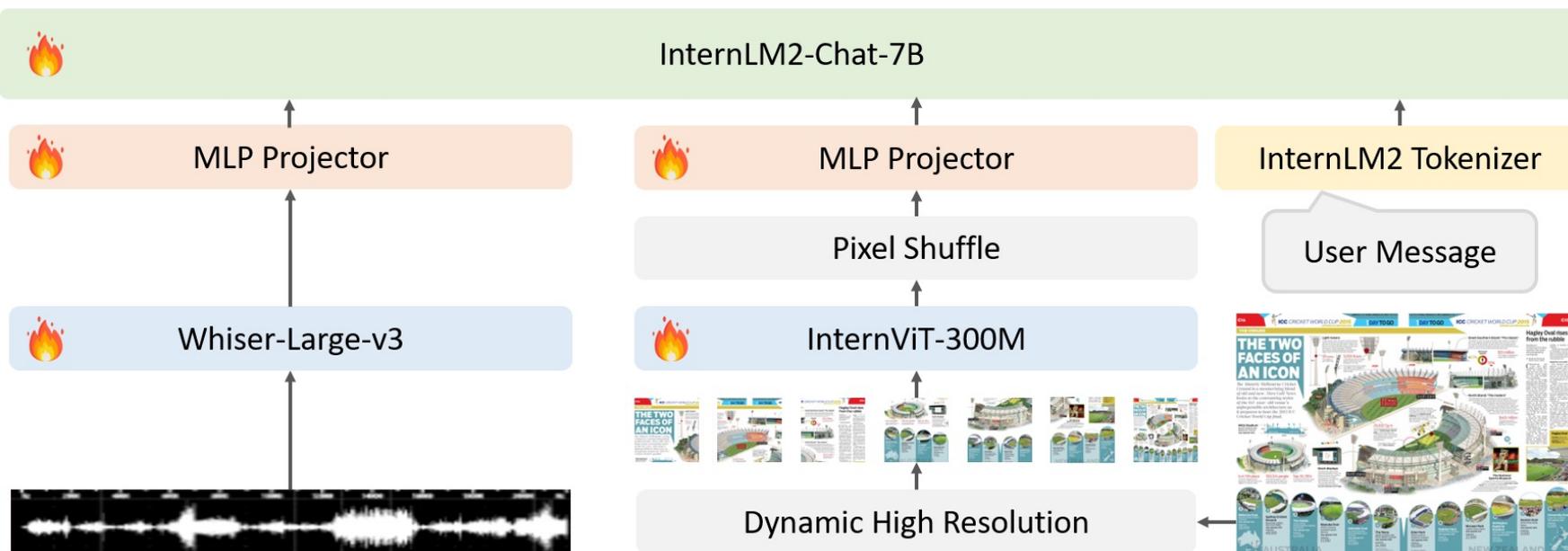


你好，我是书生·万象



按住 说话

vConsole



更多详情看blog

目录

1

多模态大模型研究背景

2

多模态大模型模型结构设计

3

多模态大模型训练算法设计

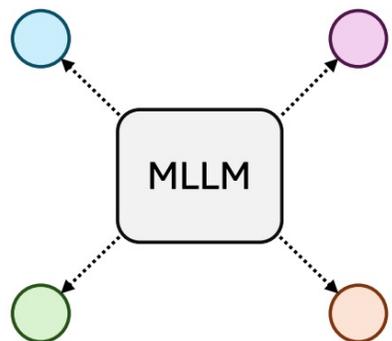
4

不止于语言输出：通专融合

不止于语言输出：通专融合

● ● ● ● downstream tools

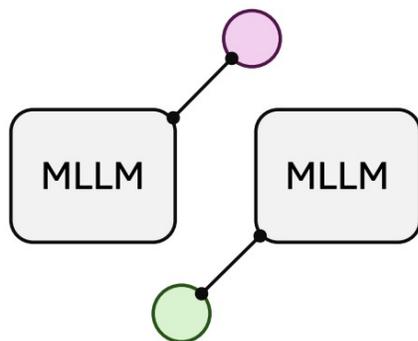
↗ text message ● embedding ↖ super link



(a) text-based method

✓ multiple tasks

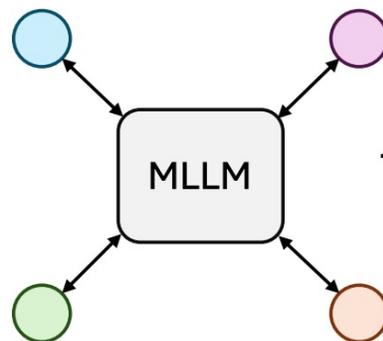
✗ efficient information transfer



(b) embedding-based method

✗ multiple tasks

✓ efficient information transfer

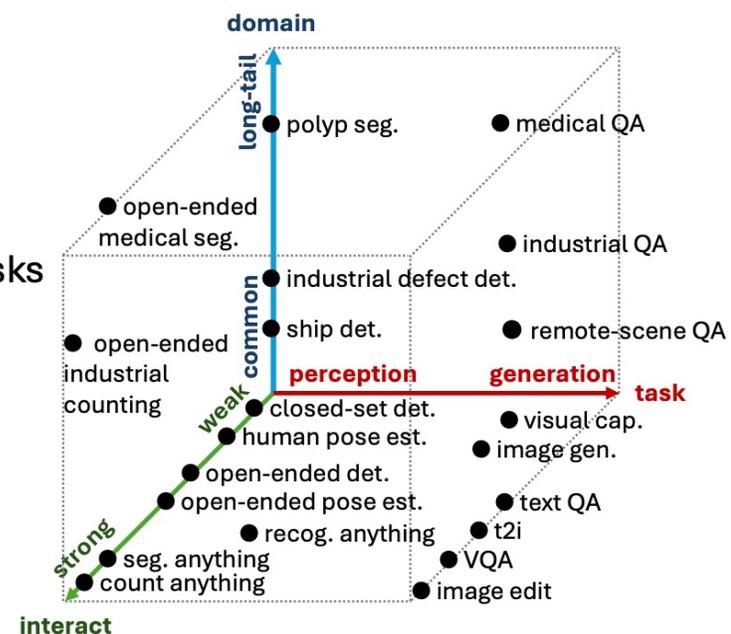


(c) our method

✓ multiple tasks

✓ efficient information transfer

100+ tasks



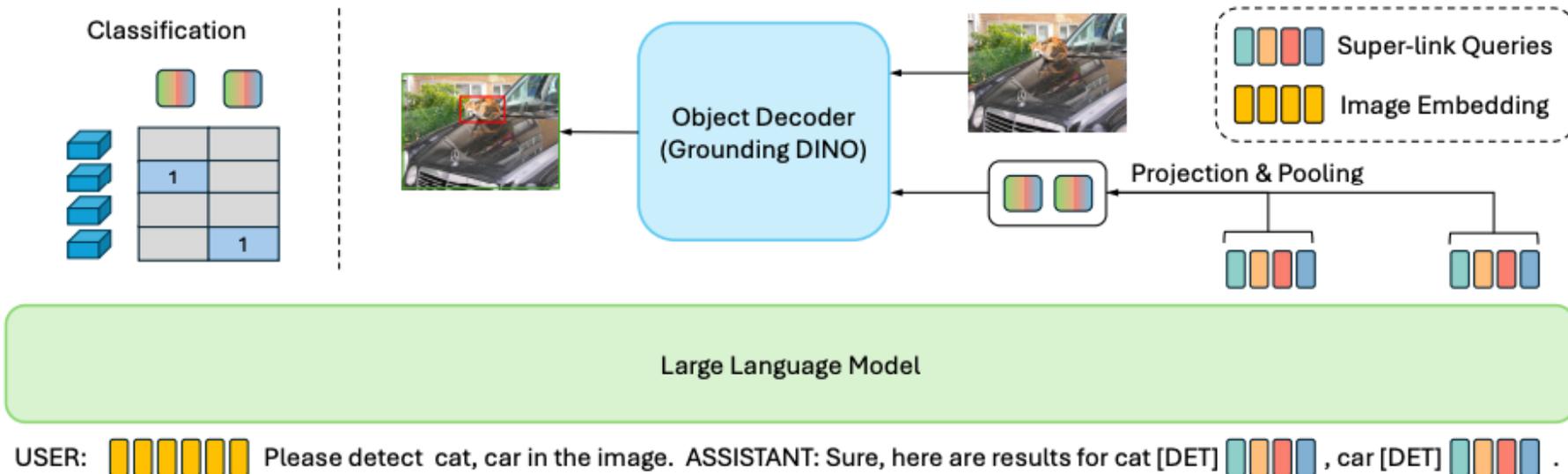


Example: Text Prompt + Visual Prompt for Interactive Segmentation.

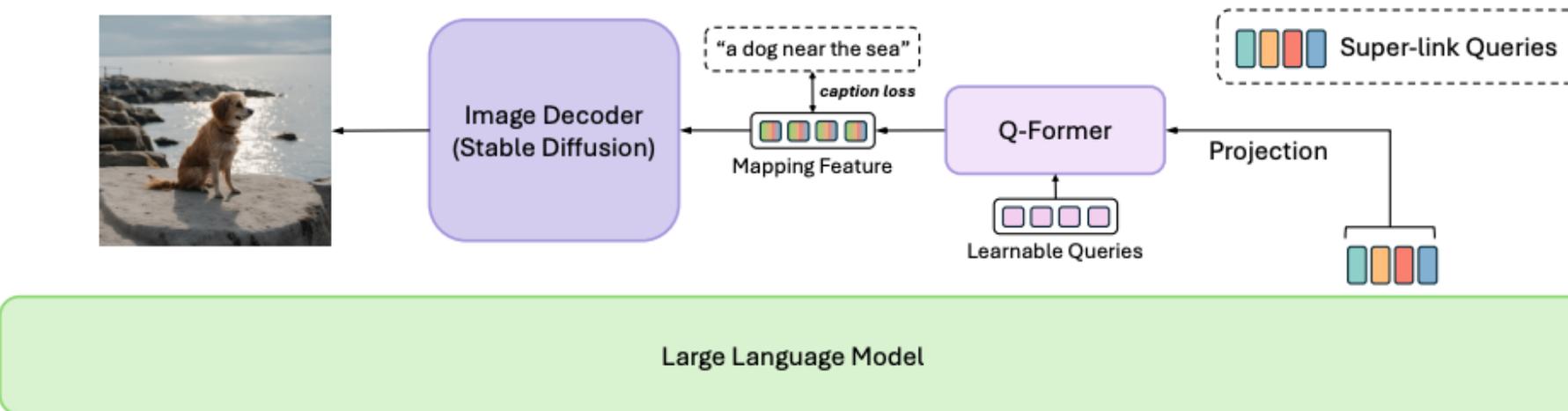
USER: <image> Could you please segment all the corresponding objects according to the visual prompts as region1 <region>, region2 <region>?

ASSISTANT: Sure, these objects are region1 [SEG], region2 [SEG].

不止于语言输出：通专融合



(a) Connecting with object decoder for visual perception.



method	query/token number	inst seg.		ground.	pose	interact seg.	
		AP _b	AP _m	P@.5	AP	mIoU	cIoU
	1	50.4	39.6	85.8	43.0	43.2	60.0
super-link queries	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.

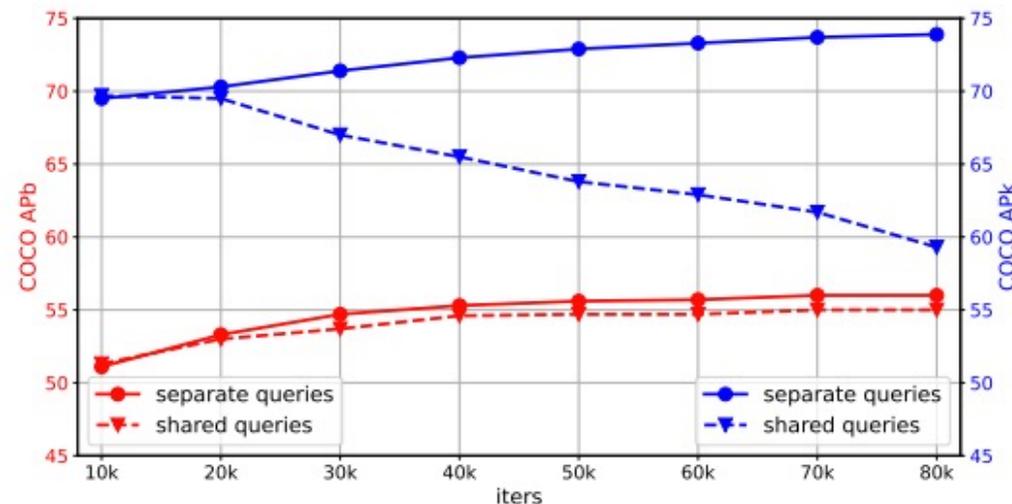
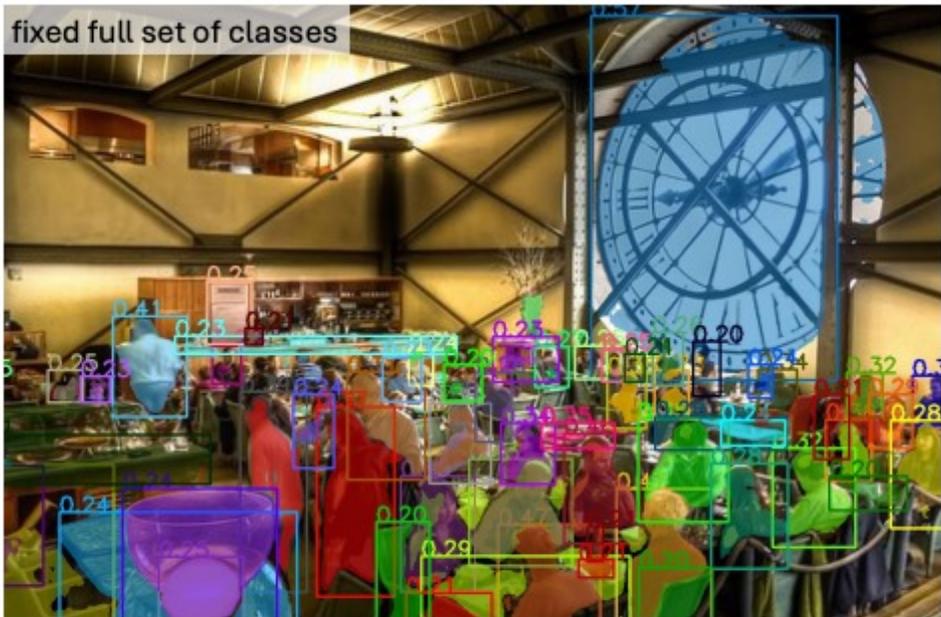


Figure 4: Shared vs. unshared super-link queries for different decoders. We report the box/keypoint AP on COCO.

- 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

不止于语言输出：通专融合

fixed full set of classes



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

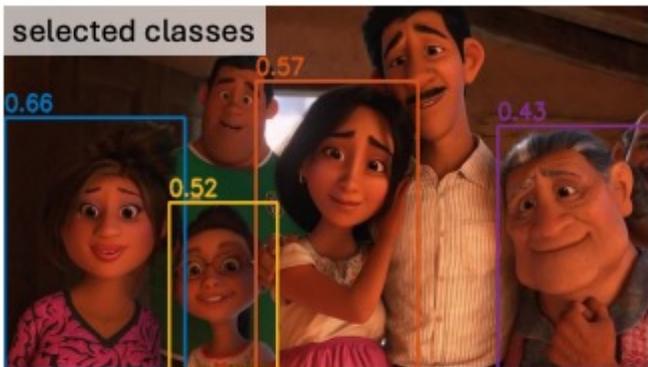
fixed full set of classes



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

← 开放检测 & 分割

selected classes



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

selected classes



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

selected classes



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

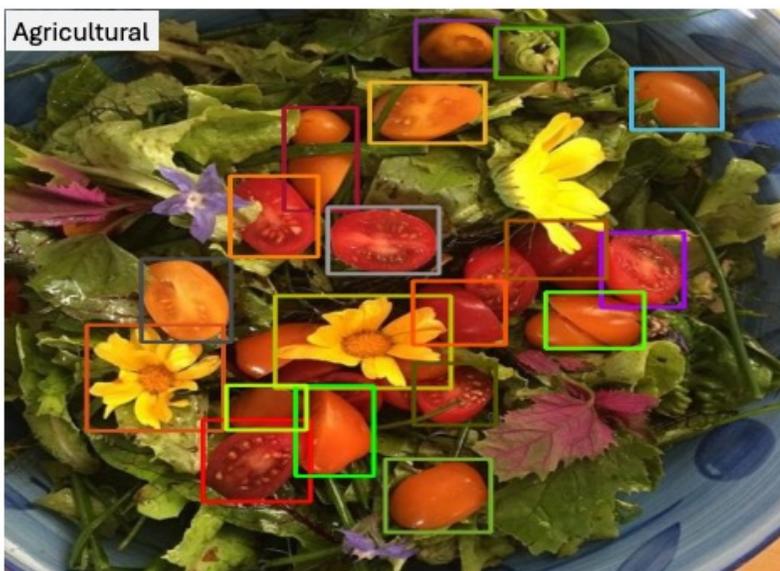
不止于语言输出：通专融合



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

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**?



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]**?

← 不同domain的开放姿态估计

不止于语言输出：通专融合



An astronaut riding a horse X, where $X \in \{“”, “in the style of van gogh”, “in the style of ink painting”, “in the style of black and white sketch”\}$



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

不止于语言输出：通专融合

Original Image

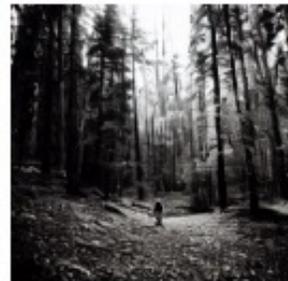
Edited Image



Change it to a painting by Vincent van Gogh.

Original Image

Edited Image



Make the image black and white.



Make it an autumn scene.



Let the sky blue.



Make his beard grow longer.



Put the penguin into the desert.

Original Image

Edited Image



Put sunglasses on him.

Original Image

Edited Image



Add some birds.



Turn the goose yellow.



Make the river a rainbow

↑
← 图像编辑

感谢观看



InternVL2.5



MPO



VisualPRM



VisionLLM v2

wangweiyun@pjlab.org.cn