Cambrian-1: A Fully Open, *Vision-Centric* Exploration of Multimodal LLMs

Shengbang Tong*, Ellis Brown*, Penghao Wu*, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, Saining Xie†

New York University

Paper Reading by Zhiying Lu

2024.07.16

- 作者介绍
- 研究背景
- 本文探索
- 数据集构建
- 实验效果
- 总结

作者介绍

Shengbang Tong

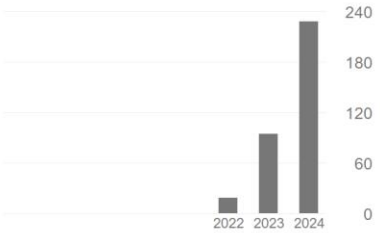
FOLLOW

NYU Courant
Verified email at berkeley.edu - Homepage
AI Computer Vision Deep Learning Representation Learning

TITLE	CITED BY	YEAR
Eyes wide shut? exploring the visual shortcomings of multimodal llms S Tong, Z Liu, Y Zhai, Y Ma, Y LeCun, S Xie CVPR 2024 (Oral)	76	2024
Investigating the catastrophic forgetting in multimodal large language models Y Zhai, S Tong, X Li, M Cai, Q Qu, YJ Lee, Y Ma CPAL 2024	69 *	2023
White-box transformers via sparse rate reduction Y Yu, S Buchanan, D Pai, T Chu, Z Wu, S Tong, B Haeffele, Y Ma NIPS 2023	39	2023

Cited by

	All	Since 2019
Citations	343	343
h-index	10	10
i10-index	10	10



Saining Xie

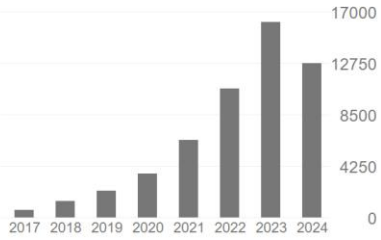
FOLLOW

Assistant Professor at the Courant Institute, New York University
Verified email at nyu.edu - Homepage
Computer Vision Machine Learning Deep Learning

TITLE	CITED BY	YEAR
Aggregated Residual Transformations for Deep Neural Networks S Xie, R Girshick, P Dollár, Z Tu, K He Computer Vision and Pattern Recognition (CVPR), 2017	12980	2017
Momentum contrast for unsupervised visual representation learning K He, H Fan, Y Wu, S Xie, R Girshick Proceedings of the IEEE/CVF conference on computer vision and pattern …	12457	2020
Masked autoencoders are scalable vision learners K He, X Chen, S Xie, Y Li, P Dollár, R Girshick Proceedings of the IEEE/CVF conference on computer vision and pattern …	6410	2022
A ConvNet for the 2020s Z Liu, H Mao, CY Wu, C Feichtenhofer, T Darrell, S Xie CVPR 2022, 11976-11986	5085	2022

Cited by VIEW ALL

	All	Since 2019
Citations	54967	52146
h-index	34	32
i10-index	46	43



Public access VIEW ALL

0 articles 14 articles

- 作者介绍
- 研究背景
- 本文探索
- 数据集构建
- 实验效果
- 总结

研究背景

- 现有MLLM的构建方式（LLAVA模式）
- 预训练的视觉encoder（CLIP为主）+adapter（Q-former/MLP/Linear）+预训练的LLM（各种）
- 两阶段训练
  - 第一阶段固定visual，利用图文对齐数据训练adapter
  - 第二阶段固定visual，利用指令微调数据训练adapter和LLM

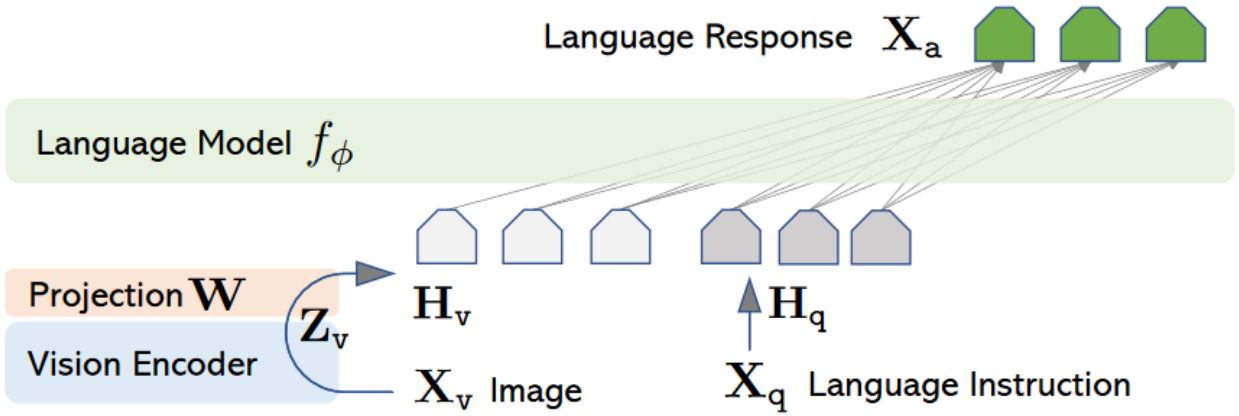


Figure 1: LLaVA network architecture.

研究背景

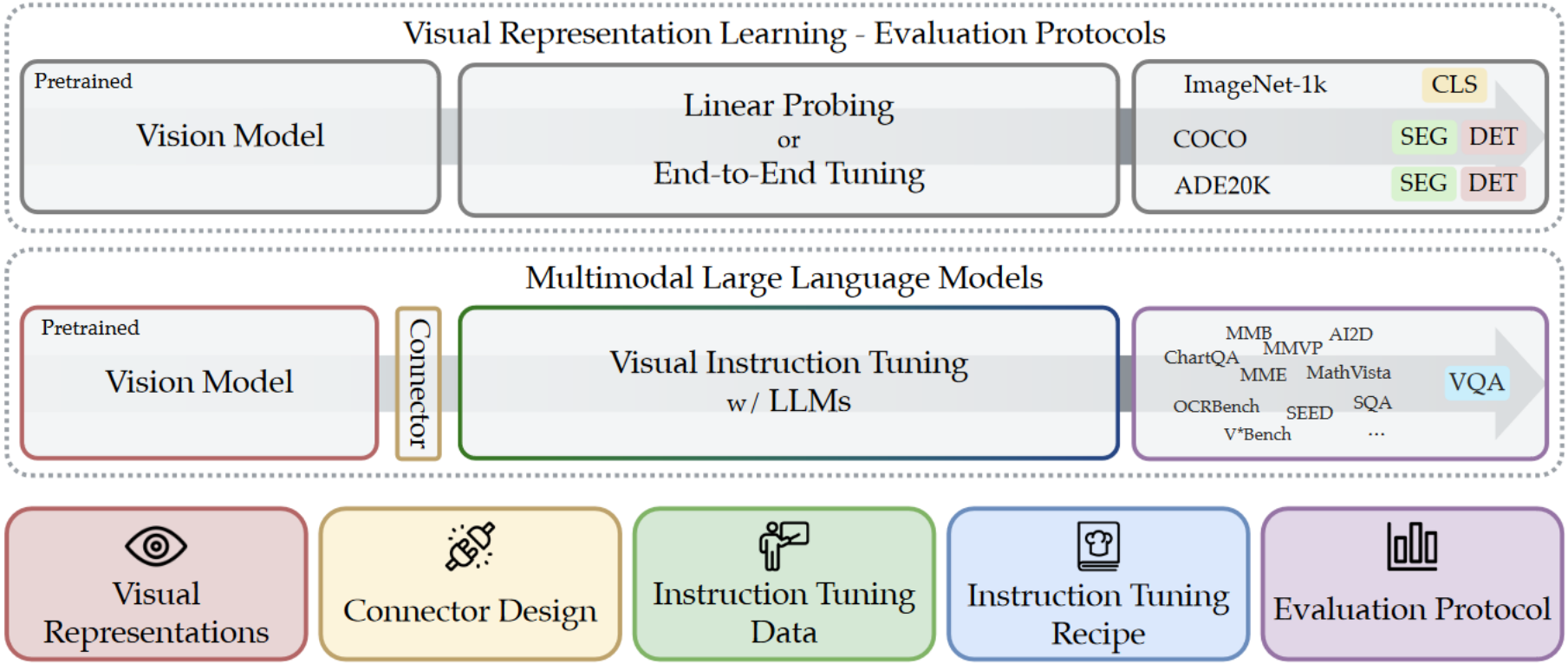
- MLLM中视觉表征学习与语言理解之争
- LLM的scaling性能使得MLLM能力提升
- 对视觉部分组件的探索缺乏系统且完整的研究，且与传统视觉研究脱离
- 采用语言监督的CLIP被广泛应用，但基于纯视觉的DINO等未被充分探索

两个主要问题：

- 现阶段MLLM研究过分依赖LLM，作为shortcut，弥补视觉部分不足
- 现有的benchmark并不具备robust real-world scenarios

研究背景

- 本文以视觉为中心进行系统性的MLLM研究设计

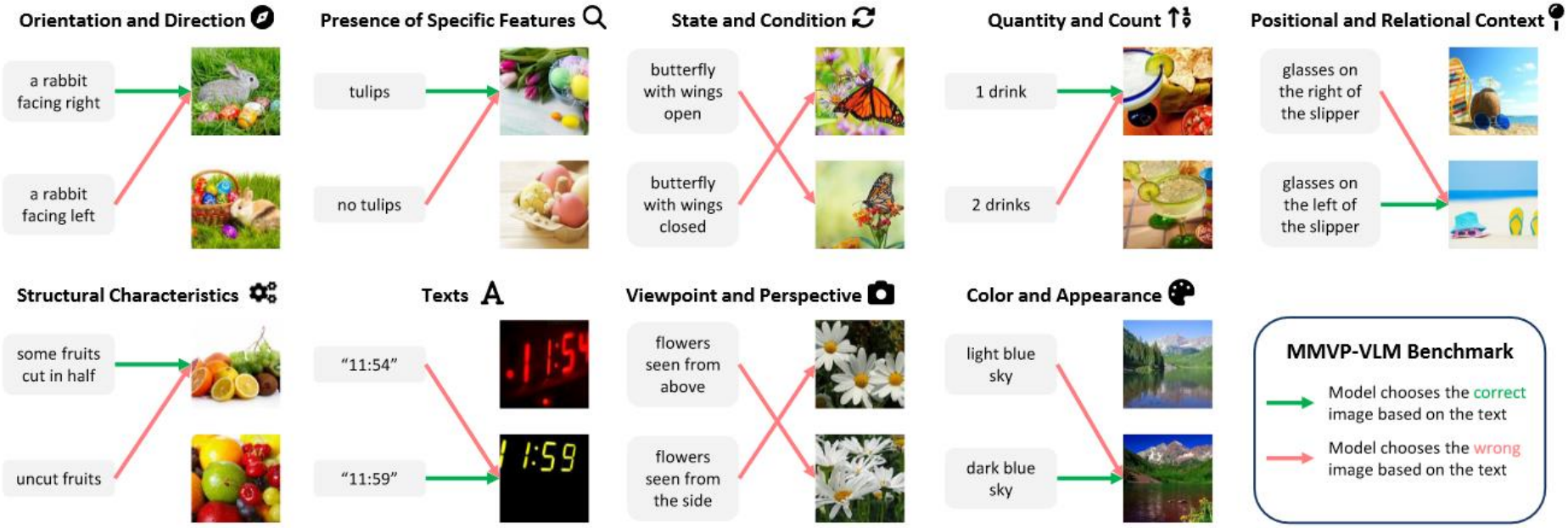


研究背景

Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs

Shengbang Tong[1] Zhuang Liu Yuexiang Zhai[2]
Yi Ma[2] Yann LeCun[1] Saining Xie[1]
[1]New York University [2]UC Berkeley

- 先行工作
- 构造视觉相关问题，评测不同MLLM的效果：方向、是否存在、视觉状态、计数、位置相关、颜色、文本……
- 基于对比学习的CLIP无法分辨视觉细节

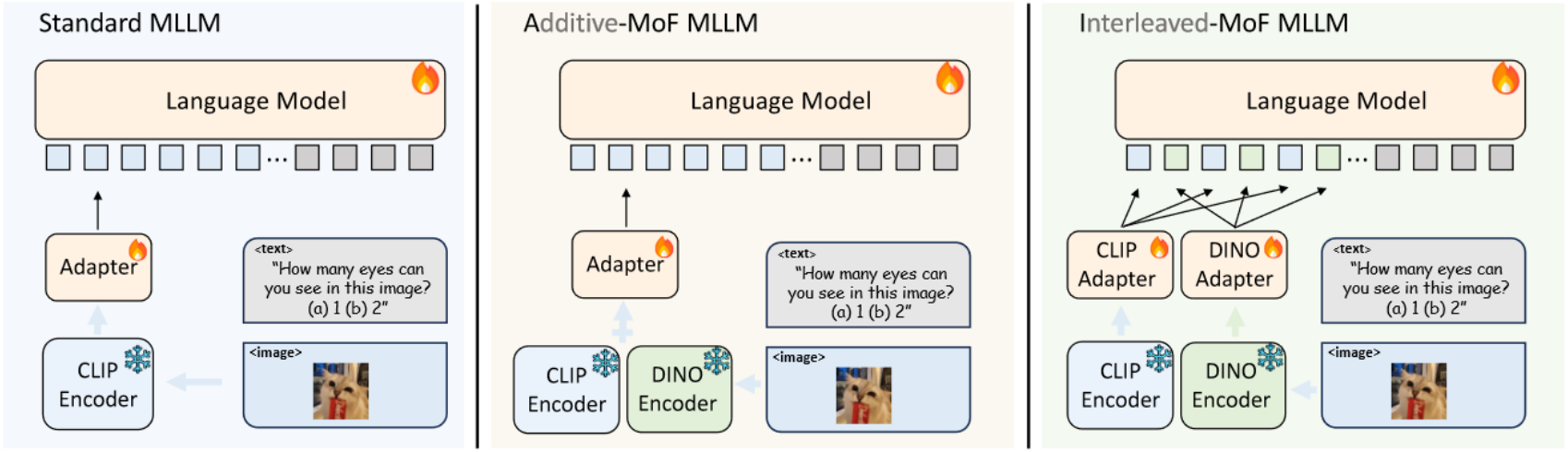


研究背景

method	res	#tokens	MMVP	LLaVA	POPE
LLaVA	224^2	256	5.5	81.8	50.0
LLaVA	336^2	576	6.0	81.4	50.1
LLaVA + I-MoF	224^2	512	16.7 (+10.7)	82.8	51.0
$\text{LLaVA}^{1.5}$	336^2	576	24.7	84.7	85.9
$\text{LLaVA}^{1.5}$ + I-MoF	224^2	512	28.0 (+3.3)		

- 利用交叉特征混合，结合CLIP和DINOv2的特征，补全视觉信息

	Image Size	Params (M)	IN-1k ZeroShot							A		MMVP Average
OpenAI ViT-L-14 [43]	224^2	427.6	75.5	13.3	13.3	20.0	20.0	13.3	53.3	20.0	6.7	13.3	19.3
OpenAI ViT-L-14 [43]	336^2	427.9	76.6	0.0	20.0	40.0	20.0	6.7	20.0	33.3	6.7	33.3	20.0
SigLIP ViT-SO-14 [66]	224^2	877.4	82.0	26.7	20.0	53.3	40.0	20.0	66.7	40.0	20.0	53.3	37.8
SigLIP ViT-SO-14 [66]	384^2	878.0	83.1	20.0	26.7	60.0	33.3	13.3	66.7	33.3	26.7	53.3	37.0
DFN ViT-H-14 [10]	224^2	986.1	83.4	20.0	26.7	73.3	26.7	26.7	66.7	46.7	13.3	53.3	39.3
DFN ViT-H-14 [10]	378^2	986.7	84.4	13.3	20.0	53.3	33.3	26.7	66.7	40.0	20.0	40.0	34.8
MetaCLIP ViT-L-14 [62]	224^2	427.6	79.2	13.3	6.7	66.7	6.7	33.3	46.7	20.0	6.7	13.3	23.7
MetaCLIP ViT-H-14 [62]	224^2	986.1	80.6	6.7	13.3	60.0	13.3	6.7	53.3	26.7	13.3	33.3	25.2
EVA01 ViT-g-14 [54]	224^2	1136.4	78.5	6.7	26.7	40.0	6.7	13.3	66.7	13.3	13.3	20.0	23.0
EVA02 ViT-bigE-14+ [54]	224^2	5044.9	82.0	13.3	20.0	66.7	26.7	26.7	66.7	26.7	20.0	33.3	33.3



- 作者介绍
- 研究背景
- 本文探索
- 数据集构建
- 实验效果
- 总结

文章结构

§3.1. Analyze the Benchmarks
§3.2. Introduce CV-Bench
§3.3. Study Instruction Tuning Recipes
§3.4. Use MLLMs as a Visual Representation Evaluator
§3.5. Investigate Combining Multiple Vision Encoders

实验设置

- 23种视觉模型+Vicuna，用shareGPT4V和LLAVA-665K分别训两个stage

Vision Backbone			General				Knowledge				OCR & Chart				Vision-Centric			
Model	Architecture	Average	MME^{P}	MMB	SEED^{I}	GQA	SQA^{I}	MMMU^{V}	MathVista^{M}	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	CV-Bench^{2D}	CV-Bench^{3D}
Language Supervised																		
OpenAI CLIP	ViT-L/14@336	48.37	1,419.43	61.45	59.85	62.26	69.87	34.50	27.80	59.82	33.73	31.70	55.39	28.00	11.33	53.46	56.44	57.40
DFN-CLIP	ViT-L/14@224	38.78	1,172.50	49.53	49.74	52.94	67.74	34.00	27.30	56.99	16.75	4.87	44.81	11.19	6.67	46.97	40.29	52.08
DFN-CLIP	ViT-H/14@378	36.79	1,091.76	41.28	44.32	50.48	65.54	33.65	26.50	56.76	15.56	2.70	43.41	10.15	4.67	47.06	39.07	52.91
EVA-CLIP-02	ViT-L/14@336	45.84	1,325.17	58.21	62.99	62.03	68.67	35.00	27.50	58.26	19.40	22.50	51.08	16.36	24.67	52.68	52.98	54.83
SigLIP	ViT-L/16@384	48.80	1,383.42	61.02	63.56	61.85	68.91	35.29	29.70	57.87	34.96	29.60	56.73	28.31	23.33	52.68	52.95	54.83
SigLIP	ViT-SO400M/14@384	47.57	1,376.75	58.76	60.59	60.92	69.01	34.40	26.50	58.35	30.72	28.60	55.10	28.31	19.33	50.71	52.33	58.67
OpenCLIP	ConvNeXt-L@512	47.38	1,404.01	57.62	61.90	60.34	69.06	33.90	29.10	58.39	28.04	25.20	55.45	28.41	24.00	54.12	53.46	48.91
OpenCLIP	ConvNeXt-L@1024	39.02	1,139.60	14.64	49.59	37.91	65.71	34.30	27.30	54.13	32.97	12.05	52.61	38.36	9.67	47.45	52.68	38.04
OpenCLIP	ConvNeXt-XXL@1024	41.83	1,219.47	48.00	49.88	55.09	66.14	35.69	27.60	56.67	16.92	5.00	46.90	40.98	16.00	47.32	43.40	52.75
Self Supervised																		
DINOv2	ViT-L/14@336	41.18	1,262.66	49.62	56.80	60.30	65.10	35.00	26.40	56.41	16.48	3.10	44.04	11.90	18.67	50.20	49.43	52.25
DINOv2	ViT-L/14@518	40.60	1,242.48	51.00	53.39	60.38	64.55	34.50	26.20	57.53	15.11	2.90	44.28	10.95	14.00	48.63	46.13	57.90
MoCo v3	ViT-B/16@224	34.94	966.45	36.77	33.00	47.35	62.96	32.80	26.20	55.05	16.04	2.60	43.81	10.31	6.67	45.36	39.03	52.83
MoCo v3	ViT-L/16@224	34.70	1010.18	34.64	41.71	47.46	64.70	33.70	26.30	55.05	16.24	2.70	42.60	10.39	4.00	45.36	44.67	35.16
MAE	ViT-L/16@224	37.69	1,114.07	42.30	35.93	55.20	63.51	34.60	26.00	56.10	16.11	2.70	43.63	10.83	14.00	44.80	45.81	55.75
MAE	ViT-H/14@224	38.58	1,083.35	41.15	50.99	55.30	64.90	34.10	26.00	56.49	15.63	3.20	43.98	11.00	12.00	46.30	47.18	54.90
I-JEPA	ViT-H/14@224	38.88	1,132.07	44.68	51.74	55.37	66.04	34.20	26.40	56.09	15.84	3.00	43.66	11.48	10.67	46.01	46.74	53.50
Other																		
SAM	ViT-L/16@1024	31.74	585.78	20.34	36.34	39.85	65.49	34.50	25.10	53.92	16.16	2.70	42.37	9.25	2.00	44.44	35.65	50.50
SAM	ViT-H/16@1024	32.37	648.96	22.30	36.31	40.52	65.20	34.10	26.00	54.44	15.56	2.40	42.39	8.75	2.00	45.36	34.83	55.25
MiDaS 3.0	ViT-L/16@384	35.65	981.36	38.57	40.93	49.04	63.41	31.80	25.70	54.72	16.36	2.60	43.19	11.24	6.67	44.97	38.78	53.40
MiDaS 3.1	ViT-L/16@518	35.44	983.34	34.79	40.20	48.53	64.60	33.90	25.00	55.18	15.64	2.60	42.76	12.08	6.66	43.66	39.63	52.58
Diffusion	SD2.1/16@512	36.59	1,044.28	37.71	42.00	48.38	64.55	33.40	25.70	56.99	15.56	3.10	43.14	10.40	9.33	45.88	44.68	52.40
SupViT	ViT-L/16@224	40.13	1,197.39	46.55	54.72	57.27	65.94	34.00	28.00	56.22	16.44	3.10	43.52	11.82	16.67	46.67	48.49	52.75
SupViT	ViT-H/14@224	37.45	1,082.43	42.61	48.45	52.98	63.51	35.29	26.50	55.78	15.16	3.30	44.16	11.49	4.66	43.79	44.55	52.91

1.Benchmark分析

- 是LLM还是MLLM回答问题?

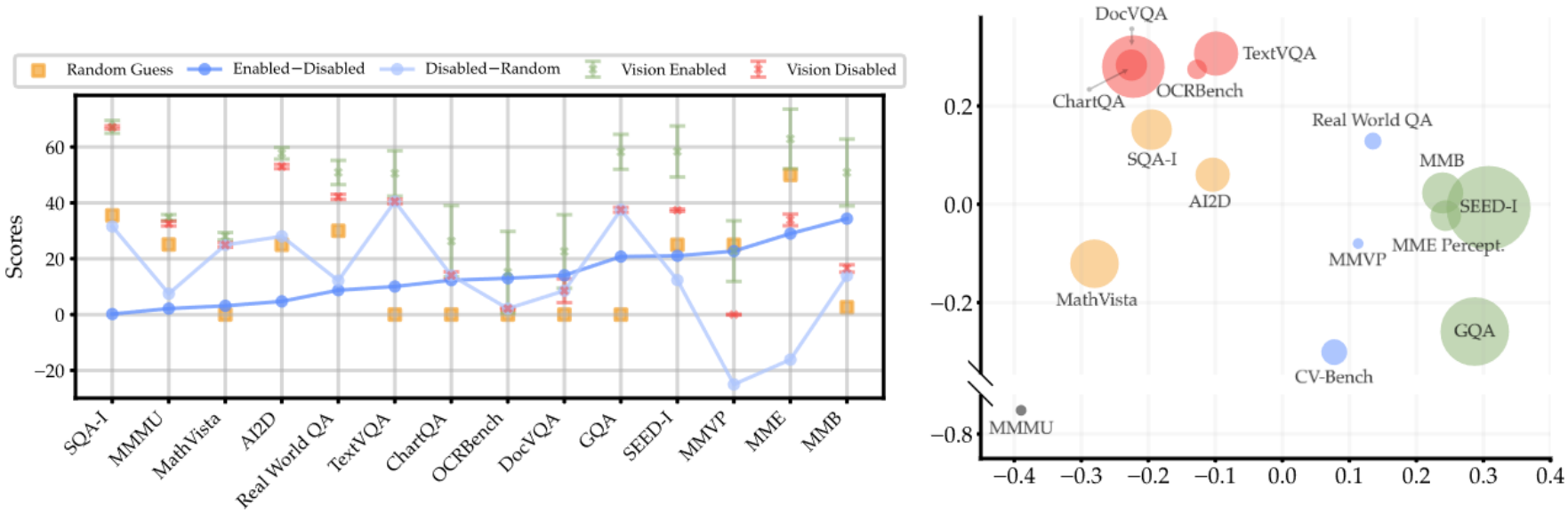
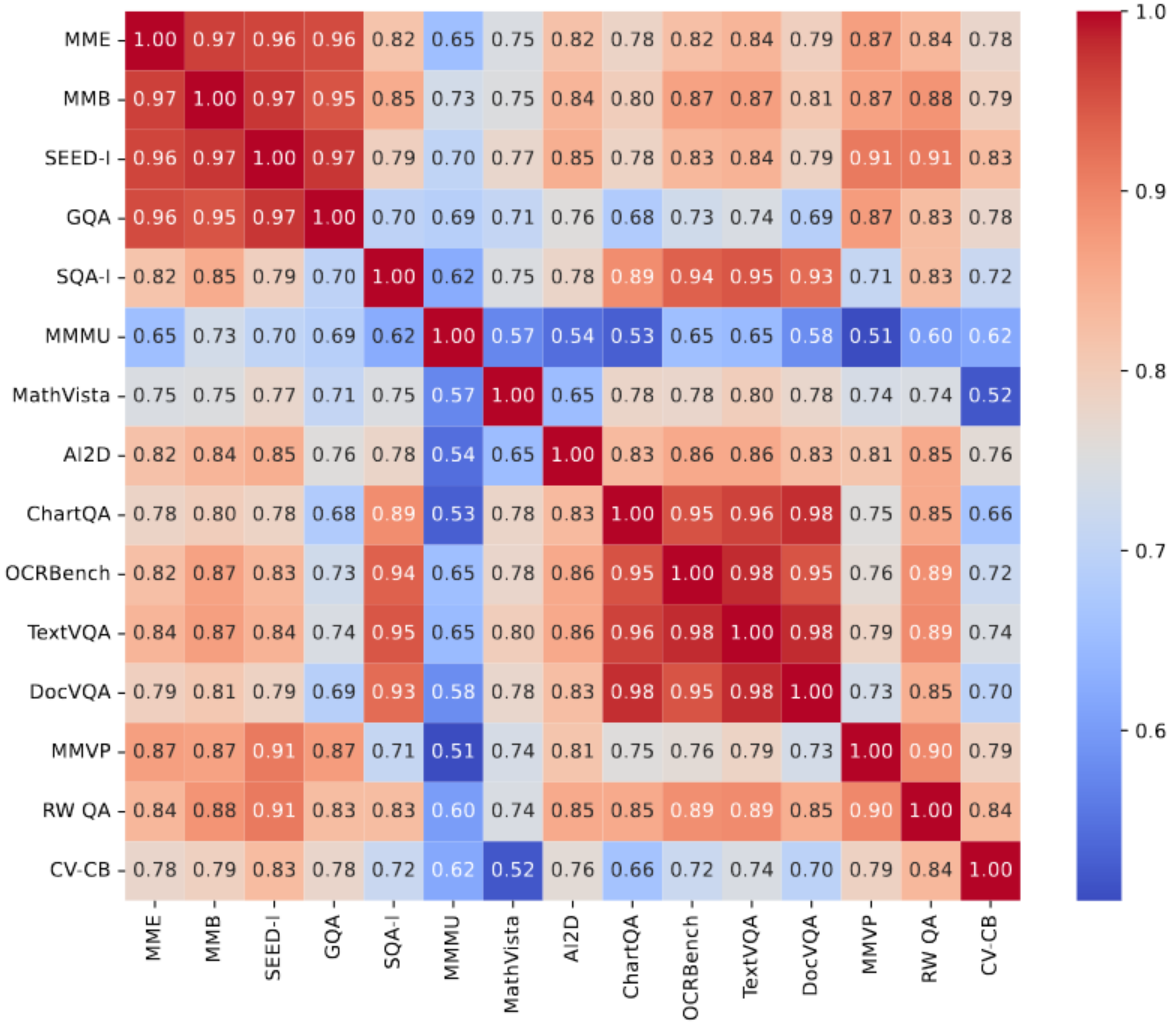


Figure 3 | **Left:** Performance comparison of MLLMs with visual input enabled and disabled across various benchmarks. Benchmarks are sorted by the difference between the average score with vision enabled and disabled. **Right:** Principal component analysis displaying clusters of benchmarks based on performance metrics, with bubble size corresponding to benchmark size. We label the clusters as "General" in green, "Knowledge" in yellow, "Chart & OCR" in red, and "Vision-Centric" in blue.

1.Benchmark分析

- Clustering the benchmarks
- 利用23个模型进行实验
- 得到每个benchmark上各个模型的分数，相当于每个模型的特征向量
- 利用这些分数去计算不同benchmark之间的关系（混淆矩阵）

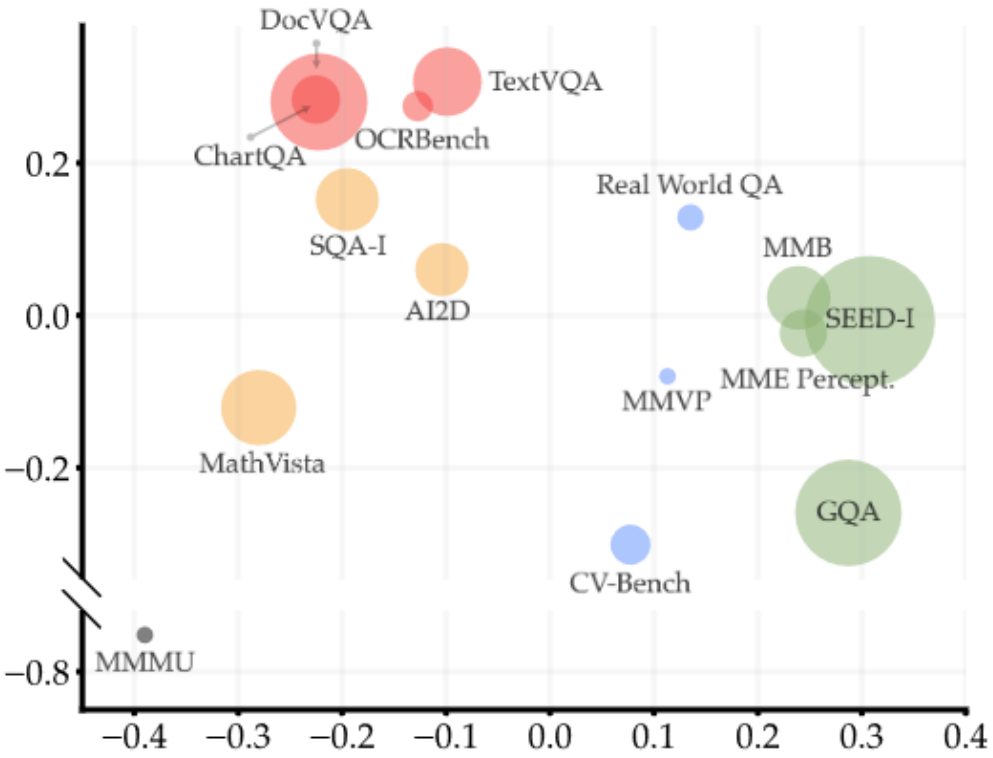


1.Benchmark分析

Finding 1: Most benchmarks do not properly measure vision-centric capabilities, and the ones that do have very few samples.

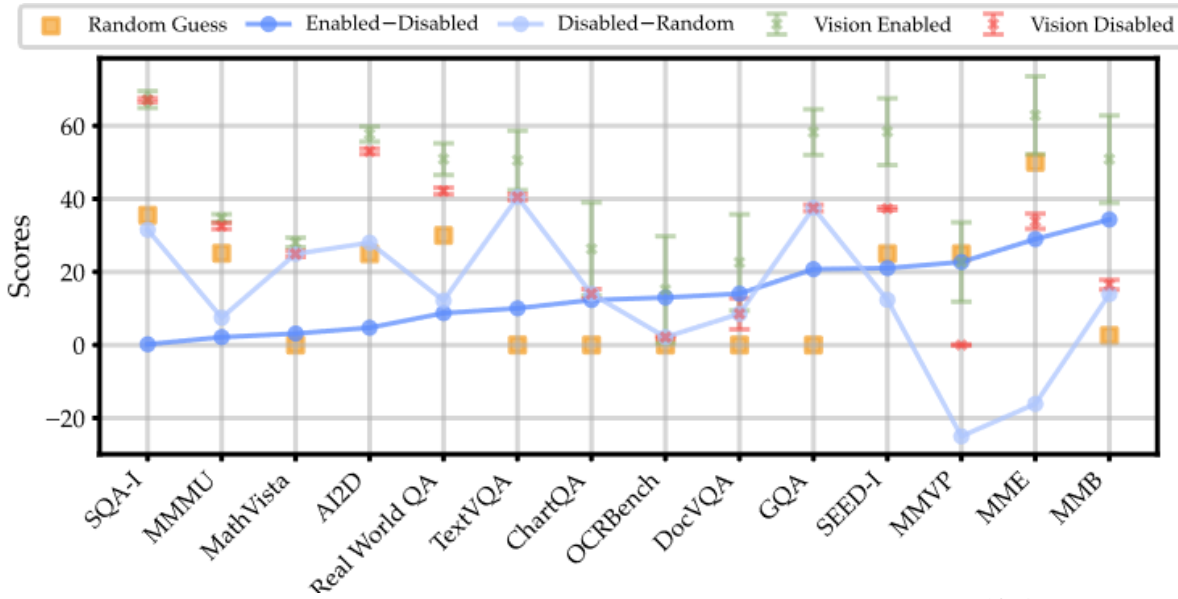
- 大部分评测集没有评估MLLM的视觉能力，以视觉为中心的评测集样本太少

- Clustering the benchmarks
- 将所有的benchmark分为四类：
  General（绿色）Knowledge（黄色）Chart&OCR （红色）Vision-Centric（蓝色）
- 视觉中心类的数据量明显不足
- MMMU最终分为知识类



1.Benchmark分析

- Random guess v.s. Vision Enabled v.s. Vision Disabled
- 科学知识类：SQA-I, MMMU, MathVista和AI2D不依赖视觉输入
- 文本类：TextVQA和GQA过于依赖语言先验（随机猜测根本猜不到）
- 视觉类：MMVP和MME中，去掉视觉效果甚至不如guess，说明visual grounding能力非常重要



2.提出Benchmark-CV Bench

Finding 2: Existing vision benchmarks can be effectively repurposed into VQA questions, enabling the assessment of vision-centric MLLM capabilities.

- 将现有纯视觉数据集构造成VQA形式的评测集，评估MLLM视觉能力

- 扩充视觉中心类的benchmark，2638个example，是MMVP的9倍
- 人工精挑细选设置问题，以VQA的格式构造

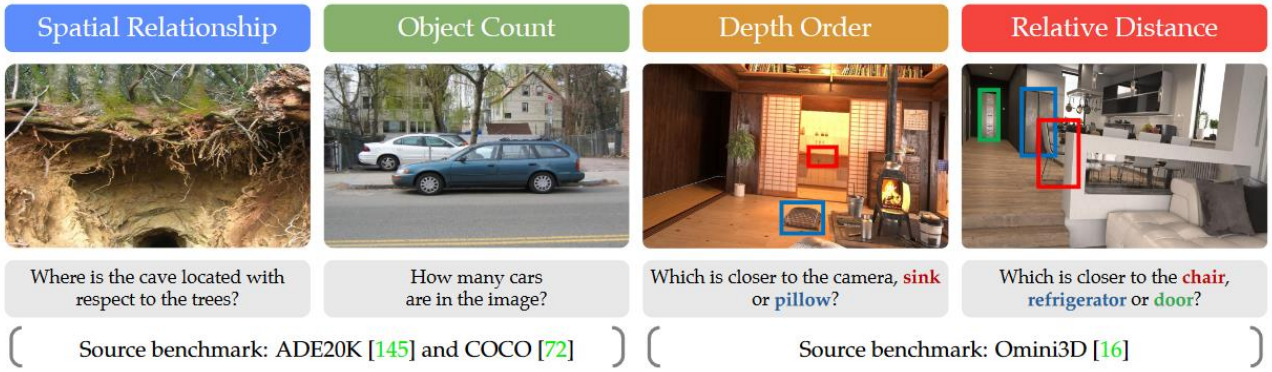


Figure 4 | **Cambrian Vision-Centric Benchmark (CV-Bench).** We repurpose standard vision benchmarks to evaluate the fundamental 2D and 3D visual understanding of MLLMs. See Section 3.2 for more details.

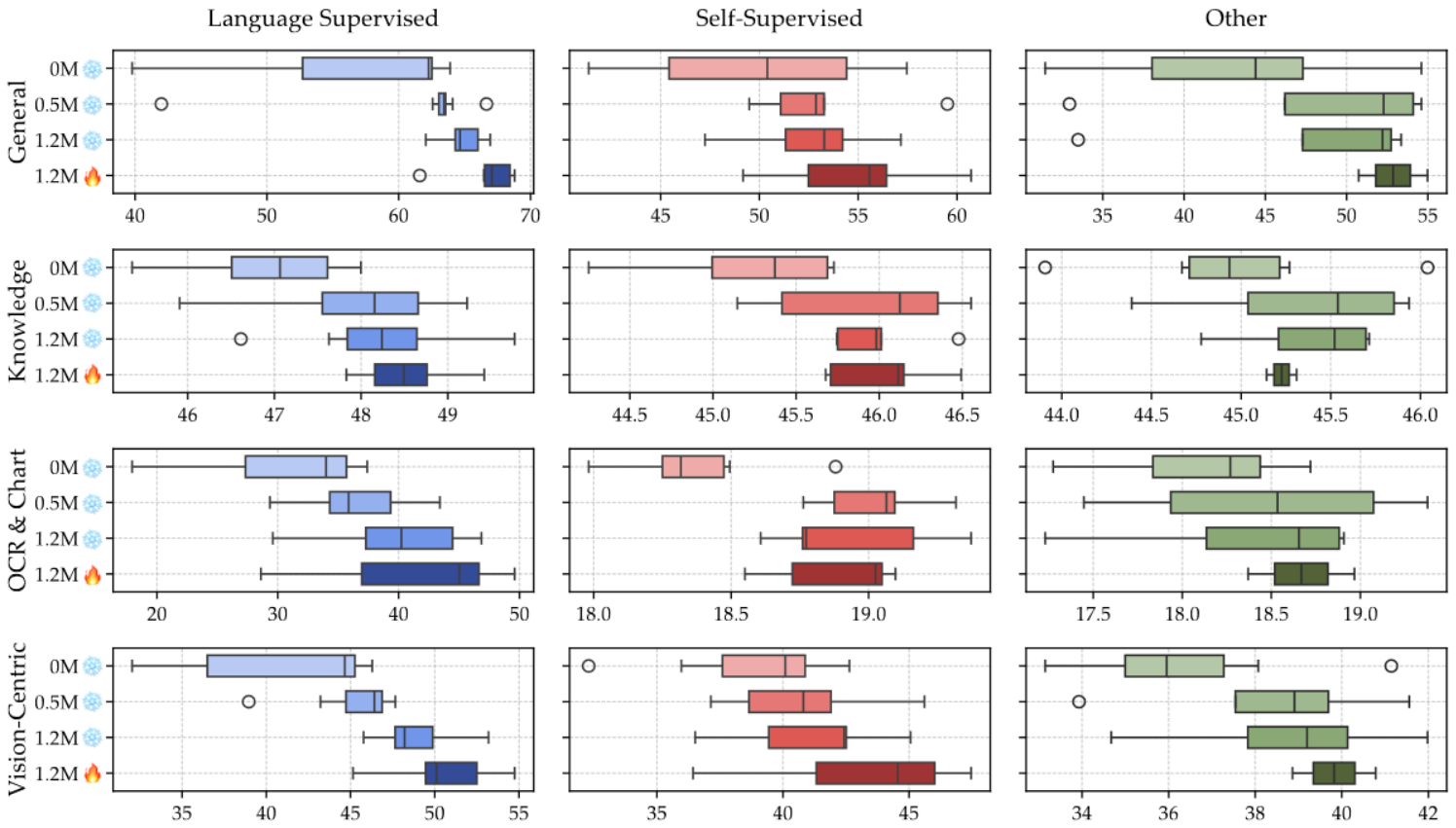
Type	Task	Description	Sources	# Samples
2D	**Spatial Relationship**	Determine the relative position of an object w.r.t. the anchor object. Consider left-right or top-bottom relationship.	ADE20K COCO	650
	Object Count	Determine the number of instances present in the image.	ADE20K COCO	788
3D	**Depth Order**	Determine which of the two distinct objects is closer to the camera.	Omni3D	600
	Relative Distance	Determine which of the two distinct objects is closer to the anchor object.	Omni3D	600

3.Instruct Tuning Recipe

Finding 3: Two-stage training is beneficial; more adapter data further improves results.

Finding 4: Unfreezing the vision encoder is widely beneficial. Language-supervised models always benefit; SSL models particularly benefit on vision-centric benchmarks.

- 两阶段训练更好，增加第一阶段数据、不固定视觉部分参数均可以提升性能



Supervision Type	Method
Language-Supervised	
Language	OpenAI CLIP
	DFN-CLIP
	DFN-CLIP
	EVA-CLIP-02
	SigLIP
	SigLIP
	OpenCLIP
	OpenCLIP
	OpenCLIP
Self-Supervised	
Contrastive	DINOv2
	DINOv2
	MoCo v3
	MoCo v3
Masked	MAE
	MAE
JEPA	I-JEPA
Other	
Segmentation	SAM
	SAM
Depth	MiDaS 3.0
	MiDaS 3.1
Diffusion	Stable Diffusion 2.1
Class Labels	SupViT
	SupViT

4.MLLM as a Visual Rep. Eval.

- 用MLLM而不是ImageNet-1k+linear prob来评估视觉模型性能

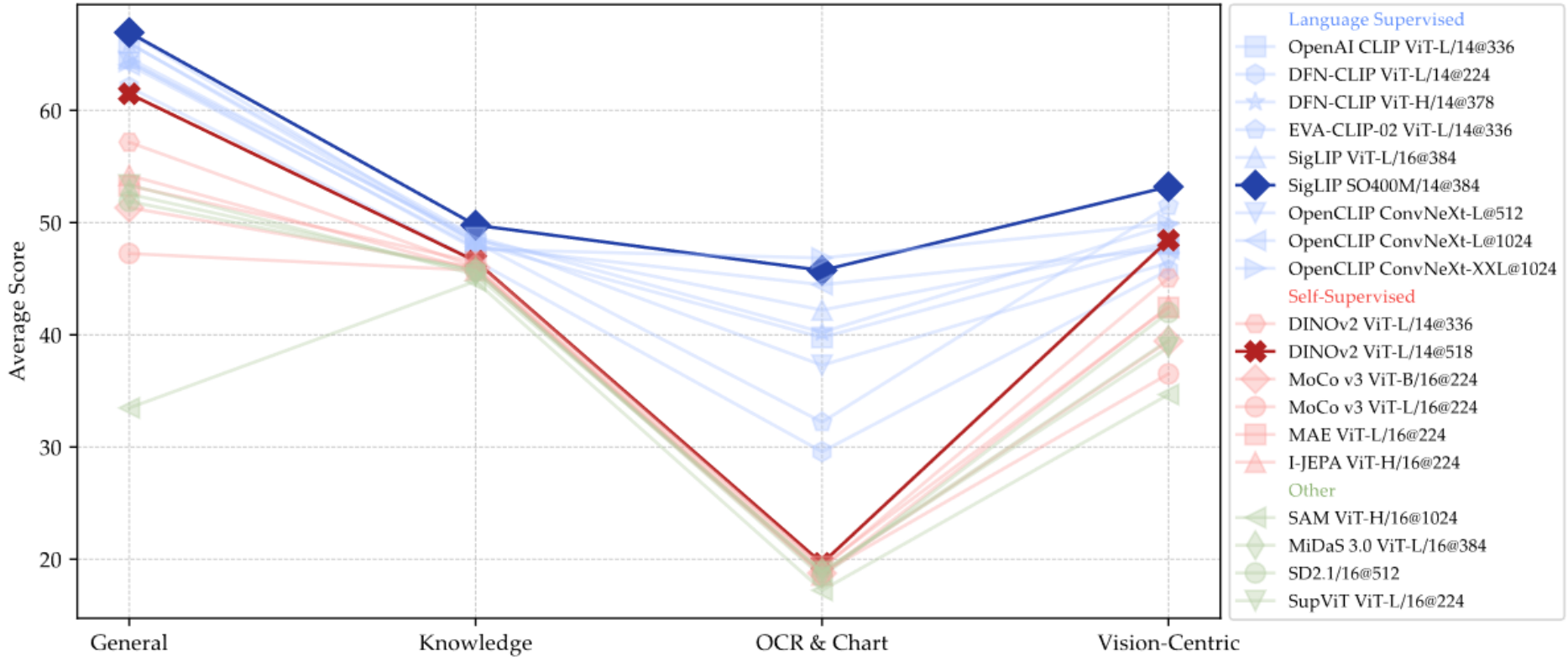


Figure 6 | **Evaluating Visual Representations with MLLMs** While language-supervised models outperform self-supervised or other models, a well-trained self-supervised model like DINOv2 can also achieve competitive performance on vision-centric tasks.

4.MLLM as a Visual Rep. Eval.

Finding 5: High-res encoders greatly enhance performance on chart & vision-centric benchmarks, and ConvNet-based architectures are inherently well-suited for such tasks.

- 高分辨率模型对图表和视觉数据集更好，卷积模型适合处理这些任务

Language Supervised

Model	Architecture	All	G	K	O	V
SigLIP	ViT-SO400M/14@384	1	1	1	2	1
OpenCLIP	ConvNeXt-XXL@1024	2	6	8	1	3
DFN-CLIP	ViT-H/14@378	3	4	2	5	4
OpenCLIP	ConvNeXt-L@1024	4	8	7	3	8
SigLIP	ViT-L/16@384	5	5	4	4	6
OpenAI CLIP	ViT-L/14@336	6	3	6	6	7
EVA-CLIP-02	ViT-L/14@336	7	2	5	8	2
OpenCLIP	ConvNeXt-L@512	8	7	3	7	9
DFN-CLIP	ViT-L/14@224	9	9	9	9	10
DINOv2*	ViT-L/14@518	10	10	10	10	5

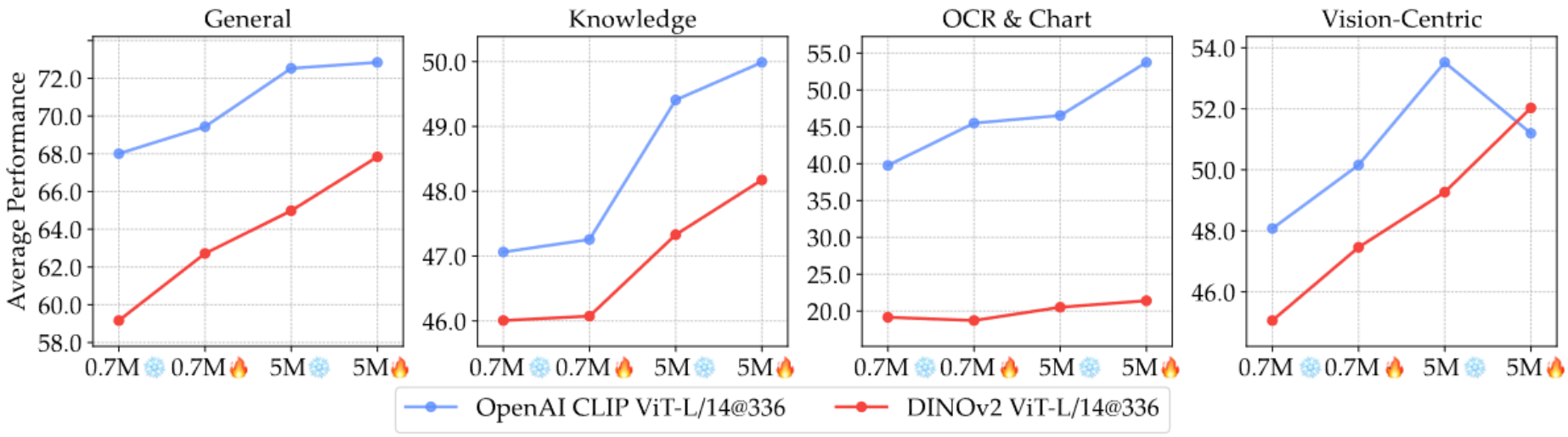
Self-Supervised & Other

Model	Architecture	All	G	K	O	V
DINOv2	ViT-L/14@518	1	1	1	1	1
DINOv2	ViT-L/14@336	2	2	3	3	2
MAE	ViT-L/16@224	3	5	2	2	4
I-JEPA	ViT-H/14@224	4	3	6	8	3
SD2.1	VAE+UNet/16@512	5	7	9	9	5
MiDaS 3.0	ViT-L/16@384	6	6	8	5	6
SupViT	ViT-L/16@224	7	4	9	4	8
MoCo v3	ViT-B/16@224	8	8	4	7	7
MoCo v3	ViT-L/16@224	9	9	5	6	9
SAM	ViT-H/16@1024	10	10	10	10	10

4.MLLM as a Visual Rep. Eval.

Finding 6: Language supervision offers strong advantages, but the performance gap can be narrowed with SSL methods given enough data and proper tuning.

- 语言监督模型比自监督要好，但是性能差距可以通过增加第二阶段数据和解冻视觉参数来弥补



5.Combine Multiple Encoders

Finding 7: Combining multiple vision encoders, including vision SSL models, enhances MLLM performance across various benchmarks, particularly in vision-centric tasks.

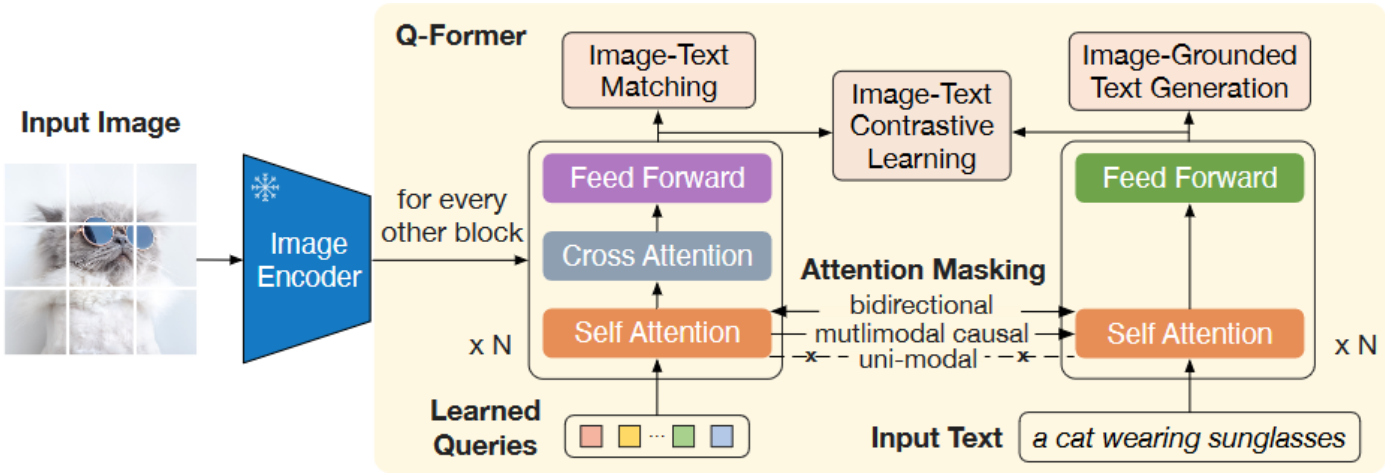
- 组合多种视觉模型可以提升性能，尤其是视觉benchmark

Vision Backbone		General				Knowledge				OCR & Chart				Vision-Centric			
Method	Average	MME^P	MMB	$SEED^I$	GQA	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	ChartQA	OCRBench	TextVQA	DocVQA	MMVP	RealWorldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
SigLIP+DINOv2	51.61	1,432.02	61.28	65.99	63.30	68.82	35.69	29.40	60.01	43.00	35.70	60.40	37.54	30.00	53.99	55.52	53.58
SigLIP+DINOv2+ConvNext	54.52	1,503.51	63.83	67.97	63.95	70.40	35.99	29.30	60.69	48.20	36.90	64.97	45.53	34.67	58.69	55.74	60.33
SigLIP+DINOv2+ConvNext+CLIP	54.74	1,479.46	63.32	67.63	64.04	71.39	35.49	29.10	59.88	50.24	39.60	64.55	46.12	32.67	58.95	58.54	60.42
SigLIP+ConvNext	54.53	1,494.97	64.60	67.98	63.58	71.05	34.90	29.80	60.85	50.64	38.00	64.53	46.52	32.00	57.91	58.83	56.58
CLIP+ConvNext	54.45	1,511.08	63.83	67.41	63.63	70.80	35.09	30.40	59.91	51.32	35.00	64.45	47.88	33.33	57.25	56.32	59.08
SigLIP+DINOv2+ConvNext	53.78	1,450.64	63.57	67.79	63.63	71.34	34.80	30.20	61.04	49.32	37.70	64.05	45.83	30.00	56.21	58.08	54.33
SigLIP+CLIP+ConvNext	54.53	1,507.28	63.23	68.64	63.63	71.10	35.89	30.90	59.97	52.36	38.50	65.40	47.92	28.67	57.25	57.66	55.92

Table 3 | **All Benchmark Results for Model Ensemble with 1.2M Adapter Data + 737K Instruction Tuning Data.** Here, "SigLIP" = ViT-SO400M/14@384, "DINOv2" = ViT-L/14@518, "ConvNext" = OpenCLIP ConvNeXt-XXL@1024, and "CLIP" = OpenAI CLIP ViT-L/14@336.

5.Combine Multiple Encoders

- 如何融合多个视觉模型的输出？采用类似Q-former设计
  - 采用一组固定数量的query，利用cross-attn，将任意数量的视觉token压缩到固定数量的query中



5.Combine Multiple Encoders

Finding 8: Spatial inductive bias and deep interaction between LLM and vision feature help to better aggregate and condense vision features.

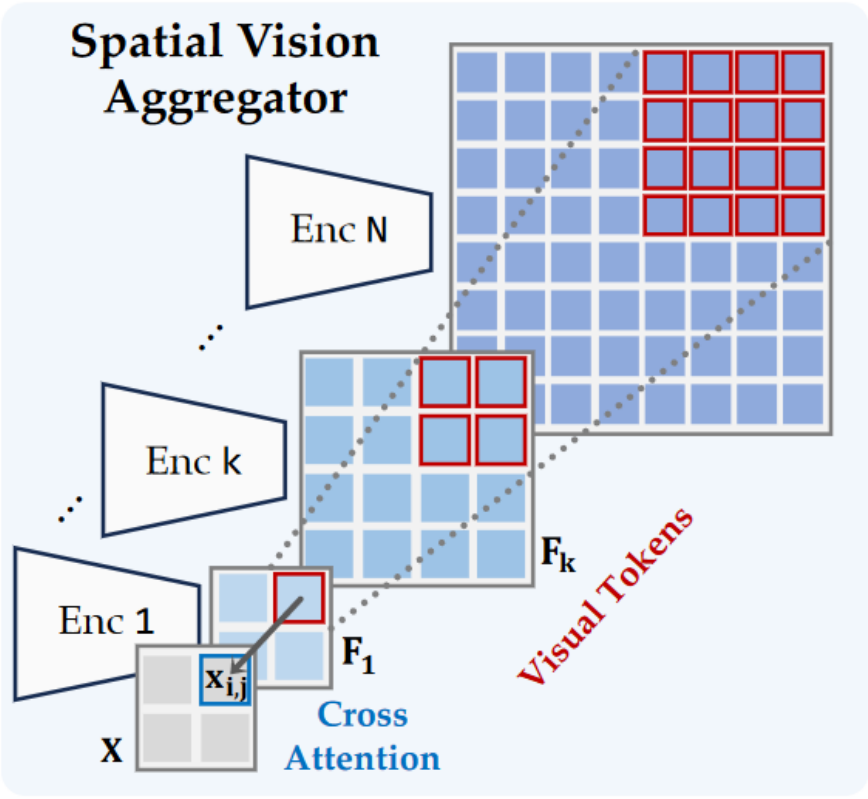
- 加入空间归纳偏置并与LLM结合可以更好聚合与压缩视觉特征

- 一个token对应不同encoder输出的对应区域，可以做到空间维度的压缩
- 利用单个query token与一串k,v进行cross-attn，只在对应局部区域内，相当于引入归纳偏置

$$\mathbf{F}_k[m_k \cdot i : m_k \cdot (i+1), m_k \cdot j : m_k \cdot (j+1)] \in \mathbb{R}^{m_k^2 \times C}$$

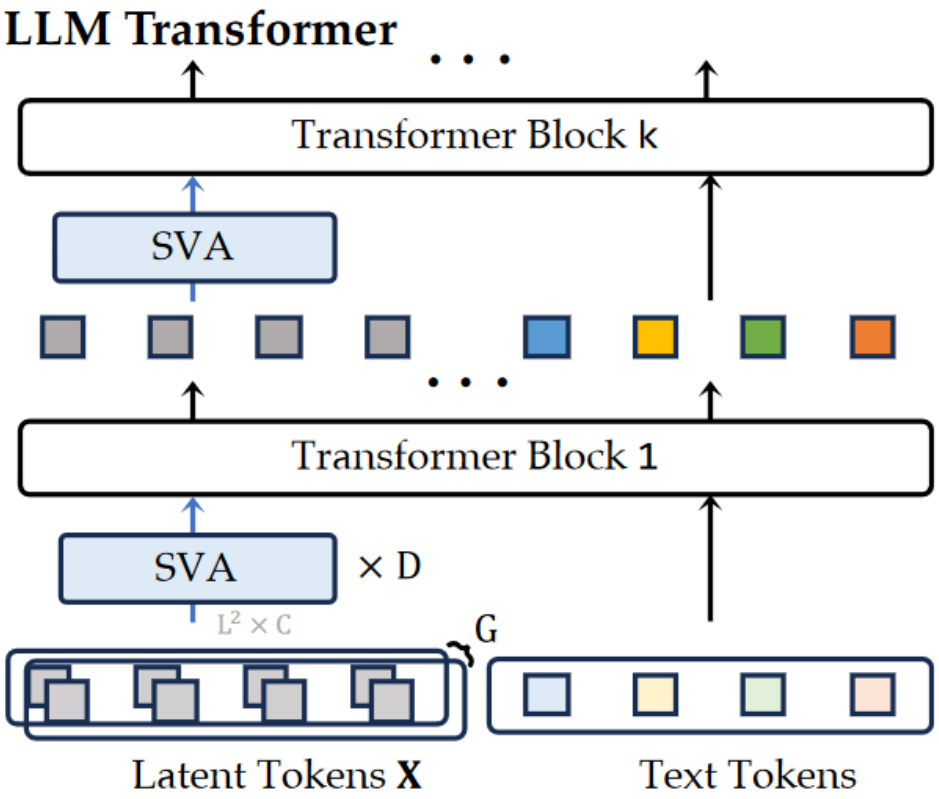
$$\mathbf{q}^*_{i,j} = \text{softmax}\left(\frac{\mathbf{q}_{i,j} \cdot [\mathbf{k}_{i,j,1}, \mathbf{k}_{i,j,2}, \ldots, \mathbf{k}_{i,j,N}]^\top}{\sqrt{C}}\right)[\mathbf{v}_{i,j,1}, \mathbf{v}_{i,j,2}, \ldots, \mathbf{v}_{i,j,N}],$$

$$\mathbf{q}_{i,j} = \mathbf{W}^Q \mathbf{x}_{i,j} \in \mathbb{R}^{1 \times C},$$
$$\mathbf{k}_{i,j,k} = \mathbf{W}_k^K \mathbf{F}_k[m_k \cdot i : m_k \cdot (i+1),\ m_k \cdot j : m_k \cdot (j+1)] \in \mathbb{R}^{m_k^2 \times C},$$
$$\mathbf{v}_{i,j,k} = \mathbf{W}_k^V \mathbf{F}_k[m_k \cdot i : m_k \cdot (i+1),\ m_k \cdot j : m_k \cdot (j+1)] \in \mathbb{R}^{m_k^2 \times C}.$$



5.Combine Multiple Encoders

- 将这样的设计应用到LLM内部，使得信息能够进行多层级注入和保留
- 每隔一些层注入一次，在LLM前面构造多组token，每次SVA可以堆叠多层



D	OCR & Chart
2	52.1
3	52.4
4	**52.8**

G	OCR & Chart
1	52.4
2	52.6
3	**53.1**

Connector	General	Knowledge	OCR & Chart	Vision-Centric
Concat. [117]	67.2	48.9	50.1	52.6
Resampler [51]	63.1	46.5	27.1	42.6
SVA-no-multi-agg	68.0	49.5	55.2	52.6
SVA	**68.5**	**49.7**	**55.5**	**53.2**

Multi-agg	OCR & Chart
No	52.4
Yes	**53.3**

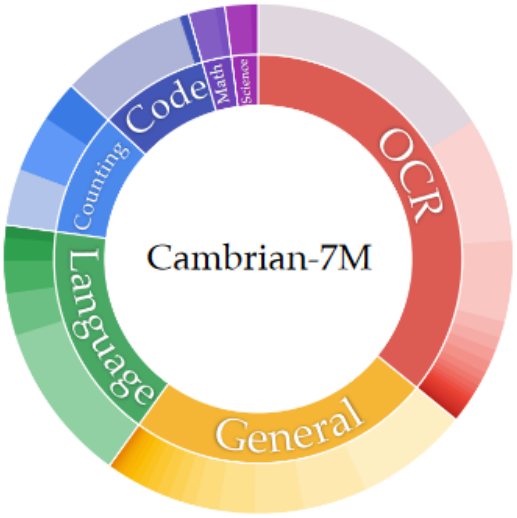
结论汇总

- 大部分评测集没有评估MLLM的视觉能力，以视觉为中心的评测集样本太少
- 将现有纯视觉数据集构造成VQA形式的评测集，评估MLLM视觉能力
- 两阶段训练更好，增加第一阶段数据、不固定视觉部分参数均可以提升性能
- 高分辨率模型对图表和视觉数据集更好，卷积模型适合处理这些任务
- 语言监督模型比自监督要好，但是性能差距可以通过增加第二阶段数据和解冻视觉参数来弥补
- 组合多种视觉模型可以提升性能，尤其是视觉benchmark
- 加入空间归纳偏置并与LLM结合可以更好聚合与压缩视觉特征

- 作者介绍
- 研究背景
- 本文探索
- 数据集构建
- 实验效果
- 总结

Instruct Tuning data

扩充Instruct Tuning data



OCR (27.6%)
- Filtered DVQA (1550.0 K)
- DVQA [54] (775.0 K)
- SynthDog [60] (500.0 K)
- ArxivQA [69] (100.0 K)
- OCRVQA [93] (80.0 K)
- ScreenQA [49] (79.0 K)
- WikiSQL [144] (74.0 K)
- Low-Level Vision [22] (50.0 K)
- DocVQA [90] (39.0 K)
- WTQ [99] (38.0 K)
- ChartQA [89] (28.0 K)
- IconQA [82] (27.0 K)
- Chart2Text [55] (26.0 K)
- TabMWP [81] (23.0 K)
- TextCaps [111] (22.0 K)
- LLAVAR [140] (20.0 K)
- ST-VQA [15] (17.0 K)
- AI2D [58] (15.0 K)
- RenderedText [125] (10.0 K)
- VisText [115] (9.0 K)
- FinQA [26] (6.0 K)
- InfoVQA [14] (2.0 K)
- TAT-QA [148] (2.0 K)
- HiTab [27] (2.0 K)

General (33.3%)
- ALLaVA [20] (700.0 K)
- Q-Instruct [126] (400.0 K)
- LNQA [101] (302.0 K)
- LVIS-Instruct4V [122] (220.0 K)
- LLaVA150K [75] (150.0 K)
- VisualGenome [62] (86.0 K)
- VQAv2 [43] (83.0 K)
- GPT4V Rewritten (77.0 K)
- GQA [50] (72.0 K)
- A-OKVQA [108] (50.0 K)
- AlfWorld [137] (45.0 K)
- ShareGPT [22] (40.0 K)
- RefCOCO [131] (30.0 K)
- VizWiz [44] (20.0 K)
- Visual7W [149] (14.0 K)
- LAION GPT-4V [63] (11.0 K)
- IDK [17] (11.0 K)
- OKVQA [88] (9.0 K)
- HatefulMemes [59] (8.0 K)
- OODVQA [120] (8.0 K)
- SketchyVQA [120] (8.0 K)
- Visualmrc [114] (3.0 K)

Language (23.8%)
- OpenOrca [71] (994.0 K)
- MathInstruct [133] (262.0 K)
- OrcaMath [92] (200.0 K)
- WizardCoder [86] (143.0 K)
- OpenCodeInterpreter [143] (66.0 K)
- Dolly [30] (11.0 K)

Counting (8.5%)
- Filtered CLEVR (350.0 K)
- CLEVR [52] (350.0 K)
- TallyQA [1] (250.0 K)

Code (0.8%)
- Filtered WebSight (790.0 K)
- WebSight [64] (10.0 K)
- DaTikz [12] (47.0 K)
- Design2Code [110] (0.5 K)

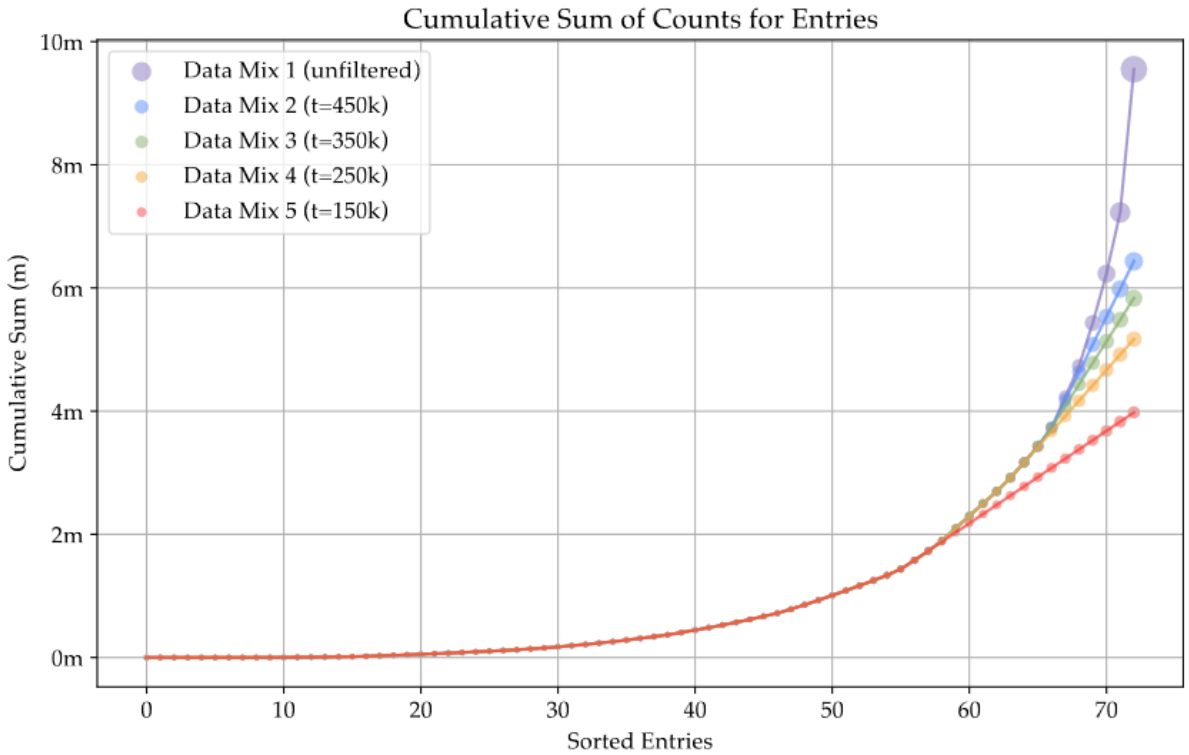
Math (3.2%)
- Geo170K [37] (170.0 K)
- RAVEN [139] (42.0 K)
- GeomVerse [57] (9.0 K)
- MathVision [123] (3.0 K)
- Inter-GPS [83] (1.0 K)
- TQA [4] (1.0 K)

Science (2.9%)
- Data Engine (161.0 K)
- PathVQA [46] (32.0 K)
- ScienceQA [84] (12.0 K)

Figure 9 | **Cambrian-7M: A Large-Scale Curated Instruction Tuning Dataset for MLLM. Left:** The inner circle shows the original distribution of Cambrian-10M. The outer circle shows the curated Cambrian-7M. **Right:** All the data sources in the Cambrian dataset as well as the ones filtered in data curation.

Instruct Tuning data

- 控制每个数据集来源的数量，实现平衡
- 从10M数据中构造出7M高质量数据

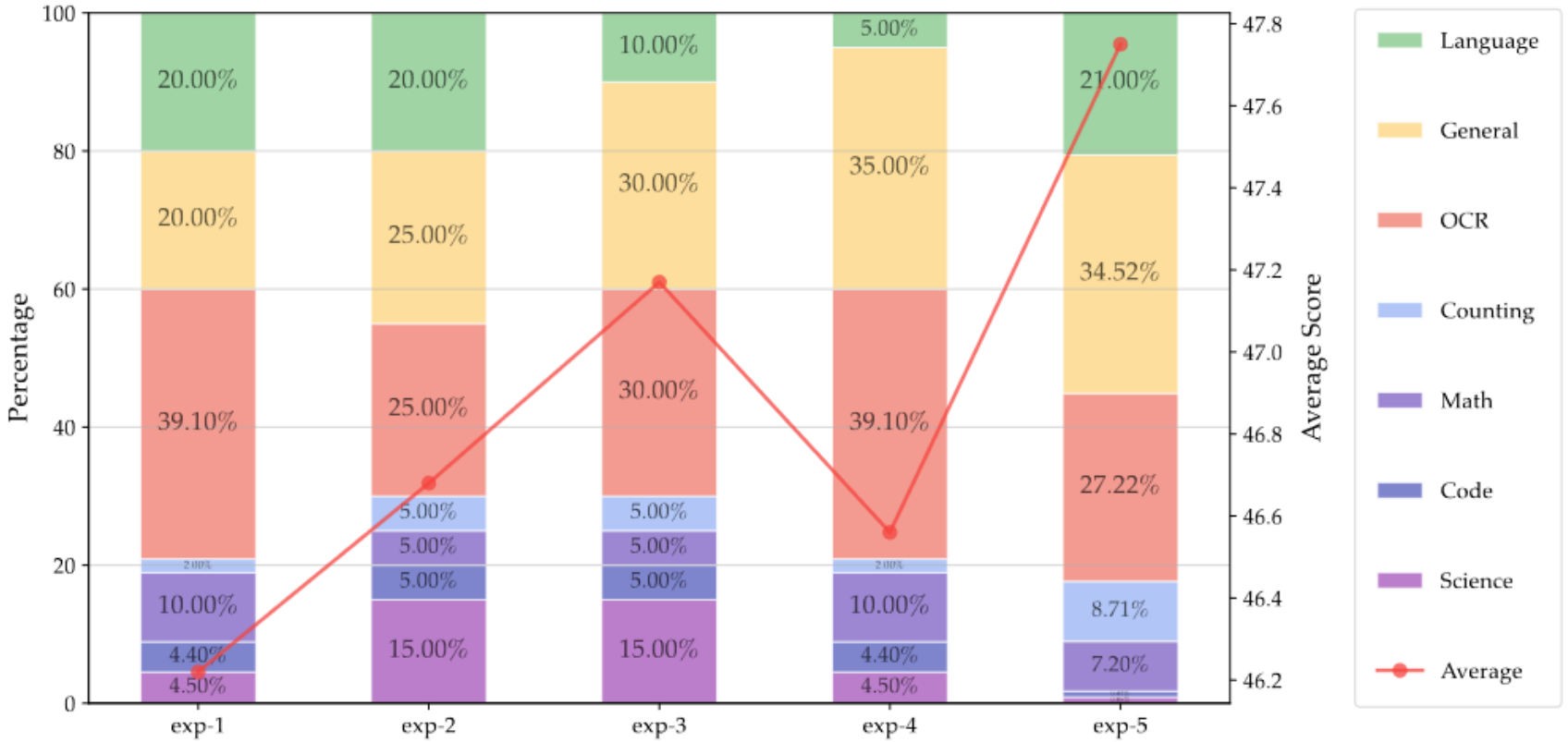


	Average	General	Knowledge	OCR & Chart	Vision-Centric
150k	53.7	68.0	51.3	45.2	50.5
250k	**54.3**	**68.1**	51.5	45.3	52.2
350k	54.3	67.4	51.4	**46.0**	**52.3**
450k	54.2	68.0	**52.2**	45.5	50.7

Instruct Tuning data

	Average	General	Knowledge	OCR & Chart	Vision-Centric
LLaVA-665K	40.7	64.7	45.2	20.8	32.0
Cambrian-10M	54.8	68.7	51.6	**47.3**	51.4
Cambrian-7M	**55.9**	**69.6**	**52.6**	**47.3**	**54.1**

- 高质量且比例均衡数据效果最好，纯文本数据防止灾难性遗忘



- 作者介绍
- 研究背景
- 本文探索
- 数据集构建
- 实验效果
- 总结

SOTA对比

Model		General					Knowledge					OCR & Chart					Vision-Centric				
Method	# Vis Tok.	Avg	MME^P	MMB	$SEED^I$	GQA	Avg	SQA^I	$MMMU^V$	$MathVista^M$	AI2D	Avg	ChartQA	OCRBench	TextVQA	DocVQA	Avg	MMVP	RealworldQA	$CV\text{-}Bench^{2D}$	$CV\text{-}Bench^{3D}$
GPT-4V	UNK.	63.0	1409.4	75.8	69.1	36.8	65.2	75.7	56.8	49.9	78.2	77.4	78.5	64.5	78.0	88.4	62.4	50.0	61.4	64.3	73.8
Gemini-1.0 Pro	UNK.	-	1496.6	73.6	70.7	-	-	79.5	47.9	45.2	-	-	-	65.9	-	-	-	-	-	-	-
Gemini-1.5 Pro	UNK.	-	-	-	-	-	-	-	58.5	52.1	80.3	-	81.3	-	73.5	86.5	-	-	67.5	-	-
Grok-1.5	UNK.	-	-	-	-	-	-	-	53.6	52.8	88.3	-	76.1	-	78.1	85.6	-	-	68.7	-	-
MM-1-8B	144	-	1529.3	72.3	69.9	-	-	72.6	37.0	35.9	-	-	-	-	-	-	-	-	-	-	-
MM-1-30B	144	-	1637.6	75.1	72.1	-	-	81.0	44.7	39.4	-	-	-	-	-	-	-	-	-	-	-
Base LLM: Llama-3-Ins-8B																					
Mini-Gemini-HD-8B	2880	72.7	**1606.0**	72.7	73.2	64.5	55.7	75.1	37.3	37.0	**73.5**	62.9	59.1	47.7	70.2	74.6	51.5	18.7	62.1	62.2	63.0
LLaVA-NeXT-8B	2880	72.5	1603.7	72.1	72.7	**65.2**	55.6	72.8	41.7	36.3	71.6	63.9	69.5	49.0	64.6	72.6	56.6	38.7	60.1	62.2	65.3
Cambrian-1-8B	576	**73.1**	1,547.1	**75.9**	**74.7**	64.6	**61.3**	**80.4**	**42.7**	**49.0**	73.0	**71.3**	**73.3**	**62.4**	**71.7**	**77.8**	**65.0**	**51.3**	**64.2**	**72.3**	**72.0**
Base LLM: Vicuna-1.5-13B																					
Mini-Gemini-HD-13B	2880	70.7	1597.0	68.6	70.6	63.7	54.1	71.9	37.3	37.0	70.1	60.8	56.6	46.6	70.2	69.8	49.4	19.3	57.5	53.6	67.3
LLaVA-NeXT-13B	2880	69.9	1575.0	70.0	65.6	**65.4**	53.7	73.5	36.2	35.1	70.0	62.9	62.2	51.4	67.1	70.9	55.9	36.0	59.1	62.7	65.7
Cambrian-1-13B	576	**73.7**	**1,610.4**	**75.7**	**74.4**	64.3	**60.2**	**79.3**	**40.0**	**48.0**	**73.6**	**71.3**	**73.8**	**61.9**	**72.8**	**76.8**	**62.2**	**41.3**	**63.0**	**72.5**	**71.8**
Base LLM: Hermes2-Yi-34B																					
Mini-Gemini-HD-34B	2880	76.2	1659.0	80.6	75.3	65.8	62.4	77.7	48.0	43.4	**80.5**	68.1	67.6	51.8	74.1	**78.9**	63.8	37.3	67.2	71.5	79.2
LLaVA-NeXT-34B	2880	76.0	1633.2	79.3	**75.9**	**67.1**	62.5	81.8	46.7	46.5	74.9	67.7	68.7	54.5	69.5	78.1	64.0	47.3	61.0	73.0	74.8
Cambrian-1-34B	576	**76.8**	**1689.3**	**81.4**	75.3	65.8	**67.0**	**85.6**	**49.7**	**53.2**	79.7	**71.9**	**75.6**	**60.0**	**76.7**	75.5	**68.5**	**52.7**	**67.8**	**74.0**	**79.7**

SOTA对比

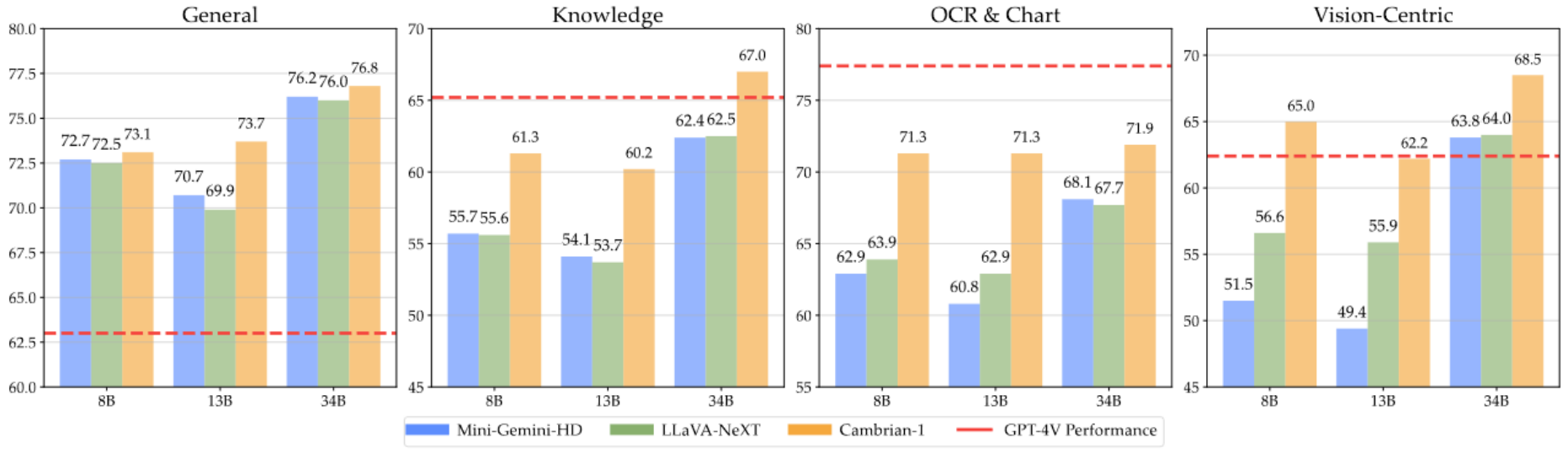


Figure 13 | **Comparison of model average performances on each category.** Cambrian-1 outperforms other open-source models across all sizes. The lead is especially large on OCR & Chart and Vision-Centric benchmarks, highlighting the advantage of our vision-centric design.

效果展示

Why is this image funny?

The humor in this image comes from the unexpected twist on the classic poem 'Roses are red, violets are blue.' Instead of a romantic or sweet sentiment, the poem is humorously altered to a more mundane and slightly unsettling scenario, which is a bug in the bed. This juxtaposition of a traditional romantic poem with a common, albeit slightly unpleasant, household occurrence creates a comedic effect.

Which character is this?

The character in the image is Homer Simpson, a central character from the animated television series 'The Simpsons.' Homer is known for his distinctive yellow skin, casual attire, and often humorous or oblivious behavior.

List all the countries in this image in the format of a json {1: 2:, etc}. Why are the flags here?

```
{
"1": "Austria",
"2": "Belgium",
"3": "France",
"4": "Germany",
"5": "Portugal",
"6": "Scotland",
"7": "Spain",
"8": "Turkey"
}
```

The flags are here to represent the participating countries in the UEFA EURO 2024 tournament, which is hosted by Germany. Each flag corresponds to a country that is either a host or a participant in the event, as indicated by the tickets below the flags. The tickets suggest that these countries are either hosting matches or have qualified to play in the tournament.

What does this image remind you of?

The image might remind you of a playful and creative culinary presentation, possibly inspired by a world map, where each piece of food represents a different continent. It could also evoke thoughts of a fun cooking challenge or a themed meal where the food is arranged to mimic the shape of the Earth.

In 1937, Chinese American screen actor Anna May Wong, who had portrayed numerous villains and secondary characters but never a heroine, finally got a starring role in Paramount Pictures' *Daughter of Shanghai*, a film that _____ "expanded the range of possibilities for Asian images on screen."

Which choice completes the text so that it conforms to the conventions of Standard English?

A) critic, Stina Chyn, claims
B) critic, Stina Chyn, claims,
C) critic Stina Chyn claims
D) critic Stina Chyn, claims,

Please solve the problem step by step:

To solve this problem, we need to determine which option correctly completes the sentence in a way that adheres to the conventions of Standard English. The sentence is:
"In 1937, Chinese American screen actor Anna May Wong, who had portrayed numerous villains and secondary characters but never a heroine, finally got a starring role in Paramount Pictures' Daughter of Shanghai, a film that _____ "expanded the range of possibilities for Asian images on screen."
The correct option should be the one that maintains the subject-verb agreement and the appropriate punctuation.
Option A: "critic, Stina Chyn, claims" - This option is incorrect because it uses a comma after the name, which is not standard in English.
Option B: "critic, Stina Chyn, claims," - This option is also incorrect because it uses a comma after the name and an extra comma at the end of the sentence, which is not standard in English.
Option C: "critic Stina Chyn claims" - This option is correct because it maintains the subject-verb agreement and does not use any unnecessary commas.
Option D: "critic Stina Chyn, claims," - This option is incorrect because it uses a comma after the name and an extra comma at the end of the sentence, which is not standard in English.
Therefore, the correct answer is Option C: "critic Stina Chyn claims."

- 作者介绍
- 研究背景
- 本文探索
- 数据集构建
- 实验效果
- 总结

总结反思

- 大部分评测集没有评估MLLM的视觉能力，以视觉为中心的评测集样本太少
- 将现有纯视觉数据集构造成VQA形式的评测集，评估MLLM视觉能力
- 两阶段训练更好，增加第一阶段数据、不固定视觉部分参数均可以提升性能
- 高分辨率模型对图表和视觉数据集更好，卷积模型适合处理这些任务
- 语言监督模型比自监督要好，但是性能差距可以通过增加第二阶段数据和解冻视觉参数来弥补
- 组合多种视觉模型可以提升性能，尤其是视觉benchmark
- 加入空间归纳偏置并与LLM结合可以更好聚合与压缩视觉特征

谢谢！