

# 迈向实用化多模态大模型





# 大模型成为人工智能前沿制高点

- 多模态大模型的能力突破，成为大模型前沿发展的新趋势和焦点
- 极大拓展大模型能力边界和应用场景，智能跃迁的下一个关键引擎

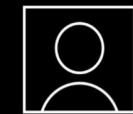
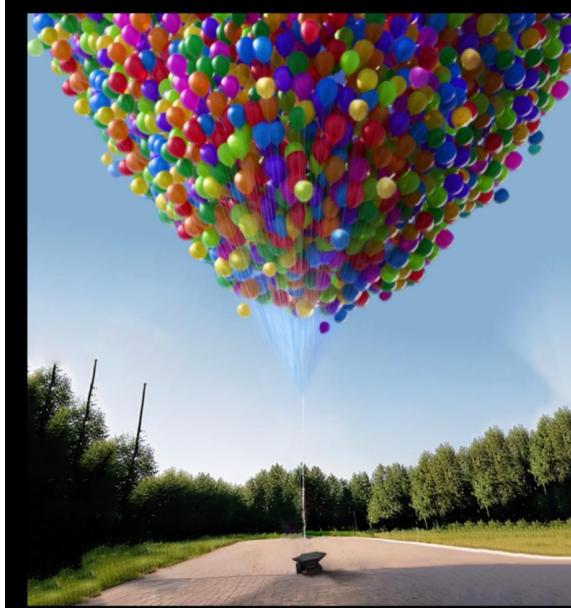
## 主要特点

基于大数据和人类反馈进行学习，与用户通过对话进行交互，能够处理多模态信息及多种任务，深层推理与常识运用能力大幅度提升

## 里程碑事件

2023/03/15: OpenAI 发布多模态对话模型 GPT-4  
2023/09/20: OpenAI 发布文生图模型 DALL·E 3  
2023/09/25: ChatGPT 全面整合多模态能力  
2023/12/06: Google 发布多模态模型 Gemini  
2024/03/04 : Anthropic发布多模态模型 Claude 3  
2024/05/13: OpenAI发布原生多模态模型 GPT-4o

## 多模态能力



What would happen if  
the strings were cut?



The balloons would  
fly away.

# 多模态大模型迈向实用的痛点问题

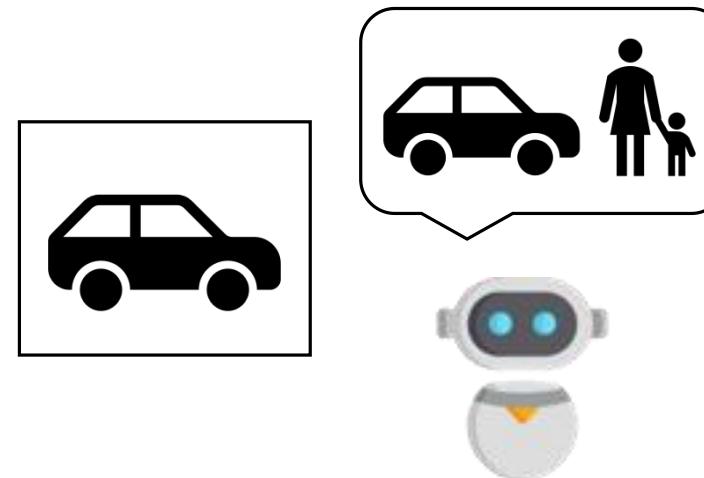
## 参数多，成本高

通常 70亿-550亿 参数量  
需要高性能服务器部署服务  
用户需上传隐私使用数据



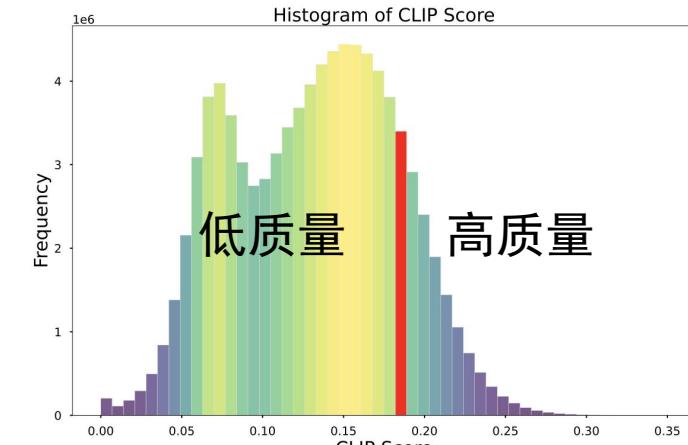
## 幻觉多，难置信

模型回复与图像内容不符  
开源模型幻觉率 60% 以上  
高风险场景危害大  
(自动驾驶、盲人助手等)



## 英文强，中文弱

多数开源模型仅支持英语  
中文多模态数据少、质量低  
难以满足中文用户需求

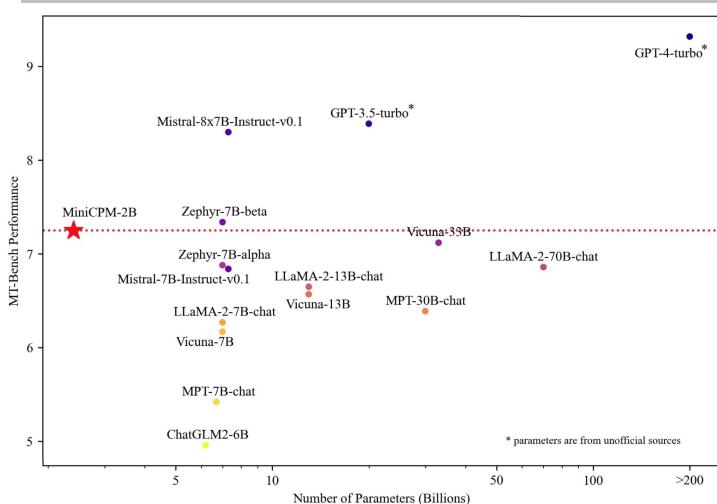


中文图文数据质量分布

# 多模态大模型迈向实用的痛点问题

## 参数多，成本高 端侧大模型

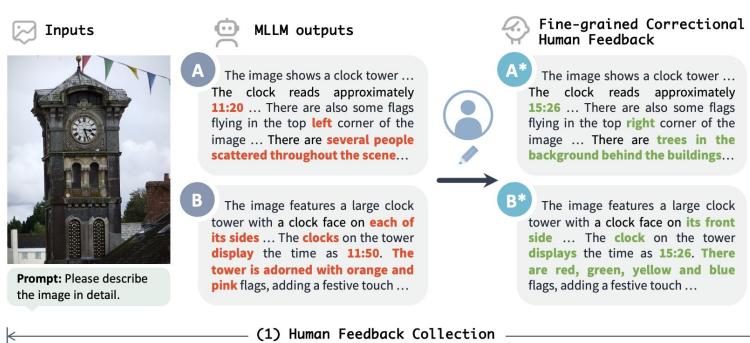
具有超出预期的巨大潜力  
大模型 Scaling Law 沙盒  
可在手机高效部署运行



MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. 2024.

## 幻觉多，难置信 多模态反馈学习

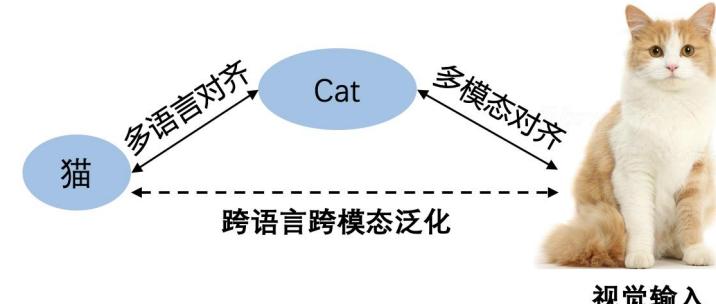
通过反馈数据对齐模型行为  
可大幅降低模型幻觉  
使高风险场景更加可信



RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. CVPR 2024.

## 英文强，中文弱 多模态跨语言泛化

基于中英双语语言基座  
仅使用英文图文数据预训练  
实现中文多模态跨语言泛化

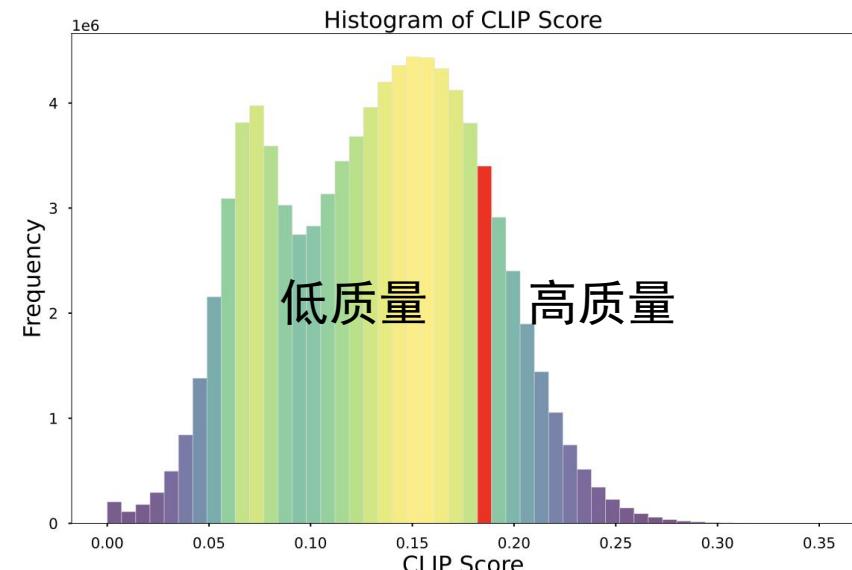
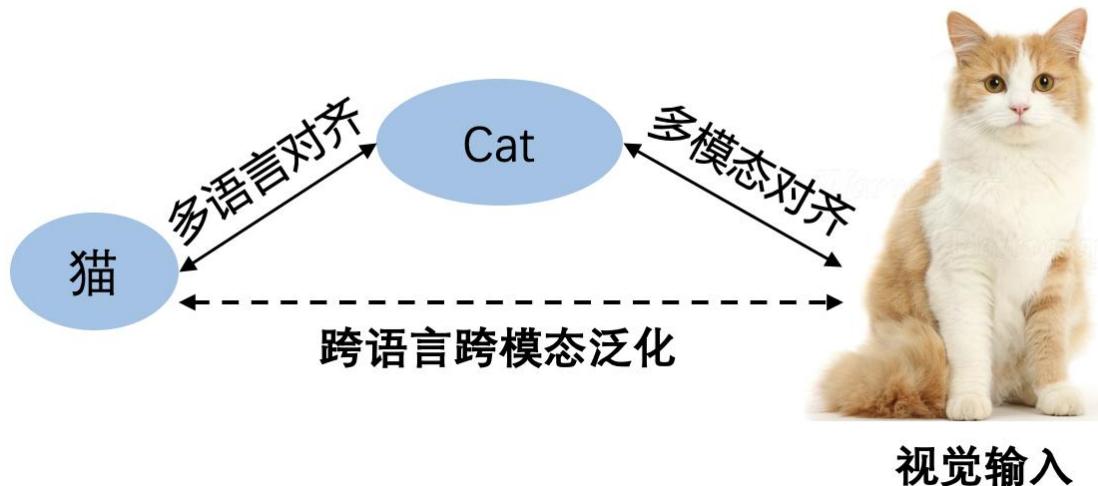


Large Multilingual Models Pivot Zero-Shot Multimodal Learning across Languages. ICLR 2024.



## 中英双语多模态能力

- 基于团队在ICLR 2024论文中提出的 **多模态能力跨语言泛化** 独家技术
- **挑战：** 中文领域缺乏高质量、大规模的多模态图文数据
- **技术：**
  - **基础能力跨语言泛化：** 基于语言模型基座 MiniCPM 的中英双语能力，可**仅通过 英文通用域 图文对数据 进行预训练**，实现多模态基础能力从英文到中文的跨语言泛化
  - **中文多模态能力高效激发：** 使用**少量高质量中文指令微调**数据，高效激发中文多模态基础能力



中文图文对数据质量分布

# 中英双语多模态能力

效果：

- 基于中英双语大语言模型，将英文多模态能力泛化到中文场景中

Was ist an diesem Bild ungewöhnlich?



An unusual aspect of the image is the man sitting on a folding chair in the back of a parked yellow taxi cab. This is not a typical scene, as one would expect passengers to be seated inside the cab or at the very least, be standing near the taxi. The man's presence in the back of the taxi, using a folding chair, adds an unconventional and unexpected element to the scene, making it stand out from regular taxi activities. Another notable aspect of the image is the presence of multiple people and vehicles, including other cars and pedestrians in the background. This urban setting contributes to the overall busy atmosphere, making the man on the chair in the taxi stand out even more.



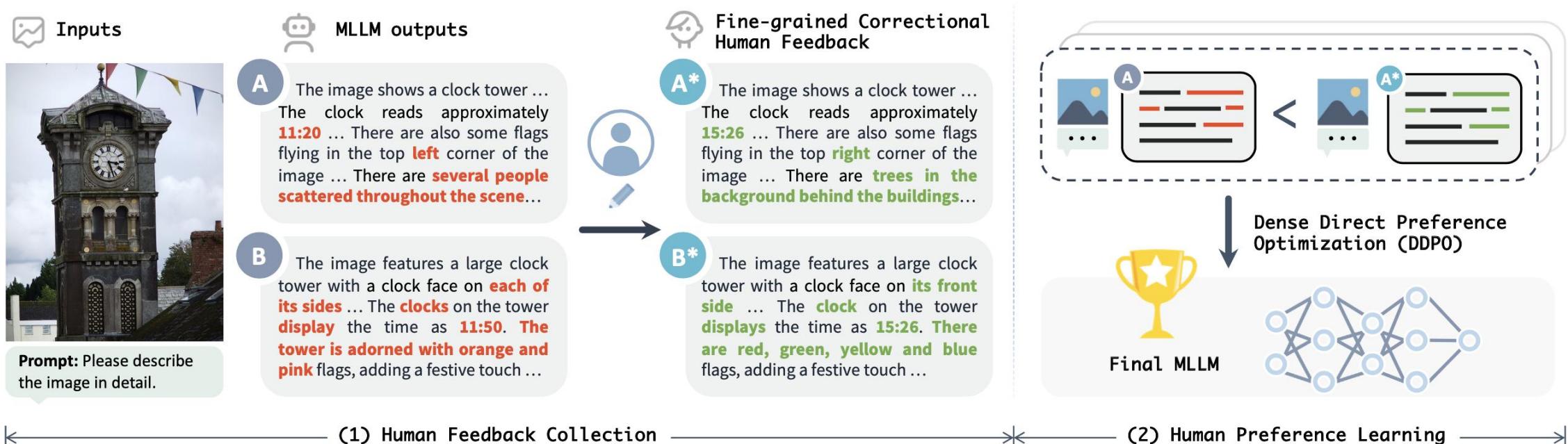
你能用苏轼的诗句描述这幅图像吗?  
(Can you describe this image with Su Shi's poems?)

在这张照片中，月亮高高地挂在天空上，照亮了古老的中国建筑和它周围的环境。这创造出一种美丽而宁静的气氛，让人想起苏轼著名的诗句：“明月几时有？把酒问青天。不知天上宫阙，今夕是何年？”  
In this photo, the moon hangs high in the sky, illuminating the ancient Chinese building and its surroundings. This creates a beautiful and serene atmosphere, reminiscent of Su Shi's famous poem: "When is the bright moon? Ask the blue sky for wine. I don't know the palace in the sky. What year is it?")



# 可信多模态行为

- 基于团队在CVPR 2024论文中提出的 **多模态偏好学习** 独家系列技术 RLHF-V
- **挑战:** 传统排序式人类反馈，标注歧义性大、反馈针对性差、学习效率低
- **技术:**
  - **人类偏好反馈:** 提出细粒度修正式人类偏好，消除标注歧义性，提升反馈针对性
  - **对齐方法:** 使用密集DPO方法，修正语段受到更大强度反馈，提升学习效率



# 可信多模态行为

- 基于团队在CVPR 2024论文中提出的 **多模态偏好学习** 独家系列技术 RLHF-V 最新方法 RLAIF-V
- **挑战：**已有反馈数据来自人工标注或私有多模态模型，规模化困难

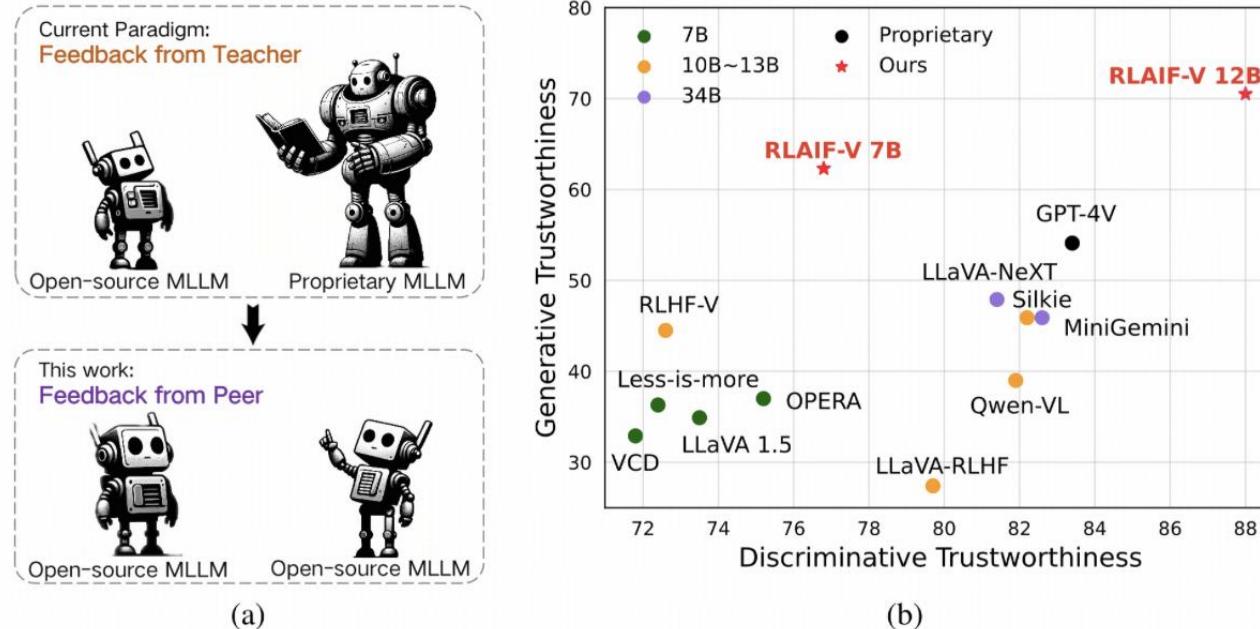
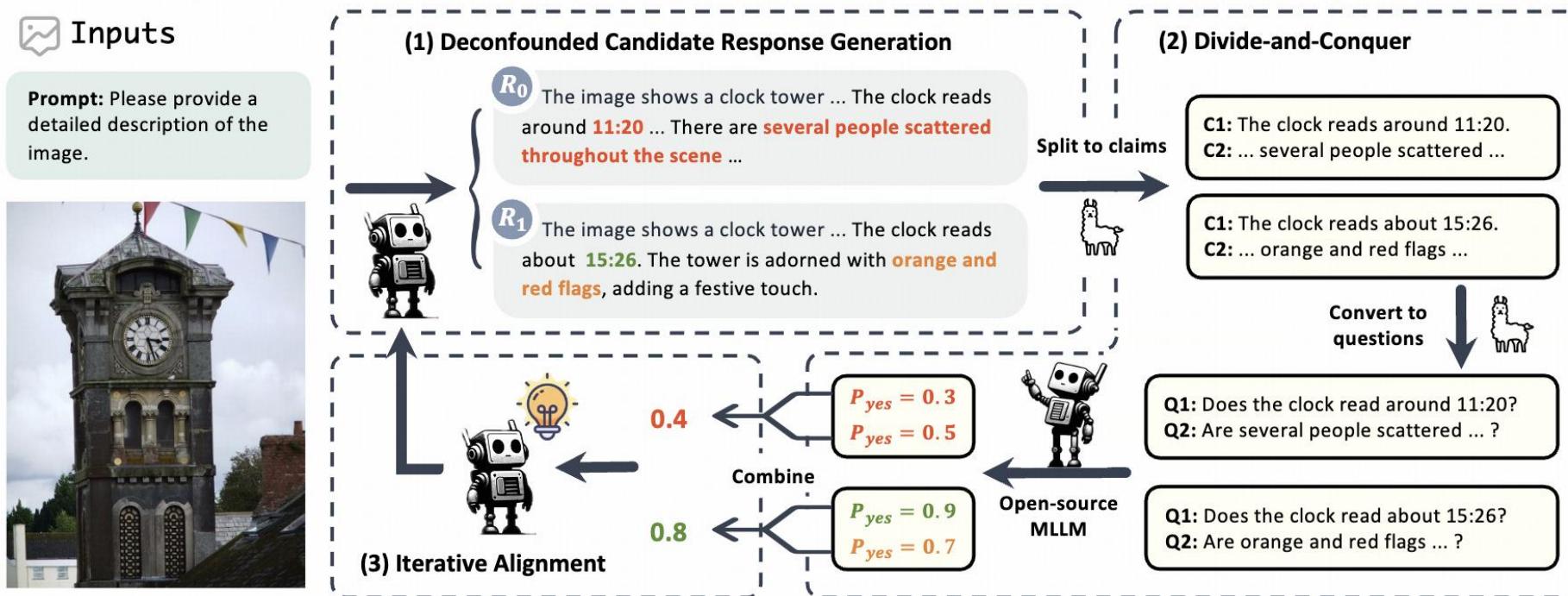


Figure 1: (a) This work aims to shift the current paradigm of aligning MLLMs with feedback from superior teachers, to align with feedback from peers exhibiting comparable or equal capabilities. (b) Trustworthiness of RLAIF-V compared to other methods. We assess the generative trustworthiness with human evaluation benchmark MHumanEval [58], and evaluate the discriminative trustworthiness with automatic evaluation benchmark AMBER [51].

# 可信多模态行为

- 基于团队在CVPR 2024论文中提出的 **多模态偏好学习** 独家系列技术 RLHF-V 最新方法 RLAIF-V
- **挑战:** 已有反馈数据来自人工标注或私有多模态模型，规模化困难
- **技术:**
  - **开源模型反馈:** 提出对比生成-拆分-反馈流程，使开源多模态模型能够提供高质量反馈
  - **迭代对齐:** 使用迭代式对齐方法，减小训练中的分布偏移，提高训练效率



# 高效端侧多模态大模型 MiniCPM-V 系列

2024/2/1

MiniCPM-V (2.8B) 发布

18:49 0.79 HD S2

- 实现同参数规模**最佳性能**, 支持**中英双语**
- 可**部署**安卓和鸿蒙系统**手机**
- 国际开源平台累计下载次数超过**4万次**
- OCR能力仍显不足, 限制进一步应用**

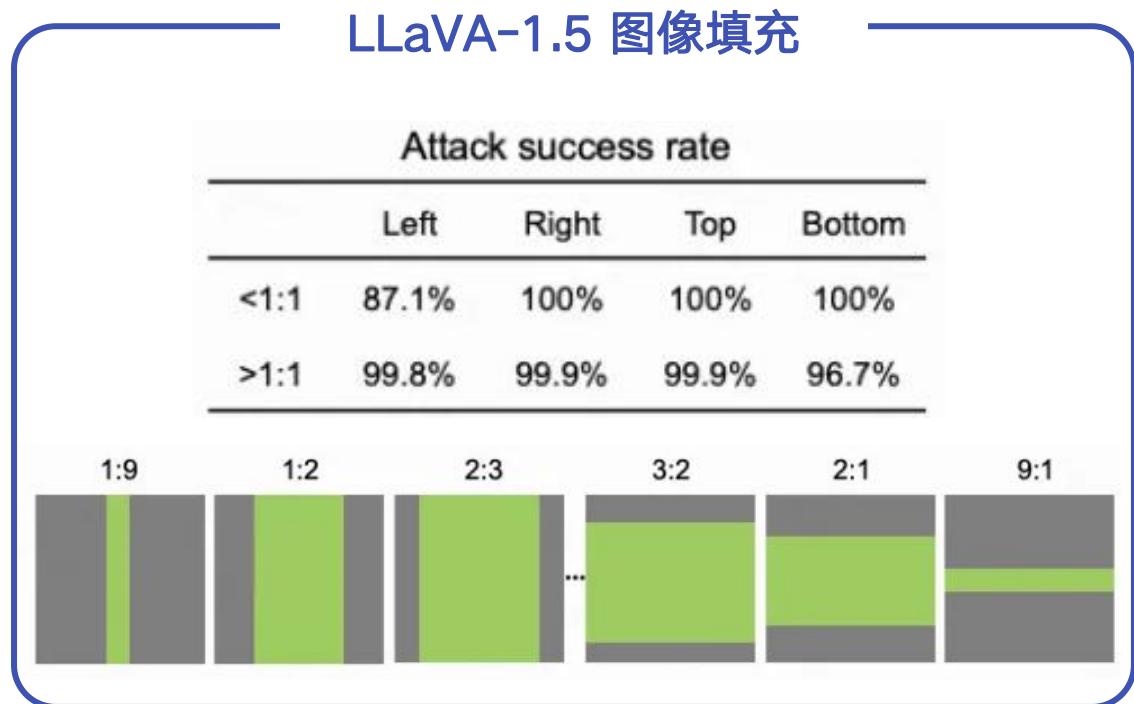
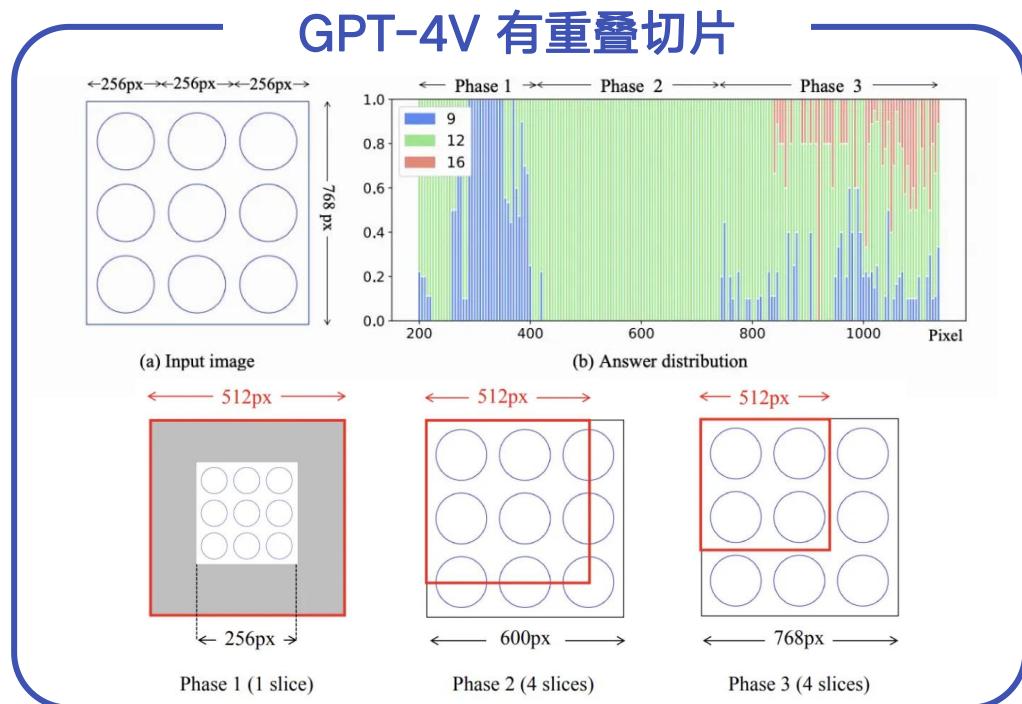
通用性能榜单

Model	Size	Visual Tokens	MME	MMB dev (en)	MMB dev (zh)	MMMU val	CMMMU val
LLaVA-Phi	3B	576	1335	59.8	-	-	-
MobileVLM	3B	144	1289	59.6	-	-	-
Imp-v1	3B	576	1434	66.5	-	-	-
Qwen-VL-Chat	9.6B	256	1487	60.6	56.7	35.9	30.7
CogVLM	17.4B	1225	1438	63.7	53.8	32.1	-
MiniCPM-V 1.0	3B	64	1452	67.9	65.3	37.2	32.1



# 高清图高效编码

- OCR能力的提升依赖于高清图像编码，团队近期提出 大模型高清图编码 独家技术 LLaVA-UHD
- 挑战：**编码高清变长宽比图像的 适应性 (adaptivity) 和效率 (efficiency)
- 视觉编码漏洞：**
  - 漏洞 1：GPT-4V 将图像进行有重叠的切片后再编码视觉压缩模块
  - 漏洞 2：LLaVA-1.5 进行大范围的图像填充 (Padding)

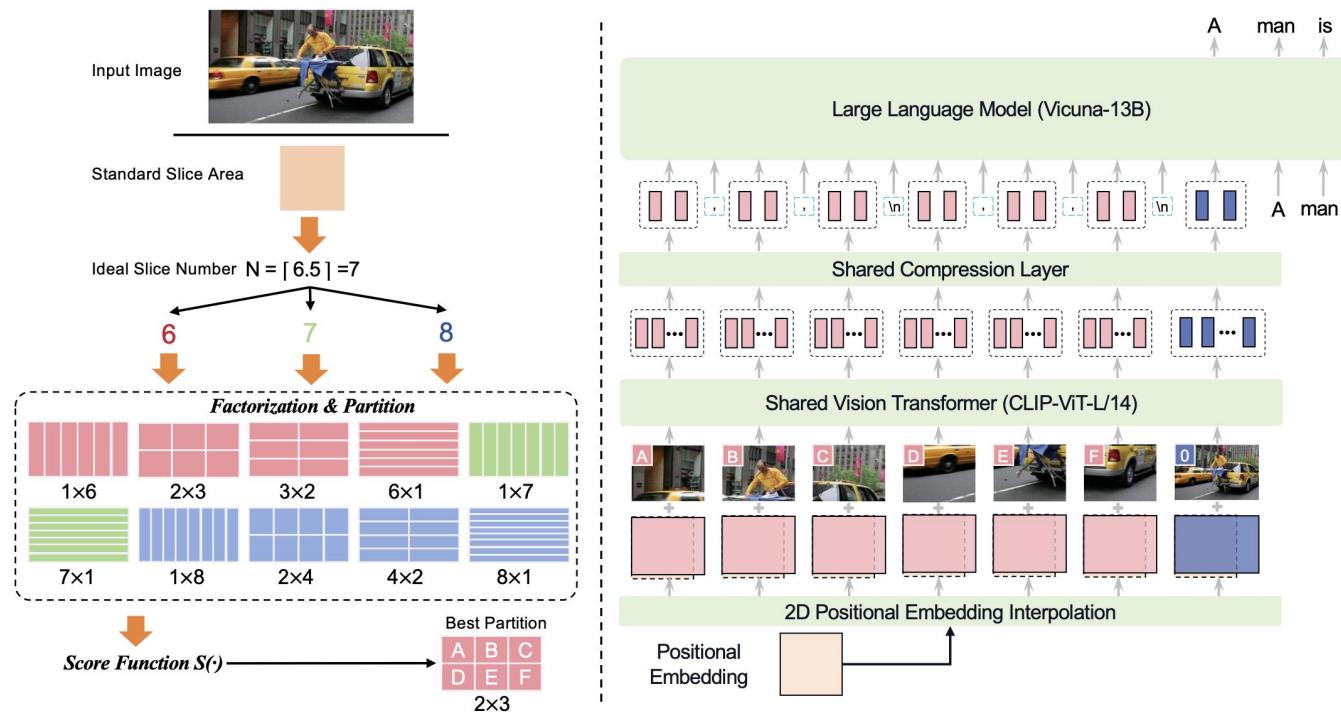




# 高清图高效编码

## 技术

- 模块化图像编码策略：将高清图切分为可变大小的切片，选择最佳切分进行高效和适应性的编码
- 视觉压缩模块：使用感知重采样器压缩来自视觉编码器的视觉符号数量，大幅提升计算效率
- 空间修饰方法：使用自然语言符号，为大语言模型编码图像切片的相对位置关系





# 高清图高效编码

## 效果：

- 图像细节感知能力显著提高：**通过简单地感知原生高分辨率图像，LLaVA-UHD **多种通用数据集、文字识别、以及多模态幻觉**方面均取得了效果提高。
- 具有优秀的计算效率：**在计算效率和可拓展性上，LLaVA-UHD **仅使用 94% 的推理计算**即支持任意宽高比的 **672×1088 分辨率图像**，能够适应更大分辨率的拓展。

Model	#Data	MaxRes.	AR.	TFLOPs	VQA <sup>v2</sup>	GQA	VQA <sup>T</sup>	POPE	SQA	VizWiz	MME	MMB
BLIP-2 [21]	129M	224×224	Fix	1.0	41.0	41.0	42.5	85.3	61.0	19.6	1293.8	-
InstructBLIP [11]	130M	224×224	Fix	1.0	-	49.5	50.7	78.9	63.1	33.4	1212.8	-
Shikra [8]	6M	224×224	Fix	8.0	77.4	-	-	-	-	-	-	58.8
Qwen-VL [5]	1.4B	448×448	Fix	9.2	78.8	59.3	63.8	-	67.1	35.2	-	38.2
SPHINX [24]	1.0B	448×448	Fix	39.7	78.1	62.6	51.6	80.7	69.3	39.9	1476.1	66.9
SPHINX-2k [24]	1.0B	762×762	Fix	69.4	80.7	63.1	61.2	87.2	70.6	44.9	1470.7	65.9
MiniGPT-v2 [7]	326M	448×448	Fix	4.3	-	60.1	-	-	-	53.6	-	-
Fuyu-8B [6]	-	1024×1024	Any	21.3	74.2	-	-	74.1	-	-	728.6	10.7
OtterHD-8B [20]	-	1024×1024	Any	21.3	-	-	-	86.0	-	-	1223.4	58.3
mPLUG-Owl2 [43]	401M	448×448	Fix	1.7	79.4	56.1	58.2	86.2	68.7	54.5	1450.2	64.5
UReader [42]	86M	896×1120	Enum	26.0	-	-	57.6	-	-	-	-	-
Monkey [23]	1.0B	896×1344	Enum	65.3	80.3	60.7	-	67.6	69.4	61.2	-	-
LLaVA-1.5 [27]	1.2M	336×336	Fix	15.5	80.0	63.3	61.3	85.9	71.6	53.6	1531.3	67.7
LLaVA-UHD (ours)	1.2M	672×1008	Any	14.6	<b>81.7</b>	<b>65.2</b>	<b>67.7</b>	<b>89.1</b>	<b>72.0</b>	<b>56.1</b>	<b>1535.0</b>	<b>68.0</b>
Δ	-	×6 times	-	-0.9	+1.7	+1.9	+6.4	+3.2	+0.4	+2.5	+3.7	+0.3



Fish & Chips

Northowram Fisheries

OPENING TIMES

Lunch Evening

Closed Closed

11:30AM - 4:00PM 7:00PM

PM from Monday to Saturday. The restaurant is closed on Sundays.

What is the opening time of this restaurant ?

LLaVA-UHD (Ours)

The image shows a sign with the opening times for the restaurant. It appears to be open from **11:30AM to 1:30 PM and from 4:00 PM to 7:00 PM** from Monday to Saturday. The restaurant is **closed on Sundays**.

---

LLaVA-1.5

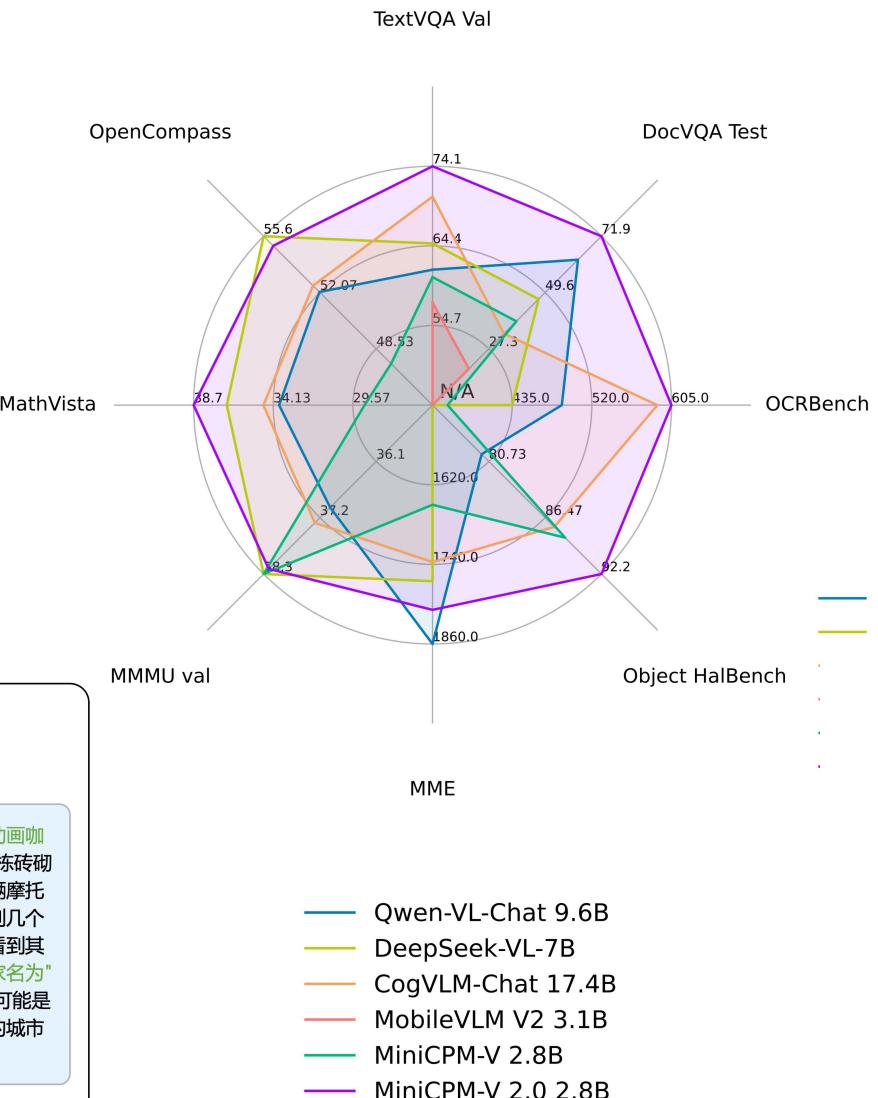
The image shows a sign with the opening times for the restaurant. It appears to be open from **12:00 to 2:00 PM and from 5:00 PM to 7:00 PM**.

# 高效端侧多模态大模型 MiniCPM-V 系列

2024/4/11

MiniCPM-V 2.0 (2.8B) 发布

- 通用能力：OpenCompass 11个常用基准综合榜单，优于 17-34B 主流模型
- OCR 能力：刷新 OCRBench **开源模型最佳成绩**；通用场景文字理解持平 Gemini Pro



# 高效端侧多模态大模型 MiniCPM-V 系列

2024/4/11

MiniCPM-V 2.0 (2.8B) 发布

- 通用能力：[OpenCompass](#) 11个常用基准综合榜单，优于 17-34B 主流模型
- OCR 能力：刷新 OCRBench [开源模型最佳成绩](#)；通用场景文字理解持平 [Gemini Pro](#)
- 可信行为：物体幻觉水平持平 [GPT-4V](#)
- 高清图理解：可支持 [180万 像素任意长宽比](#) 图像无损编码
- [中英双语，端侧部署](#)

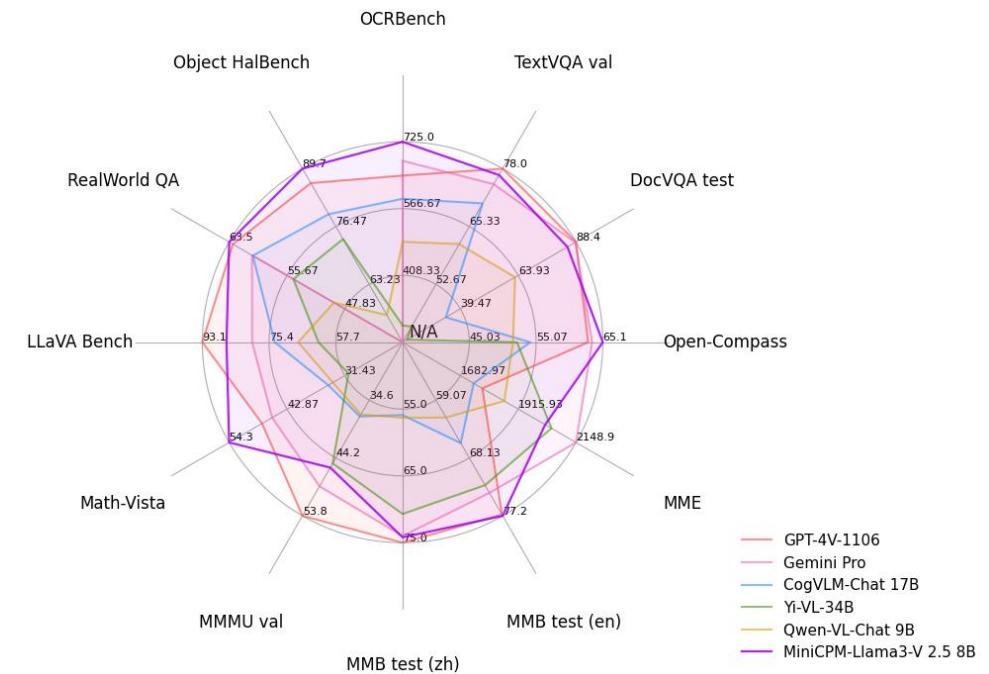


# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- **通用能力:** OpenCompass 榜单上超越闭源大模型 Gemini Pro、GPT-4V 等
- **OCR 能力:** OCRBench 超越现有开源与闭源大模型, **达到SOTA水平;** 难图长图长文本精准识别
- **可信行为:** 物体幻觉水平**超越 GPT-4V**
- **高清图理解:** 可支持 **180万 像素任意长宽比图像无损编码**
- **多语言:** 支持**30+多种语言**
- **高效部署:** 首次端侧系统级**多模态加速**, 语言解码速度 3 倍加速、图像编码 150 倍加速

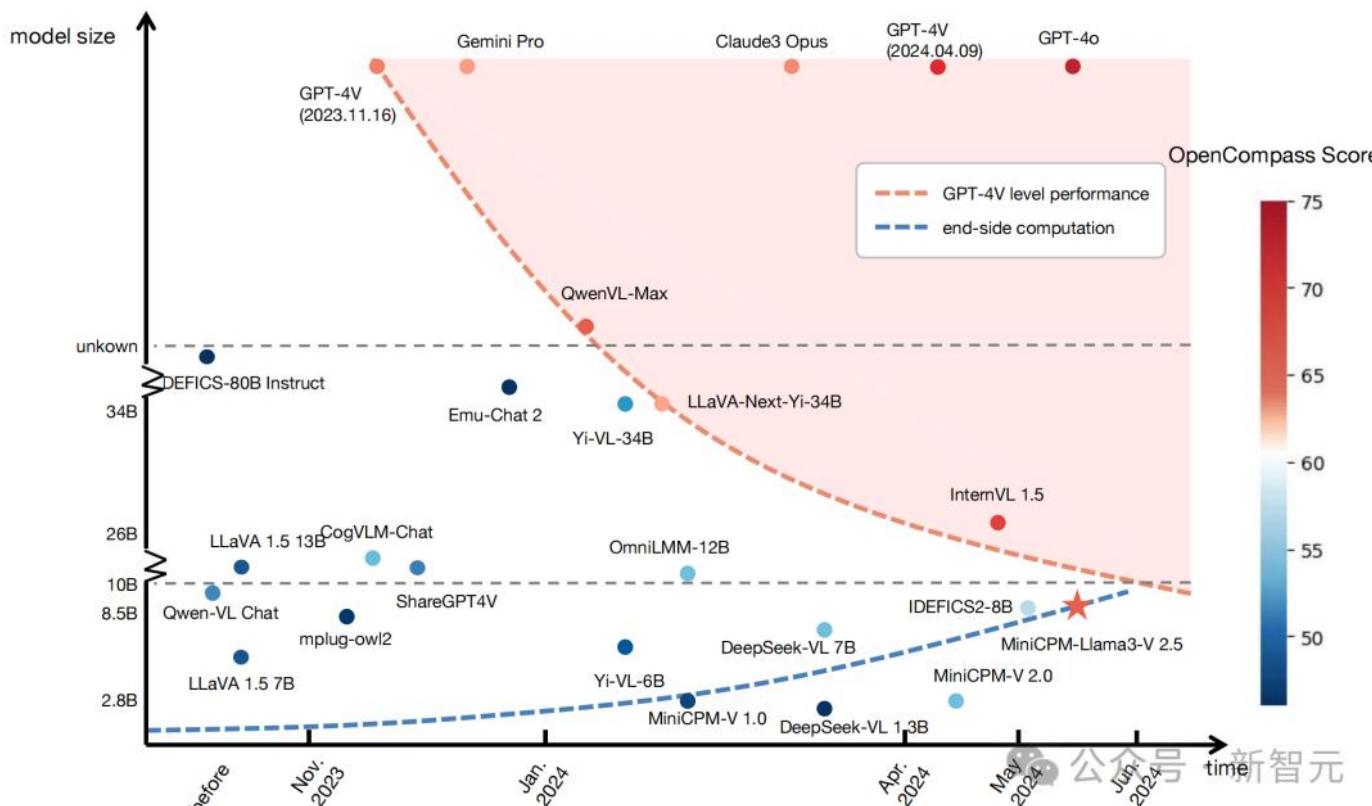


# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- 端侧算力愈益增强，**小参数、高性能的多模态大模型已经成为趋势。**
- MiniCPM-Llama3-V 2.5 采用最小参数撬动最强性能。

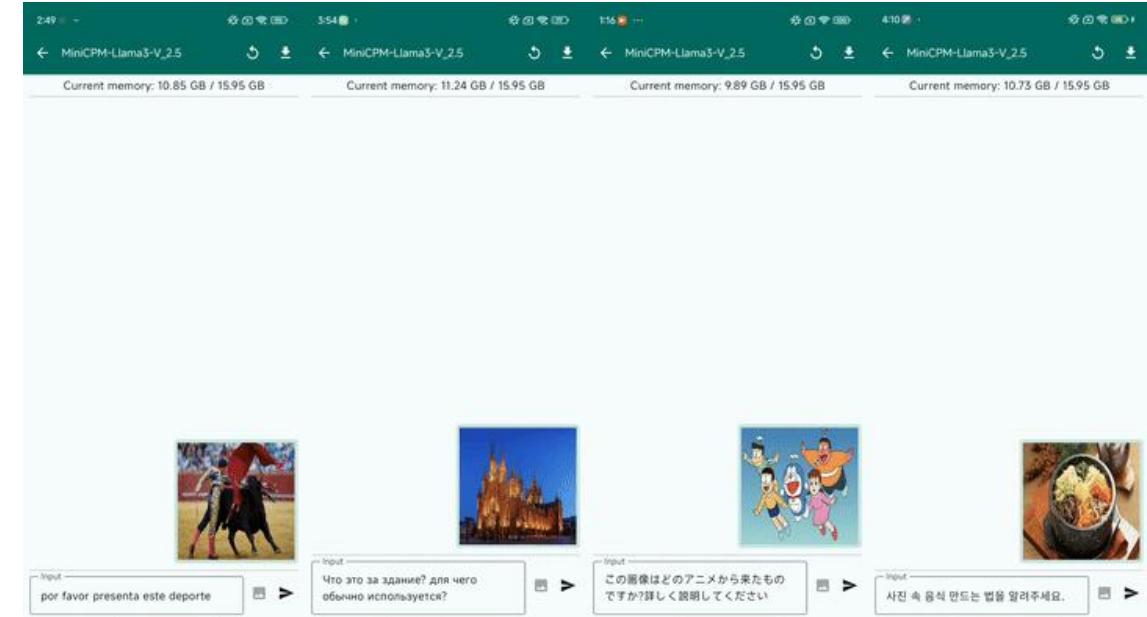
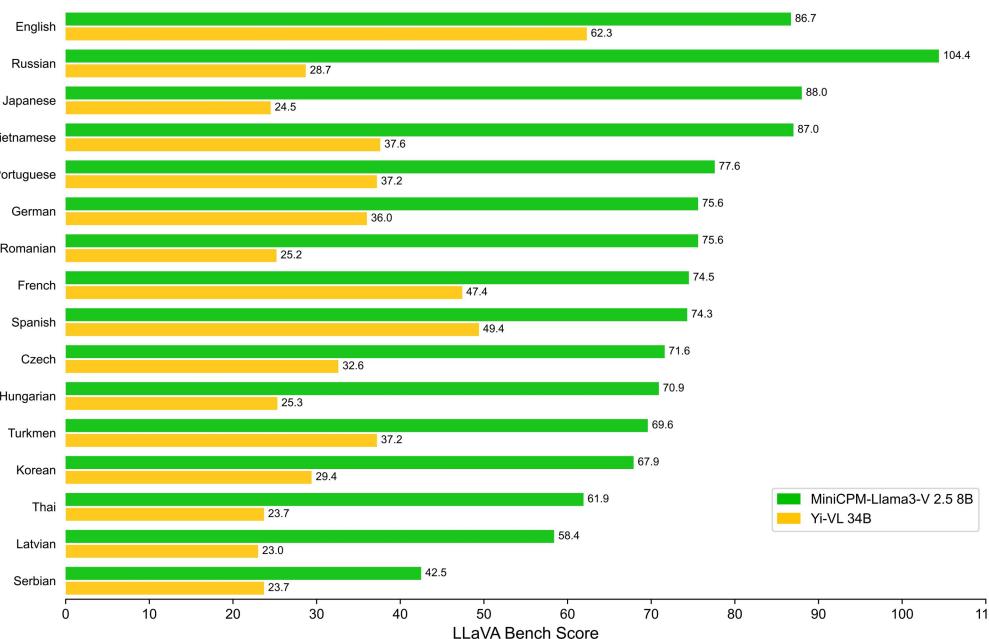


# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- **多语言性能:** 基于Llama 3 的多语言能力与跨语言泛化技术， MiniCPM-Llama3-V 2.5 通过少量翻译的多模态数据的指令微调即可支持30+多种语言



# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- 高清图像高效编码：基于自研的高清图像高效编码技术， MiniCPM-Llama3-V 2.5可以支持至多180万像素的任意长宽比图像识别，突破传统技术仅能识别20万像素小图的瓶颈。

**面壁Ultra对齐技术  
大模型上分神器！**

本次大模型「理科状元」Eurux-8x22B 的优异表现，得益于来自面壁 Ultra 对齐技术的大规模、高质量对齐数据集 UltraInteract 上新。好数据，才有好模型。此前，面壁 Ultra 对齐技术已为“领社”了全球超 200 个大模型。堪称大模型上分神器。

**GitHub 链接**  
<https://github.com/openBB/Eurus>

UltraInteract 是专门设计用于提升大模型推理能力的大规模、高质量的对齐数据集，包含了逻辑数学、代码和逻辑推理问题的12个开源数据集的86K条指令和220k个偏好对。共有五十万（条）左右数据。相比而言，Llama3-70B 模型则采用了千字量级的数据数据，这从侧面证明了 UltraInteract 数据集的优质性——数据质量胜过数据数量。

如此高质量的对齐数据是如何构建的呢？

严格质量控制和筛选，首先，我们从多个开源数据集中抽出难度较高、考察多样推理能力的 86K 复杂推理问题，并使用多个模型来采样答案。通过自动化格式检查和人工质量抽查结合的方式保证了答案格式的一致性和内容的正确性。

逐步推理，对于每条指令，模型都会按照思维链 (Ont) 格式进行逐步推理（如下图①），生成格式统一但模式多样的推理过程。

多轮交互，在模型给出推理过程之后，会自动与答案对比确定推理过程是否正确（如下图②），如果未正确，UltraInteract 会使用另一个评估模型（如下图③）指出错误并给出改进建议，生成新的逐步推理（如下图④）。再与策略

模型进行多轮交互（如下图⑤⑥），直到答案正确或达到轮数上限为止。这一步有助于模型学会反思和改错能力，在实际表现中可以更好地地进行多轮交互问答。

注意：UltraInteract尚无交互的过。

首创偏好树结构。为了深入探究如何学习在复杂推理中的作用，UltraInteract还为每个问题都构建了一棵偏好树（如下图所示），其中问题作为根节点，每一个回答作为一个子节点，每一轮生成两个节点（一对一锚相配对），所有正确推对应的节点都可以用于偏好学习。

**高效大模型工厂**

从领先的偏僻模型「小钢炮」MiniCPM，到开源模型推翻 SOTA 的 Eurux-8x22B，为什么我们总能推出同等参数、性能更优的「高效大模型」？

答案是，大模型是一项系统工程，而我们作为国内极少兼具大模型算法与 Infra 能力的团队，拥有自研的全流程高生产率线，这使得它能够快速地进行模型推理和优化。

面对 Ultra 对齐技术、Infra 工艺、独家「模型沙盒」实验和量化数据工厂，从数据、训练到调校工艺环环相扣，一条优秀的大型 Scaling Law 增长曲线由此而生。

Infra 工艺方面，我们构建了全流程优化加速工具套件平台 ModelForce，可以实现 10 倍推理加速，90% 成本降低。

**请详细总结图片中文章的要点**

**MiniCPM-Llama3-V 2.5**

这张图片是一篇关于面壁Ultra对齐技术的文章摘要。以下是文章的要点：

- 面壁Ultra介绍：**面壁Ultra是专门设计用于提升大模型推理能力的大规模、高质量对齐数据集的工具，包含了逻辑数学、代码推理问题的12个开源数据集的86k条指令和220k个偏好对。
- 面壁Ultra对齐数据集的特点：**面壁Ultra提供了一个包含多种类型对齐数据集的平台，这些数据集经过严格质量控制和筛选，确保了数据的可靠性和多样性。
- 推理过程中的挑战：**在推理过程中，面壁Ultra通过自动化格式检查和人工质量检查保证了推理的准确性和效率。
- 模型推理的优化：**面壁Ultra推理SOTA（State-Of-The-Art）的Eurux-8x22B模型，并展示了它在推理任务中的表现。
- 模型推理的复杂性和效率：**面壁Ultra使用了多轮交互的方式来提高模型推理的效率和准确性。这包括了在推理过程中不断调整参数以达到最佳性能。
- 面壁Ultra的优势：**文章提到面壁Ultra作为一个系统工程，拥有自研的全流程高生产效率线，这使得它能够快速地进行模型推理和优化。

综上所述，面壁Ultra是一个专注于推进自然语言处理领域对齐数据集的平台，旨在为模型推理提供更高效和准确的解决方案。

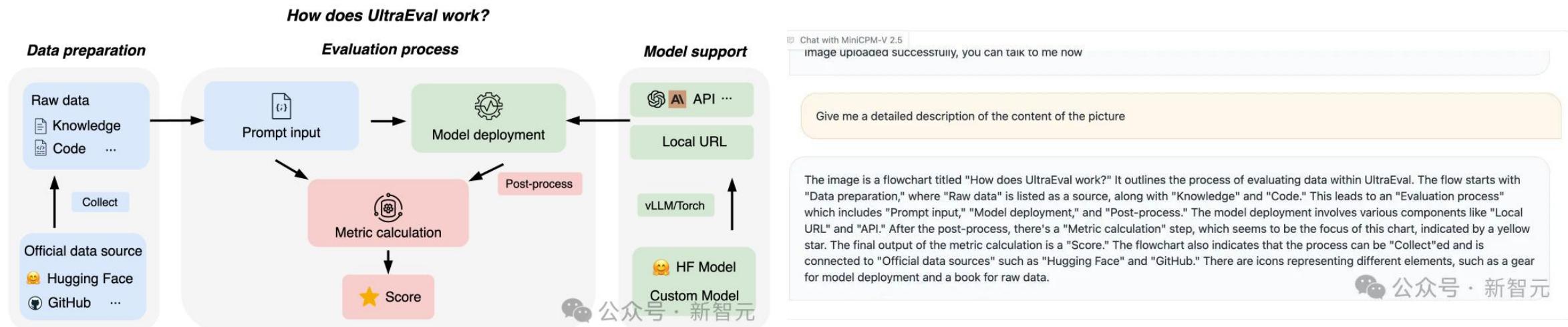
\* The colored underlines are for illustration purposes only and are not used as model input.

# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- 复杂推理能力优秀：MiniCPM-Llama3-V 2.5 具有优秀的复杂推理能力，可以理解结构化图片中的逻辑关系，如流程图中不同模块的文字、箭头之间的空间位置关系等。



# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- 超强全文OCR能力：MiniCPM-Llama3-V 2.5的结构化信息提取能力提升，对于富文字图片可以精准全面地识别出全文。

Unesco announces its newest geoparks around the world

9 April 2024  
By Lynn Brown, Features correspondent



Unesco Geoparks represent a balance of unique geological features, cultural touchpoints and a focus on sustainability (Credit: Chris Karagelis)

From dinosaur fossils in Brazil to the soaring monasteries of Meteora, these 18 new geological sites highlight a particular region's natural, cultural and intangible heritage.

Travellers interested in deep dives into geology, culture and sustainability have several new destinations to place on their bucket lists. Unesco just announced the designation of 18 new Global Geoparks, bringing the total to 213 sites representing 48 countries around the world.

From crystal-clear karst lakes to an archipelago formed during the last ice age, these parks are more than just sites with unique geological features; they also serve to highlight a particular region's natural, cultural and intangible heritage. "Unesco Global Geoparks represent landscapes where internationally significant geological heritage intertwines with human history and supports rich biodiversity," explains António Abreu, director of the Division of Ecological and Earth Sciences at Unesco. "Each new geopark tells a unique story, using geology as its lens."

Here are five of the newest geoparks on Unesco's list:

Please identify the text in the picture.

MiniCPM-Llama3-V 2.5

**Title:**  
Unesco announces its newest geoparks around the world

**Date and Author:**  
9 April 2024  
By Lynn Brown, Features correspondent

**Image Caption:**  
Unesco Geoparks represent a balance of unique geological features, cultural touchpoints and a focus on sustainability (Credit: Chris Karagelis)

**Body Text:**  
From dinosaur fossils in Brazil to the soaring meteors of Meteora, these 18 new geological sites highlight a particular region's natural, cultural and intangible heritage.  
Travellers interested in deep dives into geology, culture and sustainability have several new destinations to place on their bucket lists. Unesco just announced the designation of 18 new Global Geoparks, bringing the total to 213 sites representing 48 countries around the world.  
From crystal-clear karst lakes to an archipelago formed during the last ice age, these parks are more than just sites with unique geological features; they also serve to highlight a particular region's natural, cultural and intangible heritage. Unesco Global Geoparks represent landscapes where internationally significant geological heritage intertwines with human history and supports rich biodiversity, explains António Abreu, director of the Division of Ecological and Earth Sciences at Unesco. "Each new geopark tells a unique story, using geology as its lens."  
Here are five of the newest geoparks on Unesco's list:

# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- 社区认可度：发布 2 天以内，登上 Huggingface Trending 榜，达到全球第一，超越 LLaMA 3；目前 GitHub Star 数量达到 8000+，模型下载量达到 60 万。

## HF Trending 榜单

Models 669,453 Filter by name Full-text search ↑ Sort: Trending

\* openbmb/MiniCPM-Llama3-V-2\_5  
Visual Question Answering • Updated about 13 hours ago • ↓ 3.85k • ❤ 396

microsoft/Phi-3-vision-128k-instruct  
Text Generation • Updated 1 day ago • ↓ 2.58k • ❤ 327

meta-llama/Meta-Llama-3-8B  
Text Generation • Updated 10 days ago • ↓ 1.1M • ❤ 4.11k

microsoft/Phi-3-medium-128k-instruct  
Text Generation • Updated about 12 hours ago • ↓ 3.02k • ❤ 192

## GitHub Star

MiniCPM-V Public

MiniCPM-Llama3-V 2.5: A GPT-4V Level Multimodal LLM on Your Phone

Python ⭐ 8k 🏷 556

# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- **社区认可度:** 影响广泛，社区呼声高，端侧落地可帮助盲人解决实际困难；HuggingFace 官方提供机器，邀请开通 MiniCPM-V HF Space。

**社区反馈:** 影响广泛，社区呼声高，端侧落地可帮助盲人解决实际困难

New [comment](#) by Ibrahim Kettaneh (@ibrahimkettaneh) on [#1 - Look forward to the GGUF version of this model](#)

-

People with vision challenges such as myself with legal blindness would find a quantized version to be very helpful.

The strong OCR abilities of this model would be very helpful for making documents accessible.

This model is excellent and all your kind efforts and contributions to the community are greatly appreciated. :)

**HF Space:** HuggingFace 官方提供机器，  
邀请开通 MiniCPM-V HF Space

Hi,

I am from the Gradio team at Hugging Face, congrats on <https://github.com/OpenBMB/MiniCPM-V>, would be great to setup a gradio demo for it on Hugging Face, here is our guide to get started: <https://huggingface.co/docs/hub/spaces-sdks-gradio> and we can provide free a100 gpu for hosting: <https://huggingface.co/zero-gpu-explorers>. If you have any questions we can set up a collaboration channel on Hugging Face slack to collaborate/discuss.

Thanks,

AK

# 高效端侧多模态大模型 MiniCPM-V 系列

2024/5/20

MiniCPM-Llama3-V 2.5 (8B) 发布

- **社区影响力：** MiniCPM-Llama3-V 2.5 模型受到学界和工业界的广泛关注，引发外网呼吁重视来自中国的大模型工作。

**学术界关注：** 斯坦福人工智能实验室主任 Christopher Manning 评价 “good open-source work” 、 “helping advance science”



**工业界关注：** 谷歌 Deepmind 研究员，ViT 作者 Lucas Beyer 发声指出 MiniCPM-Llama3-V 2.5 等中国的优秀大模型应该受到更多关注。

