AiDD

2024 AI+研发数字峰会

AI+ Development Digital summit

AI驱动研发变革 促进企业**降本增效**

北京站 08/16-17

多模态文档大模型 mPLUG-DocOwl

徐海洋 阿里巴巴通义实验室

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站
K+全球软件研发行业创新峰会
时间：2024.06.21-22

K+峰会 敦煌站
K+思考周®研习社
时间：2024.10.17-19

K+峰会 香港站
K+思考周®研习社
时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2024.05.17-18

AiDD峰会 北京站
AI+研发数字峰会
时间：2024.08.16-17

AiDD峰会 深圳站
AI+研发数字峰会
时间：2024.11.08-09

AiDD峰会详情

AiDD

2024 AI+研发数字峰会

AI+ Development Digital summit

深圳站 11/08-09

AI驱动研发变革 促进企业降本增效

2024深圳站-议题设置

AI+产品线	LLM驱动产品创新	LLM驱动需求与业务分析	AI驱动设计与用户体验
AI+开发线	AI 原生应用开发框架与技术	AI Agents在研发落地实践	LLM驱动编程与单测
AI+测试线	LLM驱动测试分析与设计	基于LLM生成测试脚本与数据	LLM和AI应用的评测
AI+工程线	AI+DevOps 与工具 （LLM 时代的平台工程）	大模型对齐与安全	端侧大模型与云端协同
AI+领域线	领域大模型 SFT 与优化	知识增强与数据智能	大厂专场

扫描右侧二维码
查看更多会议详情

早鸟票限时抢购中 (截止到9月30日)

¥3680	¥2800
早鸟票	学生票

演讲嘉宾

徐海洋

阿里巴巴通义实验室-高级算法专家

负责通义多模态大模型mPLUG系列工作，包括基础多模态模型mPLUG/mPLUG-2，多模态对话大模型mPLUG-Owl，多模态文档大模型mPLUG-DocOwl，多模态智能体Mobile-Agent等，其中 mPLUG 工作在 VQA 榜单首超人类的成绩。在国际顶级期刊和会议ICML/CVPR/ICCV/ACL/EMNLP/MM/TOIS/IJCAI/AAAI等发表论文40多篇，并担任多个顶级和会议AC/PC/Reviewer。主导参与开源项目mPLUG，X-PLUG，AliceMind，DELTA。

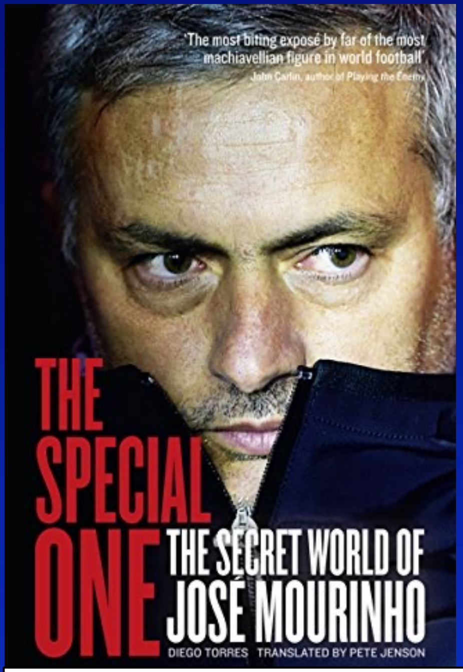
目录
CONTENTS

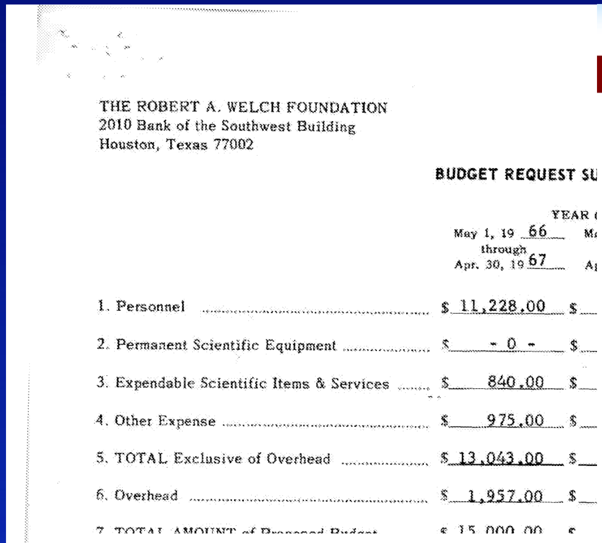
1. 多模态文档大模型背景
2. mPLUG-DocOwl系列工作
3. mPLUG-DocOwl开源实战
4. 总结与展望

PART 01

多模态文档大模型背景

多模态文档大模型背景


'The most biting exposé by far of the most machiavellian figure in world football'
John Carlin, author of Playing the Enemy
THE SPECIAL ONE
THE SECRET WORLD OF JOSÉ MOURINHO
DIEGO TORRES TRANSLATED BY PETE JENSON

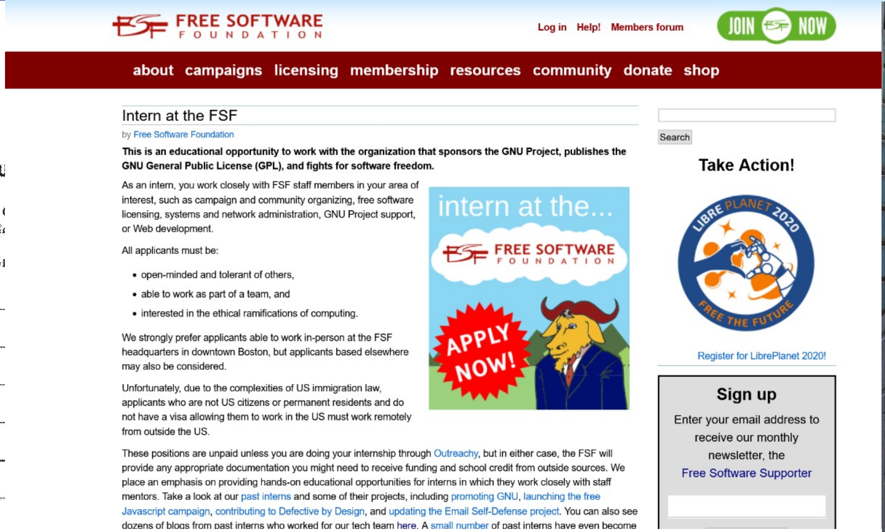


THE ROBERT A. WELCH FOUNDATION
2010 Bank of the Southwest Building
Houston, Texas 77002

BUDGET REQUEST SU

YEAR (
May 1, 19 66
through
Apr. 30, 19 67

1. Personnel	$ 11,228.00	$
2. Permanent Scientific Equipment	$ - 0 -	$
3. Expendable Scientific Items & Services	$ 840.00	$
4. Other Expense	$ 975.00	$
5. TOTAL Exclusive of Overhead	$ 13,043.00	$
6. Overhead	$ 1,957.00	$
7. TOTAL AMOUNT of Proposed Budget	$ 15,000.00	$



Intern at the FSF

by Free Software Foundation

This is an educational opportunity to work with the organization that sponsors the GNU Project, publishes the GNU General Public License (GPL), and fights for software freedom.

As an intern, you work closely with FSF staff members in your area of interest, such as campaign and community organizing, free software licensing, systems and network administration, GNU Project support, or Web development.

All applicants must be:

- open-minded and tolerant of others,
- able to work as part of a team, and
- interested in the ethical ramifications of computing.

We strongly prefer applicants able to work in-person at the FSF headquarters in downtown Boston, but applicants based elsewhere may also be considered.

Unfortunately, due to the complexities of US immigration law, applicants who are not US citizens or permanent residents and do not have a visa allowing them to work in the US must work remotely from outside the US.

These positions are unpaid unless you are doing your internship through Outreachy, but in either case, the FSF will provide any appropriate documentation you might need to receive funding and school credit from outside sources. We place an emphasis on providing hands-on educational opportunities for interns in which they work closely with staff mentors. Take a look at our past interns and some of their projects, including promoting GNU, launching the free Javascript campaign, contributing to Defective by Design, and updating the Email Self-Defense project. You can also see dozens of blogs from past interns who worked for our tech team here. A small number of past interns have even become

Take Action!

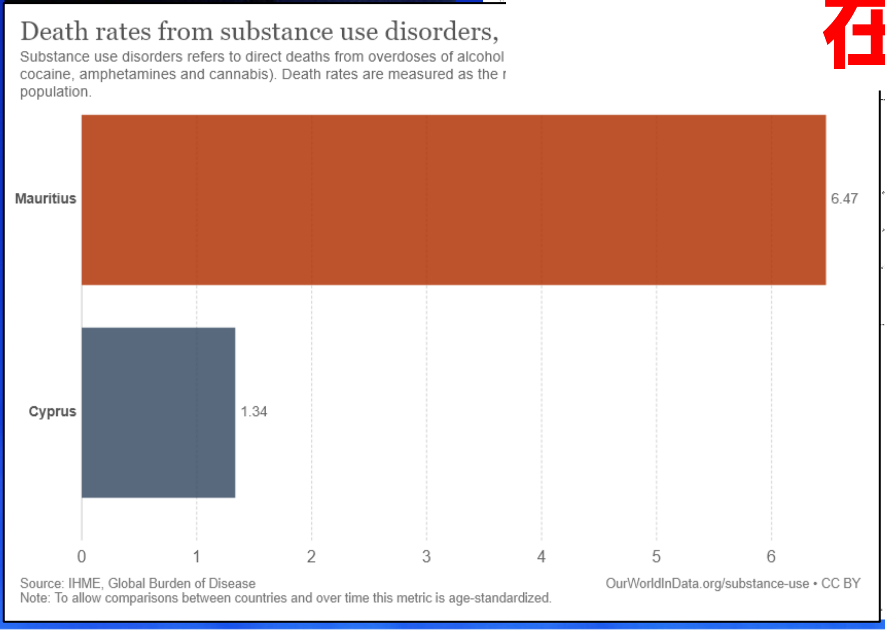
Register for LibrePlanet 2020!

Sign up

Enter your email address to receive our monthly newsletter, the Free Software Supporter

Mr. Pick
GRILL PLEZIER
BIJ UW POELIER
GEGRILDE KIPPEBOUT € 1,40 P/STUK
5 STUKS VOOR € 5,50
HELE GEGRILDE KIP € 5,10 P/STUK

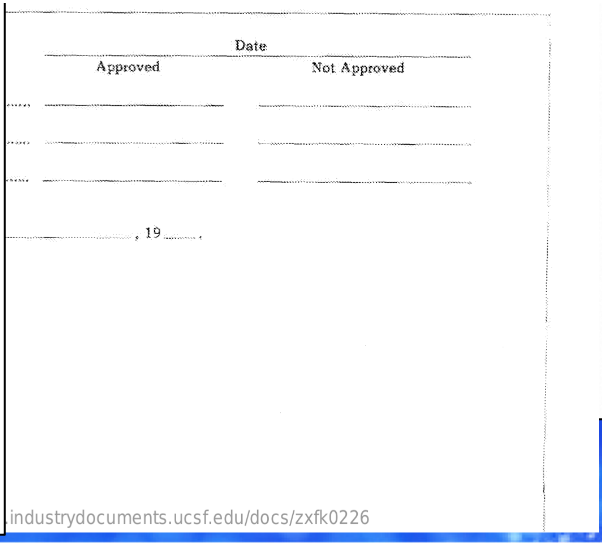
富含文字的图片（简称“文档图片”）
在生活中无处不在！



Death rates from substance use disorders,

Substance use disorders refers to direct deaths from overdoses of alcohol cocaine, amphetamines and cannabis). Death rates are measured as the r population.

Source: IHME, Global Burden of Disease
Note: To allow comparisons between countries and over time this metric is age-standardized.
OurWorldInData.org/substance-use • CC BY

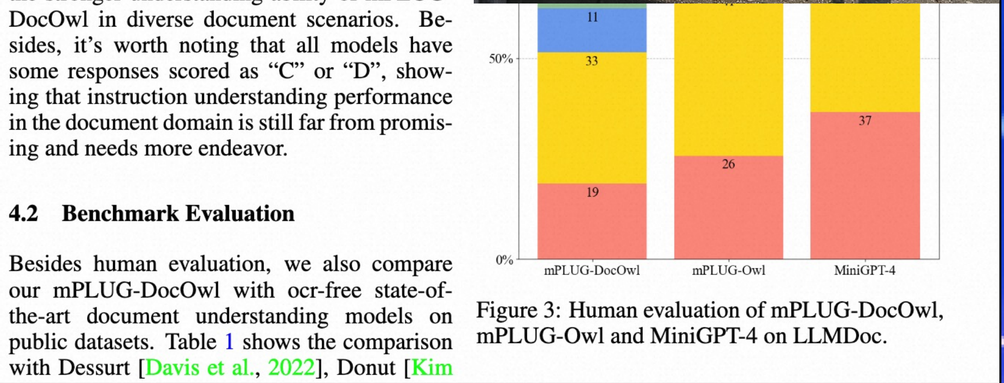


Date

Approved	Not Approved

, 19

industrydocuments.ucsf.edu/docs/zxfk0226



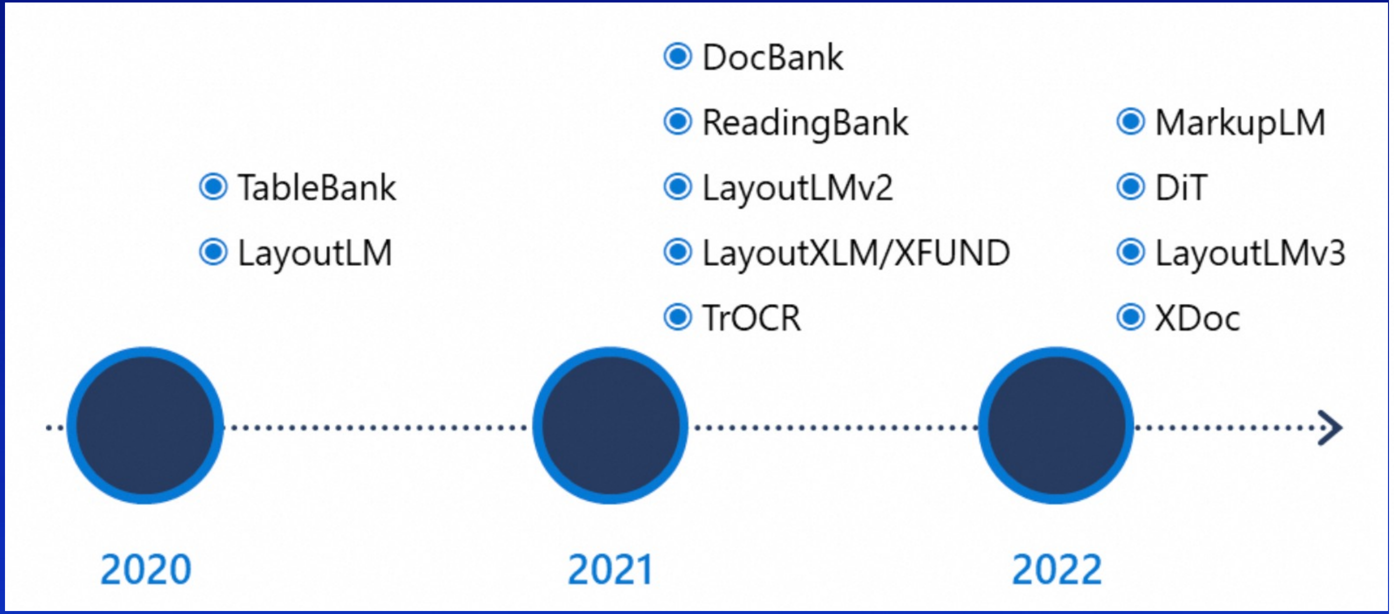
sponses being scored as "A", demonstrating the stronger understanding ability of mPLUG-DocOwl in diverse document scenarios. Besides, it's worth noting that all models have some responses scored as "C" or "D", showing that instruction understanding performance in the document domain is still far from promising and needs more endeavor.

4.2 Benchmark Evaluation

Besides human evaluation, we also compare our mPLUG-DocOwl with ocr-free state-of-the-art document understanding models on public datasets. Table 1 shows the comparison with Dessurt [Davis et al., 2022], Donut [Kim

Figure 3: Human evaluation of mPLUG-DocOwl, mPLUG-Owl and MiniGPT-4 on LLMDoc.

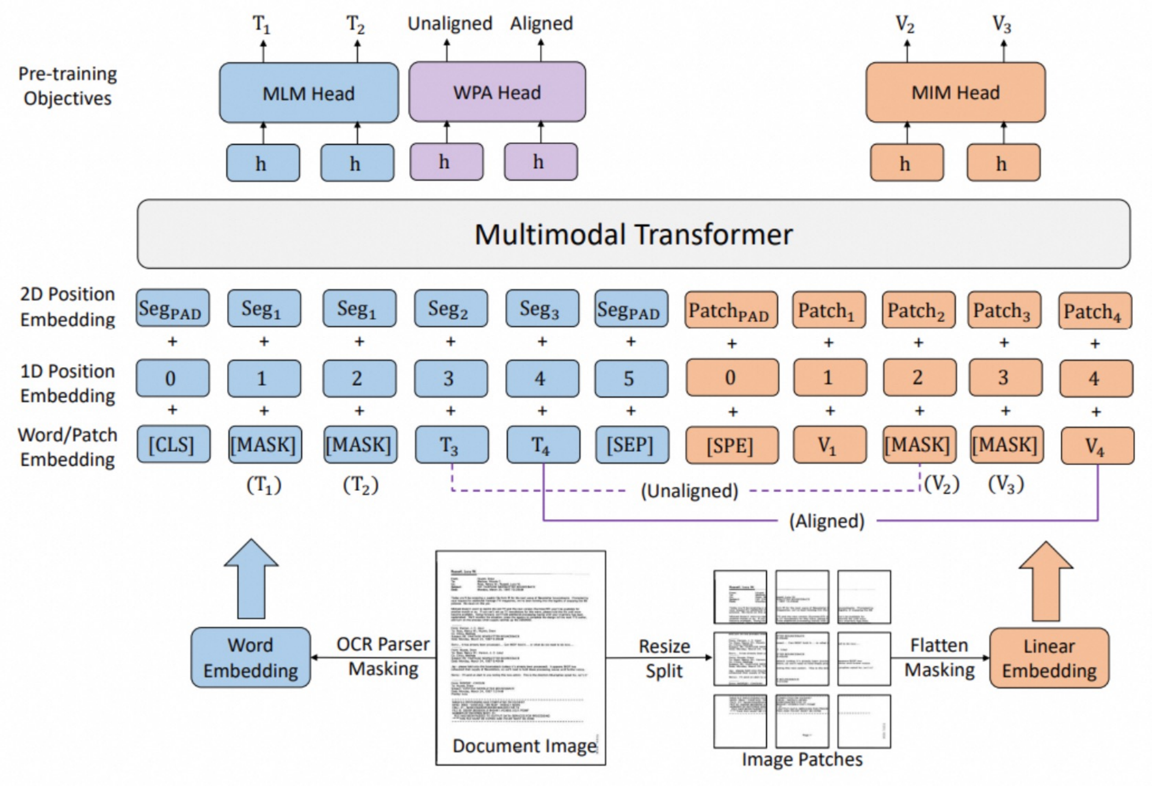
多模态文档大模型背景



2020
- TableBank
- LayoutLM

2021
- DocBank
- ReadingBank
- LayoutLMv2
- LayoutXLM/XFUND
- TrOCR

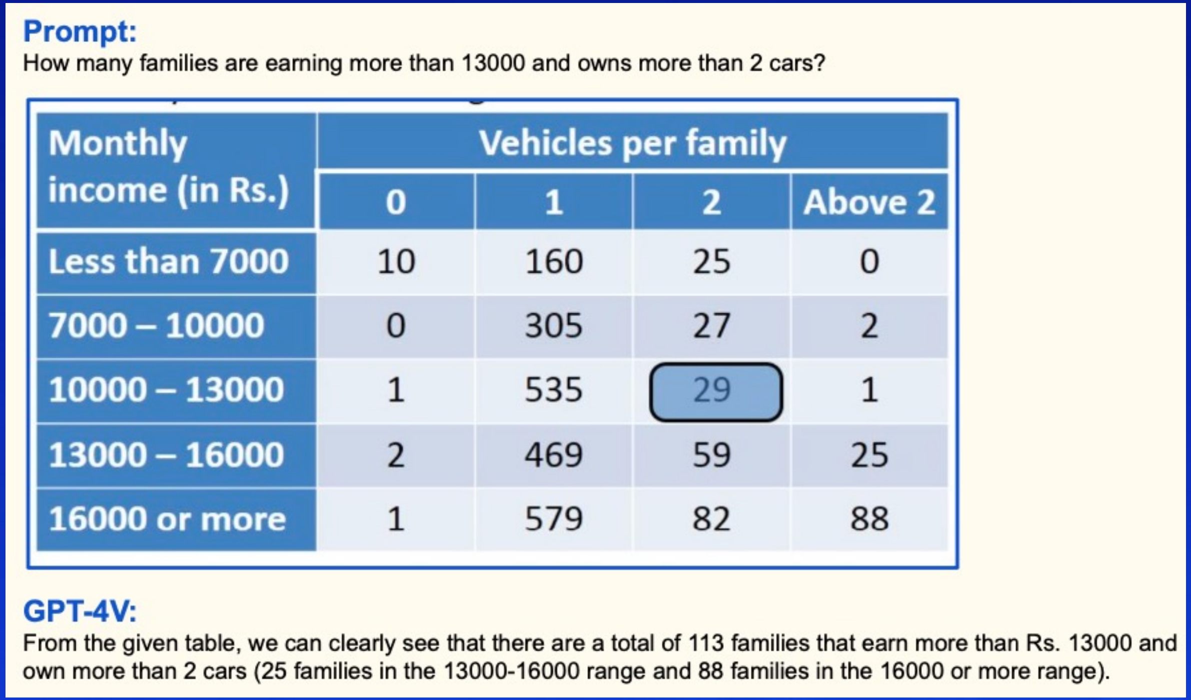
2022
- MarkupLM
- DiT
- LayoutLMv3
- XDoc



Revolutionizing Document AI with Multimodal Document Foundation Models

多模态文档大模型背景

一个通用的多模态大模型应该具备文档图片理解的能力，例如 GPT4V



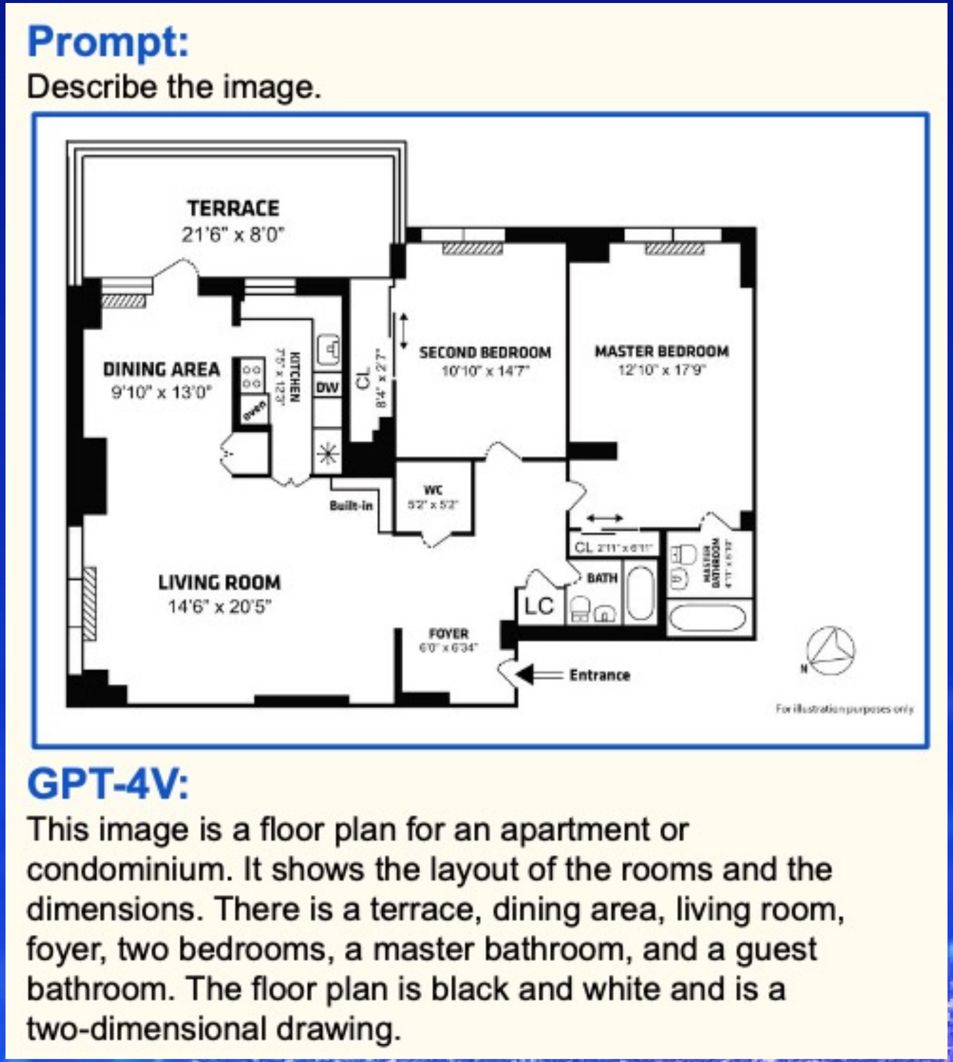
Prompt:

How many families are earning more than 13000 and owns more than 2 cars?

Monthly income (in Rs.)	Vehicles per family			
	0	1	2	Above 2
Less than 7000	10	160	25	0
7000 – 10000	0	305	27	2
10000 – 13000	1	535	29	1
13000 – 16000	2	469	59	25
16000 or more	1	579	82	88

GPT-4V:

From the given table, we can clearly see that there are a total of 113 families that earn more than Rs. 13000 and own more than 2 cars (25 families in the 13000-16000 range and 88 families in the 16000 or more range).



Prompt:

Describe the image.

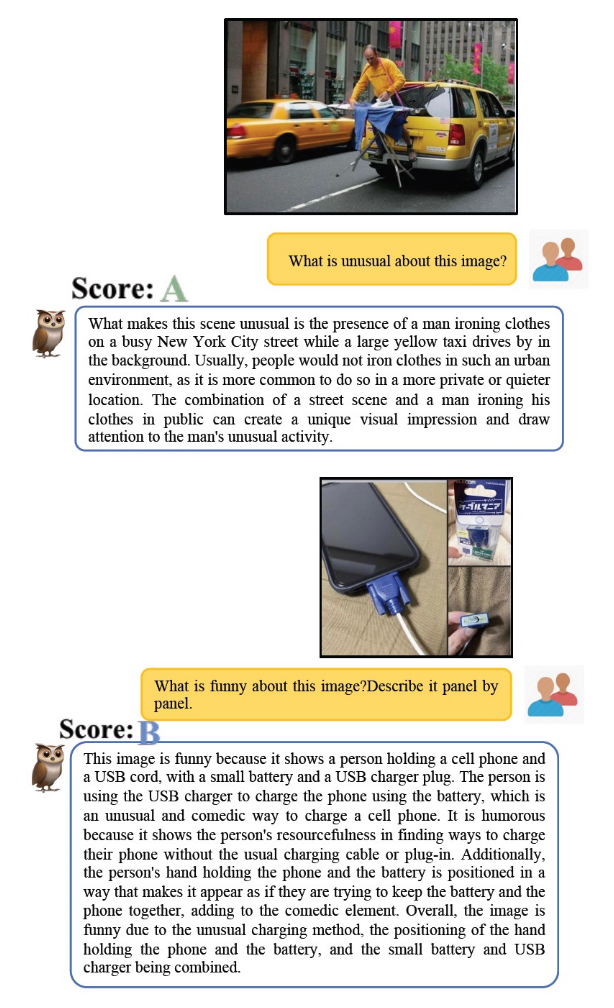
GPT-4V:

This image is a floor plan for an apartment or condominium. It shows the layout of the rooms and the dimensions. There is a terrace, dining area, living room, foyer, two bedrooms, a master bathroom, and a guest bathroom. The floor plan is black and white and is a two-dimensional drawing.

多模态文档大模型背景

多模态大模型模型（MLLM）主要关注于一般图片的对话和问答

mPLUG-Owl



What is unusual about this image?

Score: A

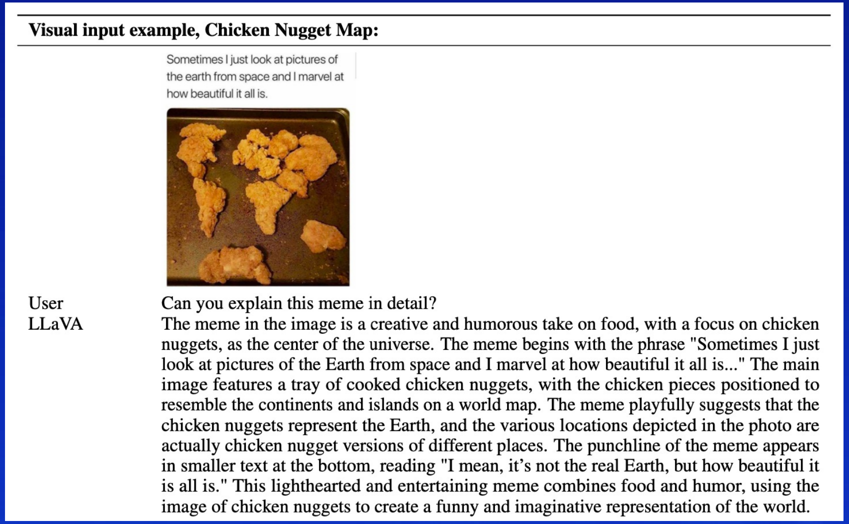
What makes this scene unusual is the presence of a man ironing clothes on a busy New York City street while a large yellow taxi drives by in the background. Usually, people would not iron clothes in such an urban environment, as it is more common to do so in a more private or quieter location. The combination of a street scene and a man ironing his clothes in public can create a unique visual impression and draw attention to the man's unusual activity.

What is funny about this image?Describe it panel by panel.

Score: B

This image is funny because it shows a person holding a cell phone and a USB cord, with a small battery and a USB charger plug. The person is using the USB charger to charge the phone using the battery, which is an unusual and comedic way to charge a cell phone. It is humorous because it shows the person's resourcefulness in finding ways to charge their phone without the usual charging cable or plug-in. Additionally, the person's hand holding the phone and the battery is positioned in a way that makes it appear as if they are trying to keep the battery and the phone together, adding to the comedic element. Overall, the image is funny due to the unusual charging method, the positioning of the hand holding the phone and the battery, and the small battery and USB charger being combined.

LLaVA

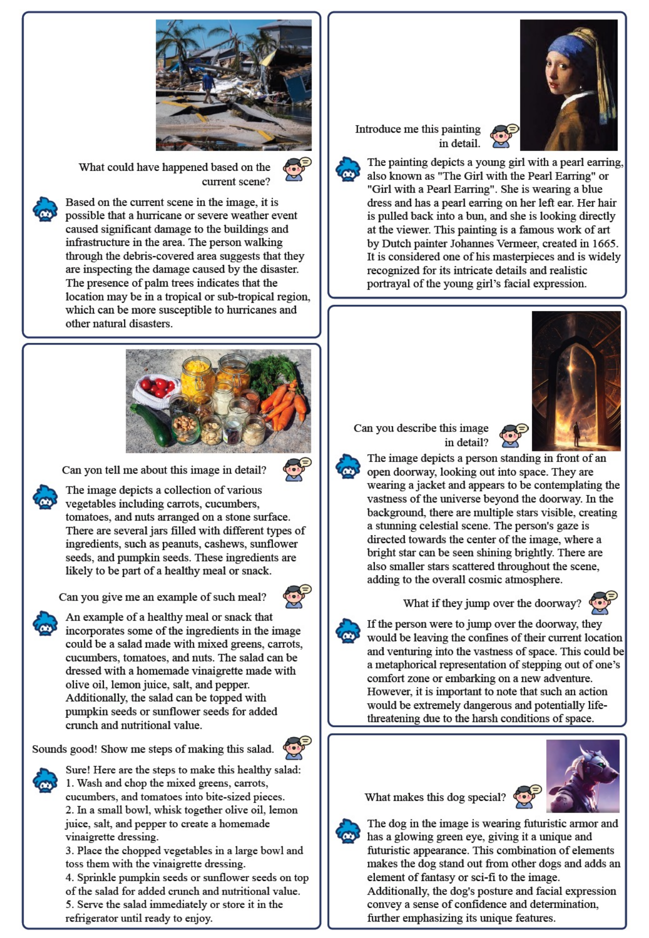


Visual input example, Chicken Nugget Map:

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.

User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.

InstructBLIP



What could have happened based on the current scene?

Based on the current scene in the image, it is possible that a hurricane or severe weather event caused significant damage to the buildings and infrastructure in the area. The person walking through the debris-covered area suggests that they are inspecting the damage caused by the disaster. The presence of palm trees indicates that the location may be in a tropical or sub-tropical region, which can be more susceptible to hurricanes and other natural disasters.

Can yon tell me about this image in detail?

The image depicts a collection of various vegetables including carrots, cucumbers, tomatoes, and nuts arranged on a stone surface. There are several jars filled with different types of ingredients, such as peanuts, cashews, sunflower seeds, and pumpkin seeds. These ingredients are likely to be part of a healthy meal or snack.

Can you give me an example of such meal?

An example of a healthy meal or snack that incorporates some of the ingredients in the image could be a salad made with mixed greens, carrots, cucumbers, tomatoes, and nuts. The salad can be dressed with a homemade vinaigrette made with olive oil, lemon juice, salt, and pepper. Additionally, the salad can be topped with pumpkin seeds or sunflower seeds for added crunch and nutritional value.

Sounds good! Show me steps of making this salad.

Sure! Here are the steps to make this healthy salad:
1. Wash and chop the mixed greens, carrots, cucumbers, and tomatoes into bite-sized pieces.
2. In a small bowl, whisk together olive oil, lemon juice, salt, and pepper to create a homemade vinaigrette dressing.
3. Place the chopped vegetables in a large bowl and toss them with the vinaigrette dressing.
4. Sprinkle pumpkin seeds or sunflower seeds on top of the salad for added crunch and nutritional value.
5. Serve the salad immediately or store it in the refrigerator until ready to enjoy.

Introduce me this painting in detail.

The painting depicts a young girl with a pearl earring, also known as "The Girl with the Pearl Earring" or "Girl with a Pearl Earring". She is wearing a blue dress and has a pearl earring on her left ear. Her hair is pulled back into a bun, and she is looking directly at the viewer. This painting is a famous work of art by Dutch painter Johannes Vermeer, created in 1665. It is considered one of his masterpieces and is widely recognized for its intricate details and realistic portrayal of the young girl's facial expression.

Can you describe this image in detail?

The image depicts a person standing in front of an open doorway, looking out into space. They are wearing a jacket and appears to be contemplating the vastness of the universe beyond the doorway. In the background, there are multiple stars visible, creating a stunning celestial scene. The person's gaze is directed towards the center of the image, where a bright star can be seen shining brightly. There are also smaller stars scattered throughout the scene, adding to the overall cosmic atmosphere.

What if they jump over the doorway?

If the person were to jump over the doorway, they would be leaving the confines of their current location and venturing into the vastness of space. This could be a metaphorical representation of stepping out of one's comfort zone or embarking on a new adventure. However, it is important to note that such an action would be extremely dangerous and potentially life-threatening due to the harsh conditions of space.

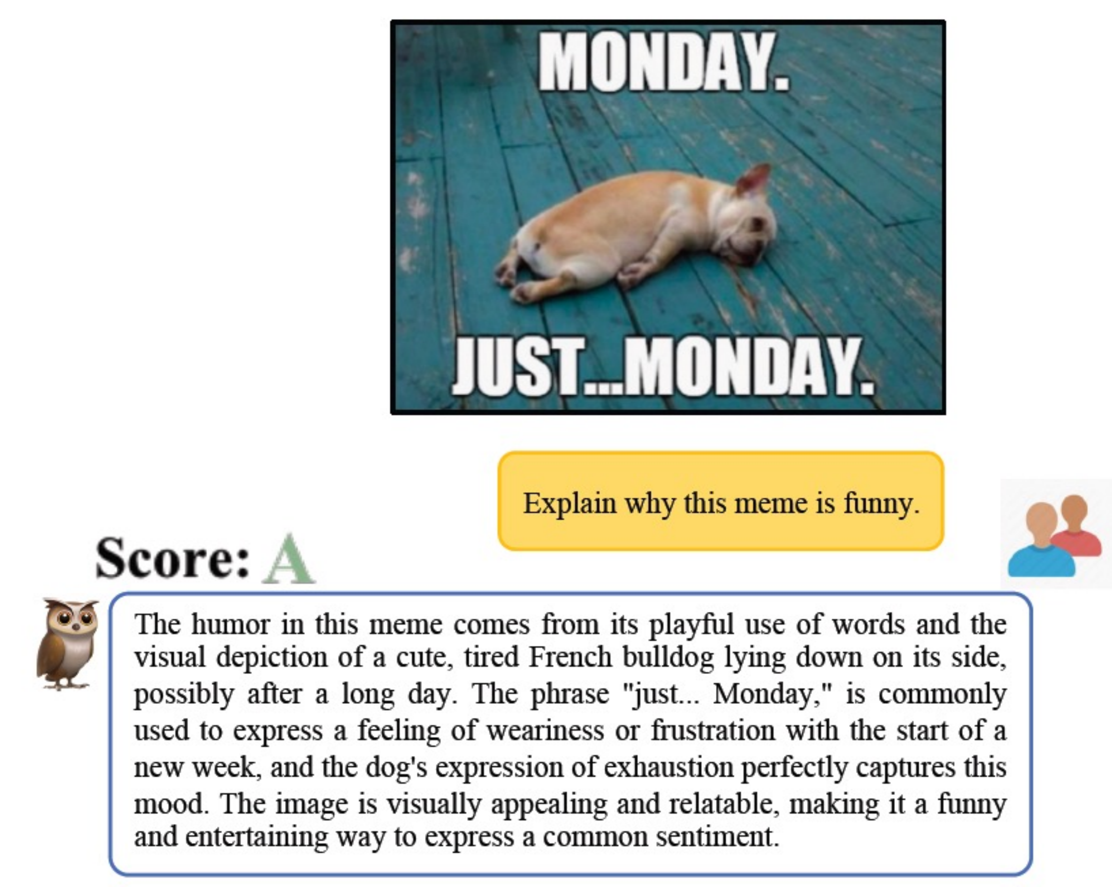
What makes this dog special?

The dog in the image is wearing futuristic armor and has a glowing green eye, giving it a unique and futuristic appearance. This combination of elements makes the dog stand out from other dogs and adds an element of fantasy or sci-fi to the image. Additionally, the dog's posture and facial expression convey a sense of confidence and determination, further emphasizing its unique features.

多模态文档大模型背景

由于预训练中存在富含文字的图片，他们也初显一定的图片文字理解能力

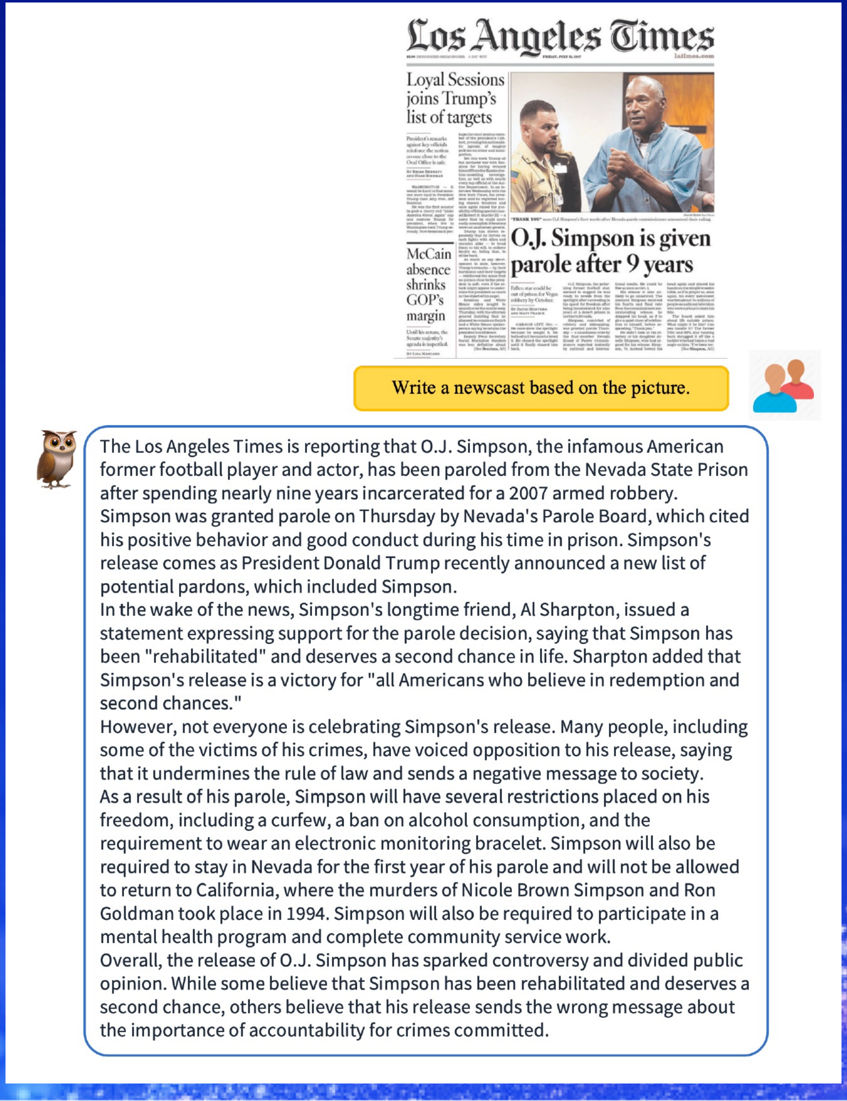
MLLM具备实现通用文档图片理解的潜力



Explain why this meme is funny.

Score: A

The humor in this meme comes from its playful use of words and the visual depiction of a cute, tired French bulldog lying down on its side, possibly after a long day. The phrase "just... Monday," is commonly used to express a feeling of weariness or frustration with the start of a new week, and the dog's expression of exhaustion perfectly captures this mood. The image is visually appealing and relatable, making it a funny and entertaining way to express a common sentiment.


Los Angeles Times
Loyal Sessions joins Trump's list of targets
McCain absence shrinks GOP's margin
O.J. Simpson is given parole after 9 years

Write a newscast based on the picture.

The Los Angeles Times is reporting that O.J. Simpson, the infamous American former football player and actor, has been paroled from the Nevada State Prison after spending nearly nine years incarcerated for a 2007 armed robbery. Simpson was granted parole on Thursday by Nevada's Parole Board, which cited his positive behavior and good conduct during his time in prison. Simpson's release comes as President Donald Trump recently announced a new list of potential pardons, which included Simpson.

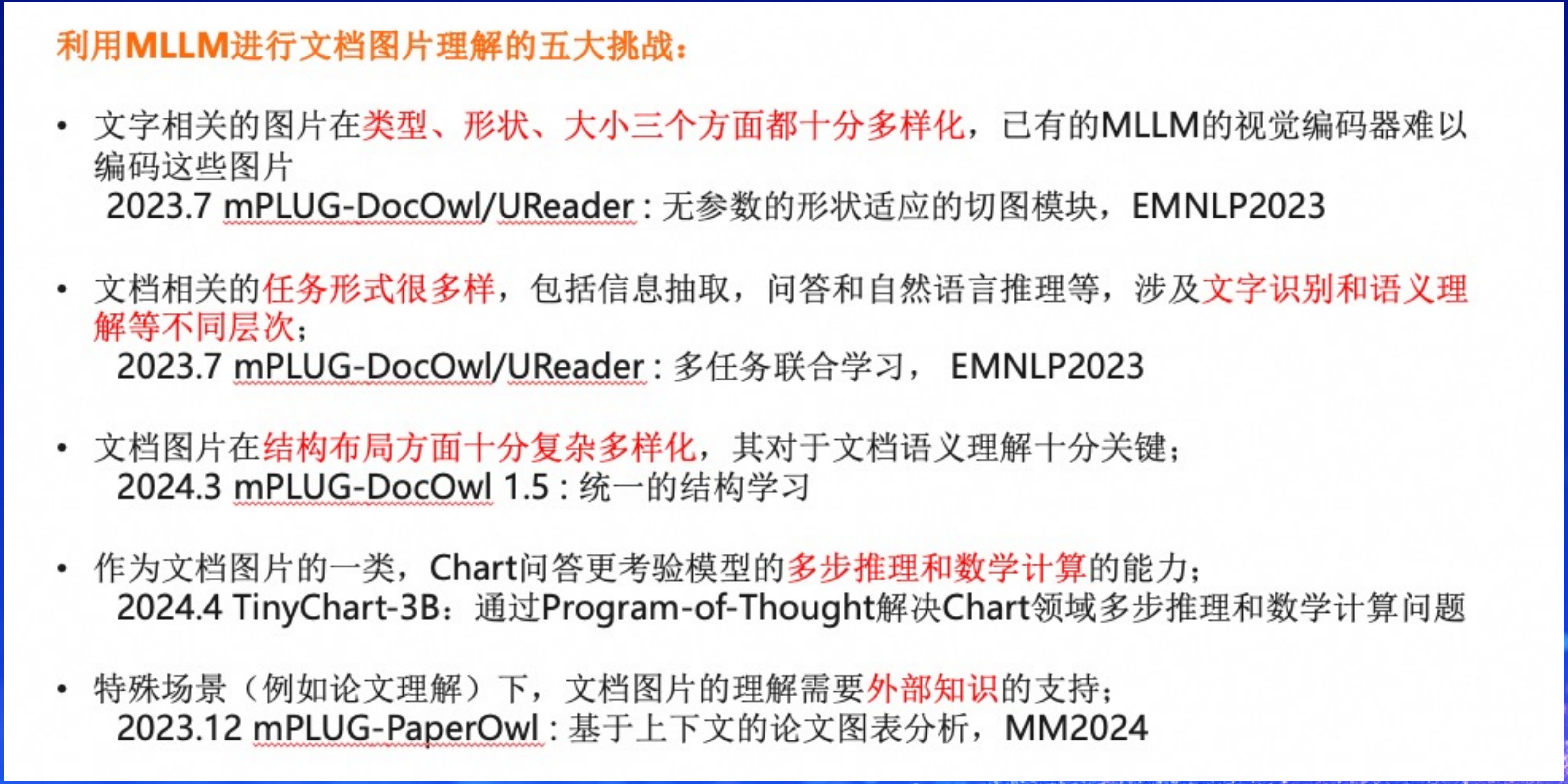
In the wake of the news, Simpson's longtime friend, Al Sharpton, issued a statement expressing support for the parole decision, saying that Simpson has been "rehabilitated" and deserves a second chance in life. Sharpton added that Simpson's release is a victory for "all Americans who believe in redemption and second chances."

However, not everyone is celebrating Simpson's release. Many people, including some of the victims of his crimes, have voiced opposition to his release, saying that it undermines the rule of law and sends a negative message to society.

As a result of his parole, Simpson will have several restrictions placed on his freedom, including a curfew, a ban on alcohol consumption, and the requirement to wear an electronic monitoring bracelet. Simpson will also be required to stay in Nevada for the first year of his parole and will not be allowed to return to California, where the murders of Nicole Brown Simpson and Ron Goldman took place in 1994. Simpson will also be required to participate in a mental health program and complete community service work.

Overall, the release of O.J. Simpson has sparked controversy and divided public opinion. While some believe that Simpson has been rehabilitated and deserves a second chance, others believe that his release sends the wrong message about the importance of accountability for crimes committed.

多模态文档大模型挑战



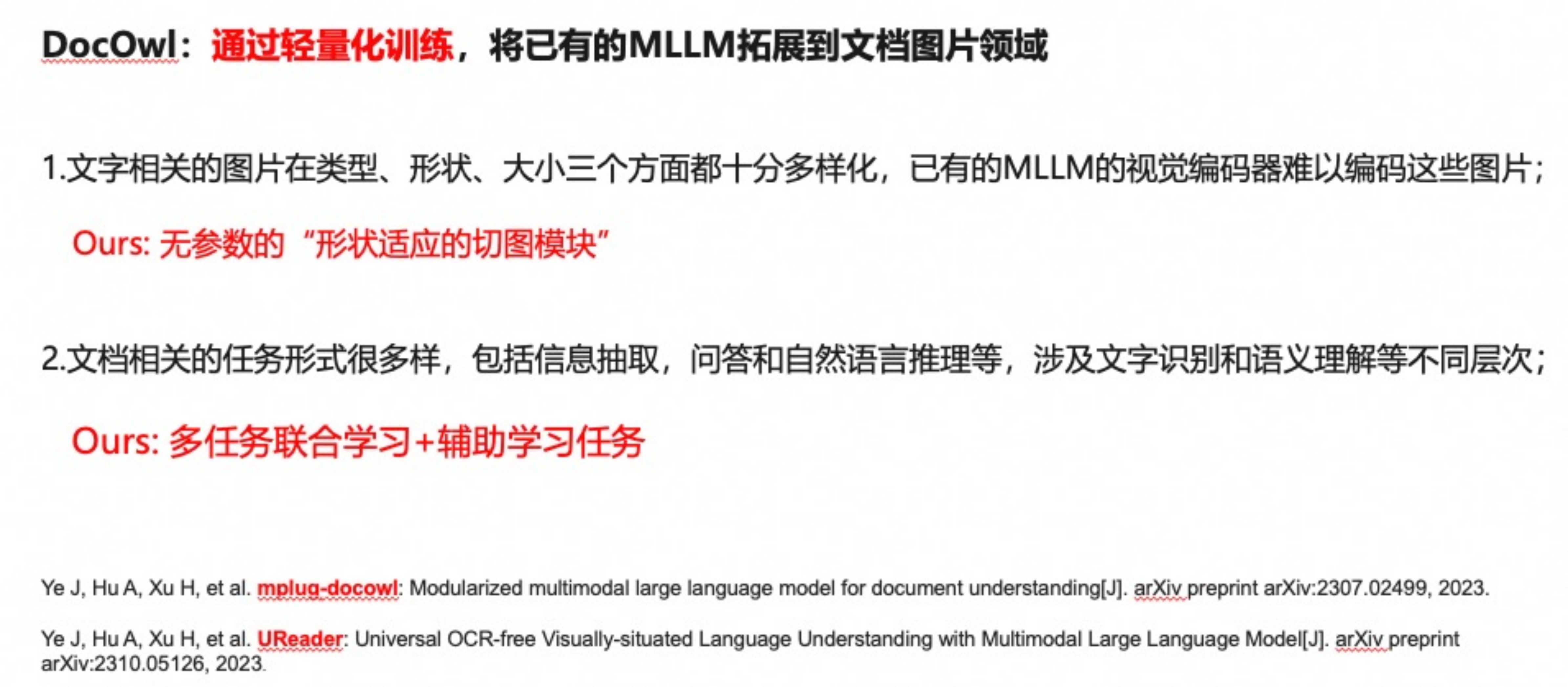
利用MLLM进行文档图片理解的五大挑战：

- 文字相关的图片在类型、形状、大小三个方面都十分多样化，已有的MLLM的视觉编码器难以编码这些图片
  2023.7 mPLUG-DocOwl/UReader : 无参数的形状适应的切图模块，EMNLP2023
- 文档相关的任务形式很多样，包括信息抽取，问答和自然语言推理等，涉及文字识别和语义理解等不同层次；
  2023.7 mPLUG-DocOwl/UReader : 多任务联合学习， EMNLP2023
- 文档图片在结构布局方面十分复杂多样化，其对于文档语义理解十分关键；
  2024.3 mPLUG-DocOwl 1.5 : 统一的结构学习
- 作为文档图片的一类，Chart问答更考验模型的多步推理和数学计算的能力；
  2024.4 TinyChart-3B：通过Program-of-Thought解决Chart领域多步推理和数学计算问题
- 特殊场景（例如论文理解）下，文档图片的理解需要外部知识的支持；
  2023.12 mPLUG-PaperOwl : 基于上下文的论文图表分析，MM2024

PART 01

多模态文档大模型背景

mPLUG-DocOwl



DocOwl：通过轻量化训练，将已有的MLLM拓展到文档图片领域

1.文字相关的图片在类型、形状、大小三个方面都十分多样化，已有的MLLM的视觉编码器难以编码这些图片；

Ours: 无参数的“形状适应的切图模块”

2.文档相关的任务形式很多样，包括信息抽取，问答和自然语言推理等，涉及文字识别和语义理解等不同层次；

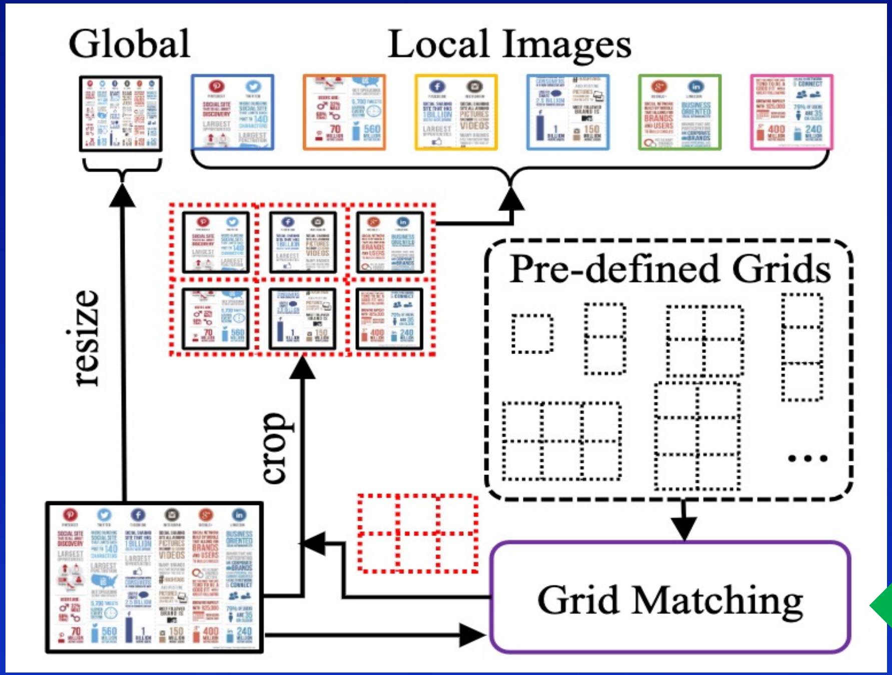
Ours: 多任务联合学习+辅助学习任务

Ye J, Hu A, Xu H, et al. mplug-docowl: Modularized multimodal large language model for document understanding[J]. arXiv preprint arXiv:2307.02499, 2023.

Ye J, Hu A, Xu H, et al. UReader: Universal OCR-free Visually-situated Language Understanding with Multimodal Large Language Model[J]. arXiv preprint arXiv:2310.05126, 2023.

mPLUG-DocOwl

- 模型结构



形状适应的切图模块

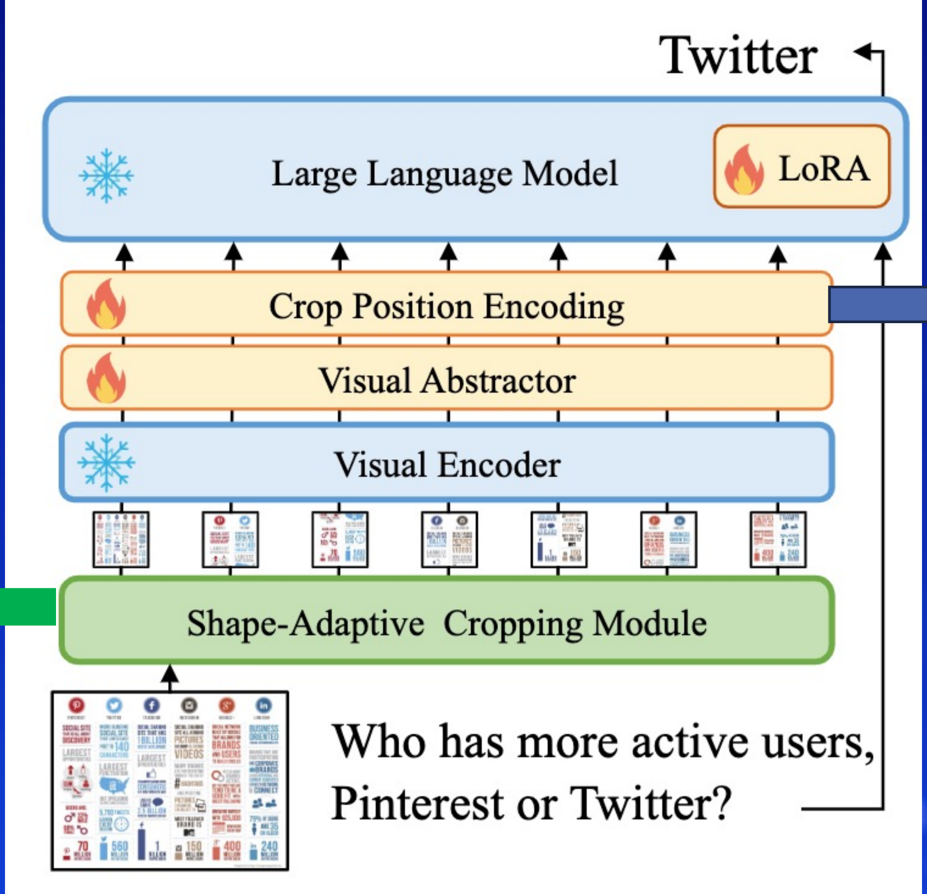
1.$Grids = \{1\times1, 1\times2, \ldots 1\times20,$

$2\times1, 2\times1, \ldots, 2\times10,$

$\ldots$

$20\times1\}$

2.同时考虑形状和大小进行grid选择）



每个子图在原图中第几行第几列

mPLUG-DocOwl

- 模型训练: 训练任务

下游任务

Visual Question Answering (问答)

Information Extraction (信息抽取)

Natural Language Inference (自然语言推理)

Image Captioning (图像描述生成)

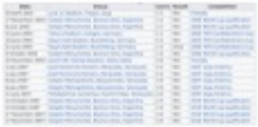
辅助任务

Text Reading (文字阅读)

Key Points Generation (要点生成)

Human: What will the presentation aim to encourage a discussion about
AI: It will aim to encourage a discussion about open data ownership rights and the differing social

Human: what is the value mentioned in the CID#:
AI: 736 00 00

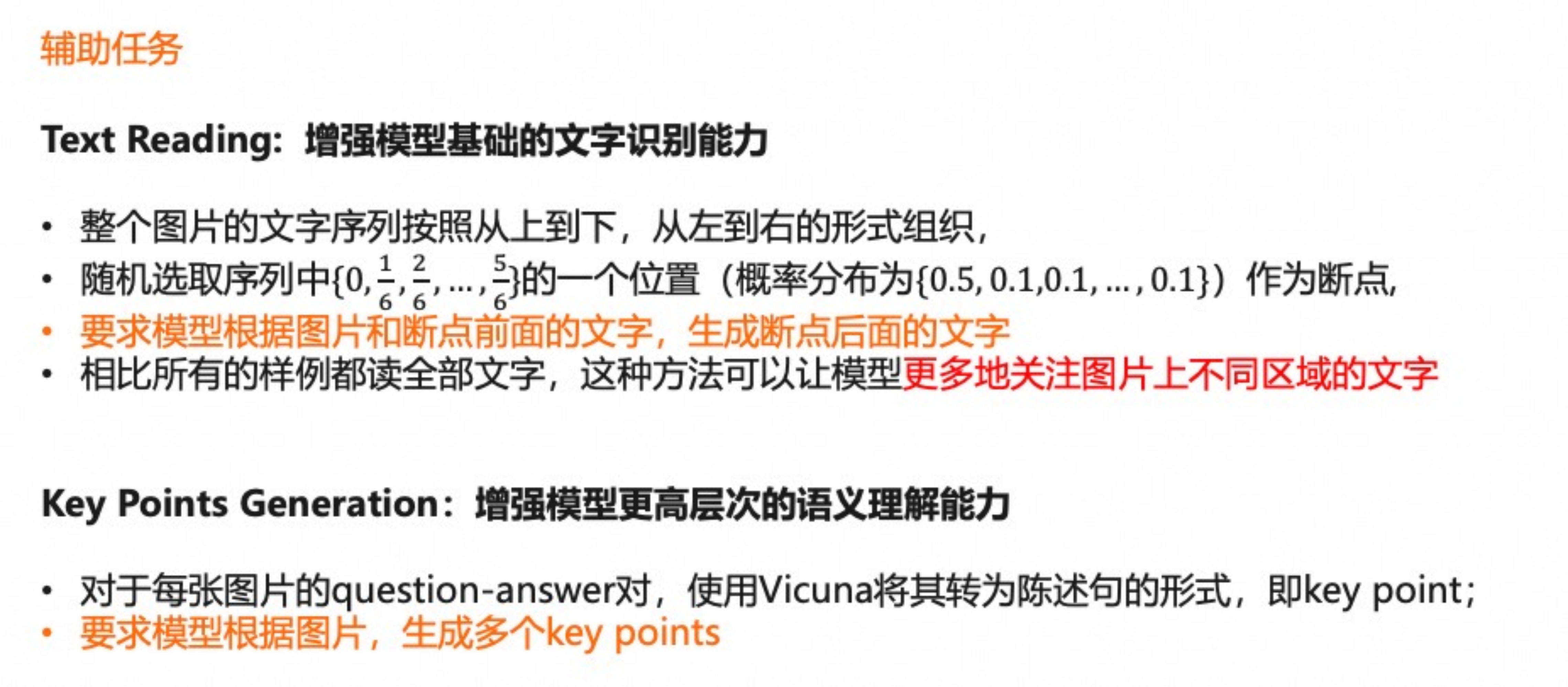


Human: juan román riquelme have the same score on july 11 , 2007 november 17 2007, Yes or No?
AI: Yes

Human : Create a compact narrative representing the image presented.
AI: Sign that says Calea Dorobantilor in blue and white

mPLUG-DocOwl

- 模型训练: 训练任务



辅助任务

Text Reading：增强模型基础的文字识别能力

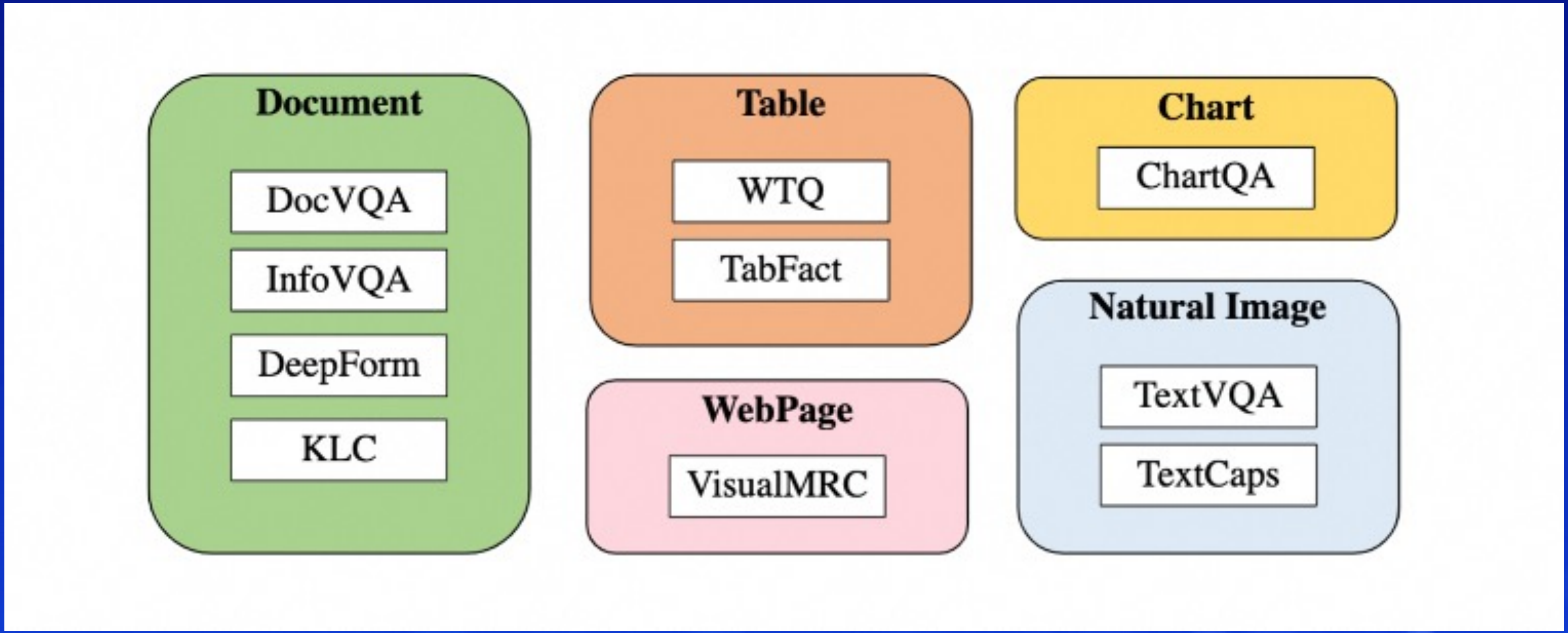
- 整个图片的文字序列按照从上到下，从左到右的形式组织，
- 随机选取序列中$\{0,\frac{1}{6},\frac{2}{6},\ldots,\frac{5}{6}\}$的一个位置（概率分布为$\{0.5,0.1,0.1,\ldots,0.1\}$）作为断点，
- 要求模型根据图片和断点前面的文字，生成断点后面的文字
- 相比所有的样例都读全部文字，这种方法可以让模型更多地关注图片上不同区域的文字

Key Points Generation：增强模型更高层次的语义理解能力

- 对于每张图片的question-answer对，使用Vicuna将其转为陈述句的形式，即key point；
- 要求模型根据图片，生成多个key points

mPLUG-DocOwl

- 模型训练: 训练数据



Document

- DocVQA
- InfoVQA
- DeepForm
- KLC

Table

- WTQ
- TabFact

Chart

- ChartQA

WebPage

- VisualMRC

Natural Image

- TextVQA
- TextCaps

mPLUG-DocOwl

- 实验结果

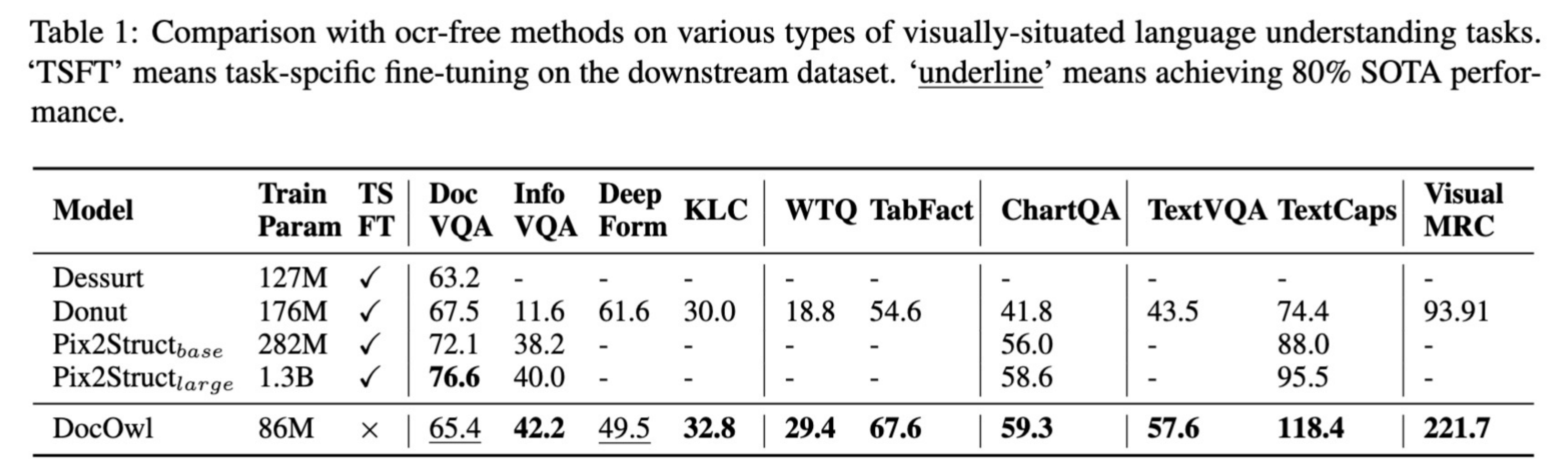
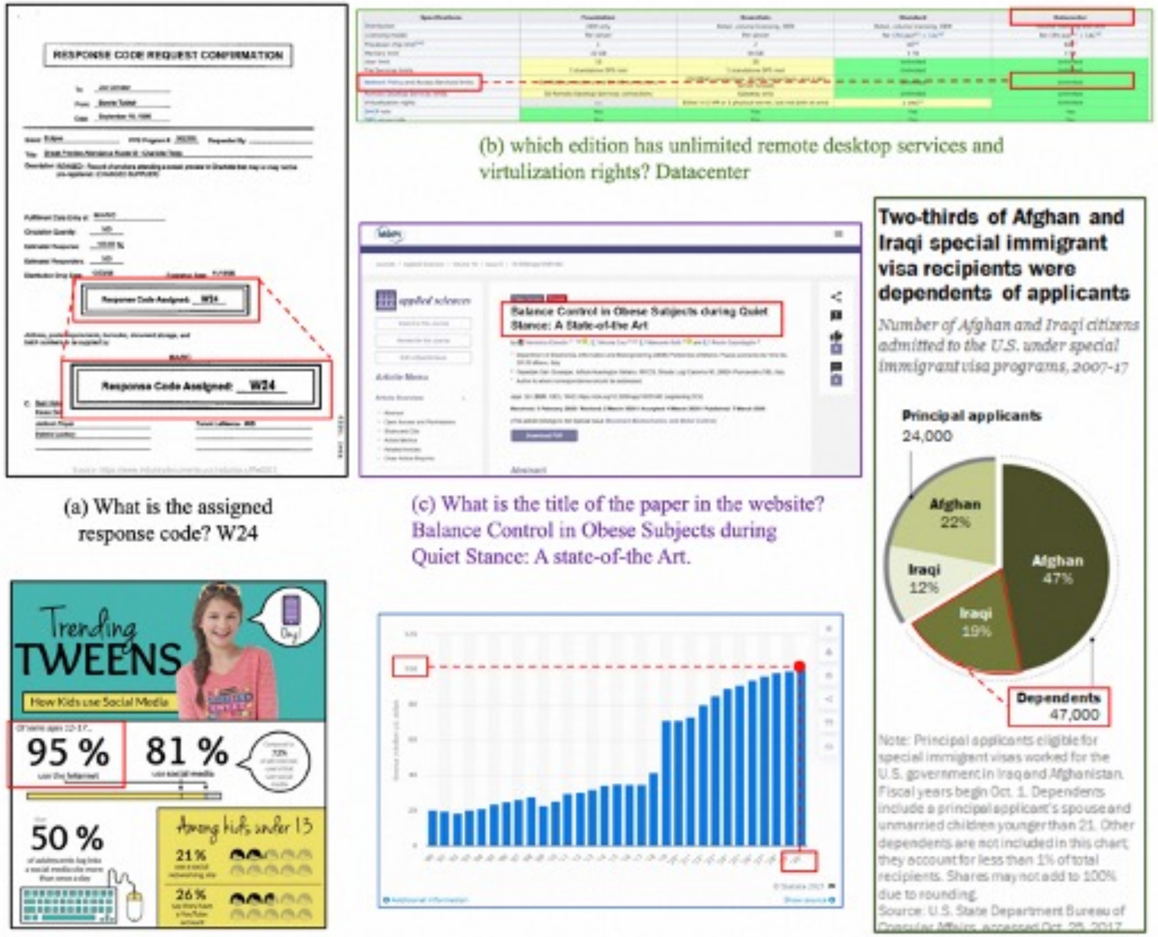


Table 1: Comparison with ocr-free methods on various types of visually-situated language understanding tasks. 'TSFT' means task-spcific fine-tuning on the downstream dataset. '<u>underline</u>' means achieving 80% SOTA performance.

Model	**Train Param**	**TS FT**	**Doc VQA**	**Info VQA**	**Deep Form**	**KLC**	**WTQ**	**TabFact**	**ChartQA**	**TextVQA**	**TextCaps**	**Visual MRC**
Dessurt	127M	✓	63.2	-	-	-	-	-	-	-	-	-
Donut	176M	✓	67.5	11.6	61.6	30.0	18.8	54.6	41.8	43.5	74.4	93.91
Pix2Struct$_{base}$	282M	✓	72.1	38.2	-	-	-	-	56.0	-	88.0	-
Pix2Struct$_{large}$	1.3B	✓	**76.6**	40.0	-	-	-	-	58.6	-	95.5	-
DocOwl	86M	×	<u>65.4</u>	**42.2**	<u>49.5</u>	**32.8**	**29.4**	**67.6**	**59.3**	**57.6**	**118.4**	**221.7**

mPLUG-DocOwl1.5

结构布局理解对于多模态文档理解十分关键


RESPONSE CODE REQUEST CONFIRMATION
Response Code Assigned: W24
Balance Control in Obese Subjects during Quiet Stance: A State-of-the Art
Trending TWEENS
How Kids use Social Media
95 %
81 %
50 %
Among kids under 13
21 %
26 %
Two-thirds of Afghan and Iraqi special immigrant visa recipients were dependents of applicants
Number of Afghan and Iraqi citizens admitted to the U.S. under special immigrant visa programs, 2007-17
Principal applicants
24,000
Afghan 22%
Iraqi 12%
Afghan 47%
Iraqi 19%
Dependents 47,000

(a) What is the assigned response code? W24

(b) which edition has unlimited remote desktop services and virtulization rights? Datacenter

(c) What is the title of the paper in the website? Balance Control in Obese Subjects during Quiet Stance: A state-of-the Art.

(d) What percentage of teenagers from the age group 12-17 didn't use the Internet? 5%

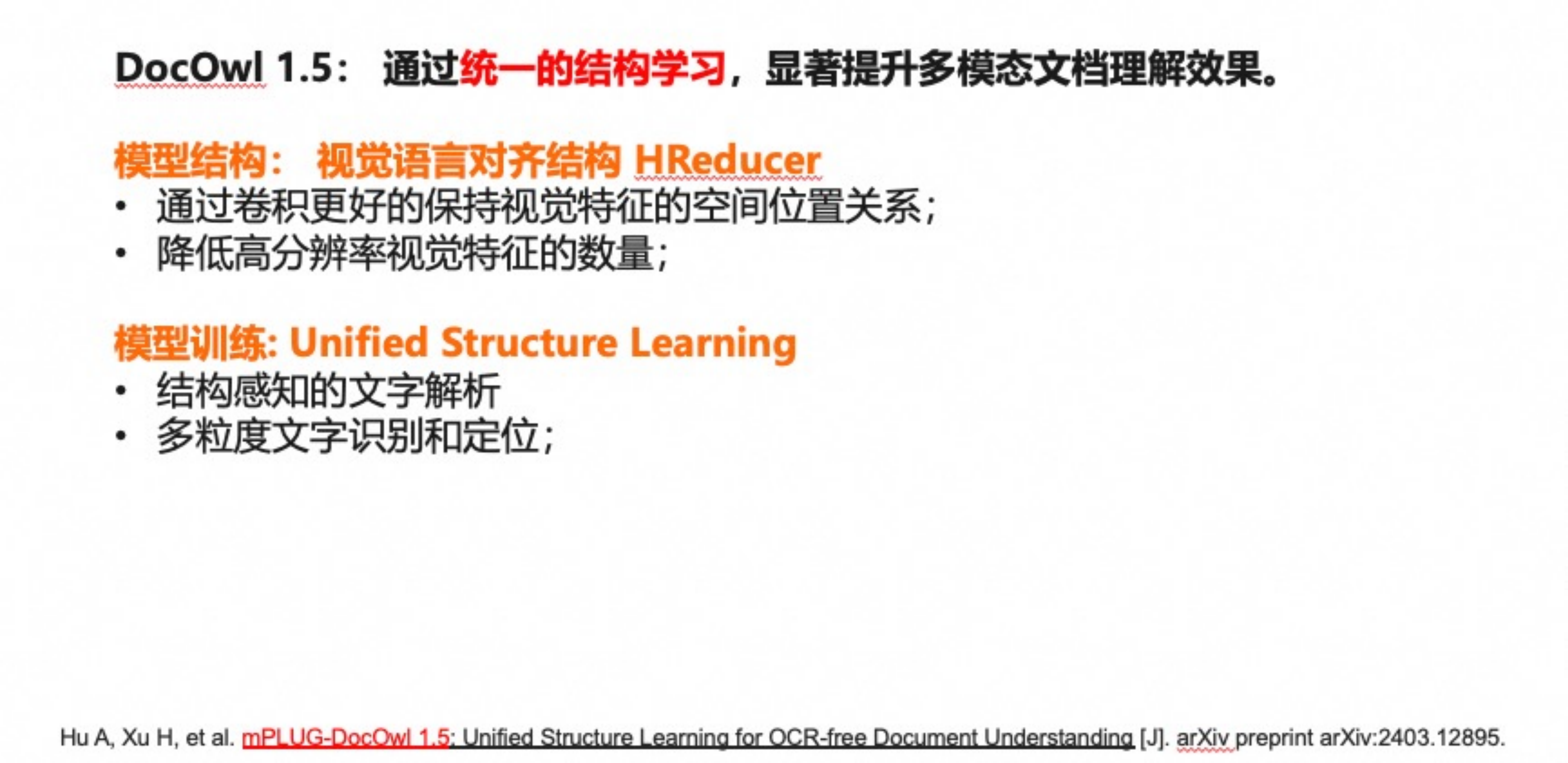
(e) What is the forecast for the increase in customs duty revenue in 2030? 100

(f) What is the percentage of Iraqi dependents citizen? 0.19.

Figure 2: Illustrations of the importance of structure information in Visual Document Understanding on documents (a), tables (b), webpages (c), infographics (d), and charts (e-f).

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5



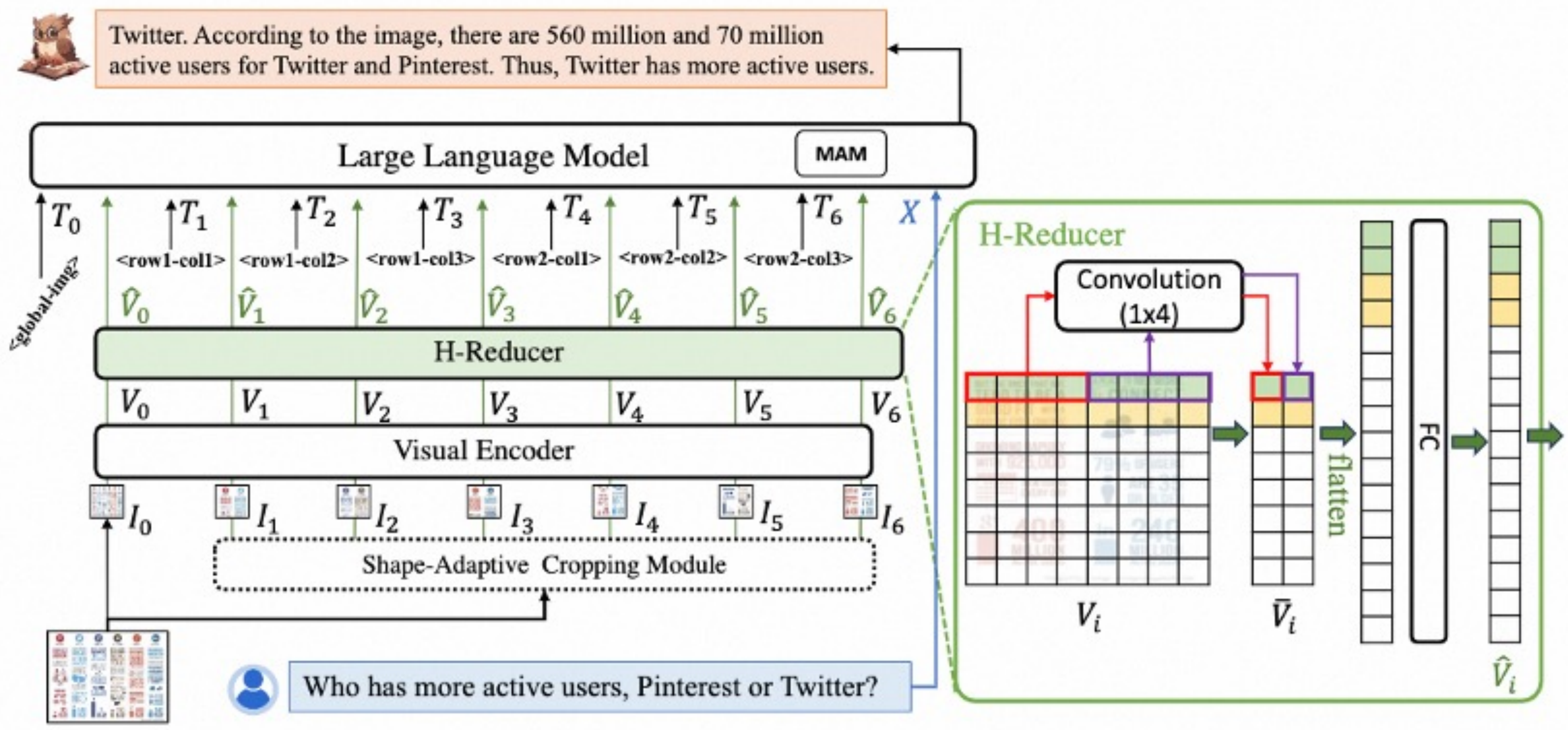
DocOwl 1.5： 通过统一的结构学习，显著提升多模态文档理解效果。

模型结构： 视觉语言对齐结构 HReducer
- 通过卷积更好的保持视觉特征的空间位置关系；
- 降低高分辨率视觉特征的数量；

模型训练: Unified Structure Learning
- 结构感知的文字解析
- 多粒度文字识别和定位；

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5

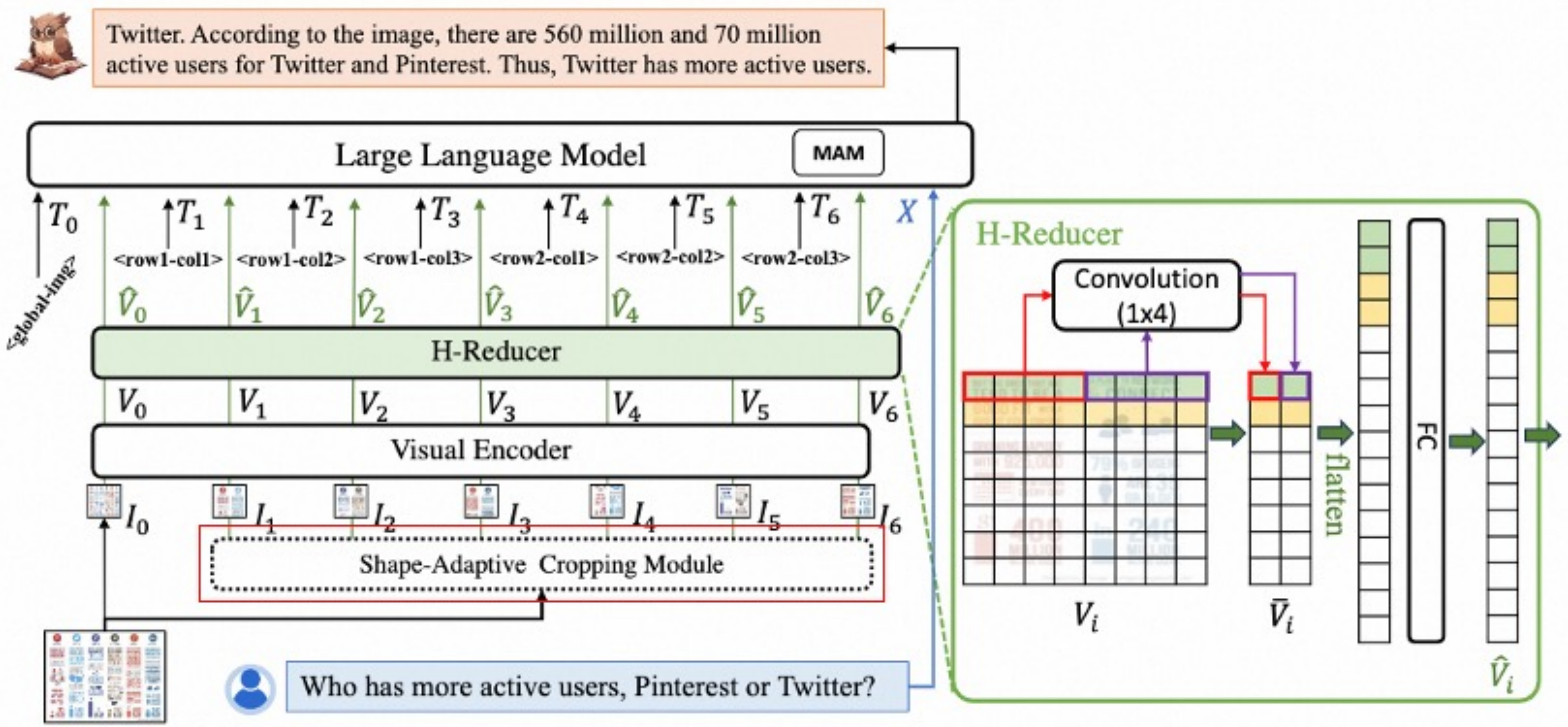


Twitter. According to the image, there are 560 million and 70 million active users for Twitter and Pinterest. Thus, Twitter has more active users.

Who has more active users, Pinterest or Twitter?

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5



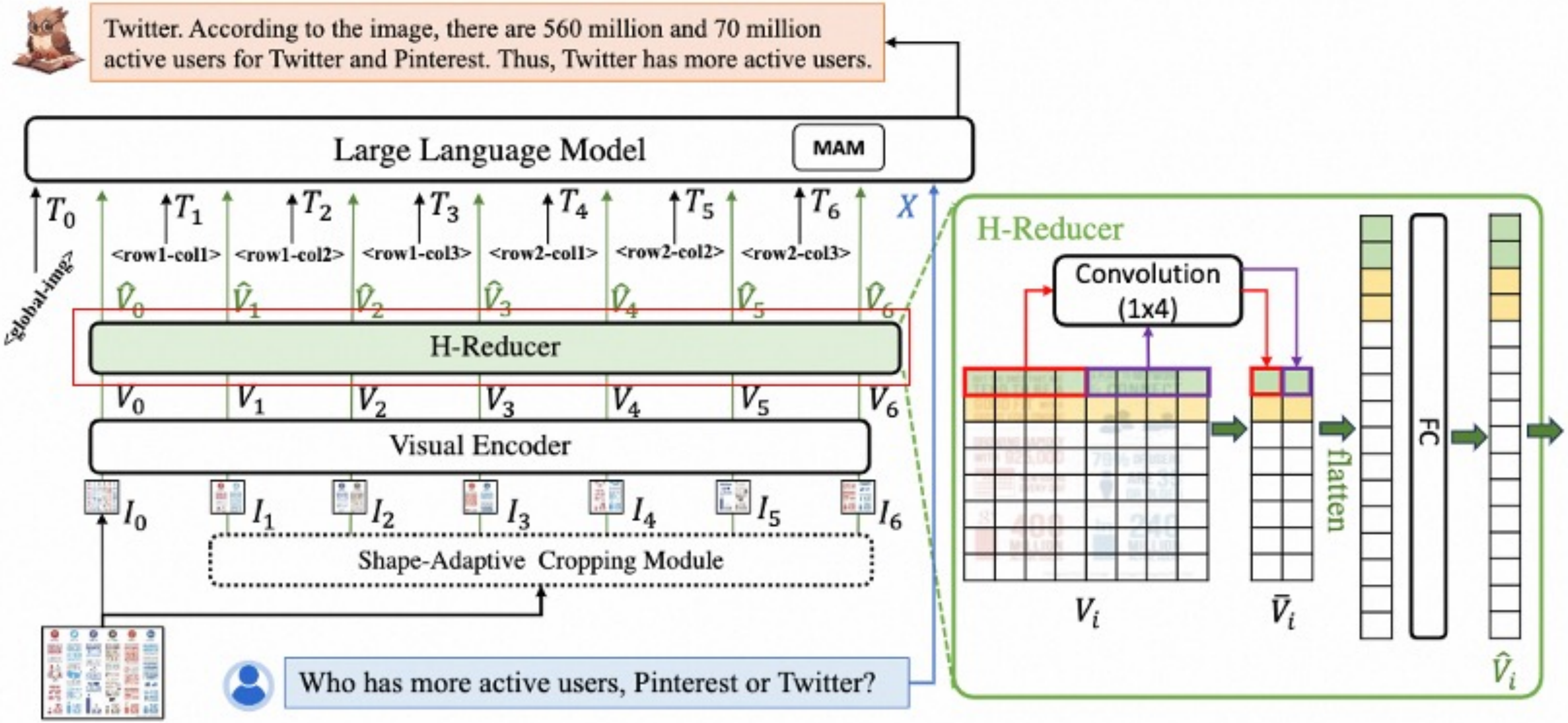
Twitter. According to the image, there are 560 million and 70 million active users for Twitter and Pinterest. Thus, Twitter has more active users.

Who has more active users, Pinterest or Twitter?

延续DocOwl，采用Shape-Adaptive Cropping Module处理高分辨率图片

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

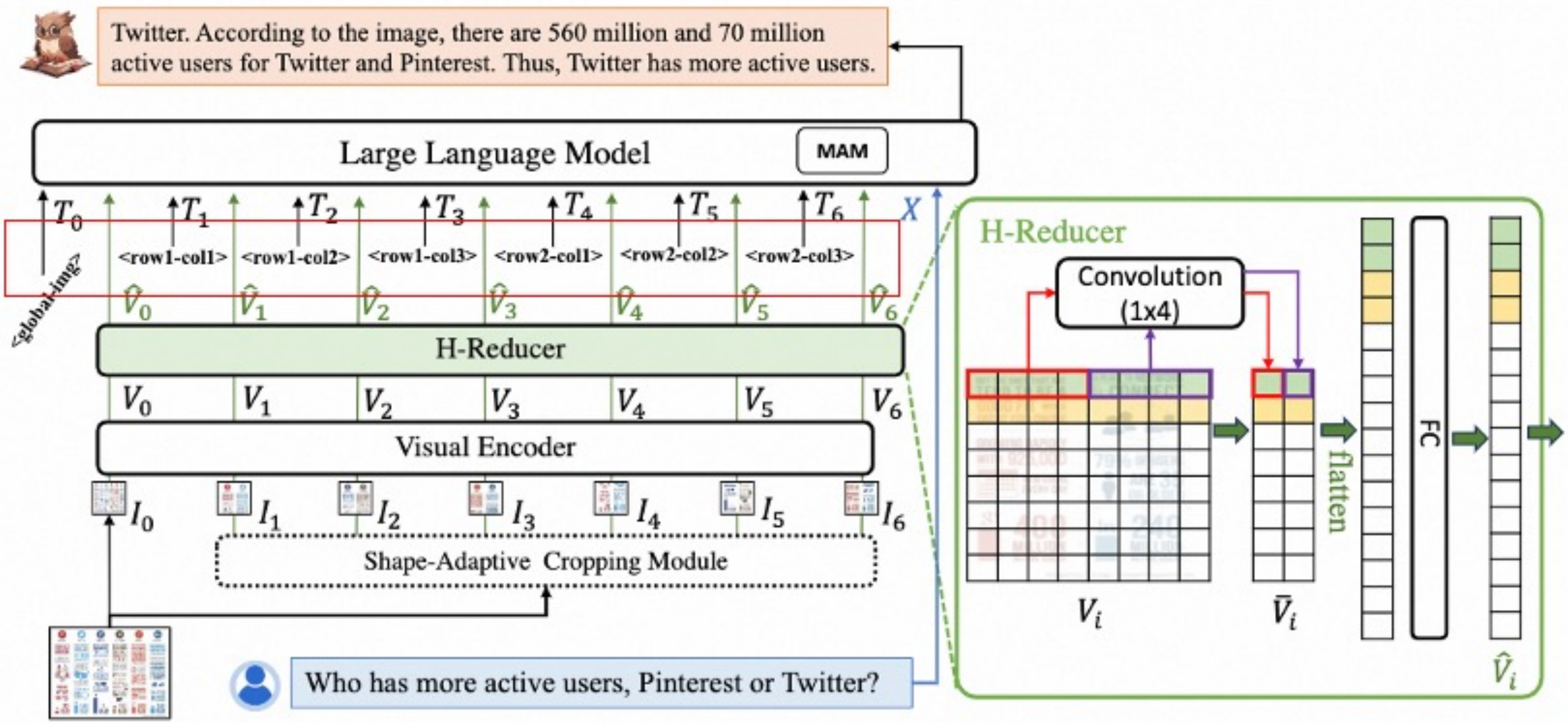
mPLUG-DocOwl1.5



- 不同于DocOwl，采用基于卷积的HReducer将视觉特征对齐到语言特征
- 考虑到文字大部分为横向排布，采用1x4的卷积

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5


Twitter. According to the image, there are 560 million and 70 million active users for Twitter and Pinterest. Thus, Twitter has more active users.

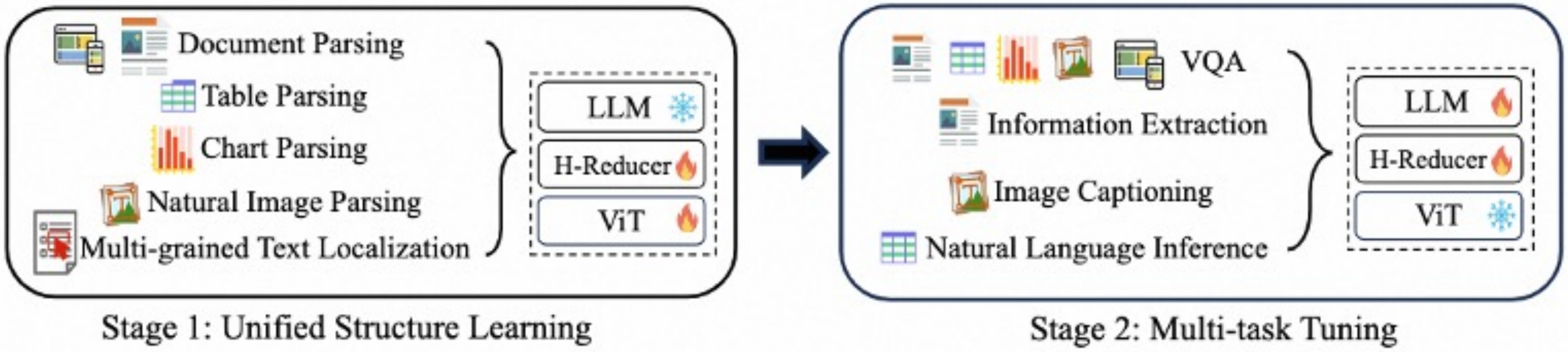
Who has more active users, Pinterest or Twitter?

- 不同于DocOwl采用学习的位置编码，DocOwl 1.5直接采用特殊的文本token表示当前子图的位置

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5

- 模型训练

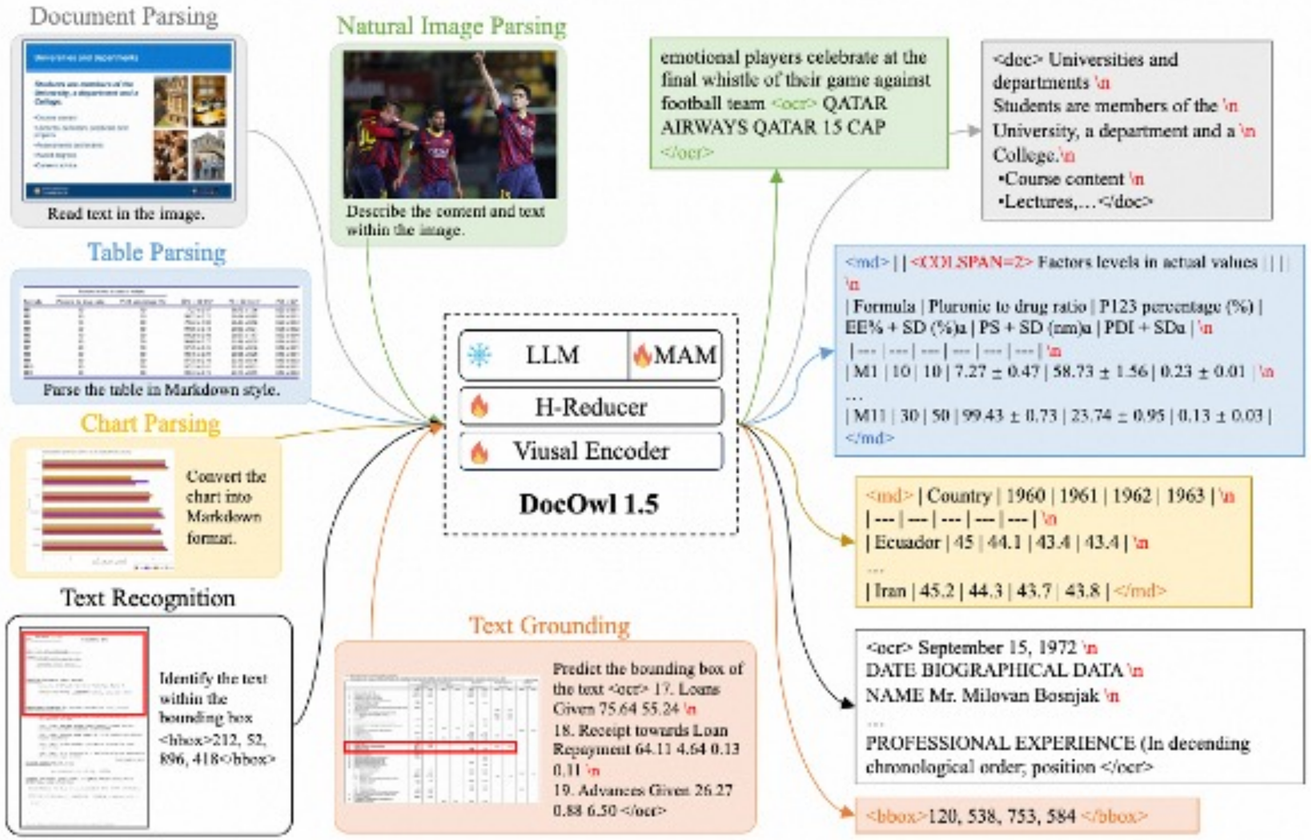


Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5

- 模型训练

- **Struct-aware Parsing**
  - Doc/Webpage Parsing
  - Table Parsing
  - Chart Parsing
  - Natural Image Parsing
- **Multi-grained Text Localization**
  - Text Recognition
  - Text Grounding


Document Parsing
Read text in the image.
Natural Image Parsing
Describe the content and text within the image.
Table Parsing
Parse the table in Markdown style.
Chart Parsing
Convert the chart into Markdown format.
Text Recognition
Identify the text within the bounding box <bbox>212, 52, 896, 418</bbox>
Text Grounding
LLM
MAM
H-Reducer
Viusal Encoder
DocOwl 1.5

Predict the bounding box of the text <ocr> 17. Loans Given 75.64 55.24 \n
18. Receipt towards Loan Repayment 64.11 4.64 0.13 0.11 \n
19. Advances Given 26.27 0.88 6.50 </ocr>

emotional players celebrate at the final whistle of their game against football team <ocr> QATAR AIRWAYS QATAR 15 CAP </ocr>

<doc> Universities and departments \n
Students are members of the \n
University, a department and a \n
College.\n
•Course content \n
•Lectures,…</doc>

<md> | | <COLSPAN=2> Factors levels in actual values | | |
\n
| Formula | Pluronic to drug ratio | P123 percentage (%) | EE% + SD (%)a | PS + SD (nm)a | PDI + SDa | \n
| --- | --- | --- | --- | --- | --- | \n
| M1 | 10 | 10 | 7.27 ± 0.47 | 58.73 ± 1.56 | 0.23 ± 0.01 | \n
…
| M11 | 30 | 50 | 99.43 ± 0.73 | 23.74 ± 0.95 | 0.13 ± 0.03 |
</md>

<md> | Country | 1960 | 1961 | 1962 | 1963 | \n
| --- | --- | --- | --- | --- | \n
| Ecuador | 45 | 44.1 | 43.4 | 43.4 | \n
…
| Iran | 45.2 | 44.3 | 43.7 | 43.8 | </md>

<ocr> September 15, 1972 \n
DATE BIOGRAPHICAL DATA \n
NAME Mr. Milovan Bosnjak \n
…
PROFESSIONAL EXPERIENCE (In decending chronological order; position </ocr>

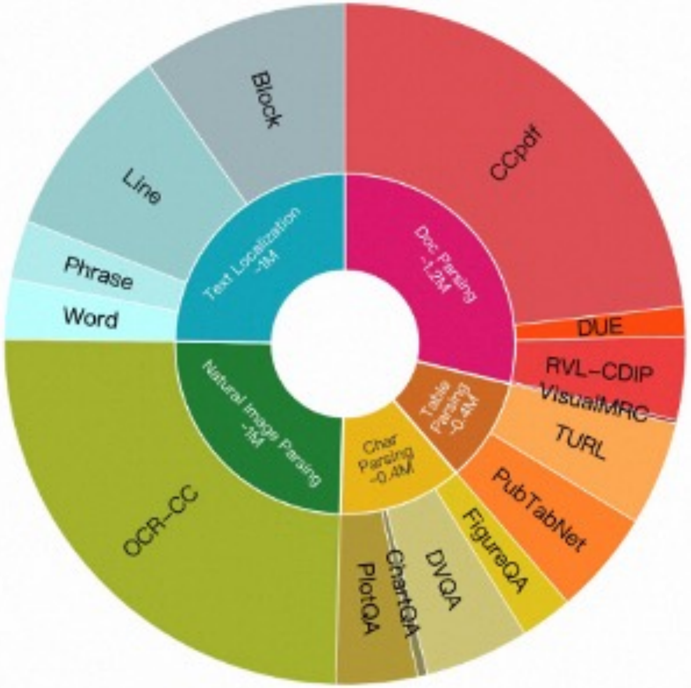
<bbox>120, 538, 753, 584 </bbox>

Figure 4: The illustration of Unified Structure Learning of DocOwl 1.5.

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5

- 模型训练



Task	Dataset	Num
Document Parsing	CCpdf	938,058
	RVL-CDIP	159,418
	DUE	56,809
	VisualMRC	7,016
Table Parsing	TURL	200,000
	PubTabNet	200,000
Chart Parsing	ChartQA	18,317
	FigureQA	99,714
	PlotQA	157,070
	DQA	200,000
Natural Image Parsing	OCR-CC	1,000,000
Multi-Grained Text Localization	Word	114,344
	Phrase	113,130
	Line	378,432
	Block	394,094
Total		4,036,402

Figure 5: Detailed statistics of DocStruct4M.

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5

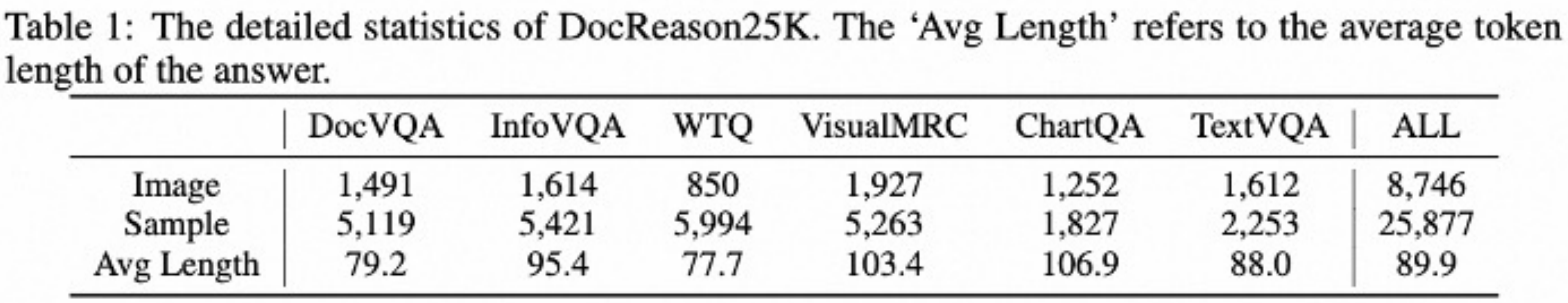


Table 1: The detailed statistics of DocReason25K. The 'Avg Length' refers to the average token length of the answer.

	DocVQA	InfoVQA	WTQ	VisualMRC	ChartQA	TextVQA	ALL
Image	1,491	1,614	850	1,927	1,252	1,612	8,746
Sample	5,119	5,421	5,994	5,263	1,827	2,253	25,877
Avg Length	79.2	95.4	77.7	103.4	106.9	88.0	89.9

已有下游任务只提供简单的词组作为答案，并不能给出详细的解释

DocReason25K：基于GPT3.5/GPT4V构建的多模态文档领域关于详细解释的微调数据集

DocOwl 1.5-Chat: 联合已有下游任务和DocReason25K进行第二阶段训练。

Hu A, Xu H, et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding [J]. arXiv preprint arXiv:2403.12895.

mPLUG-DocOwl1.5

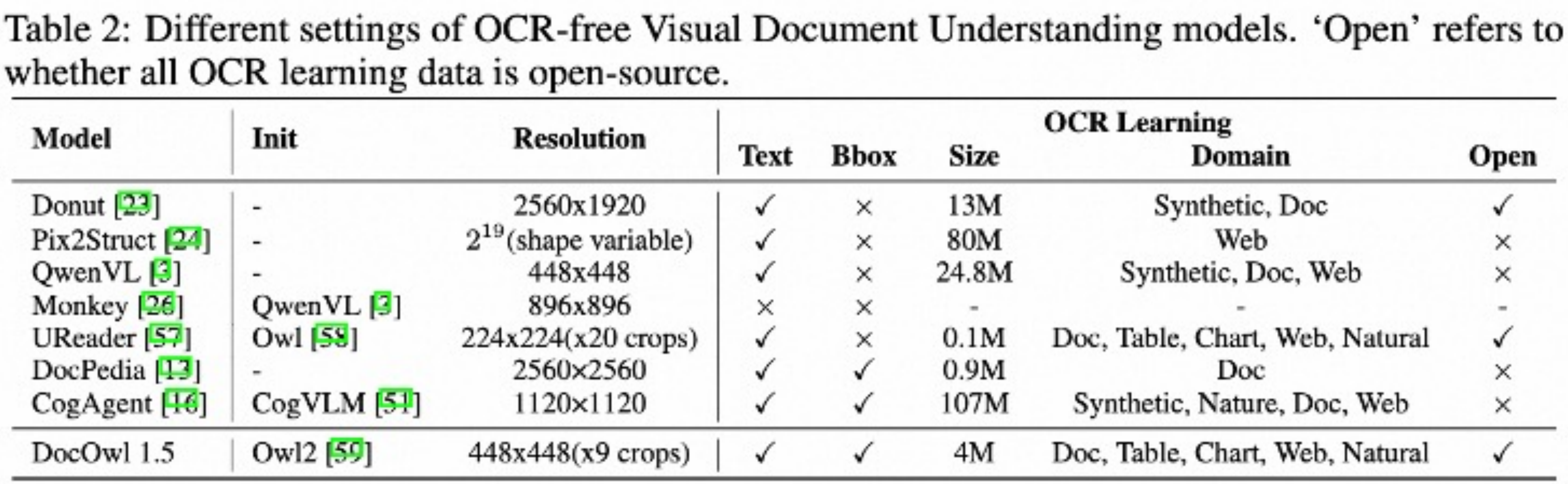


Table 2: Different settings of OCR-free Visual Document Understanding models. 'Open' refers to whether all OCR learning data is open-source.

Model	Init	Resolution	OCR Learning				
			Text	Bbox	Size	Domain	Open
Donut [23]	-	2560x1920	✓	×	13M	Synthetic, Doc	✓
Pix2Struct [24]	-	2^{19}(shape variable)	✓	×	80M	Web	×
QwenVL [3]	-	448x448	✓	×	24.8M	Synthetic, Doc, Web	×
Monkey [26]	QwenVL [3]	896x896	×	×	-	-	-
UReader [57]	Owl [58]	224x224(x20 crops)	✓	×	0.1M	Doc, Table, Chart, Web, Natural	✓
DocPedia [13]	-	2560x2560	✓	✓	0.9M	Doc	×
CogAgent [16]	CogVLM [51]	1120x1120	✓	✓	107M	Synthetic, Nature, Doc, Web	×
DocOwl 1.5	Owl2 [59]	448x448(x9 crops)	✓	✓	4M	Doc, Table, Chart, Web, Natural	✓

相比于当前SOTA文档理解大模型：

- DocOwl1.5同时进行了文字、结构以及位置的学习
- DocOwl1.5覆盖了更多的图片类型
- DocOwl1.5只利用开源数据，没有利用私有化数据，易复现

mPLUG-DocOwl1.5

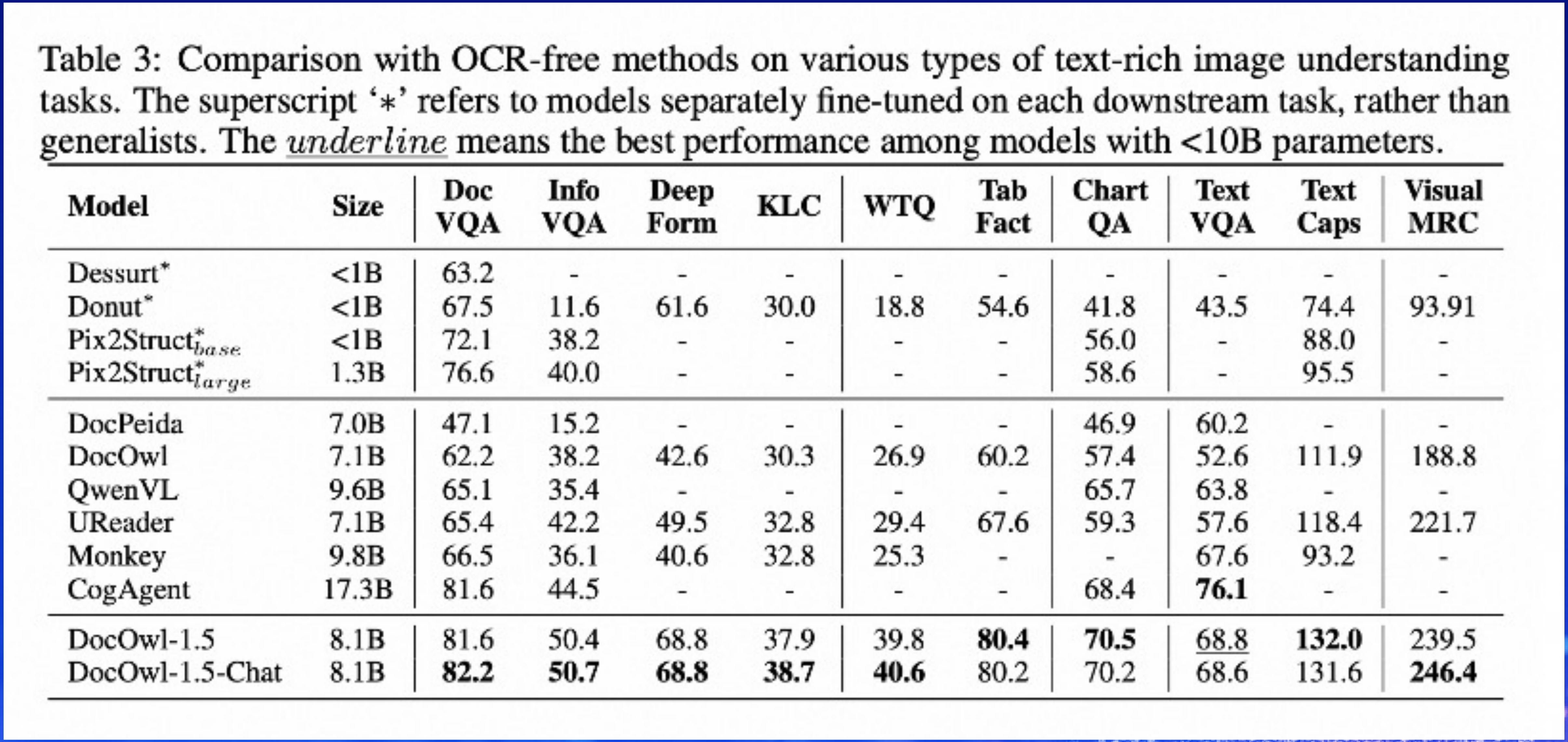
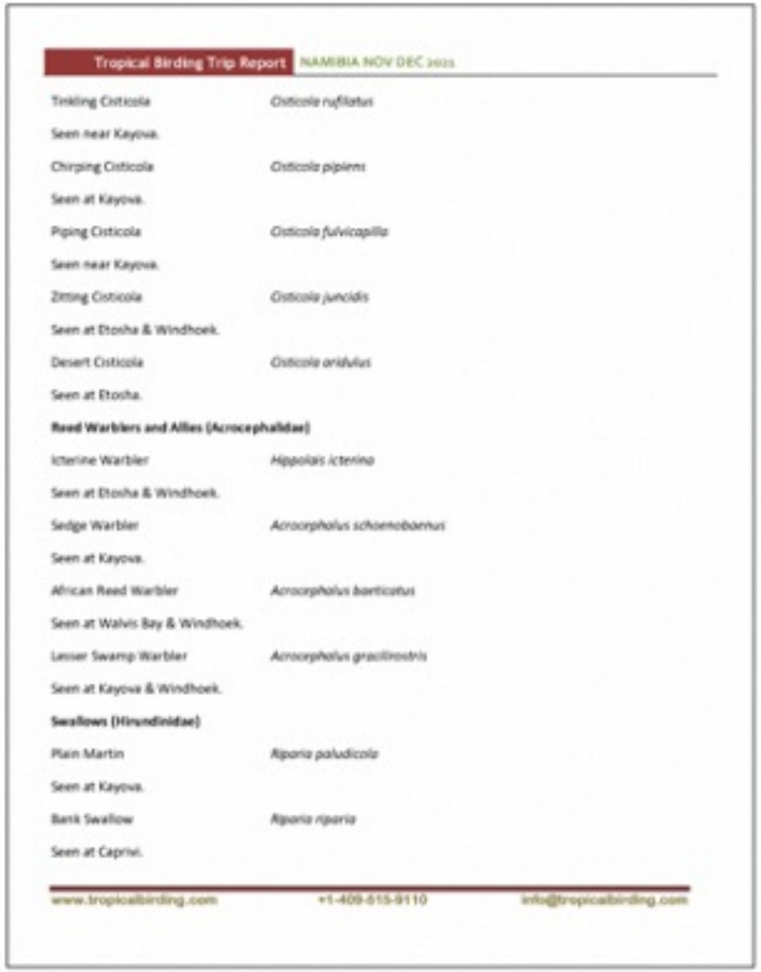


Table 3: Comparison with OCR-free methods on various types of text-rich image understanding tasks. The superscript '$*$' refers to models separately fine-tuned on each downstream task, rather than generalists. The *underline* means the best performance among models with <10B parameters.

Model	**Size**	**Doc VQA**	**Info VQA**	**Deep Form**	**KLC**	**WTQ**	**Tab Fact**	**Chart QA**	**Text VQA**	**Text Caps**	**Visual MRC**
Dessurt*	<1B	63.2	-	-	-	-	-	-	-	-	-
Donut*	<1B	67.5	11.6	61.6	30.0	18.8	54.6	41.8	43.5	74.4	93.91
Pix2Struct$^*_{base}$	<1B	72.1	38.2	-	-	-	-	56.0	-	88.0	-
Pix2Struct$^*_{large}$	1.3B	76.6	40.0	-	-	-	-	58.6	-	95.5	-
DocPeida	7.0B	47.1	15.2	-	-	-	-	46.9	60.2	-	-
DocOwl	7.1B	62.2	38.2	42.6	30.3	26.9	60.2	57.4	52.6	111.9	188.8
QwenVL	9.6B	65.1	35.4	-	-	-	-	65.7	63.8	-	-
UReader	7.1B	65.4	42.2	49.5	32.8	29.4	67.6	59.3	57.6	118.4	221.7
Monkey	9.8B	66.5	36.1	40.6	32.8	25.3	-	-	67.6	93.2	-
CogAgent	17.3B	81.6	44.5	-	-	-	-	68.4	**76.1**	-	-
DocOwl-1.5	8.1B	81.6	50.4	68.8	37.9	39.8	**80.4**	**70.5**	68.8	**132.0**	239.5
DocOwl-1.5-Chat	8.1B	**82.2**	**50.7**	**68.8**	**38.7**	**40.6**	80.2	70.2	68.6	131.6	**246.4**

mPLUG-DocOwl1.5

样例：文档解析



Tropical Birding Trip Report NAMIBIA NOV DEC 2021

Tinkling Cisticola *Cisticola rufilatus*
Seen near Kayova.
Chirping Cisticola *Cisticola pipiens*
Seen at Kayova.
Piping Cisticola *Cisticola fulvicapilla*
Seen near Kayova.
Zitting Cisticola *Cisticola juncidis*
Seen at Etosha & Windhoek.
Desert Cisticola *Cisticola aridulus*
Seen at Etosha.
Reed Warblers and Allies (Acrocephalidae)
Icterine Warbler *Hippolais icterina*
Seen at Etosha & Windhoek.
Sedge Warbler *Acrocephalus schoenobaenus*
Seen at Kayova.
African Reed Warbler *Acrocephalus baeticatus*
Seen at Walvis Bay & Windhoek.
Lesser Swamp Warbler *Acrocephalus gracilirostris*
Seen at Kayova & Windhoek.
Swallows (Hirundinidae)
Plain Martin *Riparia paludicola*
Seen at Kayova.
Bank Swallow *Riparia riparia*
Seen at Caprivi.

www.tropicalbirding.com +1-409-515-9110 info@tropicalbirding.com

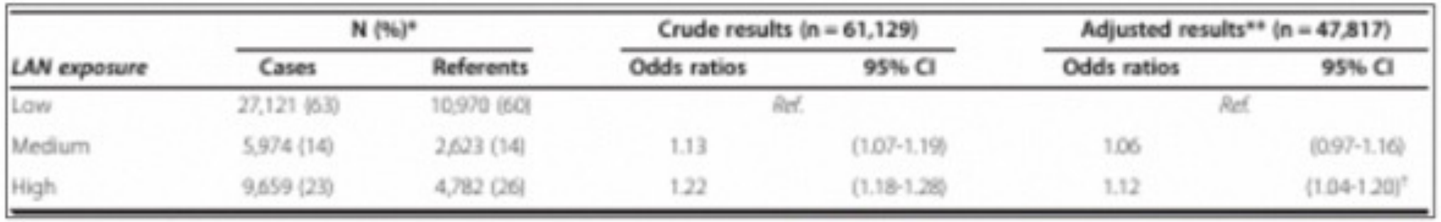
Extract words from the graphic.

DocOwl 1.5:

```
<doc> Tropical Birding Trip Report NAMIBIA NOV DEC 2021
 Tinkling Cisticola     Cisticola rufilatus
 Seen near Kayova.
 Chirping Cisticola     Cisticola pipiens
 Seen at Kayova.
 Piping Cisticola      Cisticola fulvicapilla
 Seen near Kayova.
 Zitting Cisticola     Cisticola juncidis
 Seen at Etosha & Windhoek.
 Desert Cisticola      Cistcola aridulus
 Seen at Etosha.
 Reed Warblers and Allies (Acrocephalidae)
 Icterine Warbler       Hippolais icterina
 Seen at Etosha & Windhoek.
 Sedge Warbler          Acrocephalus schoenobaenus
 Seen at Kayova.
 African Reed Warbler    Acrocephalus baeticus
 Seen at Walvis Bay & Windhoek.
 Lesser Swamp Warbler    Acrocephalus gracilirostris
 Seen at Kayova & Windhoek.
 Swallows (Hirundinidae)
 Plain Martin          Riparia paludicola
 Seen at Kayova.
 Bank Swallow           Riparia riparia
 Seen at Caprivi.
 www.tropicalbirding.com    +1-409-515-9110       info@tropicalbirding.com </doc>
```

mPLUG-DocOwl1.5

样例：图表解析

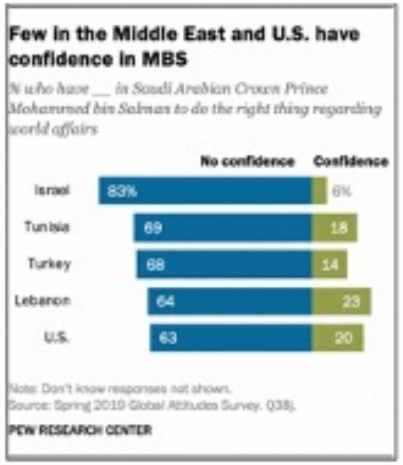


	N (%)*		Crude results (n = 61,129)		Adjusted results** (n = 47,817)	
LAN exposure	Cases	Referents	Odds ratios	95% CI	Odds ratios	95% CI
Low	27,121 (63)	10,970 (60)	*Ref.*		*Ref.*	
Medium	5,974 (14)	2,623 (14)	1.13	(1.07-1.19)	1.06	(0.97-1.16)
High	9,659 (23)	4,782 (26)	1.22	(1.18-1.28)	1.12	(1.04-1.20)†

Convert the figure into Markdown language.

DocOwl 1.5:

```
<md> | | <COLSPAN=2> N (%)** | <COLSPAN=2> Crude results (n = 61,129) | <COLSPAN=2> Adjusted results** (n = 47,817) |
| LAN exposure | Cases | Referents | Odds ratios | 95% CI | Odds ratios | 95% CI |
| --- | --- | --- | --- | --- | --- | --- |
| Low | 27,121 (63) | 10,970 (60) | | Ref. | | Ref. |
| Medium | 5,974 (14) | 2,623 (14) | 1.13 | (1.07-1.19) | 1.06 | (0.97-1.16) |
| High | 9,659 (23) | 4,782 (26) | 1.22 | (1.18-1.28) | 1.12 | (1.04-1.20)† | </md>
```



Few in the Middle East and U.S. have confidence in MBS

% who have __ in Saudi Arabian Crown Prince Mohammed bin Salman to do the right thing regarding world affairs

Note: Don't know responses not shown.
Source: Spring 2019 Global Attitudes Survey. Q38.

PEW RESEARCH CENTER

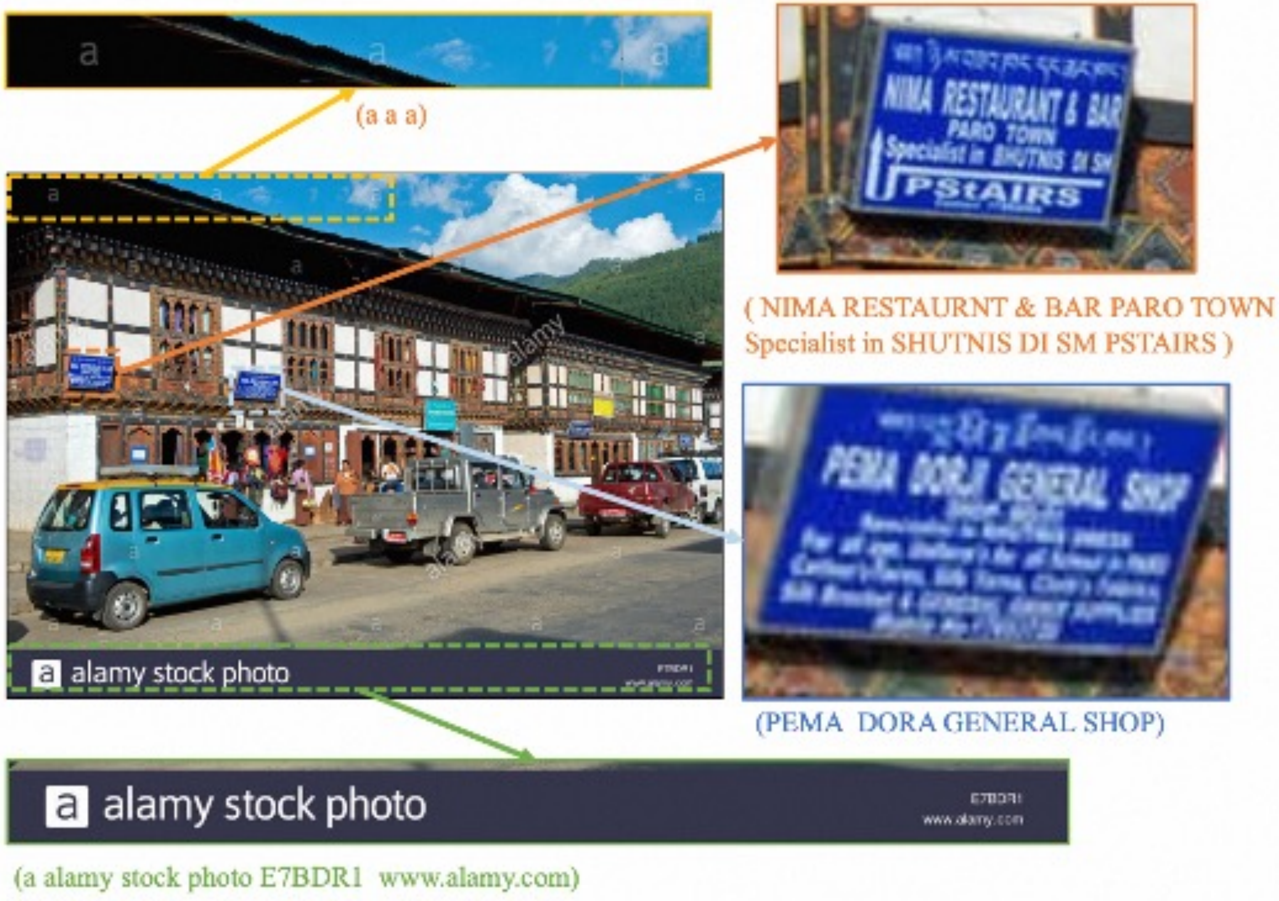
Render the figure in Markdown format.

DocOwl 1.5:

```
<md> | Entity | No confidence | Confidence |
| --- | --- | --- |
| U.S | 63 | 20 |
| Lebanon | 64 | 23 |
| Turkey | 68 | 14 |
| Tunisia | 69 | 18 |
| Israel | 83 | 6 | </md>
```

mPLUG-DocOwl1.5

样例： 自然图解析


(a a a)
(NIMA RESTAURNT & BAR PARO TOWN Specialist in SHUTNIS DI SM PSTAIRS)
(PEMA DORA GENERAL SHOP)
alamy stock photo
(a alamy stock photo E7BDR1 www.alamy.com)

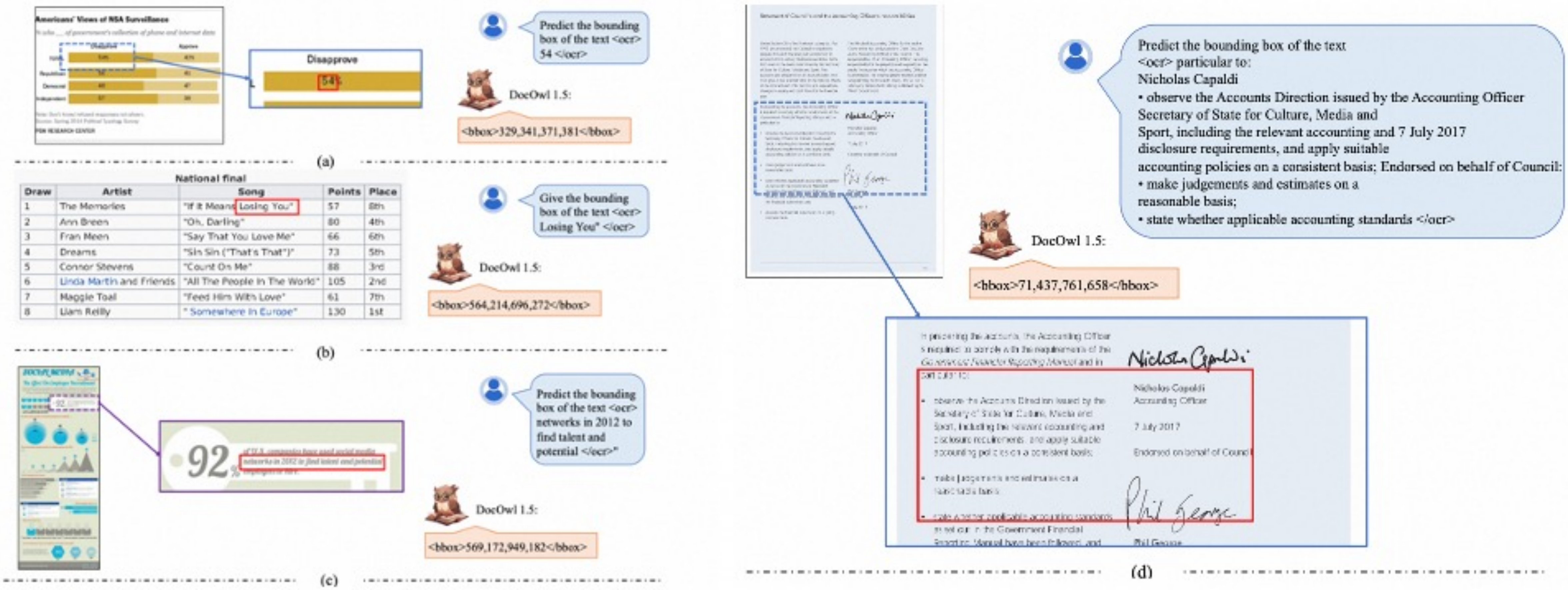
Detail the information and textual components depicted in the image.

DocOwl 1.5:

the main street in the town <ocr> a a a alamy NIMA RESTAURANT & BAR Special TOURS & DEL PSTAIRS PERIA DORA GENERAL SHOP SONAR ENTERPRISE a a a a a a alamy photo stock E7BDR1 a www.alamy.com </ocr>

mPLUG-DocOwl1.5

样例： 多粒度文字识别



Predict the bounding box of the text <ocr> 54 </ocr>

DocOwl 1.5:

<bbox>329,341,371,381</bbox>

National final

Draw	Artist	Song	Points	Place
1	The Memories	"If It Means Losing You"	57	8th
2	Ann Breen	"Oh, Darling"	80	4th
3	Fran Meen	"Say That You Love Me"	66	6th
4	Dreams	"Sin Sin ("That's That")"	73	5th
5	Connor Stevens	"Count On Me"	88	3rd
6	Linda Martin and Friends	"All The People In The World"	105	2nd
7	Maggie Toal	"Feed Him With Love"	61	7th
8	Liam Reilly	" Somewhere In Europe"	130	1st

Give the bounding box of the text <ocr> Losing You" </ocr>

DocOwl 1.5:

<bbox>564,214,696,272</bbox>

Predict the bounding box of the text <ocr> networks in 2012 to find talent and potential </ocr>"

DocOwl 1.5:

<bbox>569,172,949,182</bbox>

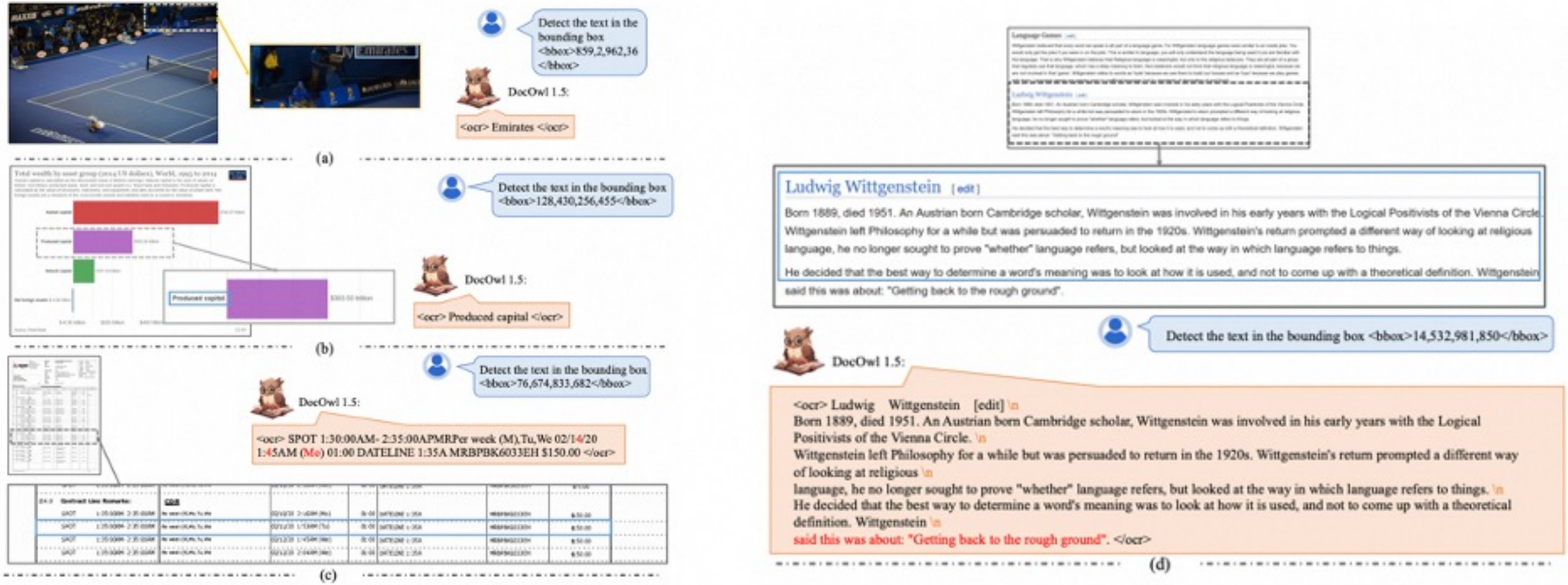
Predict the bounding box of the text <ocr> particular to: Nicholas Capaldi
• observe the Accounts Direction issued by the Accounting Officer Secretary of State for Culture, Media and Sport, including the relevant accounting and 7 July 2017 disclosure requirements, and apply suitable accounting policies on a consistent basis; Endorsed on behalf of Council:
• make judgements and estimates on a reasonable basis;
• state whether applicable accounting standards </ocr>

DocOwl 1.5:

<bbox>71,437,761,658</bbox>

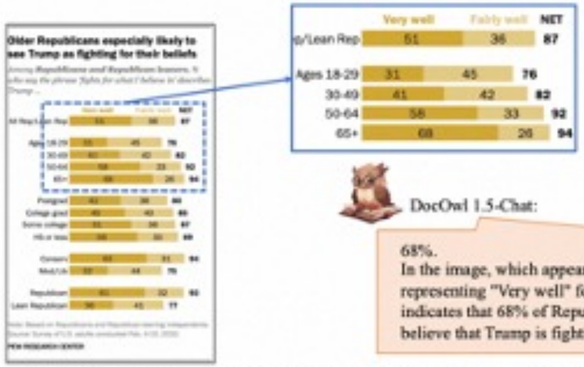
mPLUG-DocOwl1.5

样例：多粒度文字定位


Detect the text in the bounding box <bbox>859,2,962,36</bbox>
DocOwl 1.5:
<ocr> Emirates </ocr>
(a)
Detect the text in the bounding box <bbox>128,430,256,455</bbox>
Produced capital
DocOwl 1.5:
<ocr> Produced capital </ocr>
(b)
Detect the text in the bounding box <bbox>76,674,833,682</bbox>
DocOwl 1.5:
<ocr> SPOT 1:30:00AM- 2:35:00APMRPer week (M),Tu,We 02/14/20 1:45AM (Mo) 01:00 DATELINE 1:35A MRBPBK6033EH $150.00 </ocr>
(c)
Ludwig Wittgenstein [edit]
Born 1889, died 1951. An Austrian born Cambridge scholar, Wittgenstein was involved in his early years with the Logical Positivists of the Vienna Circle. Wittgenstein left Philosophy for a while but was persuaded to return in the 1920s. Wittgenstein's return prompted a different way of looking at religious language, he no longer sought to prove "whether" language refers, but looked at the way in which language refers to things.
He decided that the best way to determine a word's meaning was to look at how it is used, and not to come up with a theoretical definition. Wittgenstein said this was about: "Getting back to the rough ground".
Detect the text in the bounding box <bbox>14,532,981,850</bbox>
DocOwl 1.5:
<ocr> Ludwig Wittgenstein [edit] \n
Born 1889, died 1951. An Austrian born Cambridge scholar, Wittgenstein was involved in his early years with the Logical Positivists of the Vienna Circle. \n
Wittgenstein left Philosophy for a while but was persuaded to return in the 1920s. Wittgenstein's return prompted a different way of looking at religious \n
language, he no longer sought to prove "whether" language refers, but looked at the way in which language refers to things. \n
He decided that the best way to determine a word's meaning was to look at how it is used, and not to come up with a theoretical definition. Wittgenstein \n
said this was about: "Getting back to the rough ground". </ocr>
(d)

mPLUG-DocOwl1.5

样例：详细解释的问答

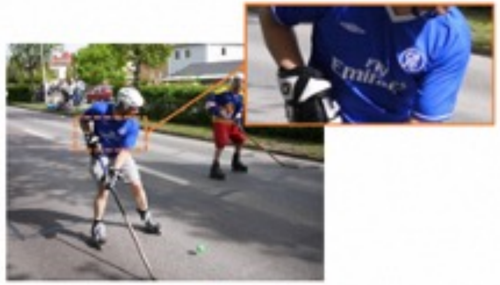


What's the value of the Very well bar in the 65+ age group? Answer the question with detailed explanation.

DocOwl 1.5-Chat:

68%.
In the image, which appears to be a chart from the Pew Research Center, the bar representing "Very well" for the age group 65 and older is marked at 68%. This indicates that 68% of Republicans and Republican leaners in this age group believe that Trump is fighting for their beliefs "Very well."

(a)

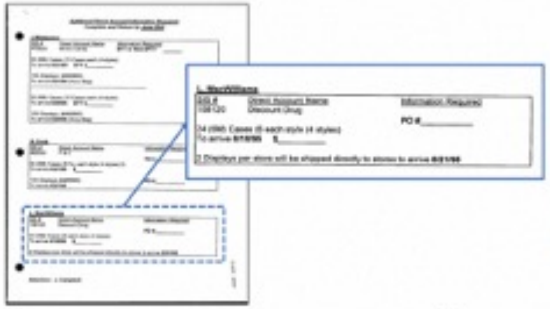


what is written on the blue shirt on the left? Answer the question with detailed explanation.

DocOwl 1.5-Chat:

Fly Emirates.
The blue shirt on the left has the word "Fly Emirates" written across the chest. The logo is in white lettering against a blue background, which is the color scheme typically used for the airline's branding.

(b)

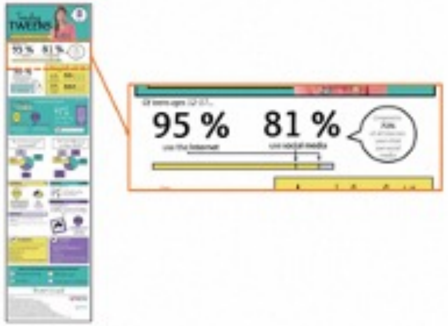

L. MacWilliams
SIS #
158120
Direct Account Name
Discount Drug
Information Required
PO #

What is the SIS # of L. MacWilliams? Answer the question with detailed explanation.

DocOwl 1.5-Chat:

SIS # 158120.
The SIS # of L. MacWilliams is 158120, as stated in the text under the section "L. MacWilliams SIS # Direct Account Name Information Required PO #".

(a)

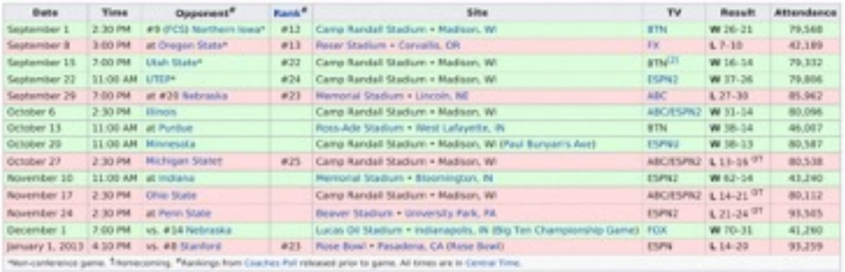


What percentage of teenagers from the age group 12-17 didn't use the Internet? Answer the question with detailed explanation.

DocOwl 1.5-Chat:

According to the text, 95% of teenagers ages 12-17 use the Internet. Therefore, the percentage of teenagers from the age group 12-17 who didn't use the Internet would be 100% - 95% = 5%.

(b)



Date	Time	Opponent#	Rank#	Site	TV	Result	Attendance
September 1	2:30 PM	#9 (FCS) Northern Iowa*	#12	Camp Randall Stadium • Madison, WI	BTN	W 26-21	79,568
September 8	3:00 PM	at Oregon State*	#13	Reser Stadium • Corvallis, OR	FX	L 7-10	42,189
September 15	7:00 PM	Utah State*	#22	Camp Randall Stadium • Madison, WI	BTN[2]	W 16-14	79,332
September 22	11:00 AM	UTEP*	#24	Camp Randall Stadium • Madison, WI	ESPN2	W 37-26	79,806
September 29	7:00 PM	at #20 Nebraska	#23	Memorial Stadium • Lincoln, NE	ABC	L 27-30	85,962
October 6	2:30 PM	Illinois		Camp Randall Stadium • Madison, WI	ABC/ESPN2	W 31-14	80,096
October 13	11:00 AM	at Purdue		Ross-Ade Stadium • West Lafayette, IN	BTN	W 38-14	46,007
October 20	11:00 AM	Minnesota		Camp Randall Stadium • Madison, WI (Paul Bunyan's Axe)	ESPN2	W 38-13	80,587
October 27	2:30 PM	Michigan State†	#25	Camp Randall Stadium • Madison, WI	ABC/ESPN2	L 13-16 OT	80,538
November 10	11:00 AM	at Indiana		Memorial Stadium • Bloomington, IN	ESPN2	W 62-14	43,240
November 17	2:30 PM	Ohio State		Camp Randall Stadium • Madison, WI	ABC/ESPN2	L 14-21 OT	80,112
November 24	2:30 PM	at Penn State		Beaver Stadium • University Park, PA	ESPN2	L 21-24 OT	93,505
December 1	7:00 PM	vs. #14 Nebraska		Lucas Oil Stadium • Indianapolis, IN (Big Ten Championship Game)	FOX	W 70-31	41,260
January 1, 2013	4:10 PM	vs. #8 Stanford	#23	Rose Bowl • Pasadena, CA (Rose Bowl)	ESPN	L 14-20	93,359

*Non-conference game. †Homecoming. #Rankings from Coaches Poll released prior to game. All times are in Central Time.

what is the last date and time on the table? Answer the question with detailed explanation.

DocOwl 1.5-Chat:

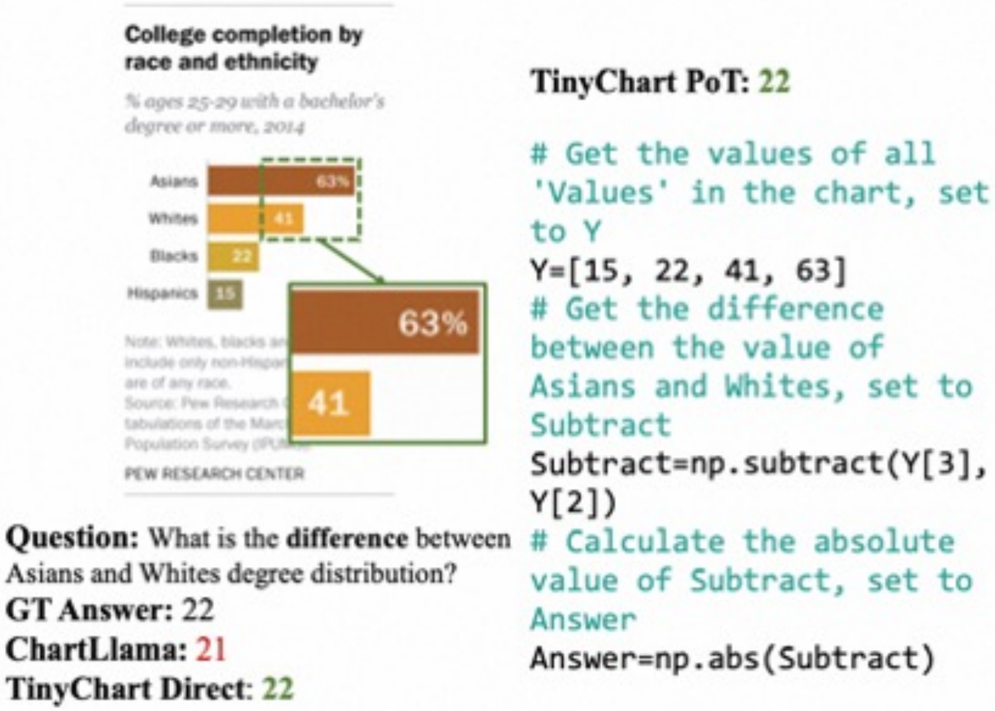
The last date and time on the chart is January 1, 2013 at 4:10 PM. The chart lists all the games played by a certain team, along with the date, time, opponent, rank, site, and result. The last game listed on the chart is against Stanford, which took place on January 1, 2013 at 4:10 PM. Therefore, the last date and time on the chart is January 1, 2013 at 4:10 PM.

(c)

TinyChart

图表（Chart）理解的主要挑战：

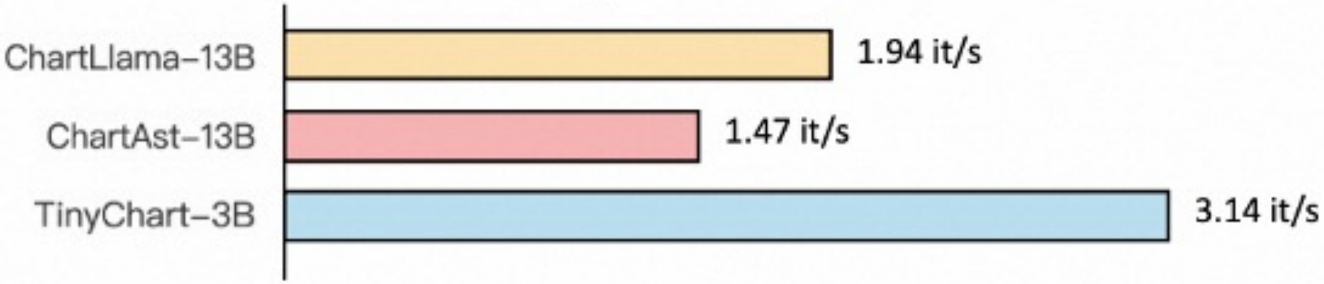
- 问答经常涉及多步数学推理
- 现有最优Chart多模态模型参数太大，推理慢（ChartLLama-13B, ChartAssistant-13B）



Question: What is the **difference** between Asians and Whites degree distribution?
GT Answer: 22
ChartLlama: 21
TinyChart Direct: 22

TinyChart PoT: 22

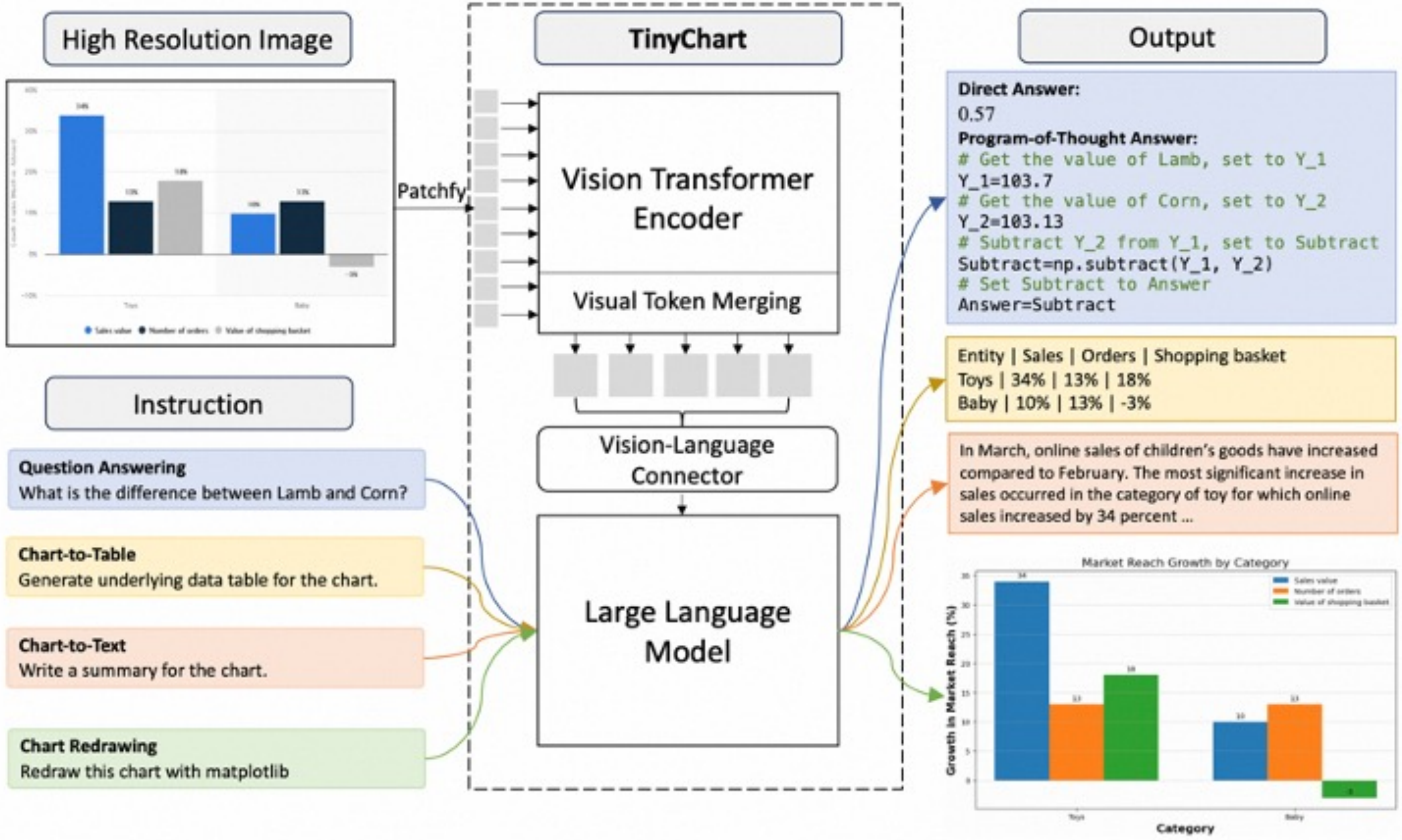
```
# Get the values of all 'Values' in the chart, set to Y
Y=[15, 22, 41, 63]
# Get the difference between the value of Asians and Whites, set to Subtract
Subtract=np.subtract(Y[3], Y[2])
# Calculate the absolute value of Subtract, set to Answer
Answer=np.abs(Subtract)
```



(b)

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

TinyChart



Question Answering
What is the difference between Lamb and Corn?

Chart-to-Table
Generate underlying data table for the chart.

Chart-to-Text
Write a summary for the chart.

Chart Redrawing
Redraw this chart with matplotlib

Direct Answer:
0.57

Program-of-Thought Answer:

```
# Get the value of Lamb, set to Y_1
Y_1=103.7
# Get the value of Corn, set to Y_2
Y_2=103.13
# Subtract Y_2 from Y_1, set to Subtract
Subtract=np.subtract(Y_1, Y_2)
# Set Subtract to Answer
Answer=Subtract
```

Entity	Sales	Orders	Shopping basket
Toys	34%	13%	18%
Baby	10%	13%	-3%

In March, online sales of children's goods have increased compared to February. The most significant increase in sales occurred in the category of toy for which online sales increased by 34 percent ...

Figure 2: Overview of TinyChart.

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

TinyChart

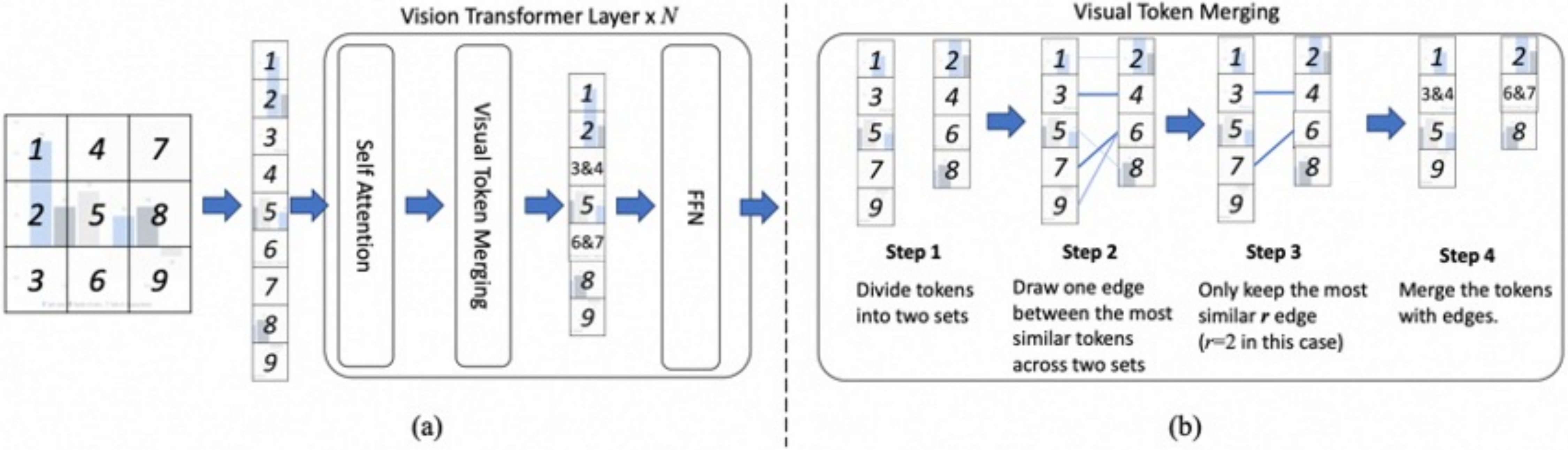
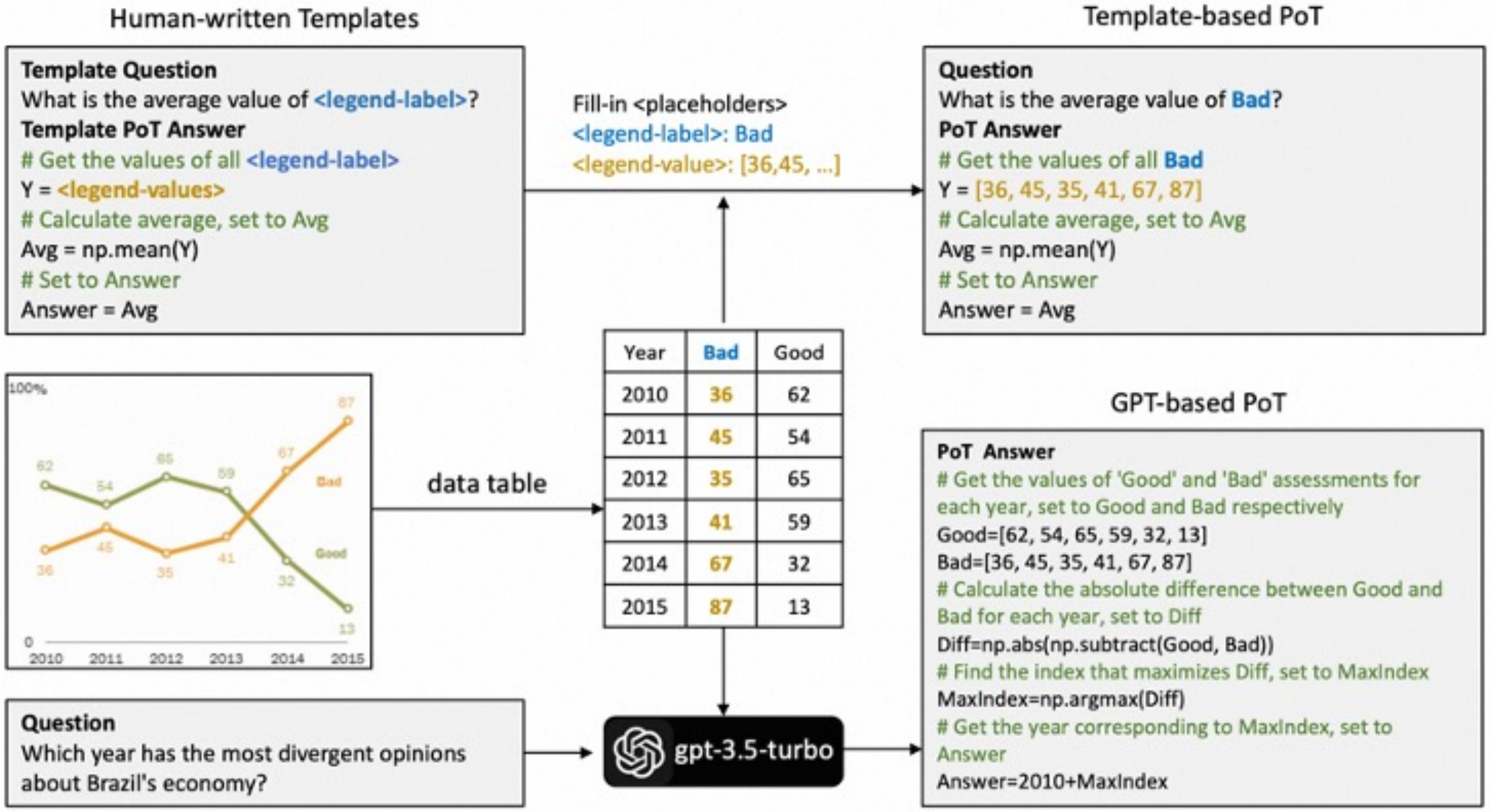

Vision Transformer Layer x *N*
1
4
7
2
5
8
3
6
9
Self Attention
Visual Token Merging
3&4
6&7
FFN
(a)
Visual Token Merging
Step 1
Divide tokens into two sets
Step 2
Draw one edge between the most similar tokens across two sets
Step 3
Only keep the most similar *r* edge (*r*=2 in this case)
Step 4
Merge the tokens with edges.
(b)

Figure 3: (a) Vision transformer layer with Visual Token Merging. (b) Process of the Visual Token Merging.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token Merging: Your ViT But Faster. In The Eleventh International Conference on Learning Representations.

TinyChart



Human-written Templates

Template Question

What is the average value of <legend-label>?

Template PoT Answer

```
# Get the values of all <legend-label>
Y = <legend-values>
# Calculate average, set to Avg
Avg = np.mean(Y)
# Set to Answer
Answer = Avg
```

Fill-in <placeholders>
<legend-label>: Bad
<legend-value>: [36,45, ...]

Template-based PoT

Question

What is the average value of Bad?

PoT Answer

```
# Get the values of all Bad
Y = [36, 45, 35, 41, 67, 87]
# Calculate average, set to Avg
Avg = np.mean(Y)
# Set to Answer
Answer = Avg
```

data table

Year	Bad	Good
2010	36	62
2011	45	54
2012	35	65
2013	41	59
2014	67	32
2015	87	13

Question

Which year has the most divergent opinions about Brazil's economy?

GPT-based PoT

PoT Answer

```
# Get the values of 'Good' and 'Bad' assessments for
each year, set to Good and Bad respectively
Good=[62, 54, 65, 59, 32, 13]
Bad=[36, 45, 35, 41, 67, 87]
# Calculate the absolute difference between Good and
Bad for each year, set to Diff
Diff=np.abs(np.subtract(Good, Bad))
# Find the index that maximizes Diff, set to MaxIndex
MaxIndex=np.argmax(Diff)
# Get the year corresponding to MaxIndex, set to
Answer
Answer=2010+MaxIndex
```

Figure 4: The demonstration of constructing Template-based PoT (upper half) and GPT-based PoT (lower half) in the ChartQA-PoT dataset.

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

TinyChart

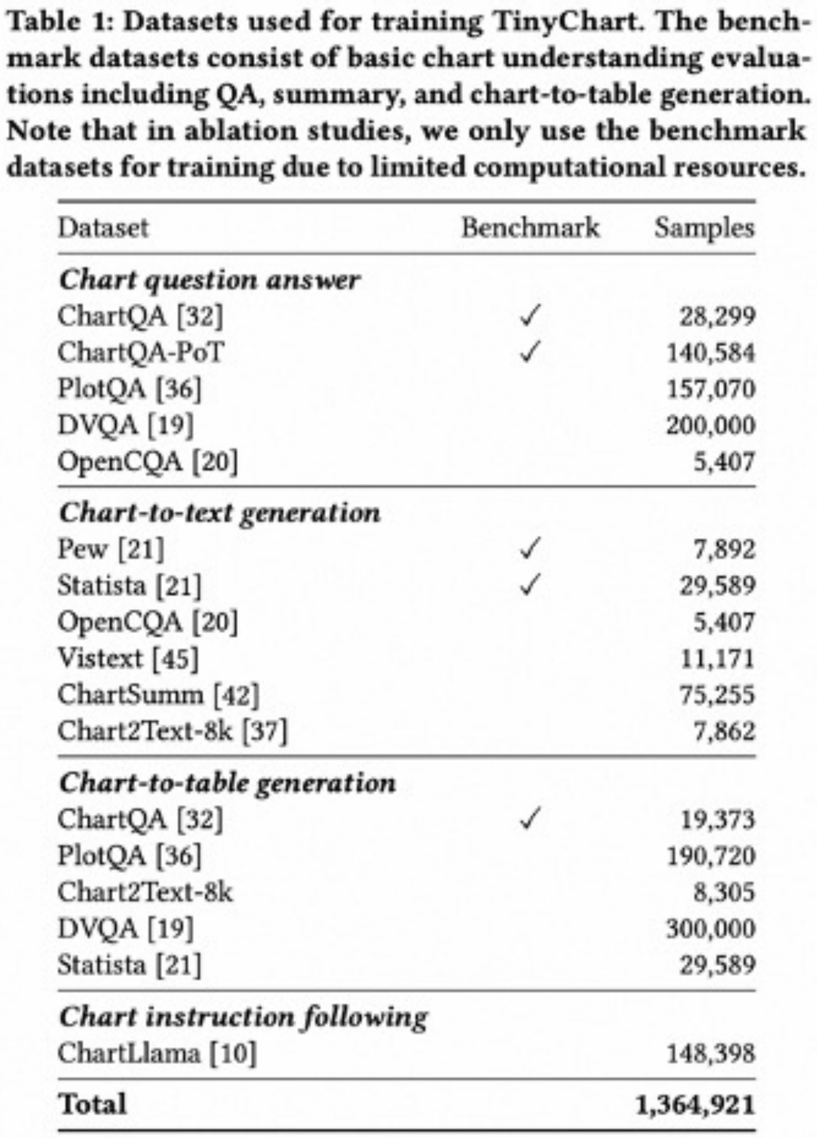


Table 1: Datasets used for training TinyChart. The benchmark datasets consist of basic chart understanding evaluations including QA, summary, and chart-to-table generation. Note that in ablation studies, we only use the benchmark datasets for training due to limited computational resources.

Dataset	Benchmark	Samples
Chart question answer		
ChartQA [32]	✓	28,299
ChartQA-PoT	✓	140,584
PlotQA [36]		157,070
DVQA [19]		200,000
OpenCQA [20]		5,407
Chart-to-text generation		
Pew [21]	✓	7,892
Statista [21]	✓	29,589
OpenCQA [20]		5,407
Vistext [45]		11,171
ChartSumm [42]		75,255
Chart2Text-8k [37]		7,862
Chart-to-table generation		
ChartQA [32]	✓	19,373
PlotQA [36]		190,720
Chart2Text-8k		8,305
DVQA [19]		300,000
Statista [21]		29,589
Chart instruction following		
ChartLlama [10]		148,398
Total		**1,364,921**

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

TinyChart

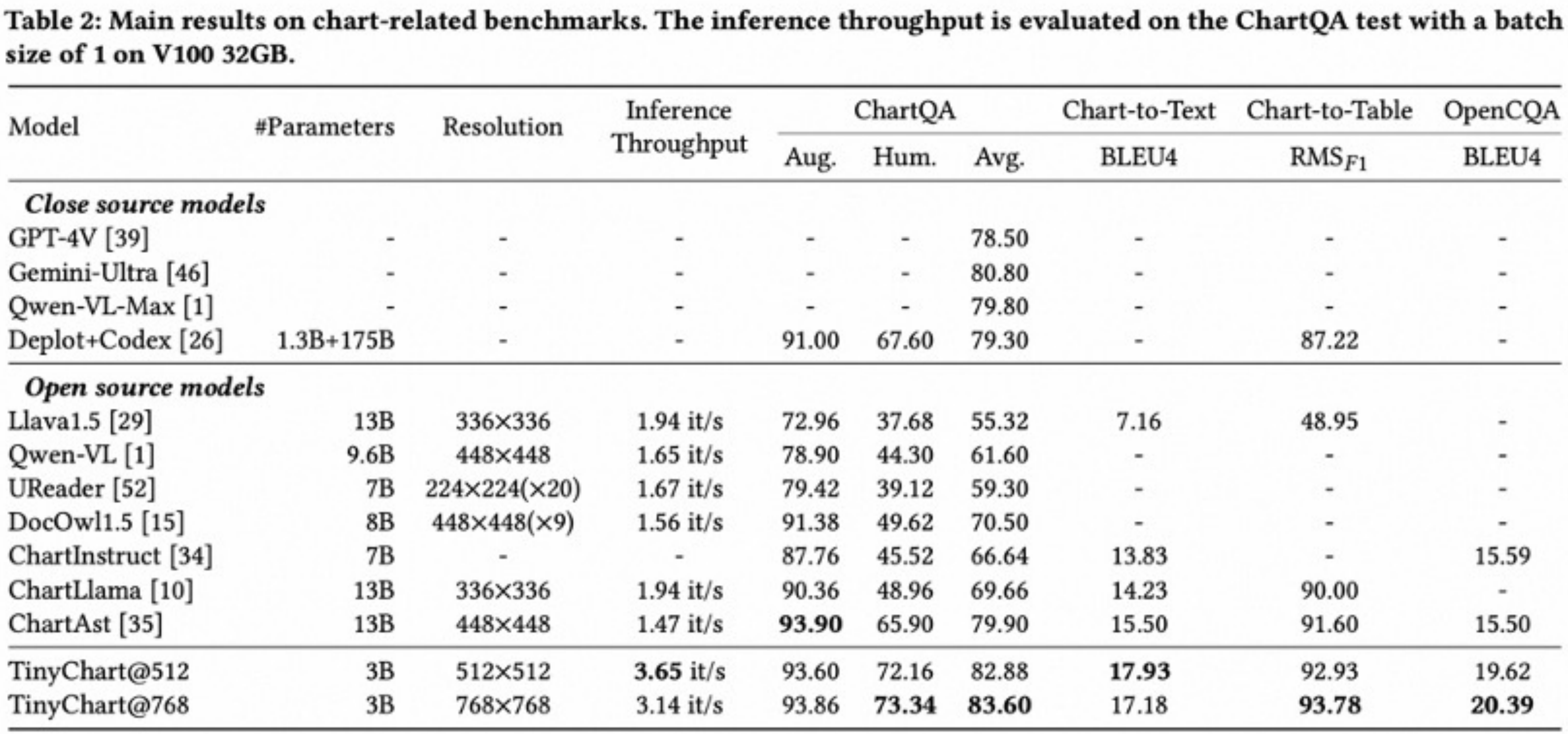
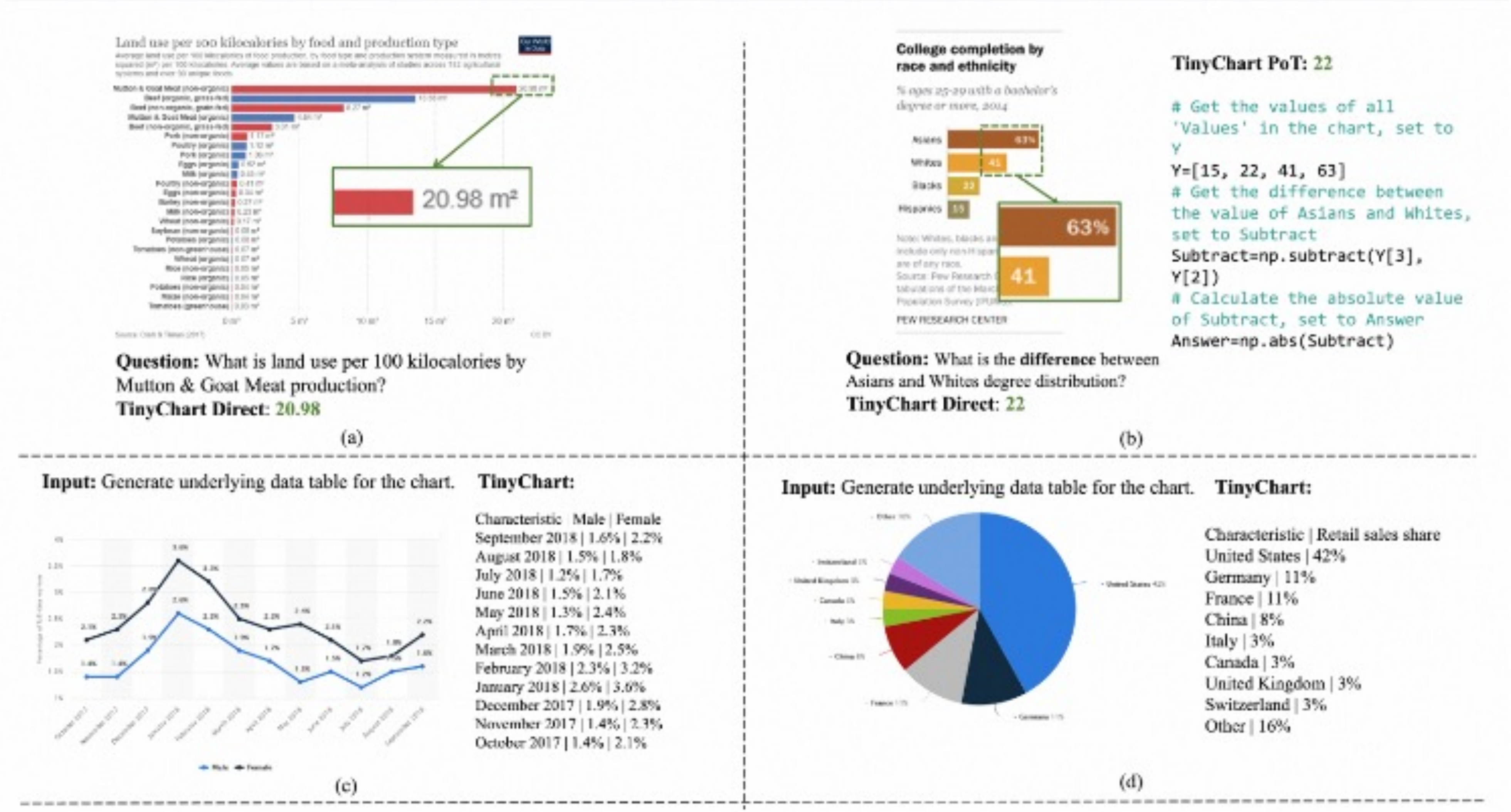


Table 2: Main results on chart-related benchmarks. The inference throughput is evaluated on the ChartQA test with a batch size of 1 on V100 32GB.

Model	#Parameters	Resolution	Inference Throughput	ChartQA			Chart-to-Text	Chart-to-Table	OpenCQA
				Aug.	Hum.	Avg.	BLEU4	RMS_{F1}	BLEU4
Close source models									
GPT-4V [39]	-	-	-	-	-	78.50	-	-	-
Gemini-Ultra [46]	-	-	-	-	-	80.80	-	-	-
Qwen-VL-Max [1]	-	-	-	-	-	79.80	-	-	-
Deplot+Codex [26]	1.3B+175B	-	-	91.00	67.60	79.30	-	87.22	-
Open source models									
Llava1.5 [29]	13B	336×336	1.94 it/s	72.96	37.68	55.32	7.16	48.95	-
Qwen-VL [1]	9.6B	448×448	1.65 it/s	78.90	44.30	61.60	-	-	-
UReader [52]	7B	224×224(×20)	1.67 it/s	79.42	39.12	59.30	-	-	-
DocOwl1.5 [15]	8B	448×448(×9)	1.56 it/s	91.38	49.62	70.50	-	-	-
ChartInstruct [34]	7B	-	-	87.76	45.52	66.64	13.83	-	15.59
ChartLlama [10]	13B	336×336	1.94 it/s	90.36	48.96	69.66	14.23	90.00	-
ChartAst [35]	13B	448×448	1.47 it/s	**93.90**	65.90	79.90	15.50	91.60	15.50
TinyChart@512	3B	512×512	**3.65 it/s**	93.60	72.16	82.88	**17.93**	92.93	19.62
TinyChart@768	3B	768×768	3.14 it/s	93.86	**73.34**	**83.60**	17.18	**93.78**	**20.39**

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

TinyChart



Question: What is land use per 100 kilocalories by Mutton & Goat Meat production?
TinyChart Direct: 20.98

(a)

Question: What is the **difference** between Asians and Whites degree distribution?
TinyChart Direct: 22

TinyChart PoT: 22

```
# Get the values of all
'Values' in the chart, set to
Y
Y=[15, 22, 41, 63]
# Get the difference between
the value of Asians and Whites,
set to Subtract
Subtract=np.subtract(Y[3],
Y[2])
# Calculate the absolute value
of Subtract, set to Answer
Answer=np.abs(Subtract)
```

(b)

Input: Generate underlying data table for the chart. **TinyChart:**

```
Characteristic  Male | Female
September 2018 | 1.6% | 2.2%
August 2018 | 1.5% | 1.8%
July 2018 | 1.2% | 1.7%
June 2018 | 1.5% | 2.1%
May 2018 | 1.3% | 2.4%
April 2018 | 1.7% | 2.3%
March 2018 | 1.9% | 2.5%
February 2018 | 2.3% | 3.2%
January 2018 | 2.6% | 3.6%
December 2017 | 1.9% | 2.8%
November 2017 | 1.4% | 2.3%
October 2017 | 1.4% | 2.1%
```

(c)

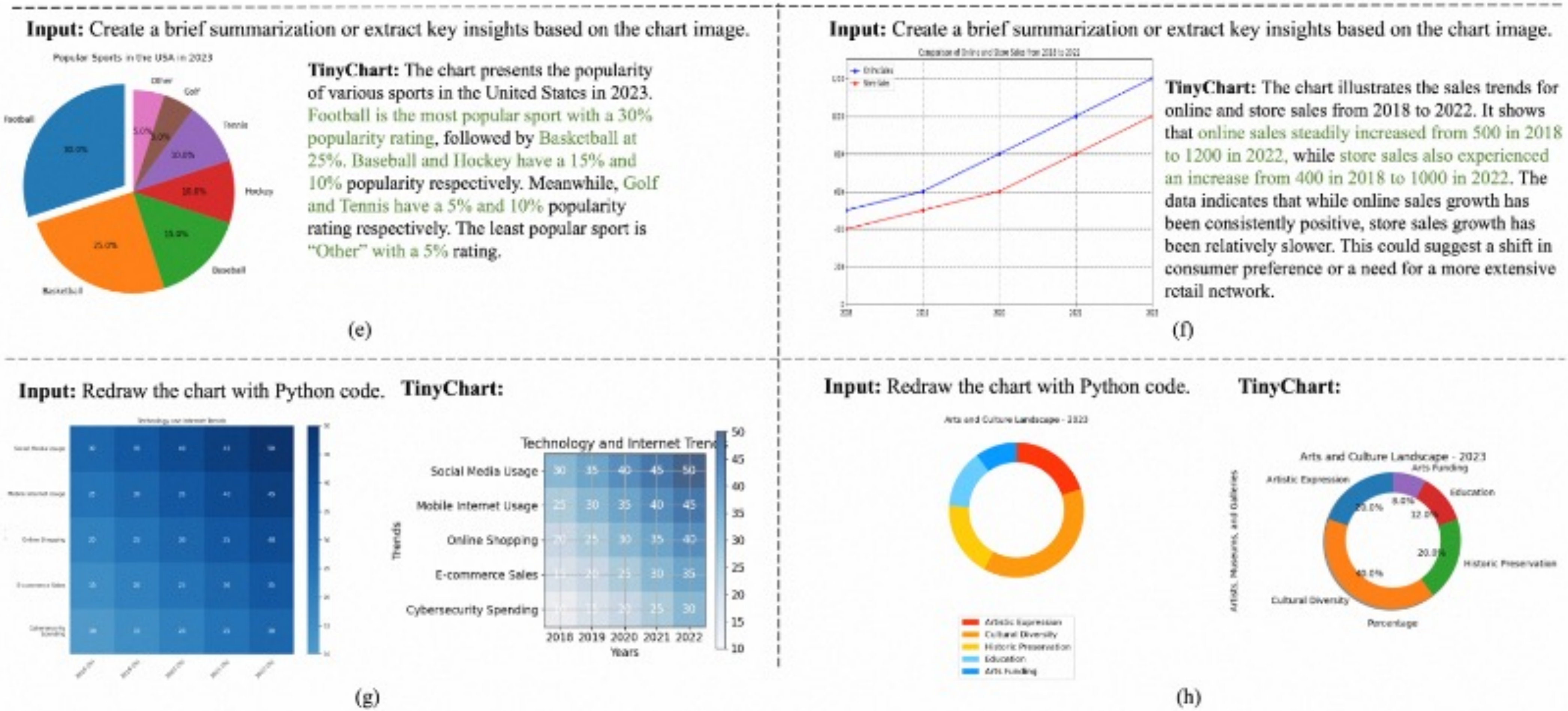
Input: Generate underlying data table for the chart. **TinyChart:**

```
Characteristic | Retail sales share
United States | 42%
Germany | 11%
France | 11%
China | 8%
Italy | 3%
Canada | 3%
United Kingdom | 3%
Switzerland | 3%
Other | 16%
```

(d)

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

TinyChart



Input: Create a brief summarization or extract key insights based on the chart image.

TinyChart: The chart presents the popularity of various sports in the United States in 2023. Football is the most popular sport with a 30% popularity rating, followed by Basketball at 25%. Baseball and Hockey have a 15% and 10% popularity respectively. Meanwhile, Golf and Tennis have a 5% and 10% popularity rating respectively. The least popular sport is "Other" with a 5% rating.

(e)

Input: Create a brief summarization or extract key insights based on the chart image.

TinyChart: The chart illustrates the sales trends for online and store sales from 2018 to 2022. It shows that online sales steadily increased from 500 in 2018 to 1200 in 2022, while store sales also experienced an increase from 400 in 2018 to 1000 in 2022. The data indicates that while online sales growth has been consistently positive, store sales growth has been relatively slower. This could suggest a shift in consumer preference or a need for a more extensive retail network.

(f)

Input: Redraw the chart with Python code. **TinyChart:**

(g)

Input: Redraw the chart with Python code. **TinyChart:**

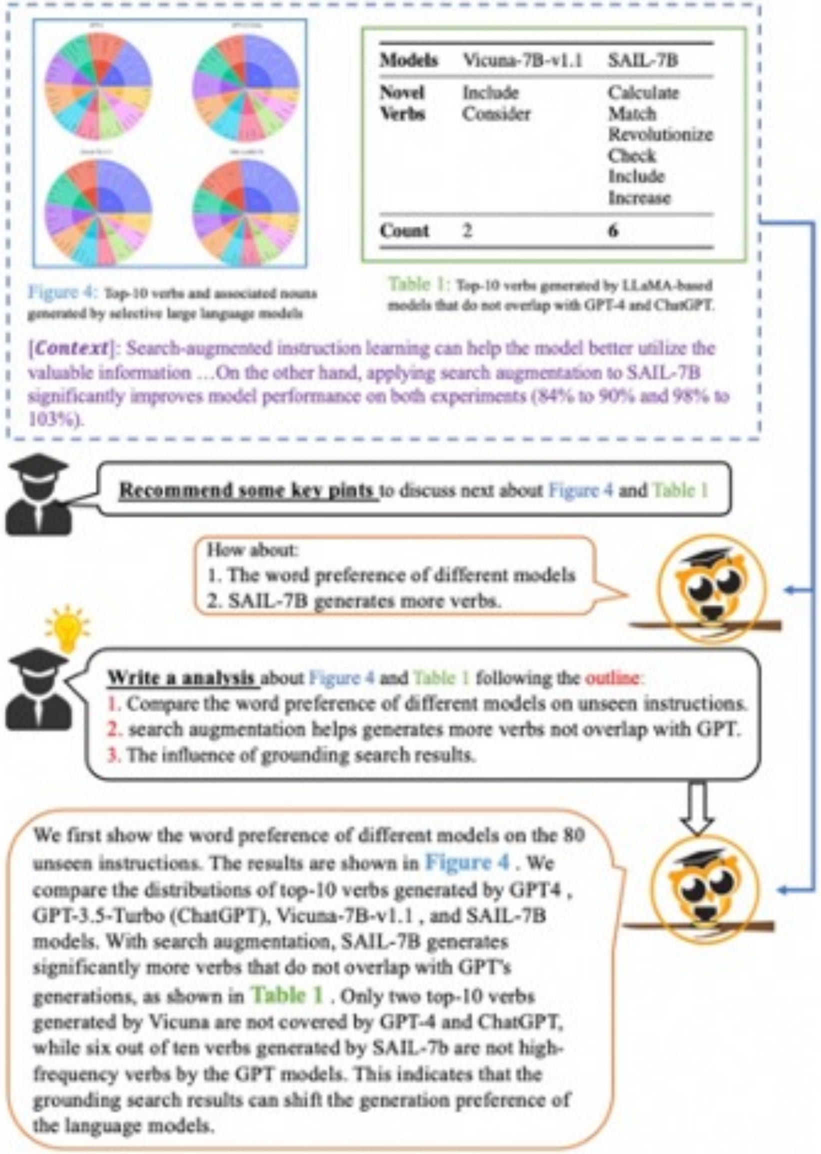
(h)

Zhang L, Hu A, Xu H, et al. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning[J]. arXiv preprint arXiv:2404.16635, 2024.

mPLUG-PaperOwl

论文图表理解的主要挑战：

1. 需要专业领域知识的辅助
2. 需要具备多图表关联分析能力



Models	Vicuna-7B-v1.1	SAIL-7B
Novel Verbs	Include Consider	Calculate Match Revolutionize Check Include Increase
Count	2	6

Figure 4: Top-10 verbs and associated nouns generated by selective large language models

Table 1: Top-10 verbs generated by LLaMA-based models that do not overlap with GPT-4 and ChatGPT.

[*Context*]: Search-augmented instruction learning can help the model better utilize the valuable information …On the other hand, applying search augmentation to SAIL-7B significantly improves model performance on both experiments (84% to 90% and 98% to 103%).

Recommend some key pints to discuss next about Figure 4 and Table 1

How about:
1. The word preference of different models
2. SAIL-7B generates more verbs.

Write a analysis about Figure 4 and Table 1 following the outline:
1. Compare the word preference of different models on unseen instructions.
2. search augmentation helps generates more verbs not overlap with GPT.
3. The influence of grounding search results.

We first show the word preference of different models on the 80 unseen instructions. The results are shown in Figure 4 . We compare the distributions of top-10 verbs generated by GPT4 , GPT-3.5-Turbo (ChatGPT), Vicuna-7B-v1.1 , and SAIL-7B models. With search augmentation, SAIL-7B generates significantly more verbs that do not overlap with GPT's generations, as shown in Table 1 . Only two top-10 verbs generated by Vicuna are not covered by GPT-4 and ChatGPT, while six out of ten verbs generated by SAIL-7b are not high-frequency verbs by the GPT models. This indicates that the grounding search results can shift the generation preference of the language models.

A Hu, Y Shi, H Xu,et al. mplug-paperowl: Scientific diagram analysis with the multimodal large language model[J] arXiv preprint arXiv:2311.18248 62023

mPLUG-PaperOwl

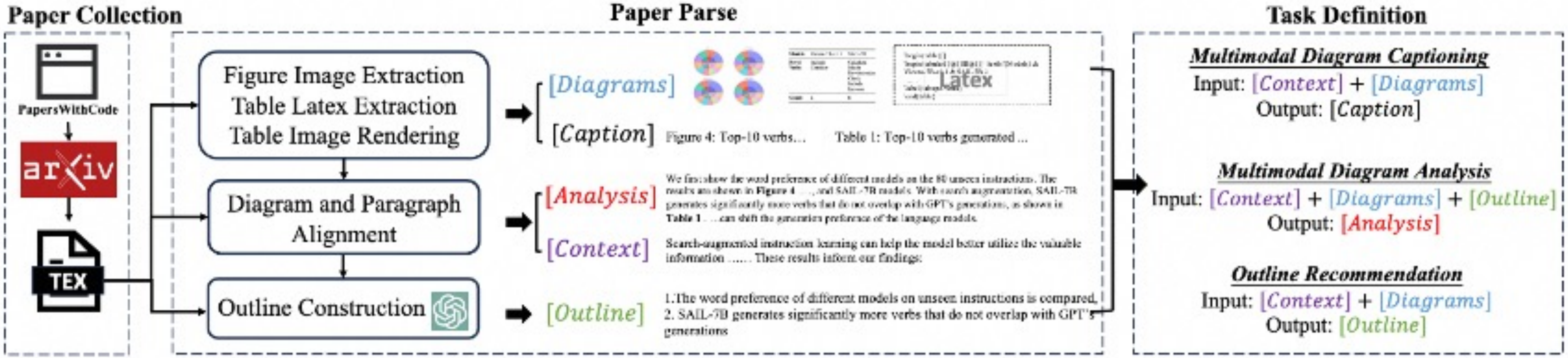

Paper Collection
PapersWithCode
arXiv
TEX
Paper Parse
Figure Image Extraction
Table Latex Extraction
Table Image Rendering
Diagram and Paragraph Alignment
Outline Construction
[Diagrams]
[Caption]
Latex
Figure 4: Top-10 verbs...
Table 1: Top-10 verbs generated ...
[Analysis]
[Context]
[Outline]
1.The word preference of different models on unseen instructions is compared.
2. SAIL-7B generates significantly more verbs that do not overlap with GPT's generations
Task Definition
Multimodal Diagram Captioning
Input: [Context] + [Diagrams]
Output: [Caption]
Multimodal Diagram Analysis
Input: [Context] + [Diagrams] + [Outline]
Output: [Analysis]
Outline Recommendation
Input: [Context] + [Diagrams]
Output: [Outline]

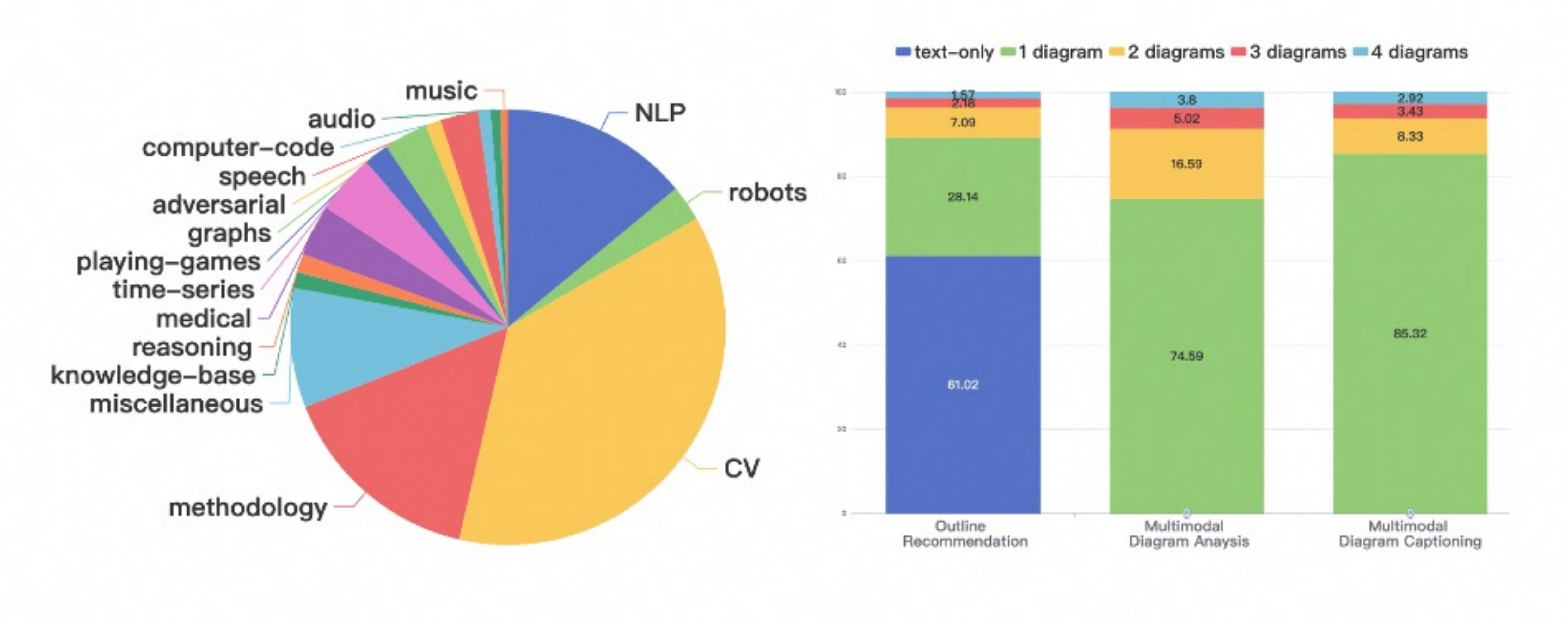
Figure 2. The pipeline of M-Paper construction and definition of our three tasks.

- 提供上下文(context)作为专业知识来源，

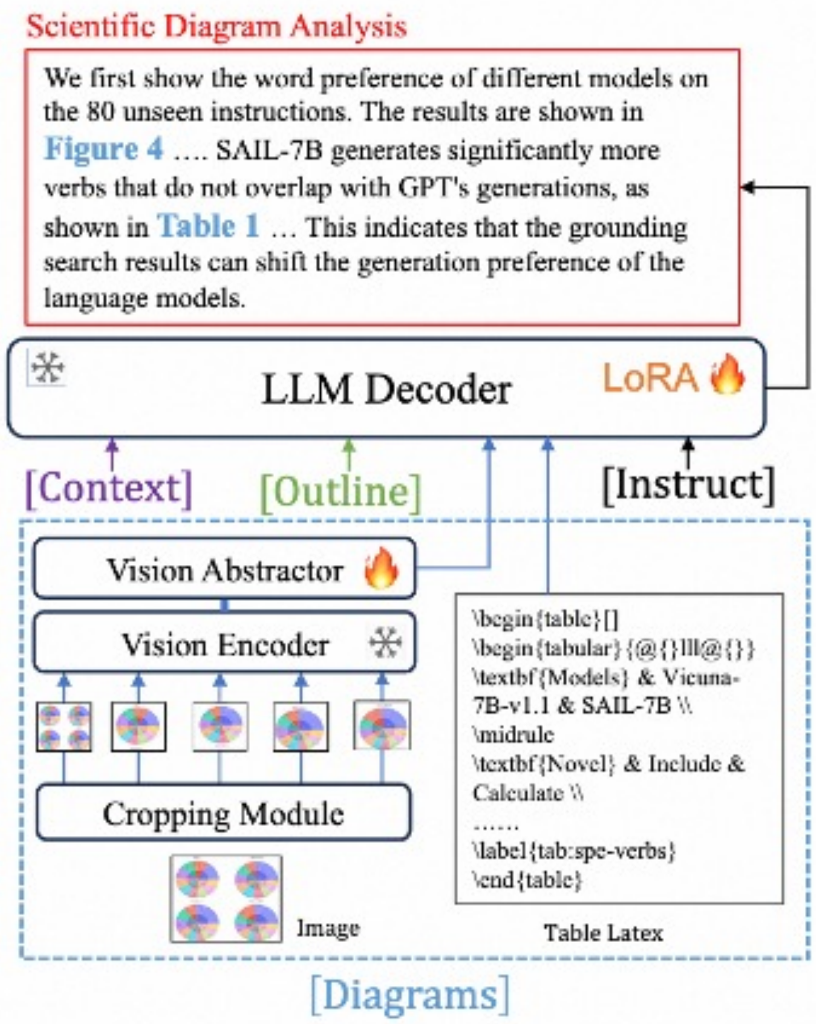
* 基于GPT3.5, 构建“要点”(outline) 作为图表分析的控制信号

M-Paper: 48k papers, 447k diagrams, 702k training samples

mPLUG-PaperOwl



mPLUG-PaperOwl

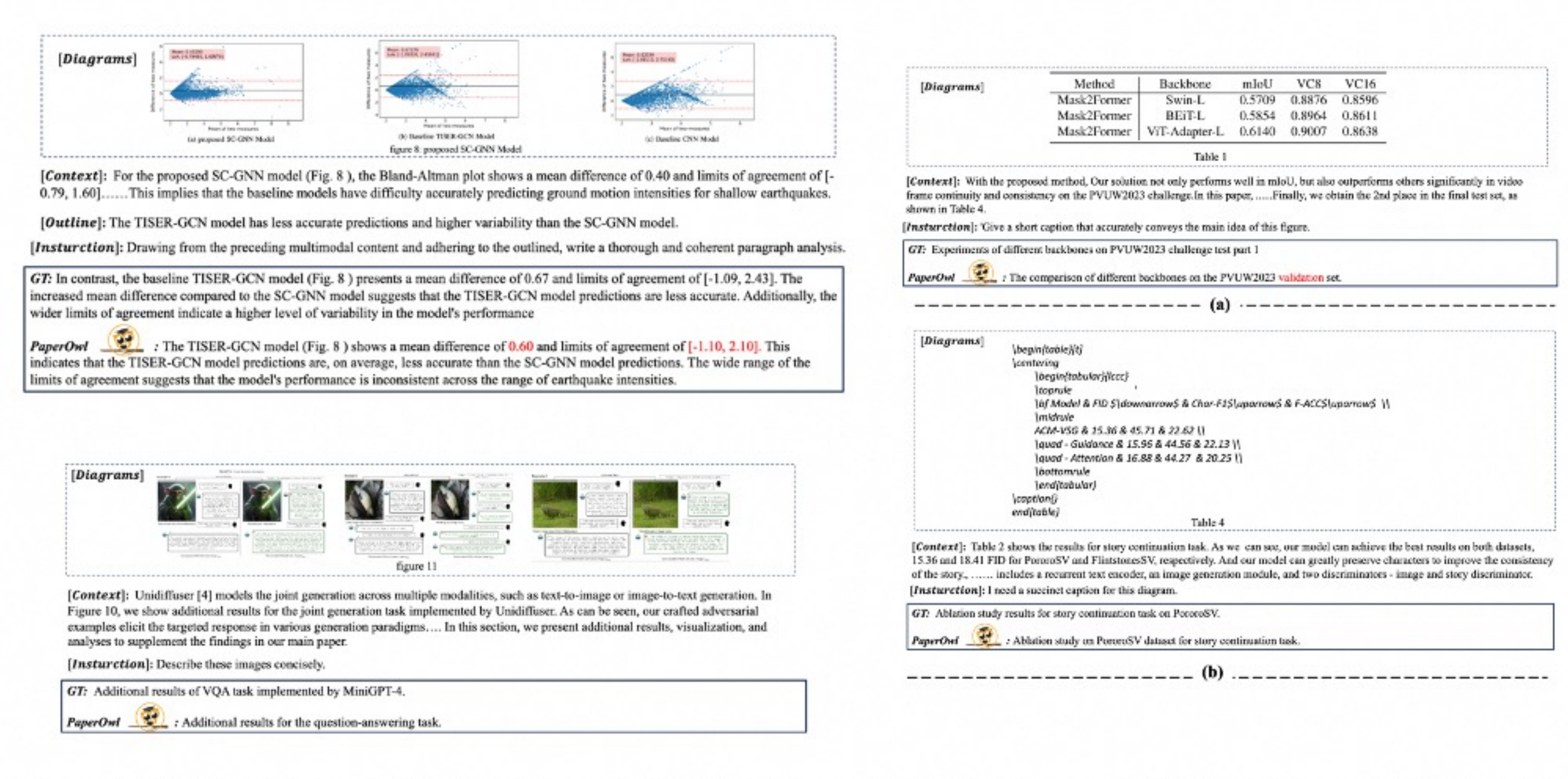


We first show the word preference of different models on the 80 unseen instructions. The results are shown in Figure 4 …. SAIL-7B generates significantly more verbs that do not overlap with GPT's generations, as shown in Table 1 … This indicates that the grounding search results can shift the generation preference of the language models.

```
\begin{table}[]
\begin{tabular}{@{}lll@{}}
\textbf{Models} & Vicuna-
7B-v1.1 & SAIL-7B \\
\midrule
\textbf{Novel} & Include &
Calculate \\
......
\label{tab:spe-verbs}
\end{table}
```

Figure 6: The overall architecture of PaperOwl.

mPLUG-PaperOwl



[*Diagrams*]

(a) proposed SC-GNN Model
(b) Baseline TISER-GCN Model
(c) Baseline CNN Model
figure 8: proposed SC-GNN Model

[***Context***]: For the proposed SC-GNN model (Fig. 8), the Bland-Altman plot shows a mean difference of 0.40 and limits of agreement of [-0.79, 1.60]……This implies that the baseline models have difficulty accurately predicting ground motion intensities for shallow earthquakes.

[***Outline***]: The TISER-GCN model has less accurate predictions and higher variability than the SC-GNN model.

[***Insturction***]: Drawing from the preceding multimodal content and adhering to the outlined, write a thorough and coherent paragraph analysis.

GT: In contrast, the baseline TISER-GCN model (Fig. 8) presents a mean difference of 0.67 and limits of agreement of [-1.09, 2.43]. The increased mean difference compared to the SC-GNN model suggests that the TISER-GCN model predictions are less accurate. Additionally, the wider limits of agreement indicate a higher level of variability in the model's performance

PaperOwl : The TISER-GCN model (Fig. 8) shows a mean difference of 0.60 and limits of agreement of [-1.10, 2.10]. This indicates that the TISER-GCN model predictions are, on average, less accurate than the SC-GNN model predictions. The wide range of the limits of agreement suggests that the model's performance is inconsistent across the range of earthquake intensities.

[*Diagrams*]

figure 11

[***Context***]: Unidiffuser [4] models the joint generation across multiple modalities, such as text-to-image or image-to-text generation. In Figure 10, we show additional results for the joint generation task implemented by Unidiffuser. As can be seen, our crafted adversarial examples elicit the targeted response in various generation paradigms…. In this section, we present additional results, visualization, and analyses to supplement the findings in our main paper.

[***Insturction***]: Describe these images concisely.

GT: Additional results of VQA task implemented by MiniGPT-4.

PaperOwl : Additional results for the question-answering task.

[*Diagrams*]

Method	Backbone	mIoU	VC8	VC16
Mask2Former	Swin-L	0.5709	0.8876	0.8596
Mask2Former	BEiT-L	0.5854	0.8964	0.8611
Mask2Former	ViT-Adapter-L	0.6140	0.9007	0.8638

Table 1

[***Context***]: With the proposed method, Our solution not only performs well in mIoU, but also outperforms others significantly in video frame continuity and consistency on the PVUW2023 challenge.In this paper, ……Finally, we obtain the 2nd place in the final test set, as shown in Table 4.

[***Insturction***]: 'Give a short caption that accurately conveys the main idea of this figure.

GT: Experiments of different backbones on PVUW2023 challenge test part 1

PaperOwl : The comparison of different backbones on the PVUW2023 validation set.

(a)

[*Diagrams*]

```
\begin{table}{t}
\centering
    \begin{tabular}{lccc}
    \toprule
    \bf Model & FID $\downarrow$ & Char-F1$\uparrow$ & F-ACC$\uparrow$  \\
    \midrule
    ACM-VSG & 15.36 & 45.71 & 22.62 \\
    \quad - Guidance & 15.96 & 44.56 & 22.13 \\
    \quad - Attention & 16.88 & 44.27  & 20.25 \\
    \bottomrule
    \end{tabular}
\caption{}
end{table}
```

Table 4

[***Context***]: Table 2 shows the results for story continuation task. As we can see, our model can achieve the best results on both datasets, 15.36 and 18.41 FID for PororoSV and FlintstonesSV, respectively. And our model can greatly preserve characters to improve the consistency of the story., …… includes a recurrent text encoder, an image generation module, and two discriminators - image and story discriminator.

[***Insturction***]: I need a succinct caption for this diagram.

GT: Ablation study results for story continuation task on PororoSV.

PaperOwl : Ablation study on PororoSV dataset for story continuation task.

(b)

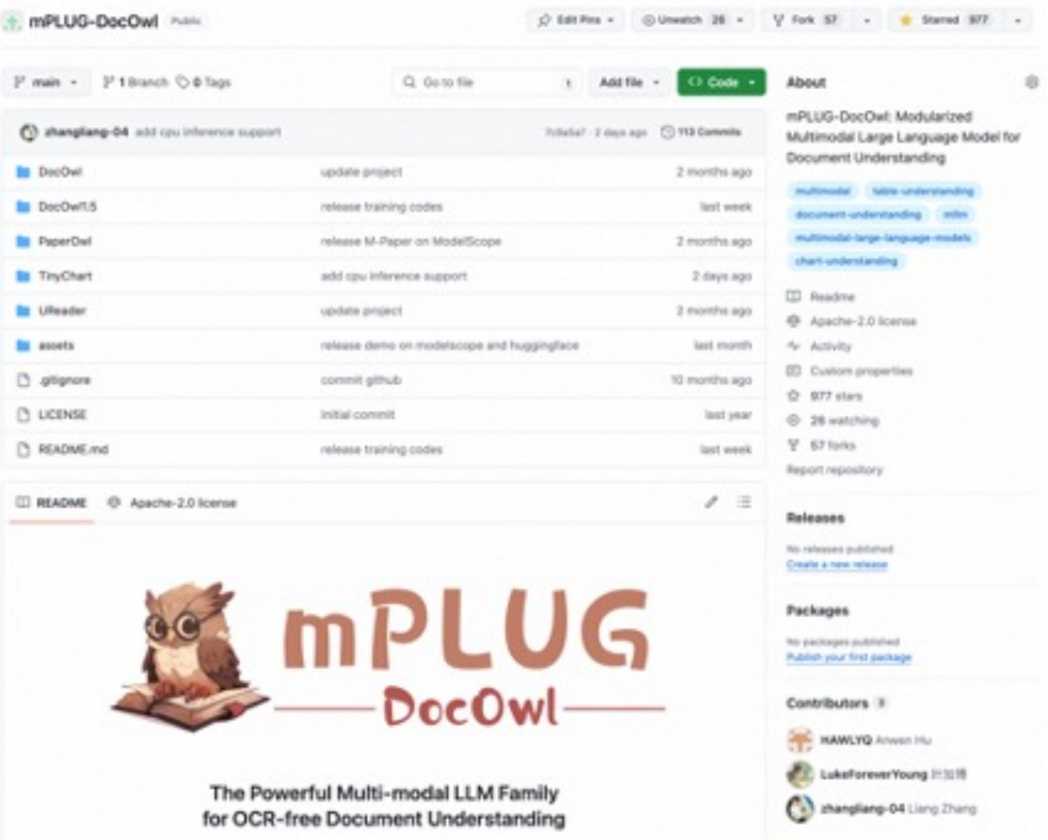
PART 03

mPLUG-DocOwl开源实战

mPLUG-DocOwl开源

数据和模型在ModelScope和HuggingFace上均可下载

https://github.com/X-PLUG/mPLUG-DocOwl



zhangliang-04 add cpu inference support · 7c8a5a7 · 2 days ago · 113 Commits

DocOwl	update project	2 months ago
DocOwl1.5	release training codes	last week
PaperOwl	release M-Paper on ModelScope	2 months ago
TinyChart	add cpu inference support	2 days ago
UReader	update project	2 months ago
assets	release demo on modelscope and huggingface	last month
.gitignore	commit github	10 months ago
LICENSE	initial commit	last year
README.md	release training codes	last week

The Powerful Multi-modal LLM Family
for OCR-free Document Understanding

About

mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding

multimodal · table-understanding · document-understanding · mllm · multimodal-large-language-models · chart-understanding

Readme
Apache-2.0 license
Activity
Custom properties
977 stars
26 watching
57 forks
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 3

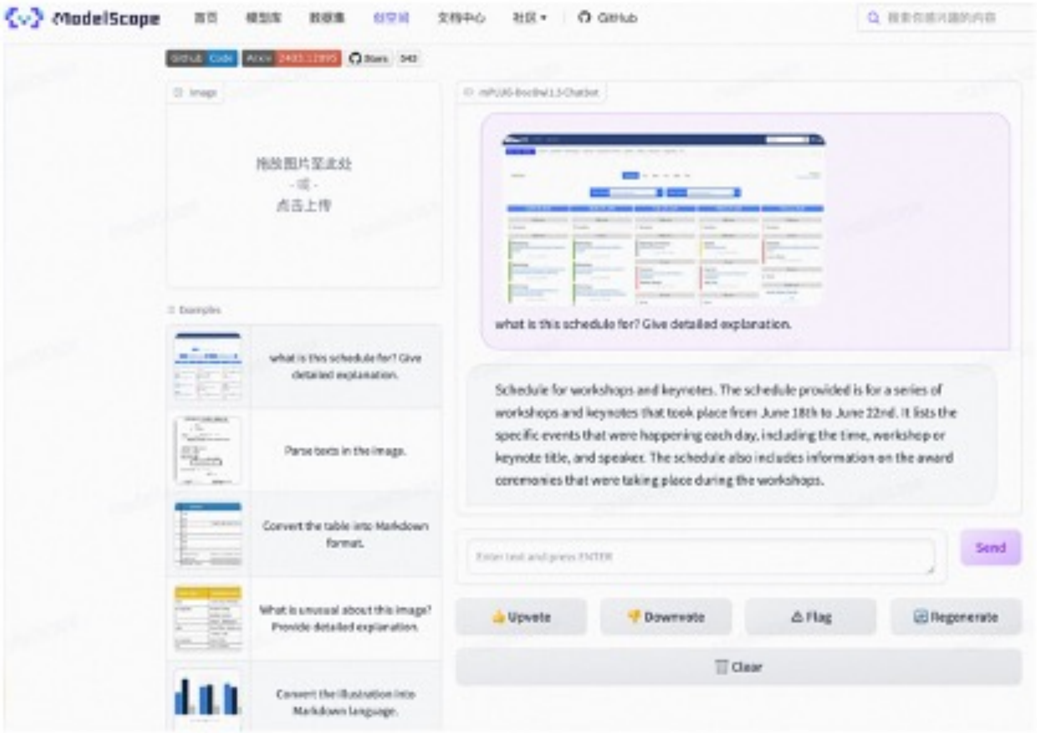
HAWLYQ Anwen Hu
LukeForeverYoung 叶加博
zhangliang-04 Liang Zhang

mPLUG-DocOwl Demo

https://huggingface.co/spaces/mPLUG/DocOwl

https://modelscope.cn/studios/iic/mPLUG-DocOwl

https://huggingface.co/spaces/mPLUG/TinyChart-3B



what is this schedule for? Give detailed explanation.

Parse texts in the image.

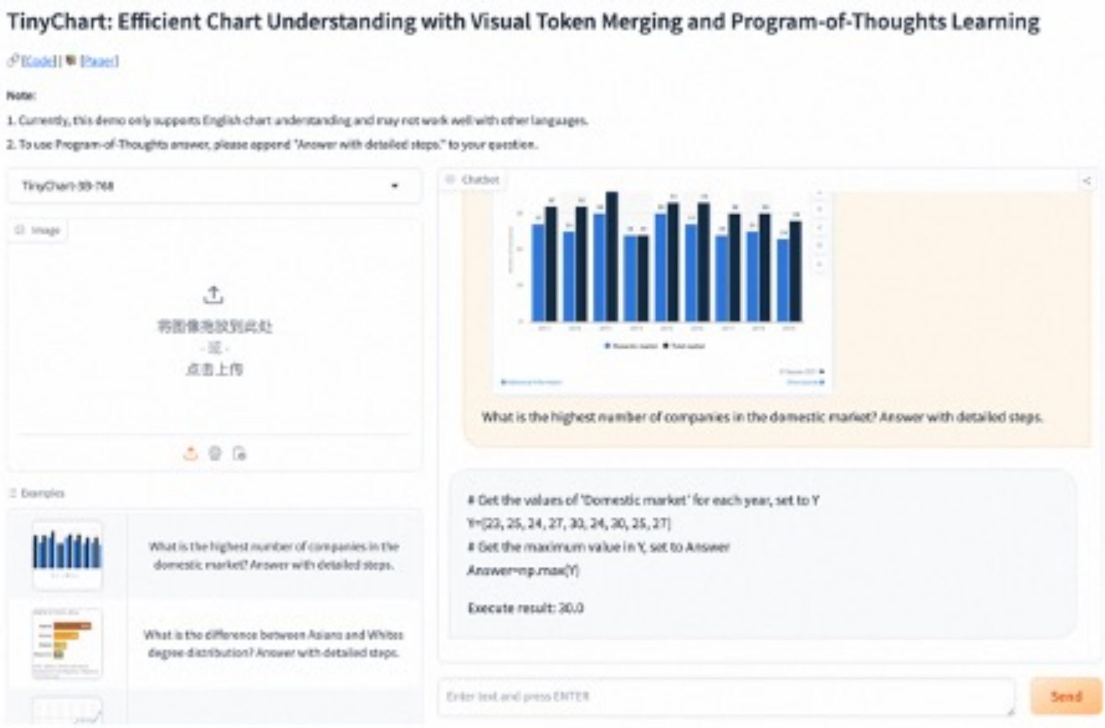
Convert the table into Markdown format.

What is unusual about this image? Provide detailed explanation.

Convert the illustration into Markdown language.

what is this schedule for? Give detailed explanation.

Schedule for workshops and keynotes. The schedule provided is for a series of workshops and keynotes that took place from June 18th to June 22nd. It lists the specific events that were happening each day, including the time, workshop or keynote title, and speaker. The schedule also includes information on the award ceremonies that were taking place during the workshops.



TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning

[Code] | [Paper]

Note:

1. Currently, this demo only supports English chart understanding and may not work well with other languages.
2. To use Program-of-Thoughts answer, please append "Answer with detailed steps." to your question.

What is the highest number of companies in the domestic market? Answer with detailed steps.

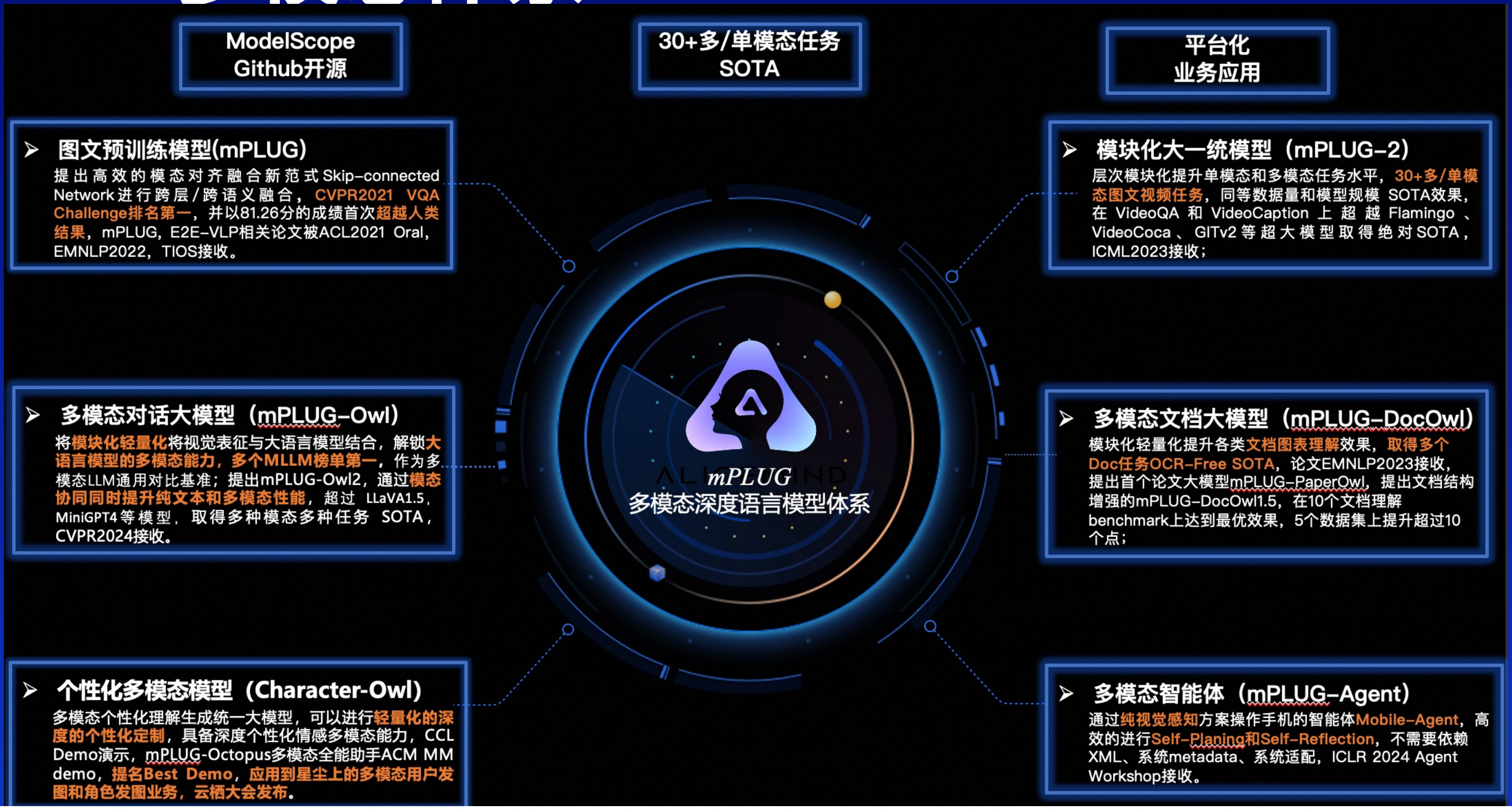
What is the difference between Asians and Whites degree distribution? Answer with detailed steps.

What is the highest number of companies in the domestic market? Answer with detailed steps.

```
# Get the values of 'Domestic market' for each year, set to Y
Y=[22, 25, 24, 27, 30, 24, 30, 25, 27]
# Get the maximum value in Y, set to Answer
Answer=np.max(Y)

Execute result: 30.0
```

通义mPLUG多模态体系



ModelScope Github开源

30+多/单模态任务 SOTA

平台化 业务应用

- **图文预训练模型(mPLUG)**
  提出高效的模态对齐融合新范式Skip-connected Network进行跨层/跨语义融合，**CVPR2021 VQA Challenge排名第一**，并以81.26分的成绩首次**超越人类结果**，mPLUG，E2E-VLP相关论文被ACL2021 Oral，EMNLP2022，TIOS接收。

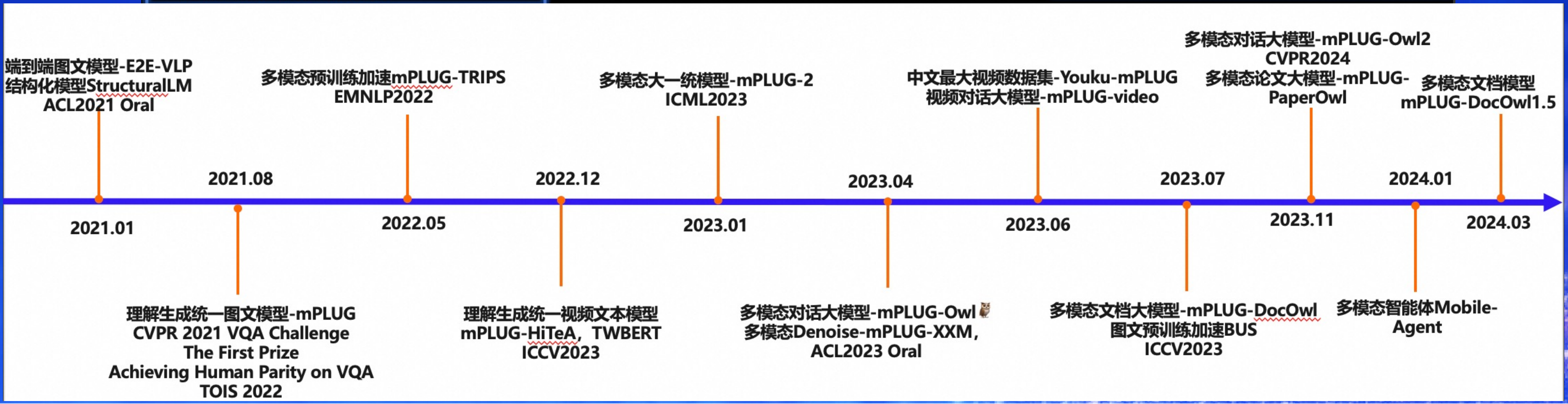
- **多模态对话大模型（mPLUG-Owl）**
  将**模块化轻量化**将视觉表征与大语言模型结合，解锁**大语言模型的多模态能力，多个MLLM榜单第一**，作为多模态LLM通用对比基准；提出mPLUG-Owl2，通过**模态协同同时提升纯文本和多模态性能**，超过 LLaVA1.5，MiniGPT4等模型，取得多种模态多种任务 SOTA，CVPR2024接收。

- **个性化多模态模型（Character-Owl）**
  多模态个性化理解生成统一大模型，可以进行**轻量化的深度的个性化定制**，具备深度个性化情感多模态能力，CCL Demo演示，mPLUG-Octopus多模态全能助手ACM MM demo，**提名Best Demo**，**应用到星尘上的多模态用户发图和角色发图业务，云栖大会发布**。

- **模块化大一统模型（mPLUG-2）**
  层次模块化提升单模态和多模态任务水平，**30+多/单模态图文视频任务**，同等数据量和模型规模 SOTA效果，在 VideoQA 和 VideoCaption 上超越 Flamingo、VideoCoca、GITv2等超大模型取得绝对SOTA，ICML2023接收；

- **多模态文档大模型（mPLUG-DocOwl）**
  模块化轻量化提升各类**文档图表理解**效果，**取得多个Doc任务OCR-Free SOTA**，论文EMNLP2023接收，提出首个论文大模型mPLUG-PaperOwl，提出文档结构增强的mPLUG-DocOwl1.5，在10个文档理解benchmark上达到最优效果，5个数据集上提升超过10个点；

- **多模态智能体（mPLUG-Agent）**
  通过**纯视觉感知**方案操作手机的智能体**Mobile-Agent**，高效的进行**Self-Planing和Self-Reflection**，不需要依赖XML、系统metadata、系统适配，ICLR 2024 Agent Workshop接收。



2021.01
端到端图文模型-E2E-VLP
结构化模型StructuralLM
ACL2021 Oral

2021.08
理解生成统一图文模型-mPLUG
CVPR 2021 VQA Challenge
The First Prize
Achieving Human Parity on VQA
TOIS 2022

2022.05
多模态预训练加速mPLUG-TRIPS
EMNLP2022

2022.12
理解生成统一视频文本模型
mPLUG-HiTeA，TWBERT
ICCV2023

2023.01
多模态大一统模型-mPLUG-2
ICML2023

2023.04
多模态对话大模型-mPLUG-Owl🦉
多模态Denoise-mPLUG-XXM，
ACL2023 Oral

2023.06
中文最大视频数据集-Youku-mPLUG
视频对话大模型-mPLUG-video

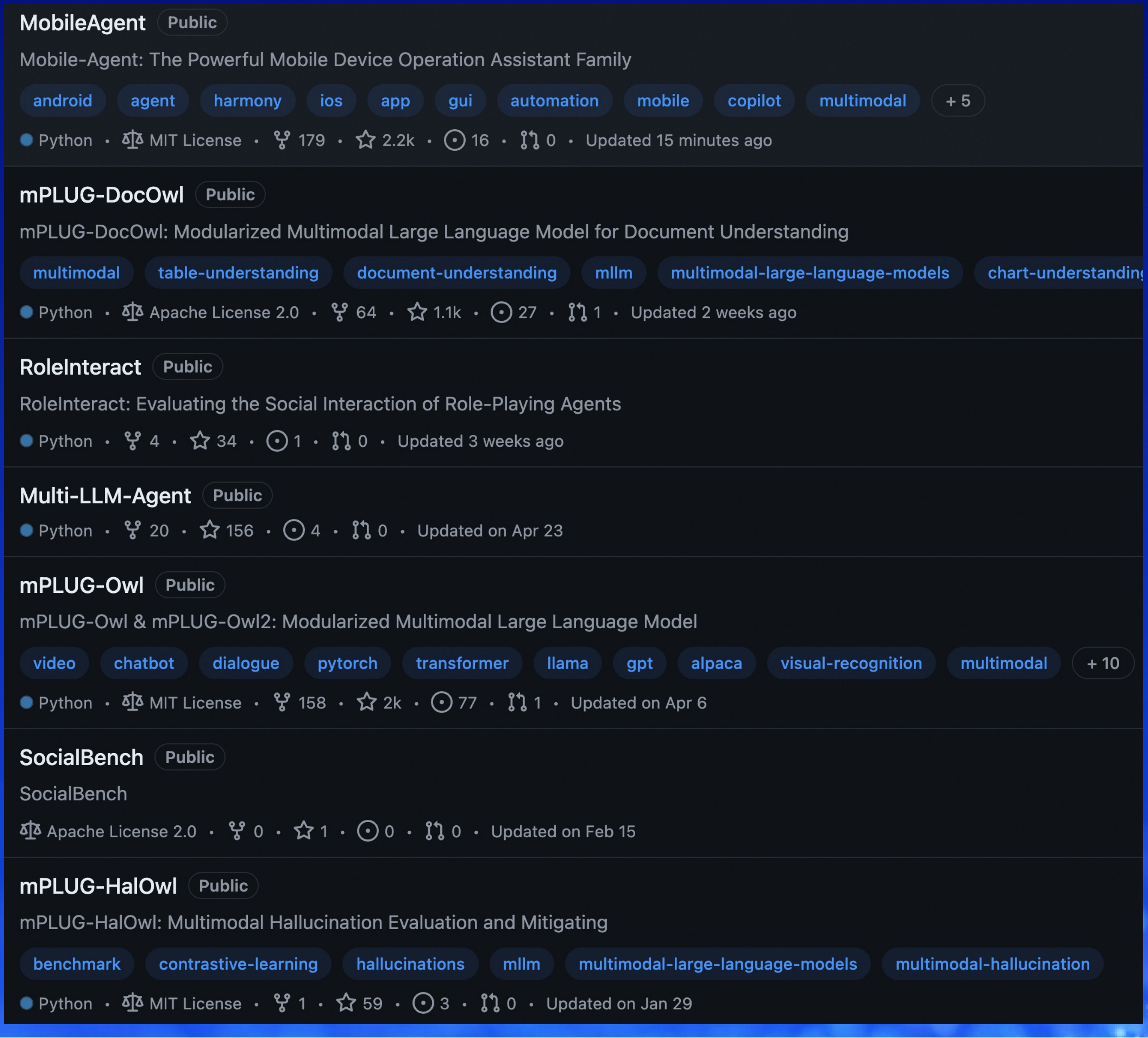
2023.07
多模态文档大模型-mPLUG-DocOwl
图文预训练加速BUS
ICCV2023

2023.11
多模态对话大模型-mPLUG-Owl2
CVPR2024
多模态论文大模型-mPLUG-PaperOwl

2024.01
多模态智能体Mobile-Agent

2024.03
多模态文档模型
mPLUG-DocOwl1.5

通义mPLUG多模态体系



MobileAgent Public

Mobile-Agent: The Powerful Mobile Device Operation Assistant Family

android · agent · harmony · ios · app · gui · automation · mobile · copilot · multimodal · + 5

● Python · ⚖ MIT License · ⑂ 179 · ☆ 2.2k · ⊙ 16 · ⇅ 0 · Updated 15 minutes ago

mPLUG-DocOwl Public

mPLUG-DocOwl: Modularized Multimodal Large Language Model for Document Understanding

multimodal · table-understanding · document-understanding · mllm · multimodal-large-language-models · chart-understandin

● Python · ⚖ Apache License 2.0 · ⑂ 64 · ☆ 1.1k · ⊙ 27 · ⇅ 1 · Updated 2 weeks ago

RoleInteract Public

RoleInteract: Evaluating the Social Interaction of Role-Playing Agents

● Python · ⑂ 4 · ☆ 34 · ⊙ 1 · ⇅ 0 · Updated 3 weeks ago

Multi-LLM-Agent Public

● Python · ⑂ 20 · ☆ 156 · ⊙ 4 · ⇅ 0 · Updated on Apr 23

mPLUG-Owl Public

mPLUG-Owl & mPLUG-Owl2: Modularized Multimodal Large Language Model

video · chatbot · dialogue · pytorch · transformer · llama · gpt · alpaca · visual-recognition · multimodal · + 10

● Python · ⚖ MIT License · ⑂ 158 · ☆ 2k · ⊙ 77 · ⇅ 1 · Updated on Apr 6

SocialBench Public

SocialBench

⚖ Apache License 2.0 · ⑂ 0 · ☆ 1 · ⊙ 0 · ⇅ 0 · Updated on Feb 15

mPLUG-HalOwl Public

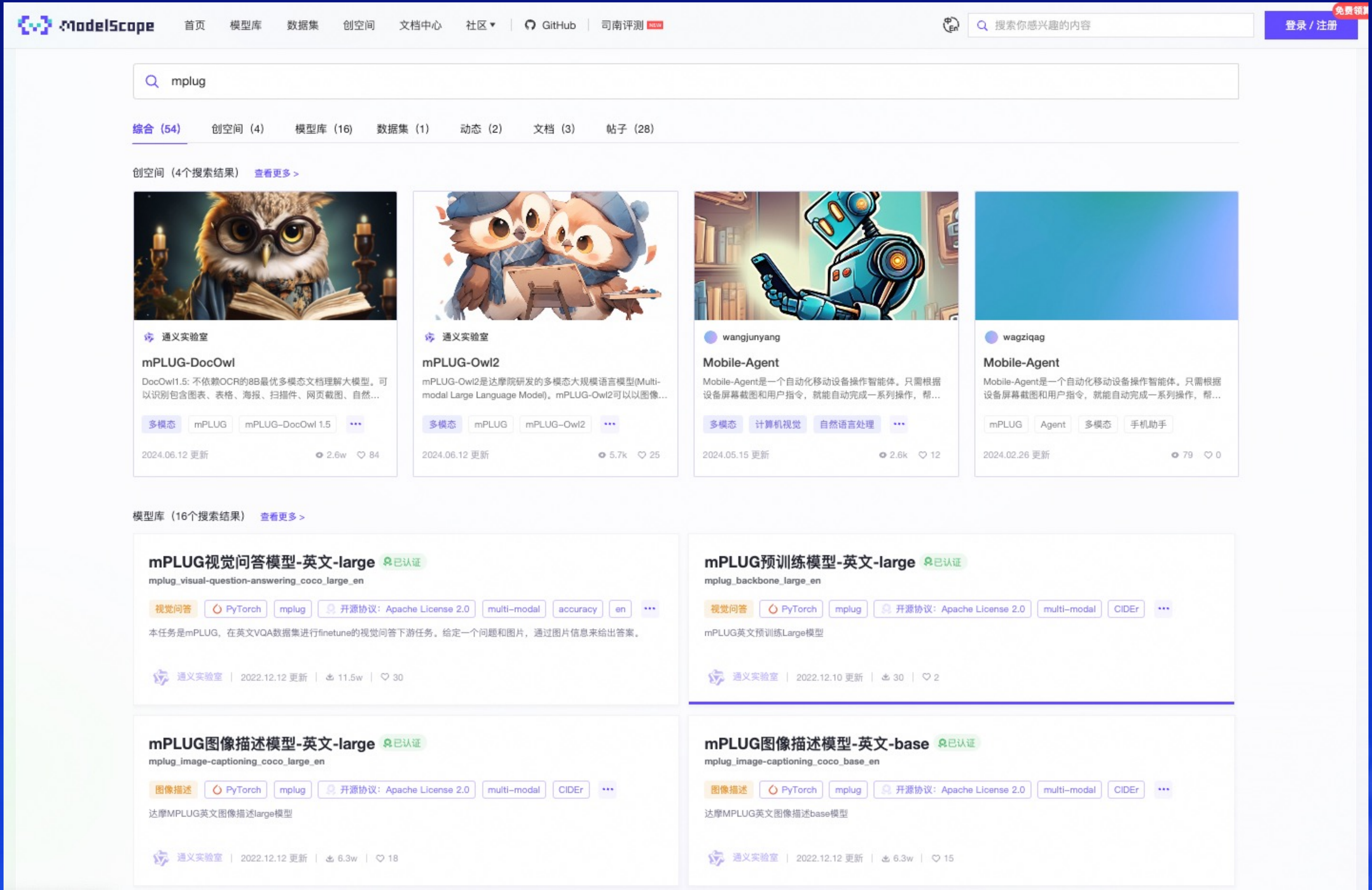
mPLUG-HalOwl: Multimodal Hallucination Evaluation and Mitigating

benchmark · contrastive-learning · hallucinations · mllm · multimodal-large-language-models · multimodal-hallucination

● Python · ⚖ MIT License · ⑂ 1 · ☆ 59 · ⊙ 3 · ⇅ 0 · Updated on Jan 29

https://github.com/orgs/X-PLUG/repositories

通义mPLUG多模态体系



创空间（4个搜索结果）

通义实验室
mPLUG-DocOwl
DocOwl1.5: 不依赖OCR的8B最优多模态文档理解大模型。可以识别包含图表、表格、海报、扫描件、网页截图、自然...
多模态 · mPLUG · mPLUG-DocOwl 1.5 · ···
2024.06.12 更新 · 2.6w · 84

通义实验室
mPLUG-Owl2
mPLUG-Owl2是达摩院研发的多模态大规模语言模型(Multi-modal Large Language Model)。mPLUG-Owl2可以以图像...
多模态 · mPLUG · mPLUG-Owl2 · ···
2024.06.12 更新 · 5.7k · 25

wangjunyang
Mobile-Agent
Mobile-Agent是一个自动化移动设备操作智能体。只需根据设备屏幕截图和用户指令，就能自动完成一系列操作，帮...
多模态 · 计算机视觉 · 自然语言处理 · ···
2024.05.15 更新 · 2.6k · 12

wagziqag
Mobile-Agent
Mobile-Agent是一个自动化移动设备操作智能体。只需根据设备屏幕截图和用户指令，就能自动完成一系列操作，帮...
mPLUG · Agent · 多模态 · 手机助手
2024.02.26 更新 · 79 · 0

模型库（16个搜索结果）

mPLUG视觉问答模型-英文-large 已认证
mplug_visual-question-answering_coco_large_en
视觉问答 · PyTorch · mplug · 开源协议：Apache License 2.0 · multi-modal · accuracy · en · ···
本任务是mPLUG，在英文VQA数据集进行finetune的视觉问答下游任务。给定一个问题和图片，通过图片信息来给出答案。
通义实验室 | 2022.12.12 更新 | 11.5w | 30

mPLUG预训练模型-英文-large 已认证
mplug_backbone_large_en
视觉问答 · PyTorch · mplug · 开源协议：Apache License 2.0 · multi-modal · CIDEr · ···
mPLUG英文预训练Large模型
通义实验室 | 2022.12.10 更新 | 30 | 2

mPLUG图像描述模型-英文-large 已认证
mplug_image-captioning_coco_large_en
图像描述 · PyTorch · mplug · 开源协议：Apache License 2.0 · multi-modal · CIDEr · ···
达摩MPLUG英文图像描述large模型
通义实验室 | 2022.12.12 更新 | 6.3w | 18

mPLUG图像描述模型-英文-base 已认证
mplug_image-captioning_coco_base_en
图像描述 · PyTorch · mplug · 开源协议：Apache License 2.0 · multi-modal · CIDEr · ···
达摩MPLUG英文图像描述base模型
通义实验室 | 2022.12.12 更新 | 6.3w | 15

https://modelscope.cn/search?search=mplug

PART 04

总结与展望

总结与展望

- 多模态文档大模型展示了Promising效果，但仍然存在通用性、鲁棒性、效率等问题，仍然还有很多提升的空间；
- 针对高分辨率，多页，探索更高效的视觉编码，图文融合结构；
- 增加更多Domain数据，探索通用的、鲁棒的多模态文档大模型；
- 大小模型协同，能够结合OCR、Chart等专业小模型；

https://github.com/X-PLUG/mPLUG-DocOwl

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

K+峰会

K+峰会 上海站

K+全球软件研发行业创新峰会

时间：2024.06.21-22

K+峰会 敦煌站

K+思考周®研习社

时间：2024.10.17-19

K+峰会 香港站

K+思考周®研习社

时间：2024.11.10-12

K+峰会详情

AiDD峰会

AiDD峰会 上海站

AI+研发数字峰会

时间：2024.05.17-18

AiDD峰会 北京站

AI+研发数字峰会

时间：2024.08.16-17

AiDD峰会 深圳站

AI+研发数字峰会

时间：2024.11.08-09

AiDD峰会详情

AiDD

2024 AI+研发数字峰会

AI+ Development Digital summit

深圳站 11/08-09

AI驱动研发变革 促进企业降本增效

2024深圳站-议题设置

AI+产品线	LLM驱动产品创新	LLM驱动需求与业务分析	AI驱动设计与用户体验
AI+开发线	AI 原生应用开发框架与技术	AI Agents在研发落地实践	LLM驱动编程与单测
AI+测试线	LLM驱动测试分析与设计	基于LLM生成测试脚本与数据	LLM和AI应用的评测
AI+工程线	AI+DevOps 与工具 （LLM 时代的平台工程）	大模型对齐与安全	端侧大模型与云端协同
AI+领域线	领域大模型 SFT 与优化	知识增强与数据智能	大厂专场

扫描右侧二维码
查看更多会议详情

早鸟票限时抢购中 (截止到9月30日)

¥3680
早鸟票

¥2800
学生票

THANKS