



2024 AI+研发数字峰会

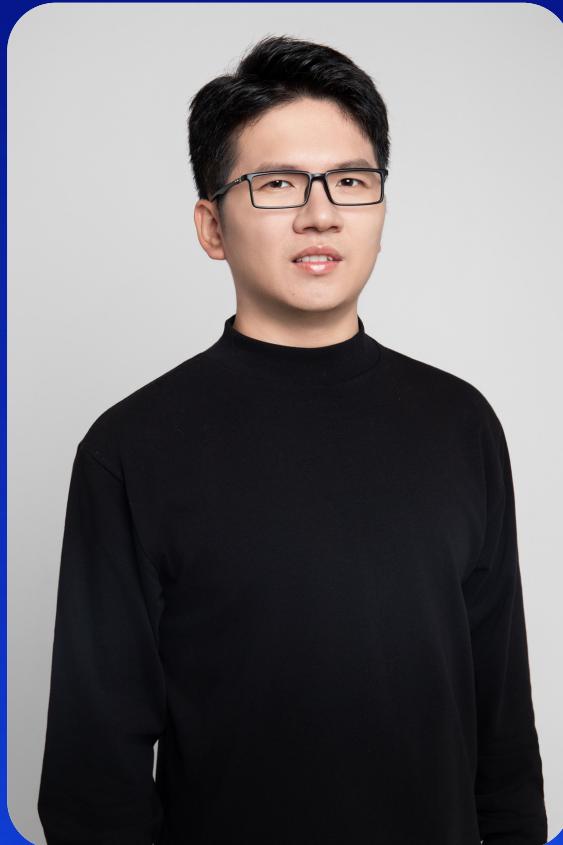
AI+ Development Digital summit

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



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

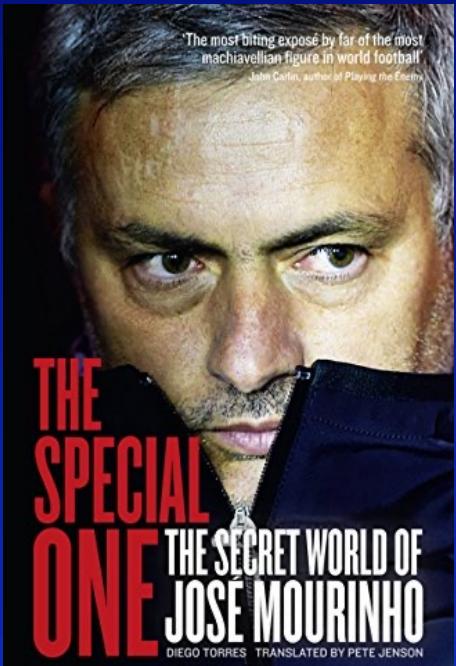
CONTENTS

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

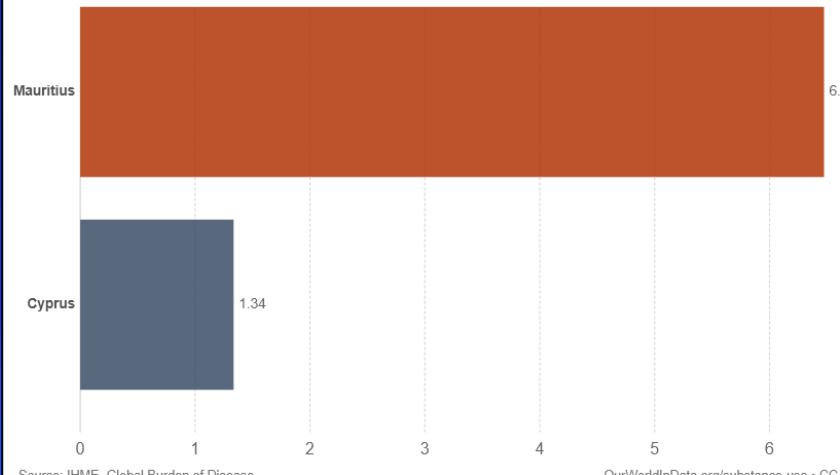
PART 01

多模态文档大模型背景

► 多模态文档大模型背景



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 rate per 100,000 population.



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

BUDGET REQUEST SUMMARY

YEAR : May 1, 1966 - Apr. 30, 1967

1. Personnel \$ 11,228.00

2. Permanent Scientific Equipment \$ 0.00

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

富含文字的图片（简称“文档图片”）在生活中无处不在！

FSF FREE SOFTWARE FOUNDATION

about campaigns licensing membership resources community donate shop

Intern at the FSF

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 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 bios from past interns who worked for our tech team here. A small number of past interns have even become dozens of bios from past interns who worked for our tech team here. A small number of past interns have even become

Take Action!

intern at the... LIBRE PLANET 2020 FREE THE FUTURE

APPLY NOW! Register for LibrePlanet 2020!

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

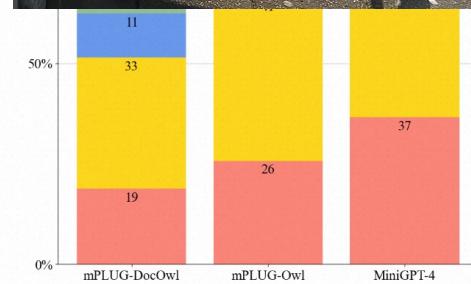


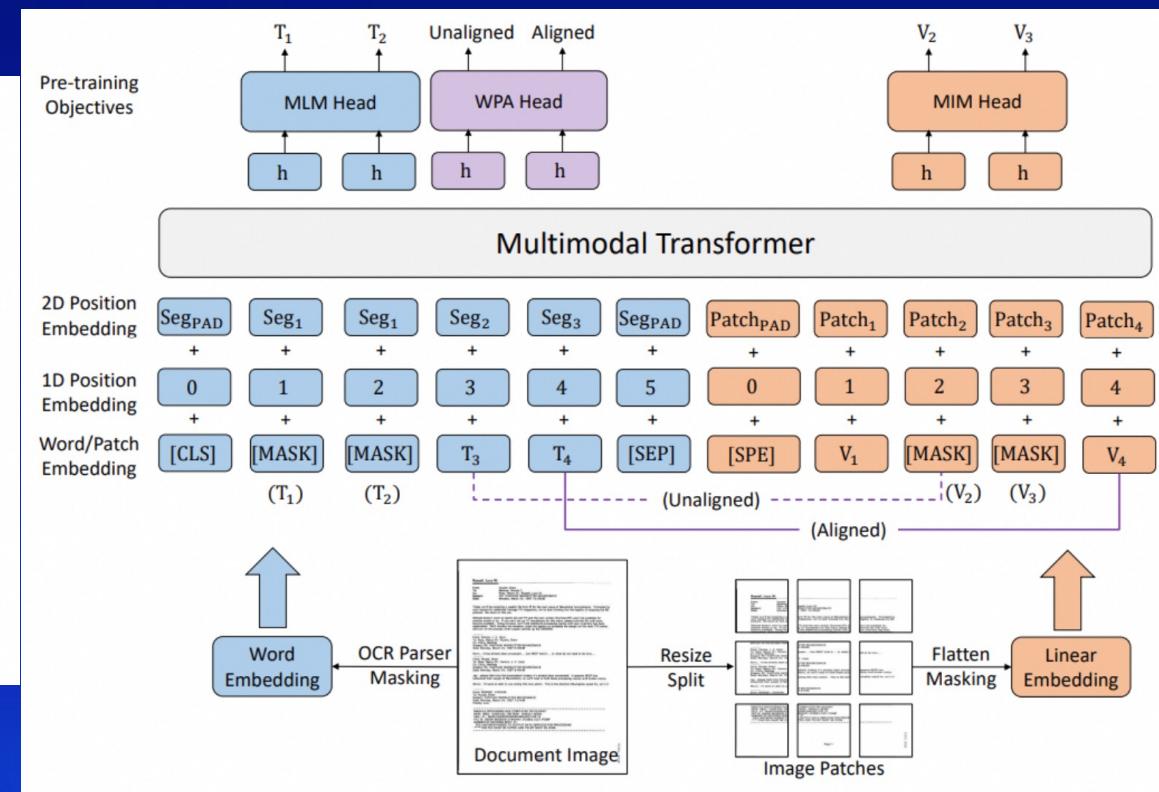
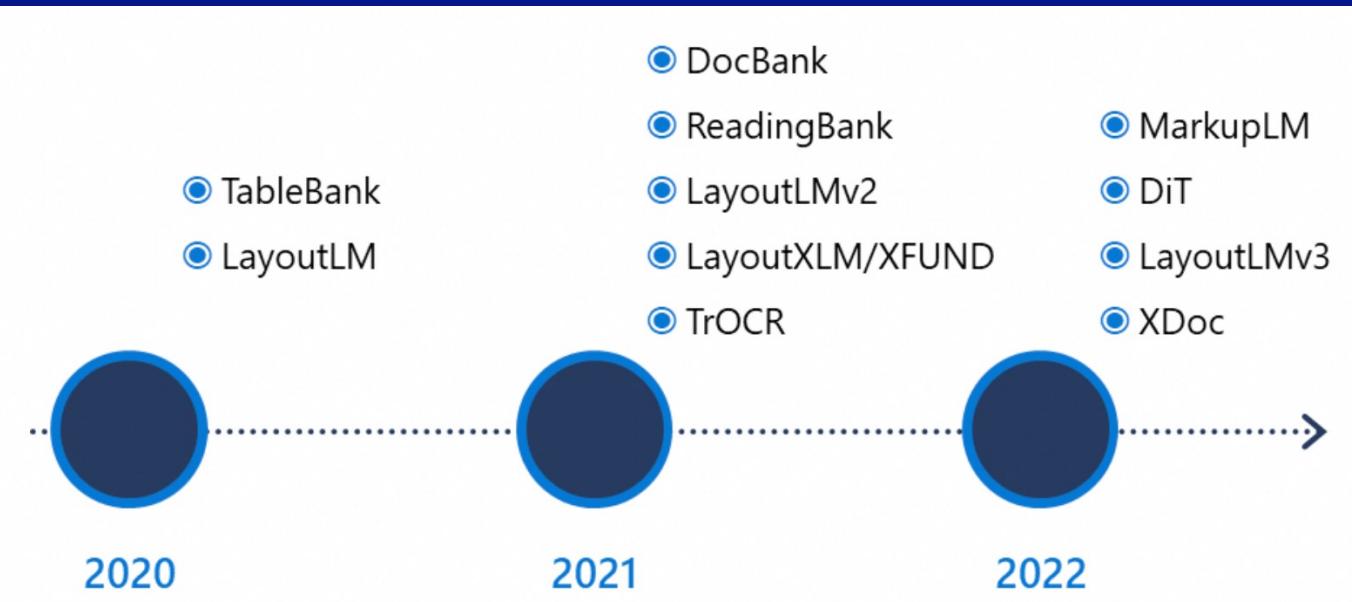
Figure 3: Human evaluation of mPLUG-DocOwl, mPLUG-Owl and MiniGPT-4 on LLMDoc.

spurious claims 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 Dessert [Davis et al., 2022], Donut [Kim

► 多模态文档大模型背景



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?

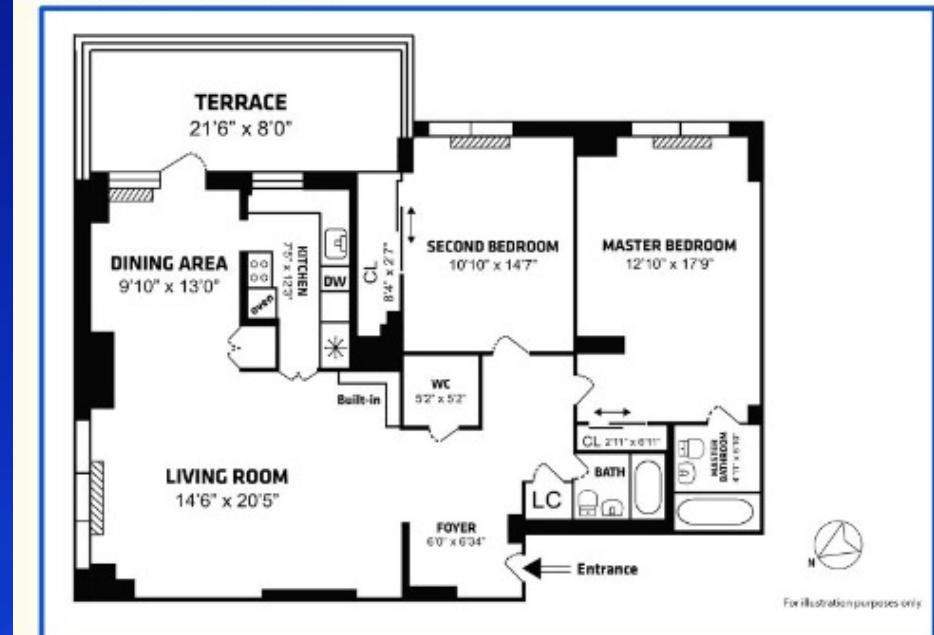
Monthly income (in Rs.)	Vehicles per family			
	0	1	2	Above 2
Less than 7000	10	160	25	0
7000 – 10000	0	305	27	2
10000 – 13000	1	535	29	1
13000 – 16000	2	469	59	25
16000 or more	1	579	82	88

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 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:



Can you explain this meme in detail?

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.

User
LLaVA

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.



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, capturing the 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.



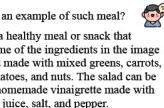
Can you describe this image in detail?

The image depicts a 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 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.



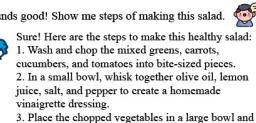
Can you give me an example of such meat?

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.



What makes this dog special?

The dog in the image is wearing futuristic armor and has a pair of glowing green eyes. 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.



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 washed 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.

► 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

President considers action against key officials, including Sessions, in Oval Office battle for House majority

McCain absence shrinks GOP's margin

O.J. Simpson is given parole after 9 years

TRAVIS SHARP - O.J. Simpson's two brothers after Nevada parole commissioners announced his release

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；

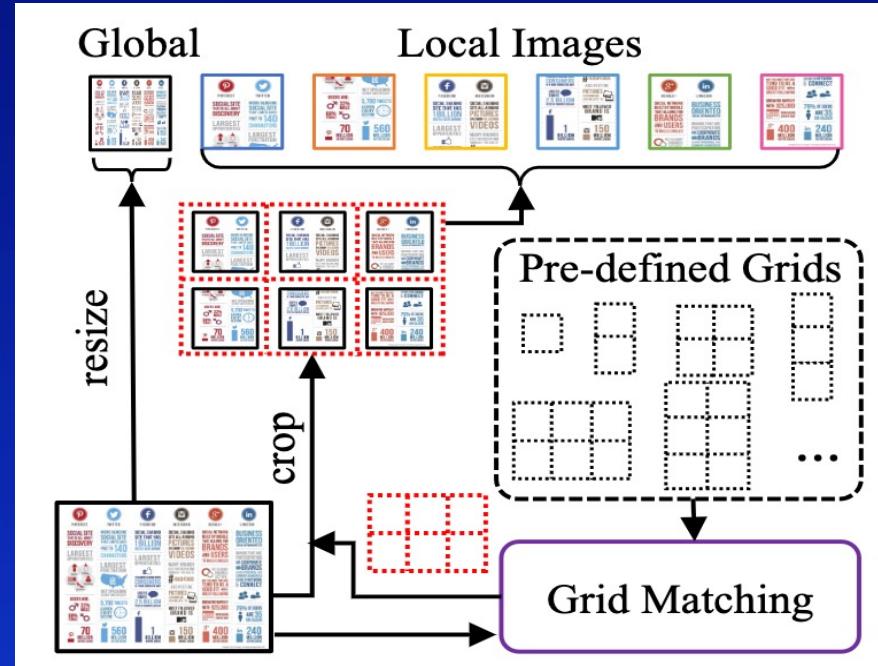
Ours: 多任务联合学习+辅助学习任务

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. [URReader](#): 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

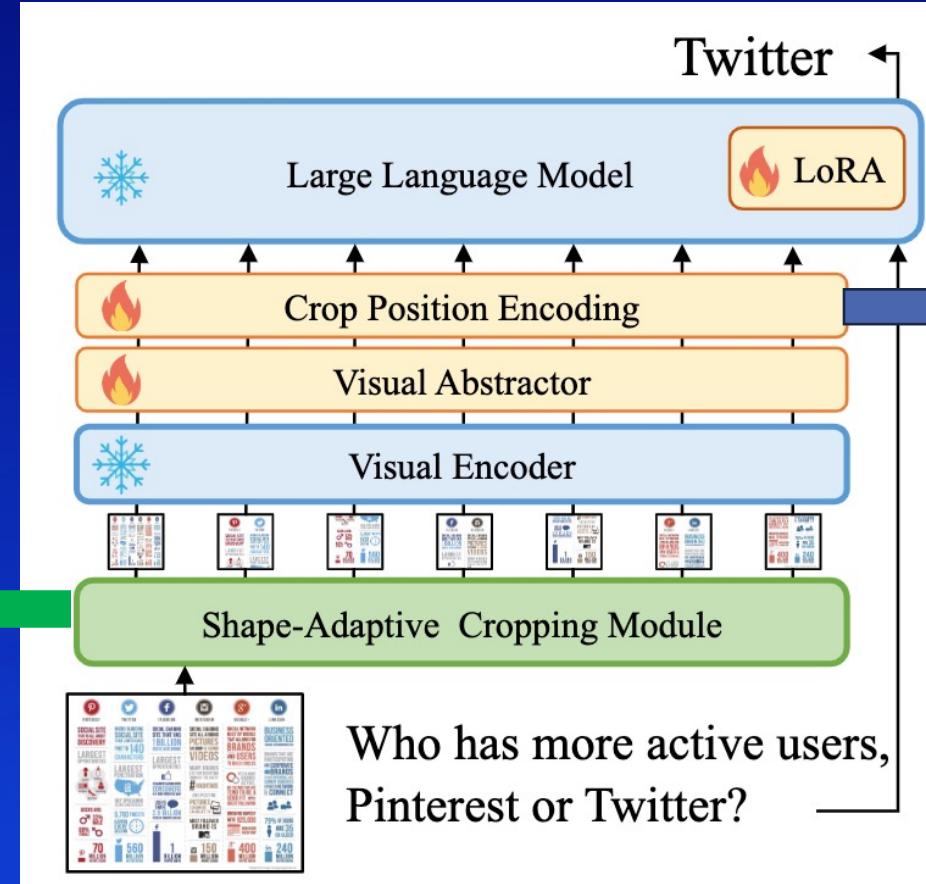
1. $Grids = \{1 \times 1, 1 \times 2, \dots, 1 \times 20,$

$2 \times 1, 2 \times 1, \dots, 2 \times 10,$

\dots

$20 \times 1\}$

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)



Category	Sub-Categories	Value
Category A	Sub-Cat A1, Sub-Cat A2	Value A1, Value A2
Category B	Sub-Cat B1, Sub-Cat B2	Value B1, Value B2
Category C	Sub-Cat C1, Sub-Cat C2	Value C1, Value C2
Category D	Sub-Cat D1, Sub-Cat D2	Value D1, Value D2



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

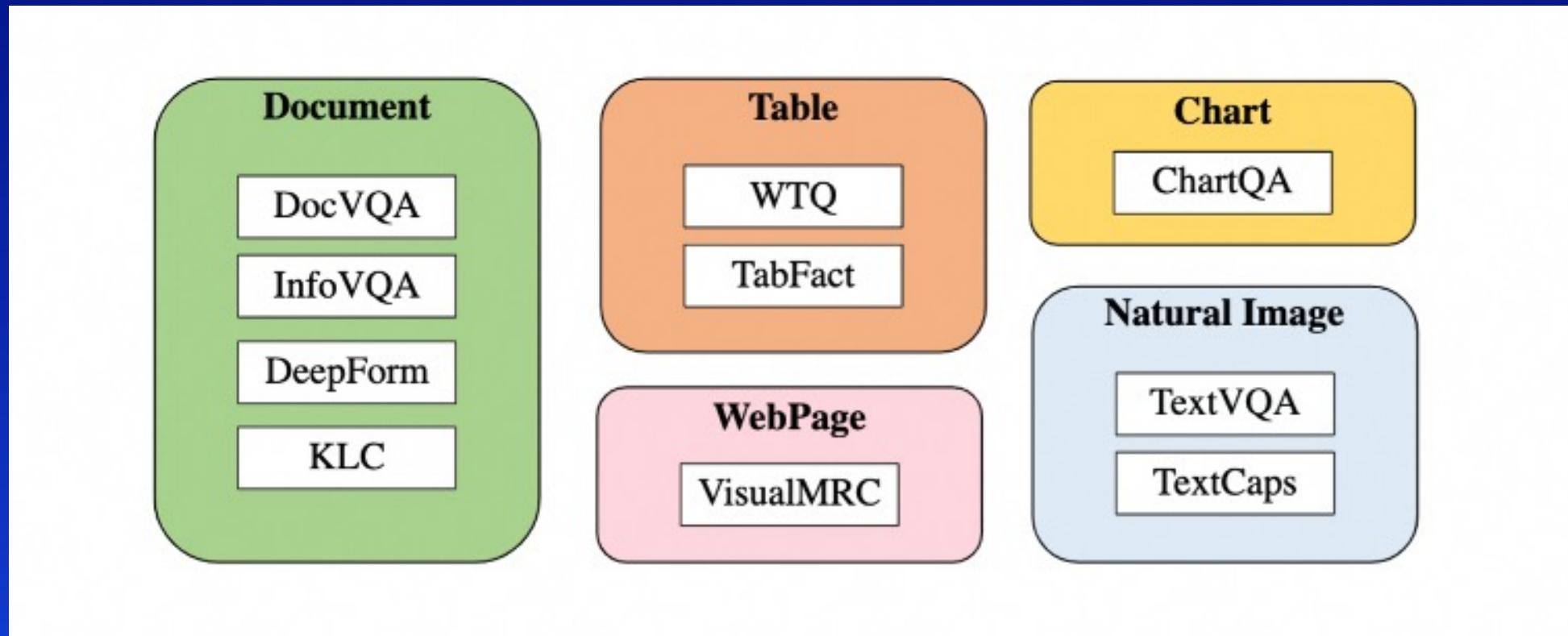
- 整个图片的文字序列按照从上到下，从左到右的形式组织，
- 随机选取序列中 $\{0, \frac{1}{6}, \frac{2}{6}, \dots, \frac{5}{6}\}$ 的一个位置（概率分布为 $\{0.5, 0.1, 0.1, \dots, 0.1\}$ ）作为断点，
- 要求模型根据图片和断点前面的文字，生成断点后面的文字
- 相比所有的样例都读全部文字，这种方法可以让模型更多地关注图片上不同区域的文字

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



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-specific fine-tuning on the downstream dataset. ‘underline’ 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>	<u>42.2</u>	<u>49.5</u>	<u>32.8</u>	<u>29.4</u>	<u>67.6</u>	<u>59.3</u>	<u>57.6</u>	<u>118.4</u>	<u>221.7</u>

mPLUG-DocOwl1.5

结构布局理解对于多模态
文档理解十分关键

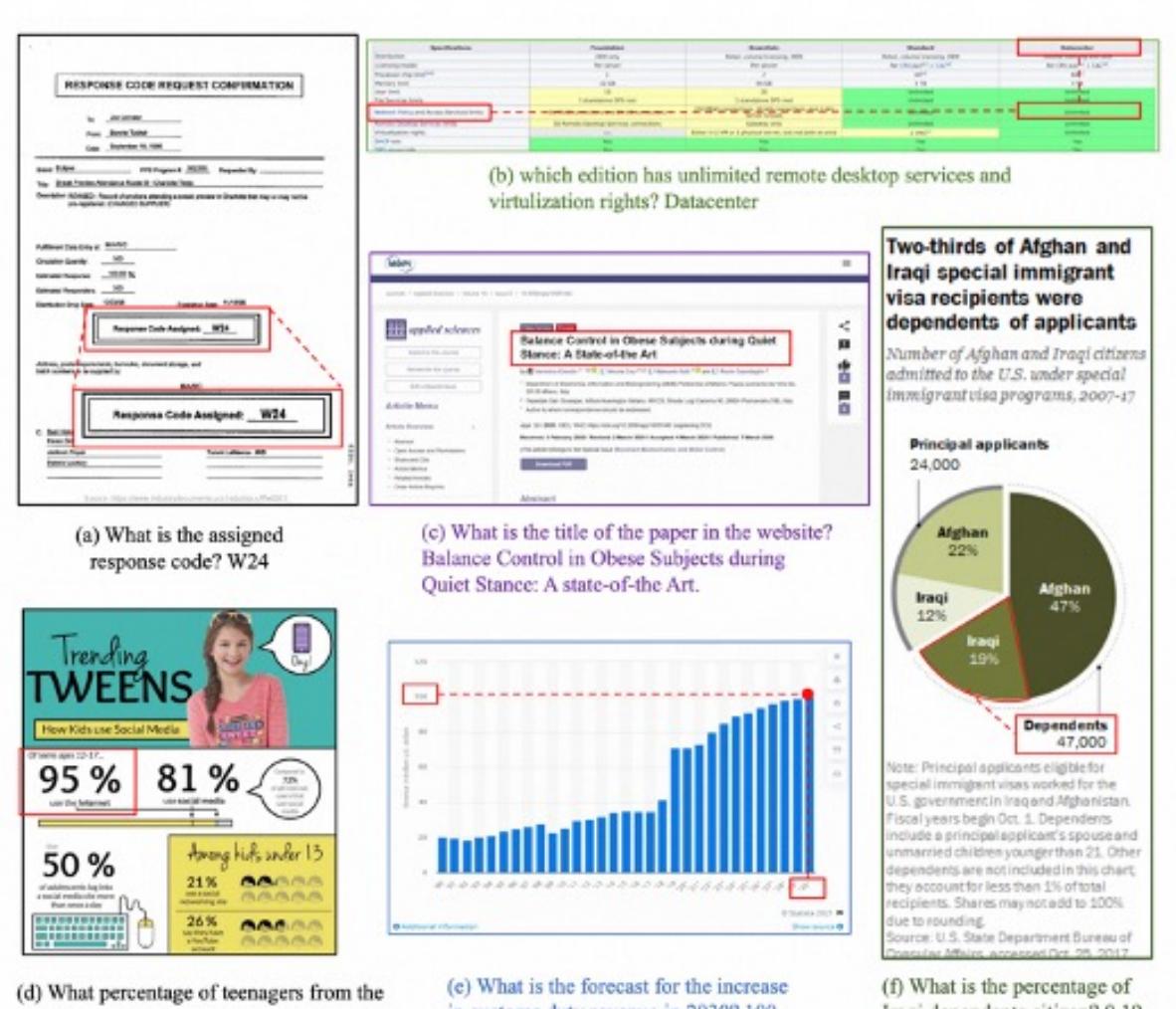


Figure 2: Illustrations of the importance of structure information in Visual Document Understanding on documents (a), tables (b), webpages (c), infographics (d), and charts (e-f).

► 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；

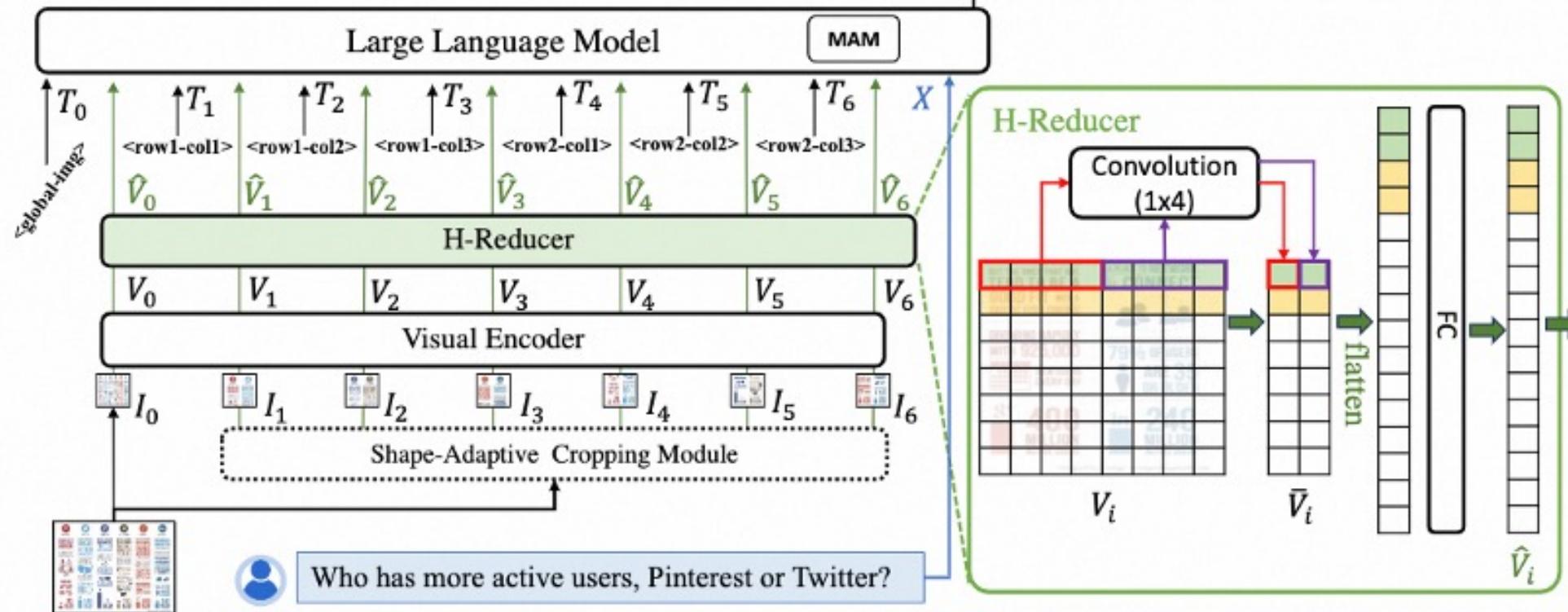
模型训练: Unified Structure Learning

- 结构感知的文字解析
- 多粒度文字识别和定位；

mPLUG-DocOwl1.5



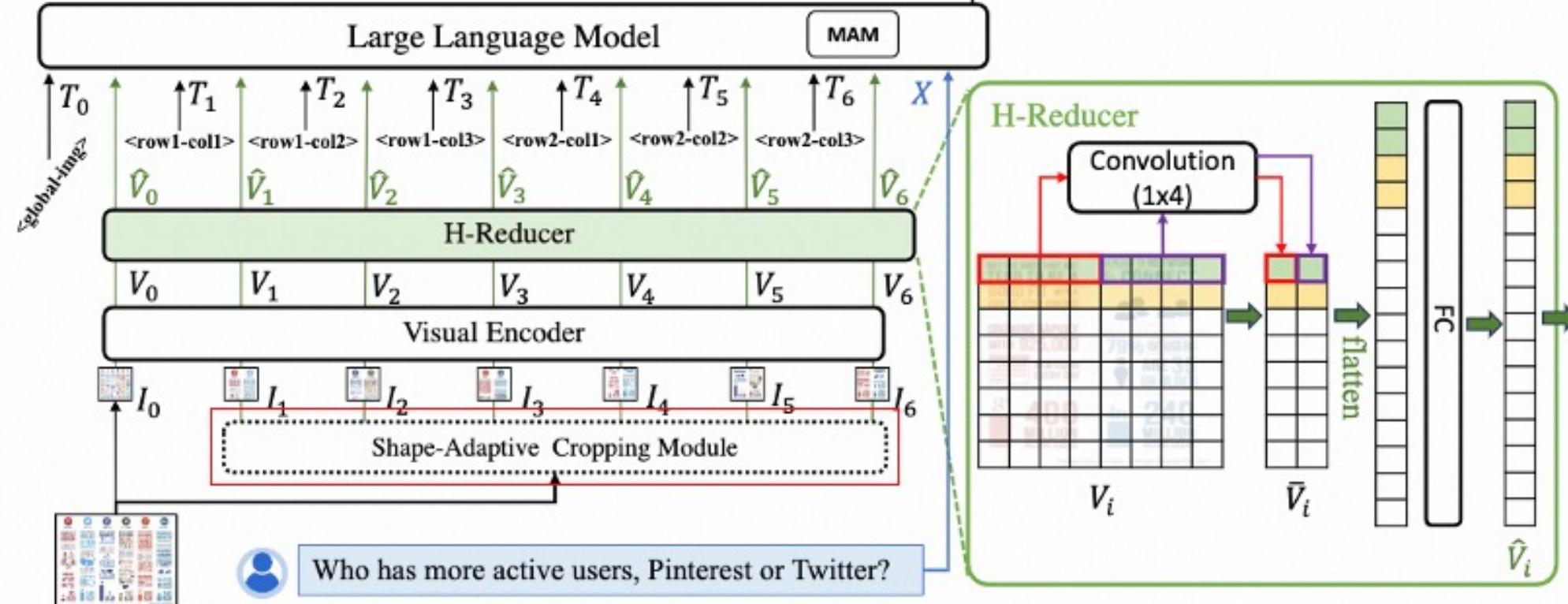
Twitter. According to the image, there are 560 million and 70 million active users for Twitter and Pinterest. Thus, Twitter has more active users.



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.

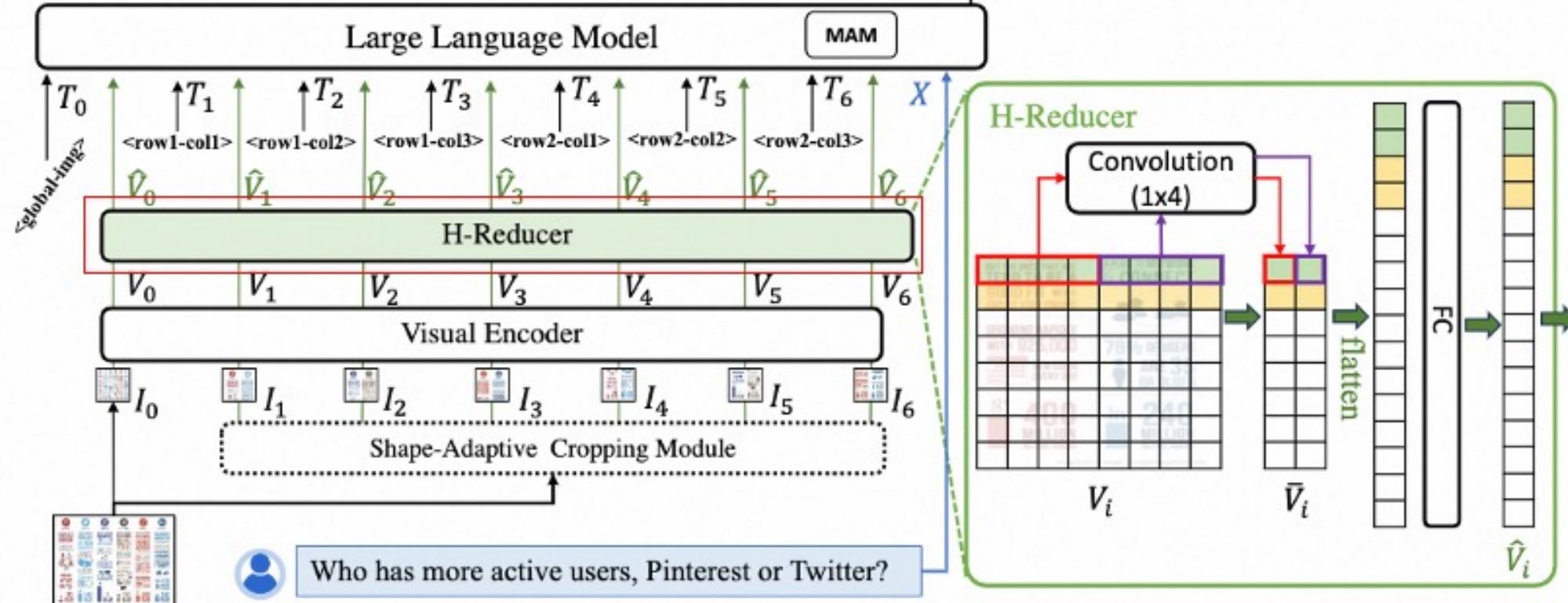


延续DocOwl, 采用Shape-Adaptive Cropping Module处理高分辨率图片

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.

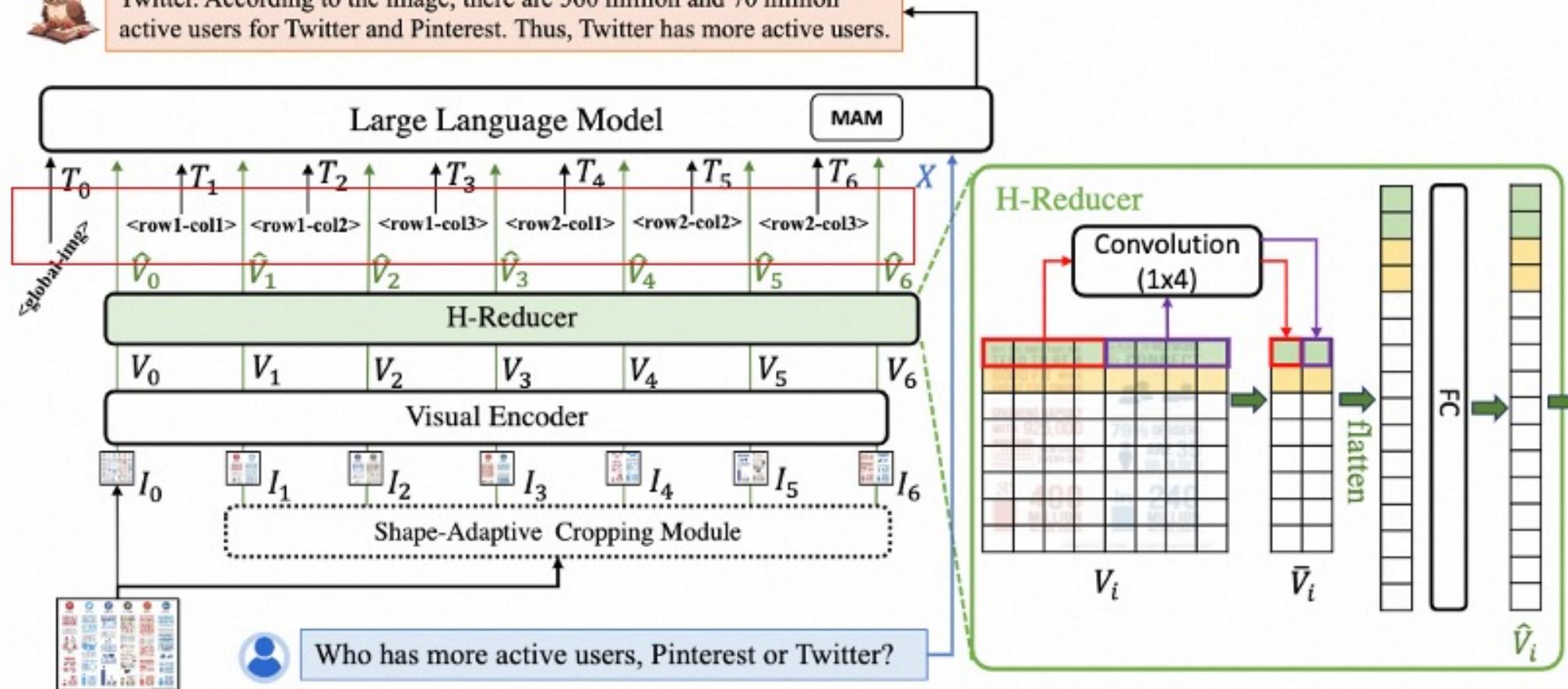


- 不同于DocOwl, 采用基于卷积的HReducer将视觉特征对齐到语言特征
- 考虑到文字大部分为横向排布, 采用1x4的卷积

mPLUG-DocOwl1.5



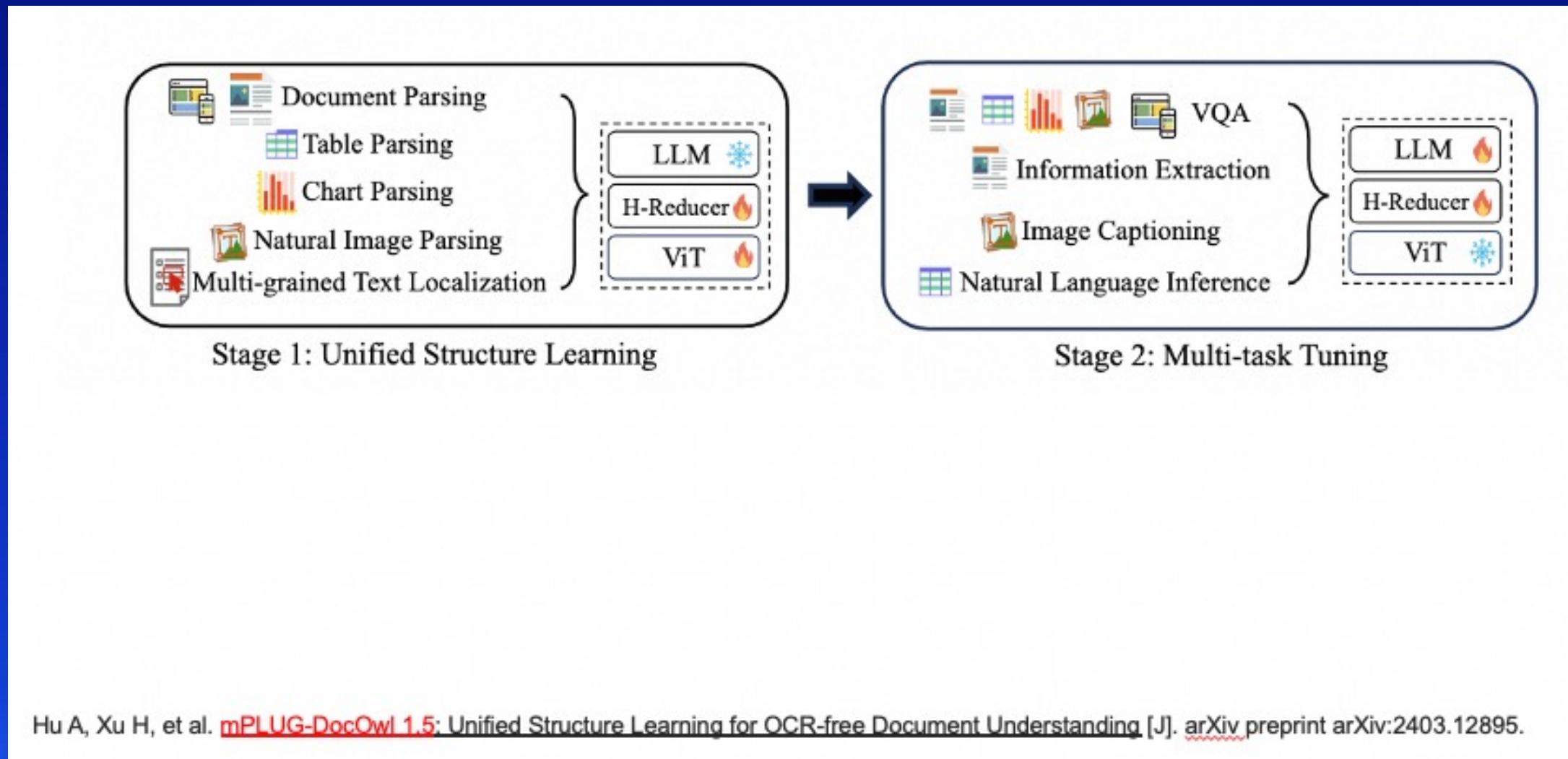
Twitter. According to the image, there are 560 million and 70 million active users for Twitter and Pinterest. Thus, Twitter has more active users.



- 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

► 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

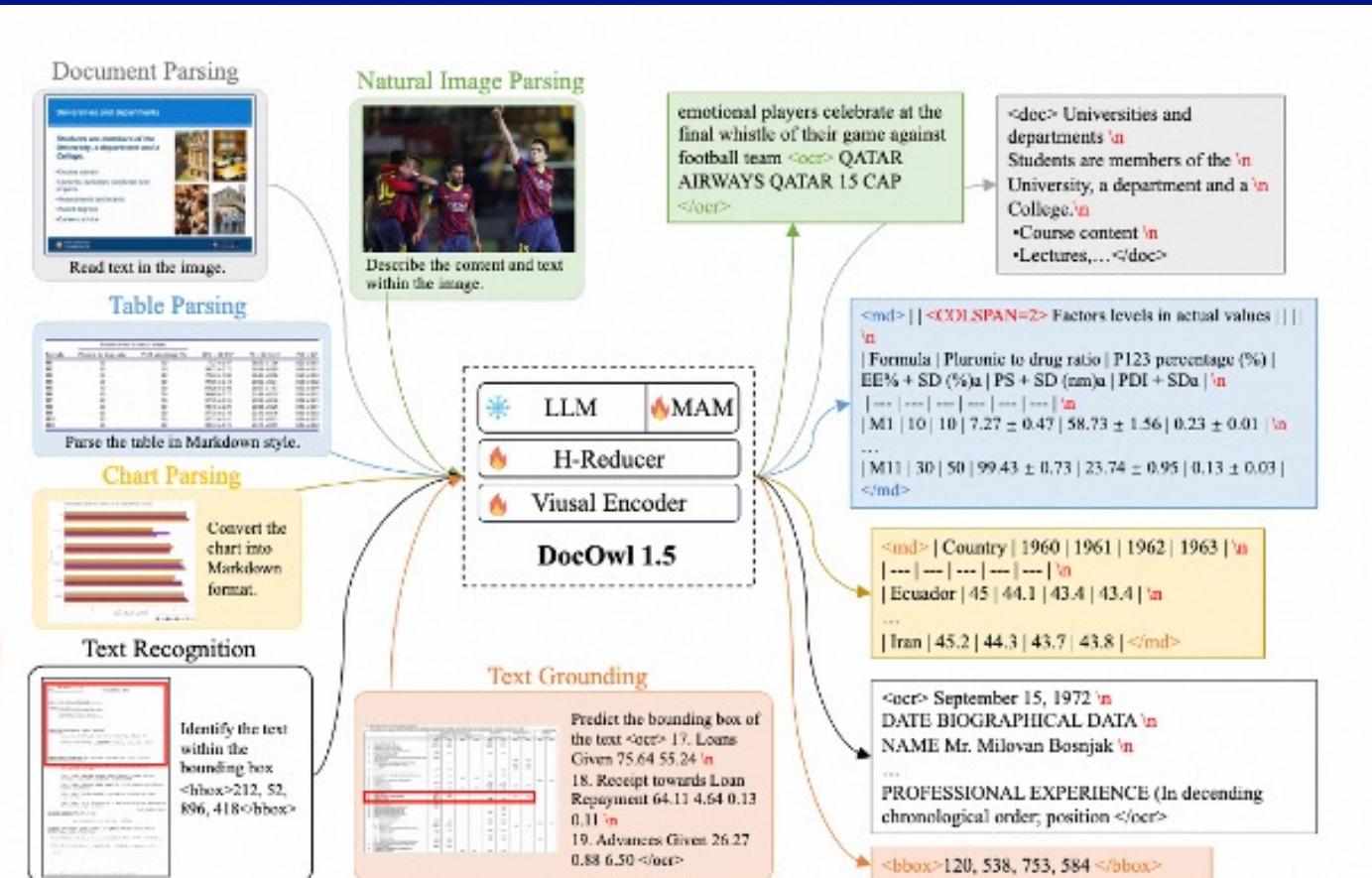


Figure 4: The illustration of Unified Structure Learning of DocOwl 1.5.

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.

► 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 Sample	1,491 5,119	1,614 5,421	850 5,994	1,927 5,263	1,252 1,827	1,612 2,253	8,746 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。

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			Open
			Text	Bbox	Size	
Donut [23]	-	2560x1920	✓	✗	13M	Synthetic, Doc
Pix2Struct [24]	-	2^{19} (shape variable)	✓	✗	80M	Web
QwenVL [9]	-	448x448	✓	✗	24.8M	Synthetic, Doc, Web
Monkey [26]	QwenVL [9]	896x896	✗	✗	-	-
UReader [57]	Owl [59]	224x224(x20 crops)	✓	✗	0.1M	Doc, Table, Chart, Web, Natural
DocPedia [13]	-	2560x2560	✓	✓	0.9M	Doc
CogAgent [46]	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	<u>68.8</u>	132.0	239.5
DocOwl-1.5-Chat	8.1B	82.2	50.7	68.8	38.7	40.6	80.2	70.2	<u>68.6</u>	131.6	246.4

mPLUG-DocOwl1.5

样例：文档解析

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



DocOwl 1.5:



Extract words from the graphic.

```
<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       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 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

LAN exposure	N (%)*		Crude results (n = 61,129)		Adjusted results** (n = 47,817)	
	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)†

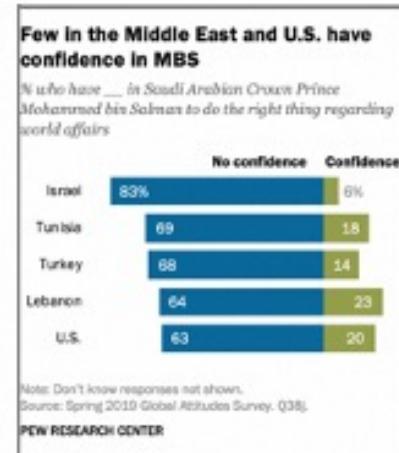


DocOwl 1.5:



Convert the figure into Markdown language.

```
<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>
```



DocOwl 1.5:



Render the figure in Markdown format.

```
<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



DocOwl 1.5:



Detail the information and textual components depicted in the image.

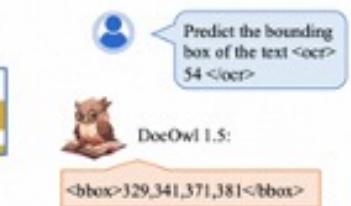
the main street in the town <ocr> a a alamy NIMA RESTAURANT & BAR Special TOURS & DEL PSTAIRS PERIA DORA GENERAL SHOP SONAR ENTERPRISE a a a a a a alamy photo stock E7BDR1 a www.alamy.com </ocr>

mPLUG-DocOwl1.5

样例：多粒度文字识别

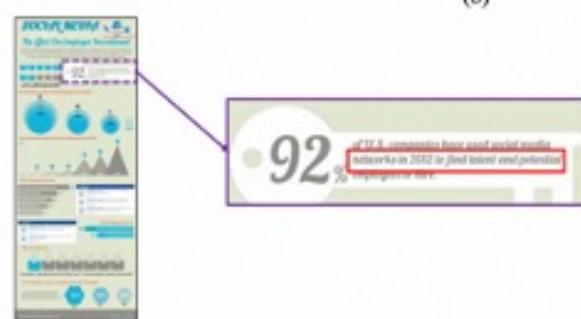
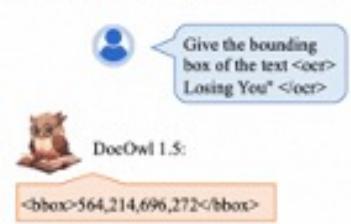


(a)

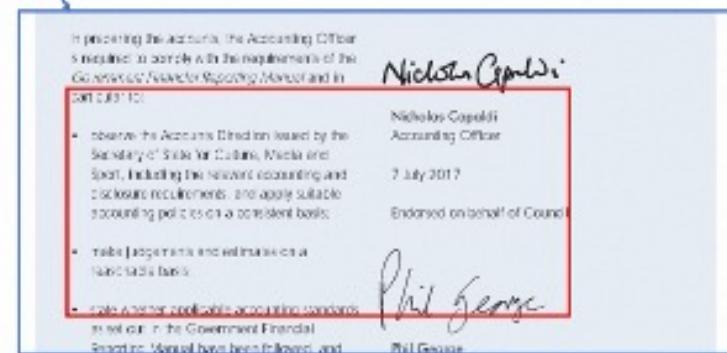
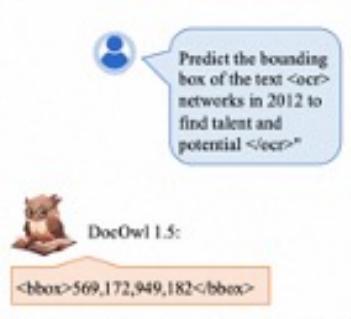


National final				
Draw	Artist	Song	Points	Place
1	The Memories	"If It Means Losing You"	57	6th
2	Ann Breen	"Oh, Darling"	80	4th
3	Fran Neen	"Say That You Love Me"	66	6th
4	Dreams	"Sis Sis ('That's That')"	73	5th
5	Connor Stevens	"Count On Me"	88	3rd
6	Linda Martin and Friends	"All The People In The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

(b)



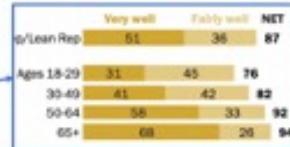
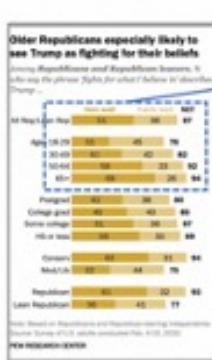
(c)



(d)

mPLUG-DocOwl1.5

样例：详细解释的问答



DocOwl 1.5-Chat:

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

(a)



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)

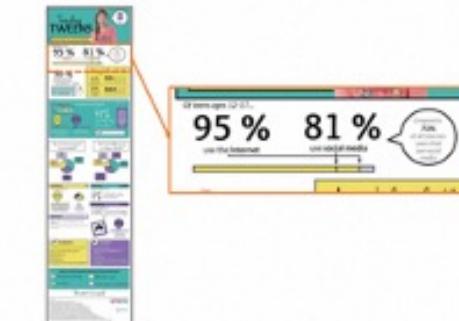
(a)

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 #".



(b)

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 $100\% - 95\% = 5\%$.

Date	Time	Opponent ^a	Rank ^b	Site	TV	Result	Attendance
September 1	2:30 PM	#9 NC State Northern Illinois	#12	Camp Randall Stadium • Madison, WI	BTN	W 26-21	79,368
September 3	3:00 PM	at Oregon State*	#13	Rosie Stadium • Corvallis, OR	FX	L 7-10	41,189
September 13	7:00 PM	Utah State*	#22	Camp Randall Stadium • Madison, WI	BTM ^c	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 #30 Nebraska	#23	Memorial Stadium • Lincoln, NE	ABC	L 27-39	85,942
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 36-14	46,007
October 29	11:00 AM	Minnesota		Camp Randall Stadium • Madison, WI (Paul Bunyan's Ave)	ESPNW	W 36-13	80,387
October 27	2:30 PM	Michigan State	#25	Camp Randall Stadium • Madison, WI	ABC/ESPN2	L 13-26 OT ^d	80,538
November 10	11:00 AM	at Indiana		Memorial Stadium • Bloomington, IN	ESPN2	W 62-14	41,240
November 17	2:30 PM	Ohio State		Camp Randall Stadium • Madison, WI	ABC/ESPN2	L 14-21 OT	80,112
December 1	2:30 PM	vs. #34 Nebraska		Beaver Stadium • University Park, PA	ESPN2	L 21-24 OT	93,345
January 1, 2013	4:30 PM	vs. #8 Stanford	#23	Lucas Oil Stadium • Indianapolis, IN (Big Ten Championship Game)	Fox	W 70-31	41,260
Non-conference game. *Homecoming. ^a Rankings from Coaches Poll released prior to game. All times are in Central Time.							

(c)

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



DocOwl 1.5-Chat:

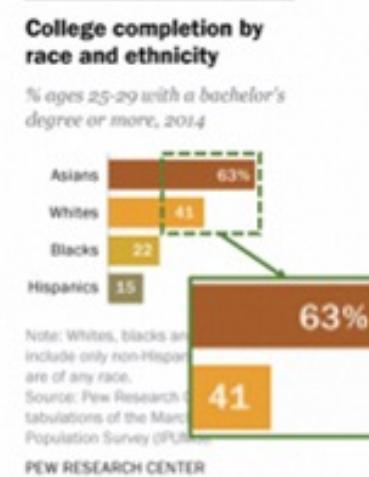
The last date and time on the chart is January 1, 2013 at 4:10 PM. The chart lists all the games played by a certain team, along with the date, time, opponent, rank, site, and result. The last game listed on the chart is against Stanford, which took place on January 1, 2013 at 4:10 PM. Therefore, the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

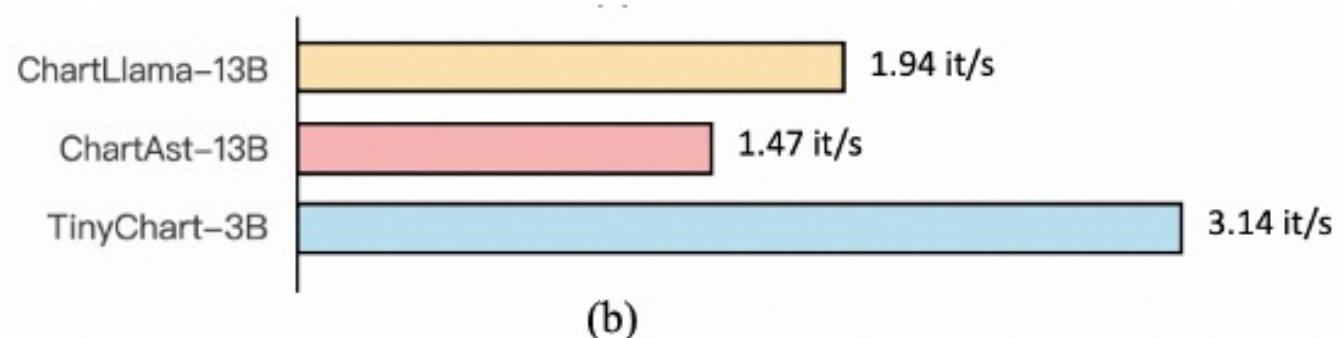
GT Answer: 22

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)
```



TinyChart

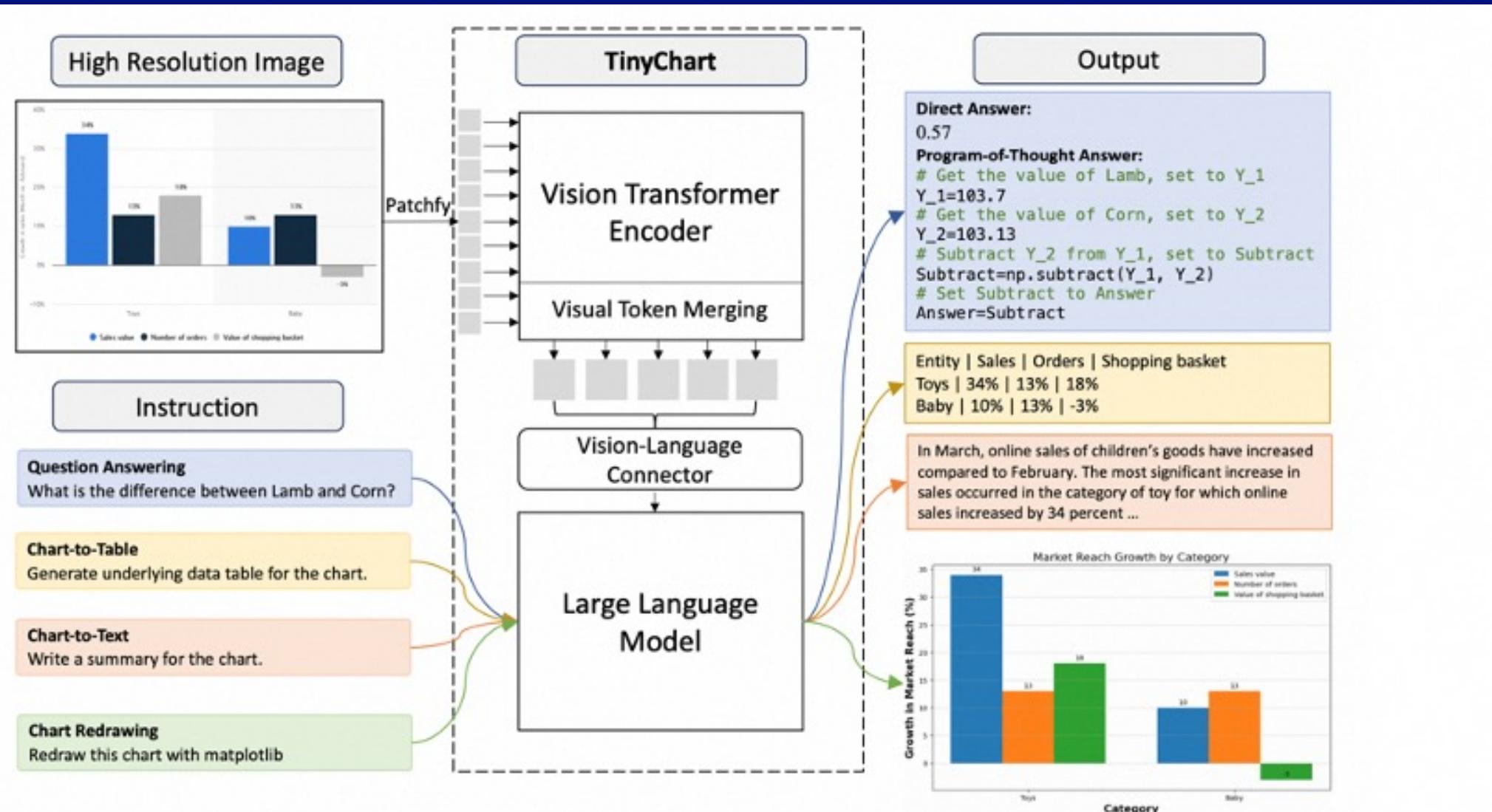


Figure 2: Overview of TinyChart.

TinyChart

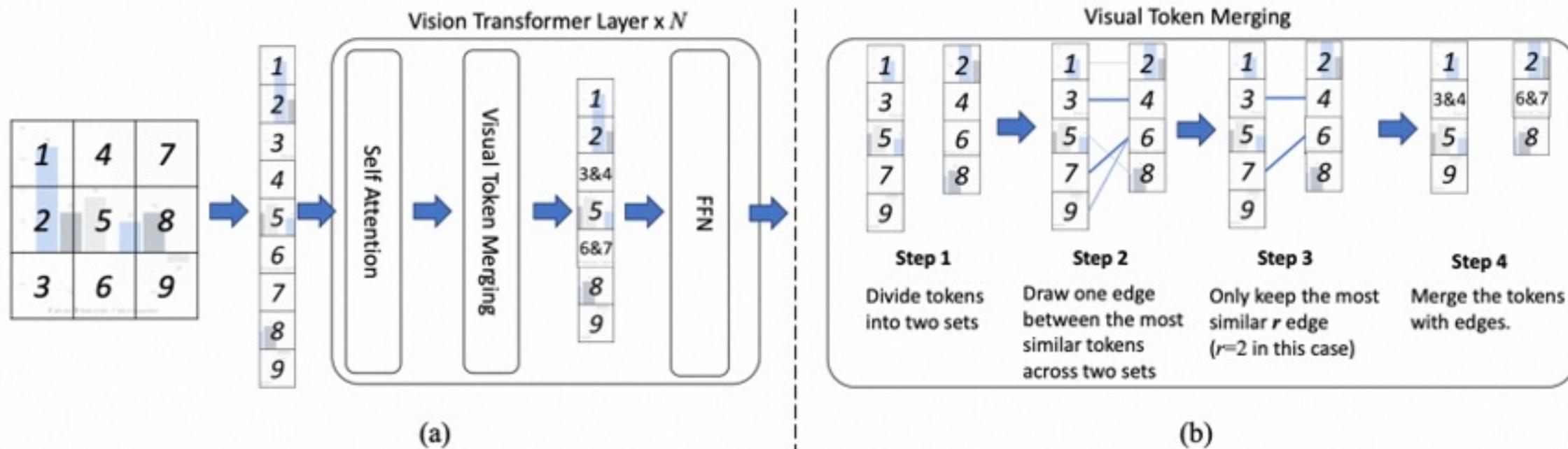


Figure 3: (a) Vision transformer layer with Visual Token Merging. (b) Process of the Visual Token Merging.

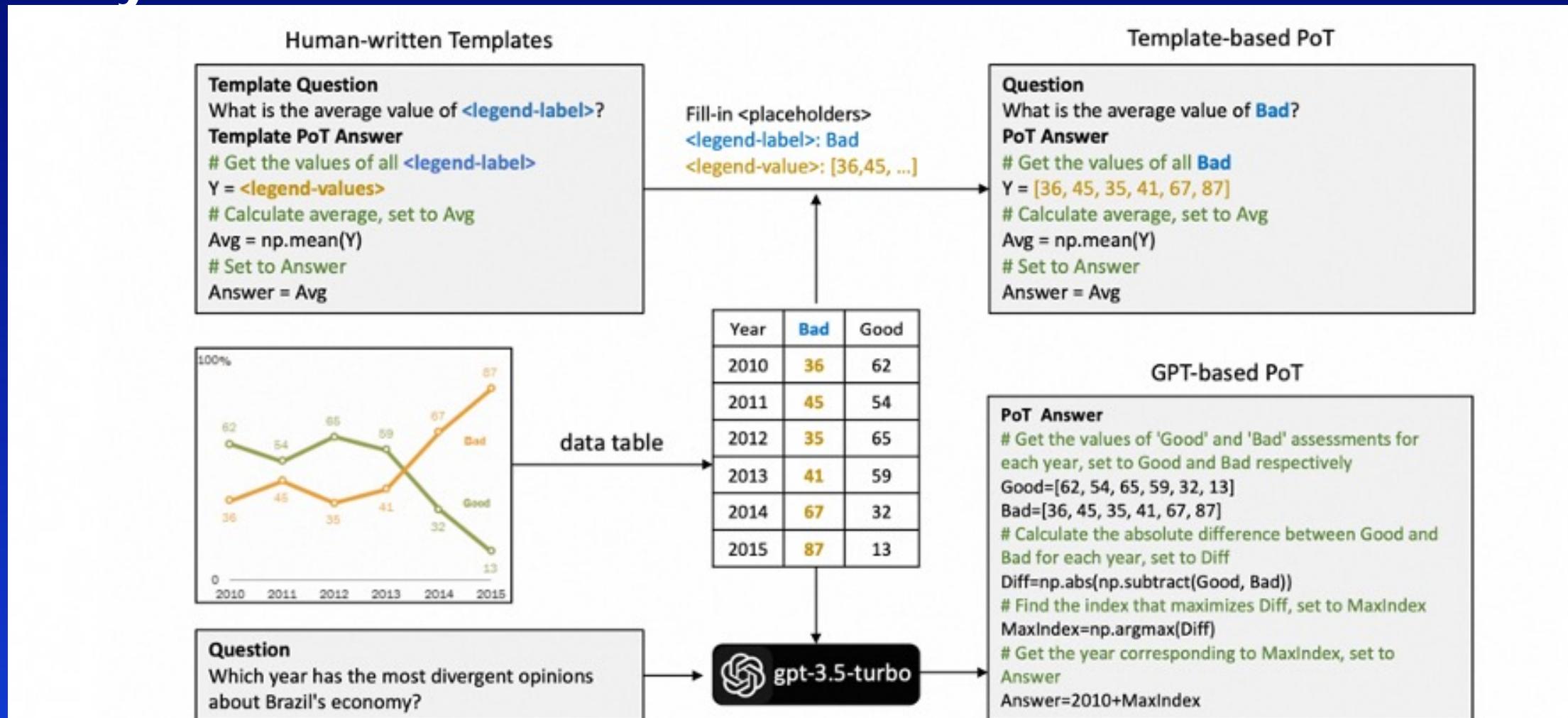


Figure 4: The demonstration of constructing Template-based PoT (upper half) and GPT-based PoT (lower half) in the ChartQA-PoT dataset.

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.

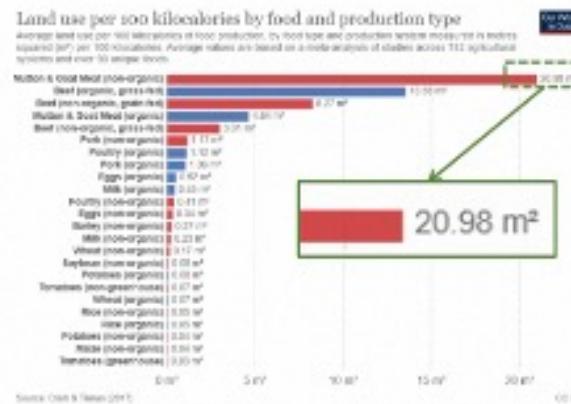
Dataset	Benchmark	Samples
<i>Chart question answer</i>		
ChartQA [32]	✓	28,299
ChartQA-PoT	✓	140,584
PlotQA [36]		157,070
DVQA [19]		200,000
OpenCQA [20]		5,407
<i>Chart-to-text generation</i>		
Pew [21]	✓	7,892
Statista [21]	✓	29,589
OpenCQA [20]		5,407
Vistext [45]		11,171
ChartSumm [42]		75,255
Chart2Text-8k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

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.			
<i>Close source models</i>									
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	-
<i>Open source models</i>									
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

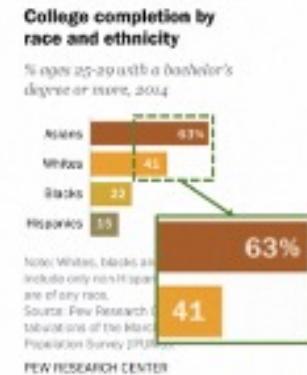
TinyChart



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

TinyChart Direct: 20.98

(a)

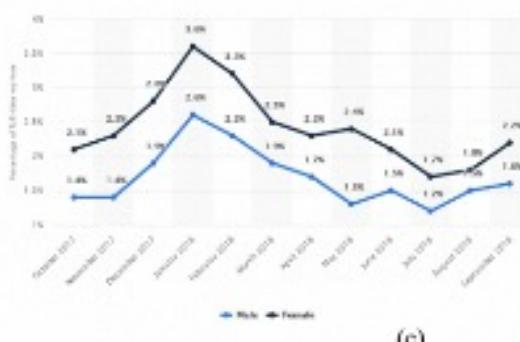


Question: What is the difference between Asians and Whites degree distribution?

TinyChart Direct: 22

(b)

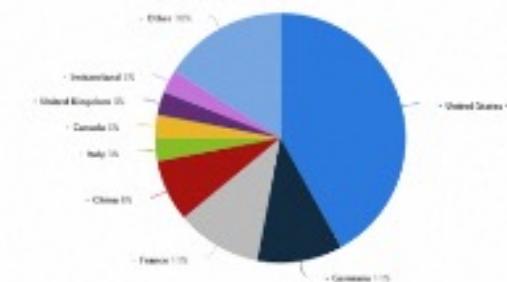
Input: Generate underlying data table for the chart. **TinyChart:**



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% | 3.0%
 December 2017 | 1.9% | 2.8%
 November 2017 | 1.4% | 2.3%
 October 2017 | 1.4% | 2.1%

(c)

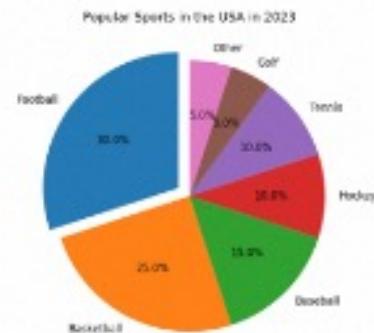
Input: Generate underlying data table for the chart. **TinyChart:**



(d)

TinyChart

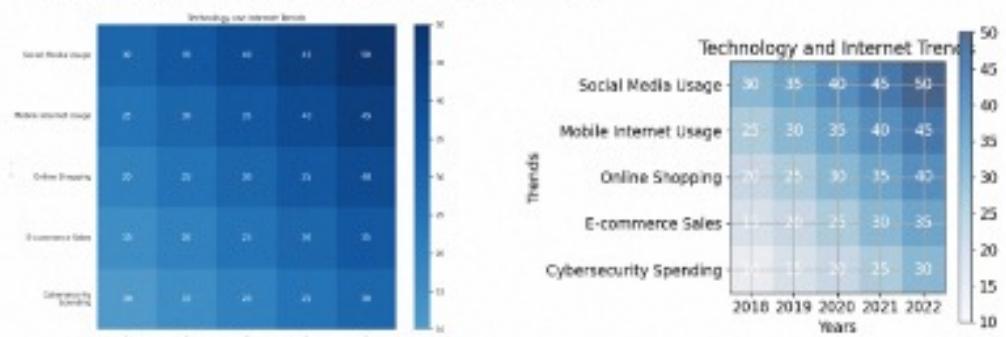
Input: Create a brief summarization or extract key insights based on the chart image.



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.

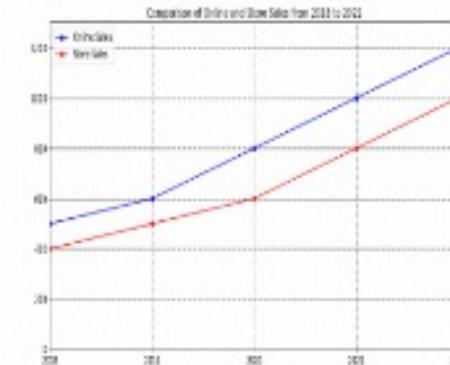
(e)

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



(g)

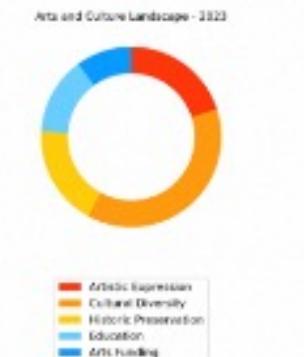
Input: Create a brief summarization or extract key insights based on the chart image.



(f)

Input: Redraw the chart with Python code.

TinyChart:



(h)

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

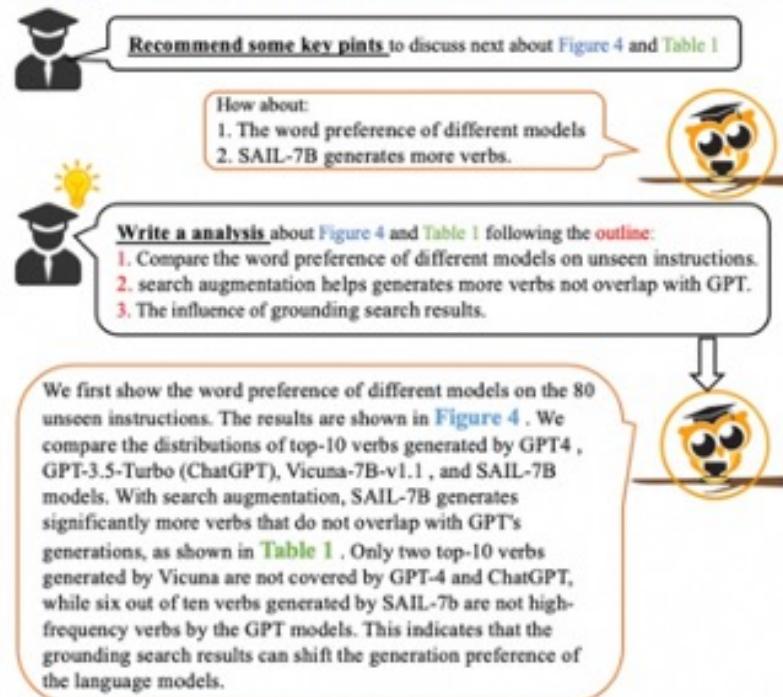


Figure 4: Top-10 verbs and associated nouns generated by selective large language models

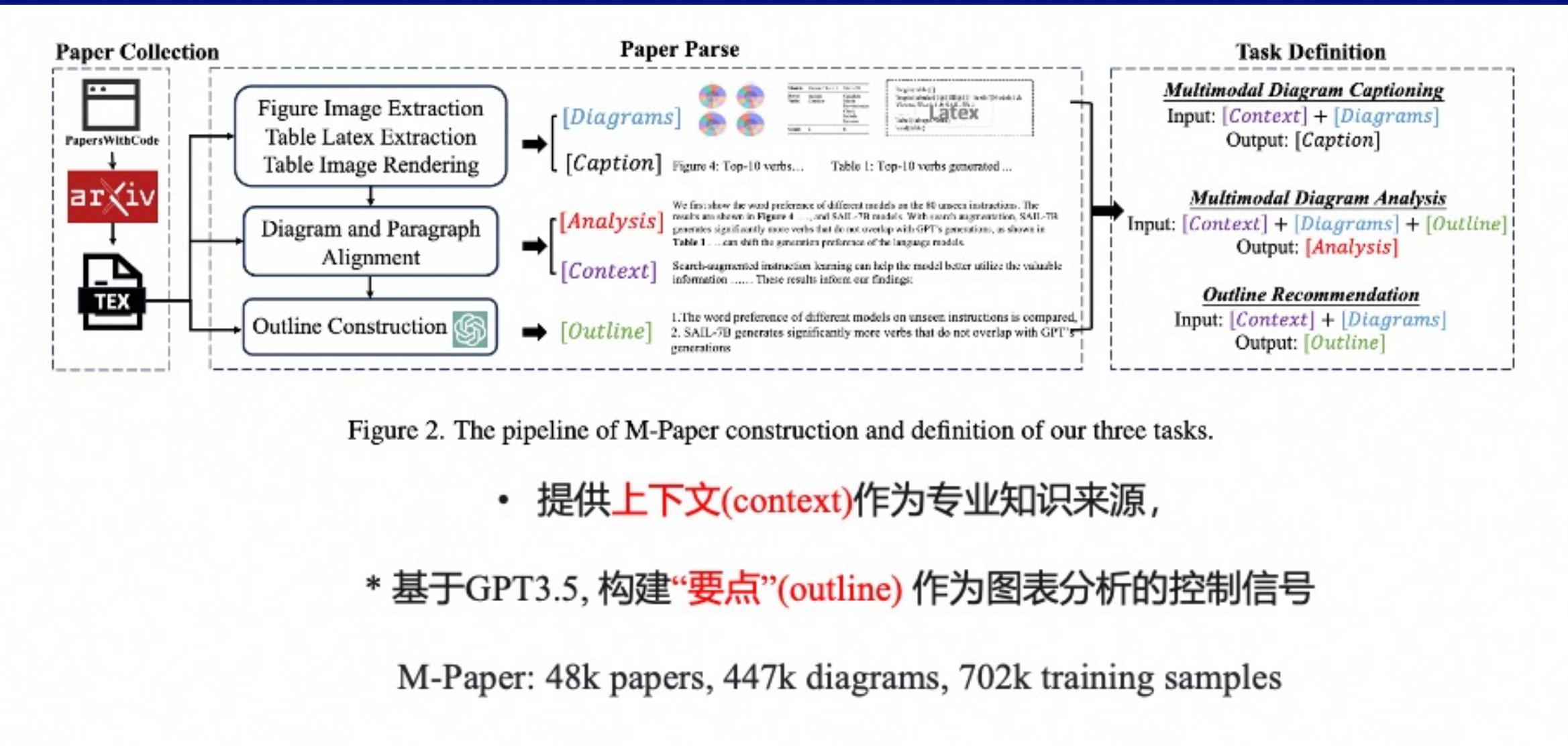
Models	Vicuna-7B-v1.1	SAIL-7B
Novel Verbs	Include Consider	Calculate Match Revolutionize Check Include Increase
Count	2	6

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

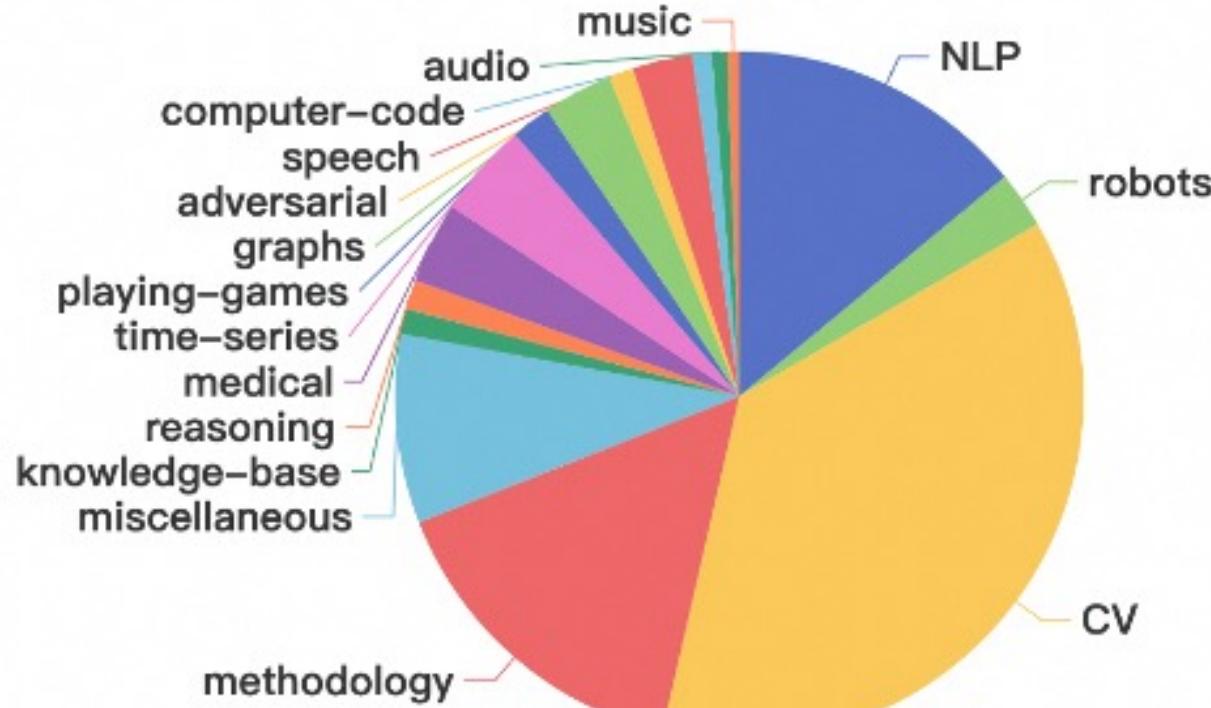
[Context]: Search-augmented instruction learning can help the model better utilize the valuable information ... On the other hand, applying search augmentation to SAIL-7B significantly improves model performance on both experiments (84% to 90% and 98% to 103%).



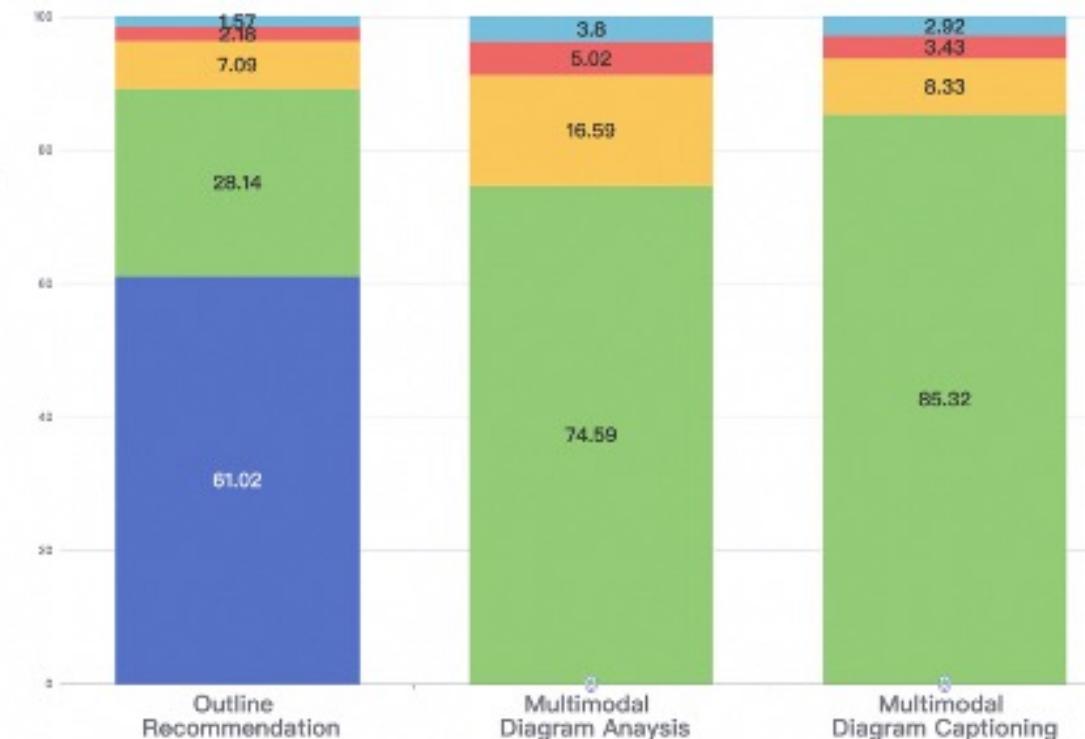
mPLUG-PaperOwl



► mPLUG-PaperOwl



text-only 1 diagram 2 diagrams 3 diagrams 4 diagrams



mPLUG-PaperOwl

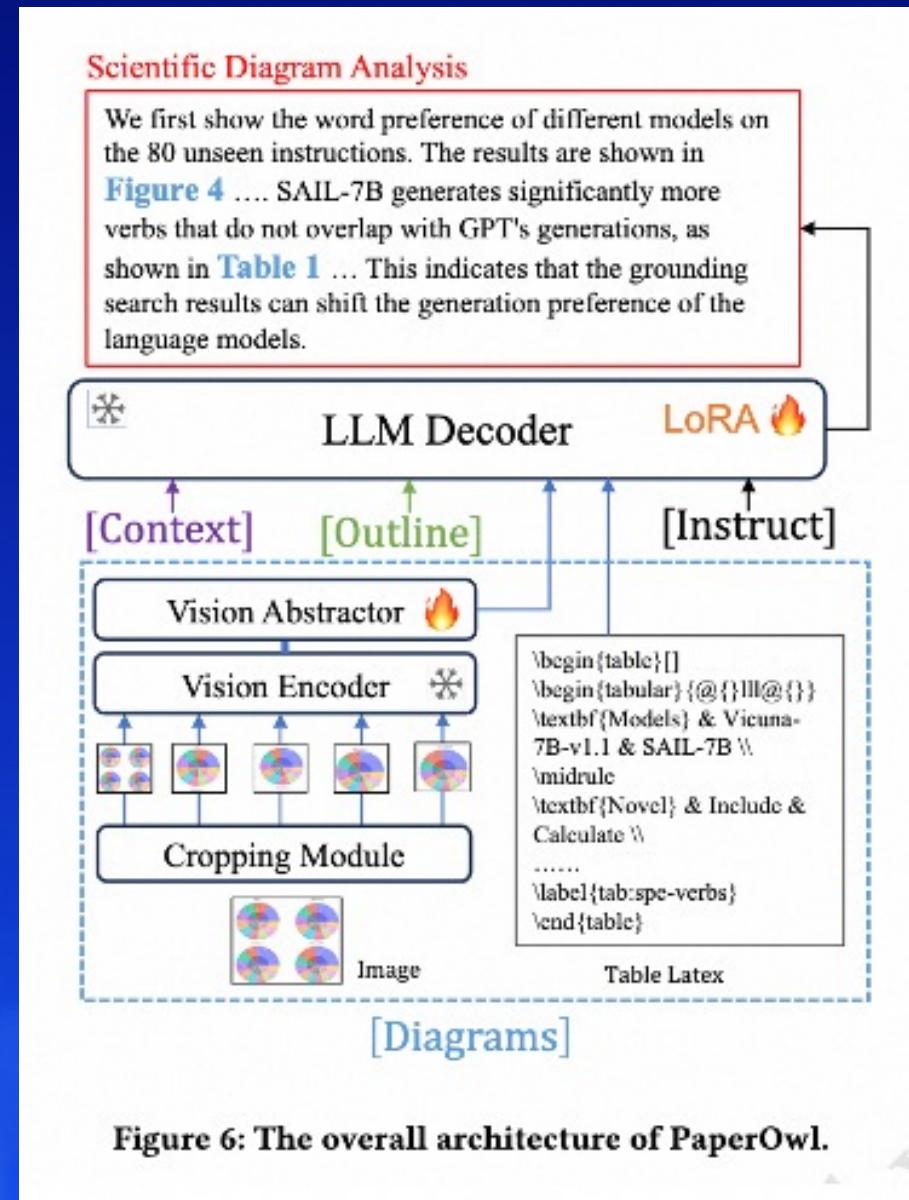
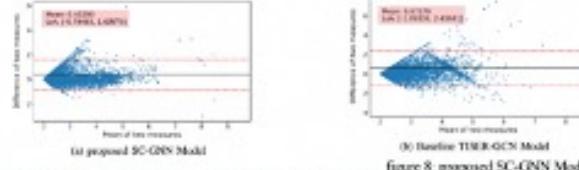


Figure 6: The overall architecture of PaperOwl.

► mPLUG-PaperOwl

[Diagrams]



[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.

[Instruction]: 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

[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 complement the findings in our main paper.

[Instruction]: Describe these images concisely.

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

AnswerOut Additional results for the question-answering task

[Diagrams]

Method	Backbone	mIoU	VC8	VC16
Mask2Former	Swin-L	0.5709	0.8876	0.8596
Mask2Former	BET-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.

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

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

PaperTool : The comparison of different backbones on the PVLM2023 validation set

(a)

Diagrams

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 PokmonSV and FlatstonesSV, 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.

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

GT: Ablation study results for story continuation task on ParrotSY

Powerful : Ablation study on PowerSV dataset for story continuation task

(b)

PART 03

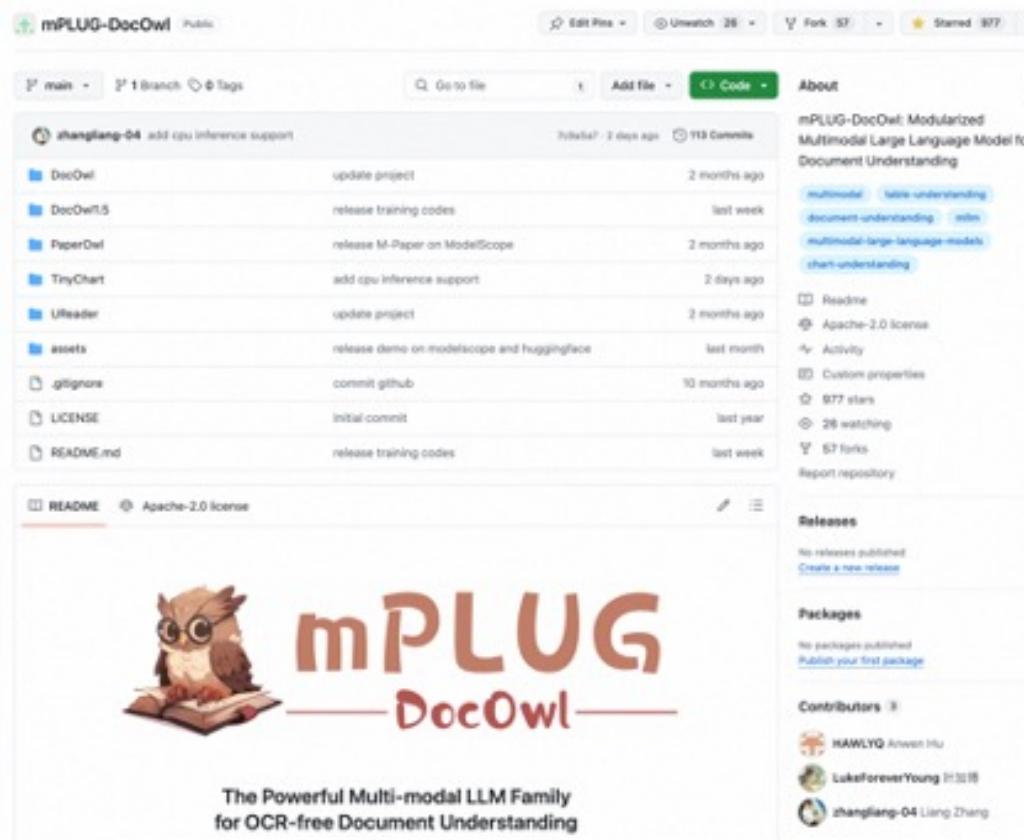
mPLUG-DocOwl开源实战

► 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>



The screenshot shows the GitHub repository page for 'mPLUG-DocOwl'. The repository has 913 commits, 877 forks, and 28 watchers. It features a modularized multimodal Large Language Model for Document Understanding, with support for note-understanding, document-understanding, multi-modal-large-language-models, and chart-understanding. The README file is Apache-2.0 licensed. The repository also includes a Readme, Apache-2.0 license, Activity, Custom properties, 917 stars, 57 forks, and a Report repository. There are no releases published, and users can Create a new release. Packages are available for publishing, and contributors include HAWLYQ, Anwen Hu, LukeForeverYoung, and Liang Zhang.

mPLUG-DocOwl

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

► mPLUG-DocOwl Demo

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

<https://modelscope.cn/studios/jic/mPLUG-DocOwl>

The screenshot shows the ModelScope platform's interface. On the left, there are several examples of documents and tables. In the center, a detailed schedule for workshops and keynotes is displayed. Below the schedule, a text box asks for a detailed explanation of what it is. A large input field contains the schedule text, and a button labeled "Enter text and press ENTER" is visible. At the bottom, there are buttons for "Update", "Download", "Flag", "Regenerate", and "Clear".

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

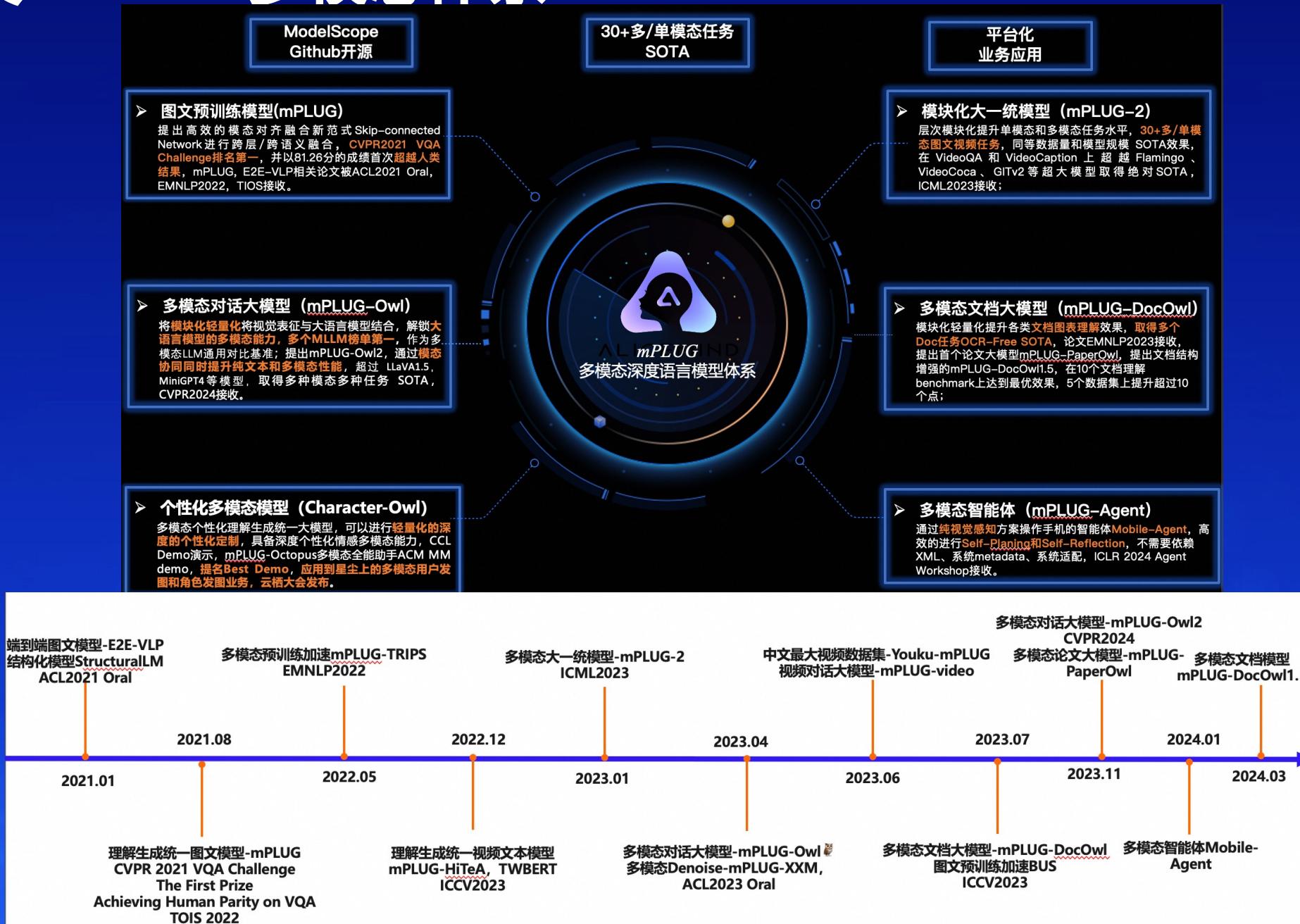
TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning

The screenshot shows the TinyChart demo interface. It features a bar chart with two series: 'Domestic market' (blue bars) and 'Overseas market' (black bars). The x-axis represents years from 2013 to 2017. A text box at the top asks for the highest number of companies in the domestic market. Below the chart, a question is displayed: "What is the highest number of companies in the domestic market? Answer with detailed steps." A code editor window shows the following Python code:

```
# Get the values of 'Domestic market' for each year, set to Y
Y=[23, 25, 24, 27, 30, 24, 30, 25, 27]
# Get the maximum value in Y, set to Answer
Answer=max(Y)
```

The result of the execution is shown as "Execute result: 30.0".

▶ 通义mPLUG多模态体系



▶ 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-understanding

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

ModelScope 首页 模型库 数据集 创空间 文档中心 社区 GitHub 司南评测

搜索你感兴趣的内容 免费试用 登录 / 注册

综合 (54) 创空间 (4) 模型库 (16) 数据集 (1) 动态 (2) 文档 (3) 贴子 (28)

创空间 (4个搜索结果) 查看更多 >



通义实验室
mPLUG-DocOwl
DocOwl1.5: 不依赖OCR的BB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) 查看更多 >



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；



THANKS

