

AppAgent
能够操作任何手机App的Agent模型
张驰

自我介绍

- 南洋理工大学博士毕业，腾讯公司高级研究员
- 博士期间从事计算机视觉和机器学习（小样本、弱监督、长尾学习，终身学习）研究
- 目前就职于腾讯QQ影像中心，从事AIGC领域的研究，专注于大模型解决问题。

提纲

- 通用多模态大模型的研究
- 专注于图表理解的多模态大模型
- 3D+大模型
- 基于多模态大模型的Agent应用

大语言模型的革命



大语言模型with Eyes

User What is funny about this image? Describe it panel by panel. 这张图片有什么好玩的？逐一描述它的板块。



Source: [hmmm \(Reddit\)](#)

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

该图片显示了一个Lightning Cable适配器的包装，有三个面板

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

面板1：一个带有VGA接口（通常用于电脑显示器的蓝色大15针接口）的智能手机。

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

面板2：“Lightning Cable适配器”的包装，上面有一个VGA接口的图片。

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

面板3：VGA连接器的特写，末端是一个小的Lightning连接器。用于给iPhone和其他苹果设备充电）的特写。

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

这张图片的幽默感来自于将一个大的、过时的VGA连接器插入一个小的、现代的智能手机充电端口的荒谬性。

User Can you explain this meme? 能解释一下这张meme吗？

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



这张meme是一个笑话，它结合了两个不相关的东西：从太空看地球的照片和鸡块。

该备忘录的文字表明，下面的图片是从太空拍摄的地球的美丽照片。然而，该图片实际上是鸡块，被排列得隐隐约约像一张世界地图。

这个meme的幽默感来自于文字和图片的意外并置。文字设置了一个对地球雄伟图像的期望，但图像实际上是一些平凡和愚蠢的东西。

GPT-4 This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.

The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.

The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

GPT4-V的应用

Sec. 4.6 Coding Capability with Vision

Prompt: $a^2 + b^2 = c^2$	Prompt: $\sum_{i=1}^N (i^2 + i + 1)^{0.5}$	Prompt: $y = \int f(x, y) dx$
GPT-4V: $a^2 + b^2 = c^2$	GPT-4V: $\sum_{i=0}^N (i^2 + i + 1)^{0.5}$	GPT-4V: $y = \int f(x, y) dx$
<hr/>		
Prompt: $a^2 + b^2 = c^2 + \sum_{i=1}^N (i^2 + i + 1)^{0.5} + y = \int f(x, y) dx$		
GPT-4V: $\P(x \neq 6) = c \times 5 \sum_{i=1}^N (i \times (i + 1)) + y_s \int_a^b f(x, y) dx$		

Multilingual Scene Text Recognition

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is an Arabic phrase "بسم الله الرحمن الرحيم" (Bismillah al-Rahman al-Rahim). It translates to "In the name of Allah, the Most Merciful, the Most Compassionate". This phrase is commonly used by Muslims at the beginning of any action or speech as a way to seek blessings and guidance from God.

Prompt:

What are all the scene text in the image?



GPT-4V:

The scene text in the image is "MUSEE DU LOUVRE" which is the name of the famous art museum in Paris, France.

Prompt:

What are all the scene text in the image?



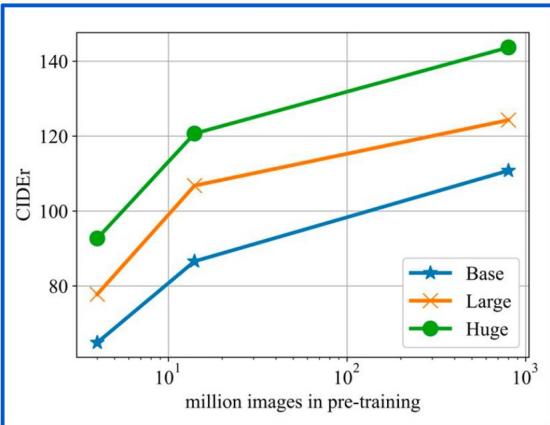
GPT-4V:

The scene text in the image are:
- "尖沙咀" (Chinese characters)
- "Tsim Sha Tsui" (English text)

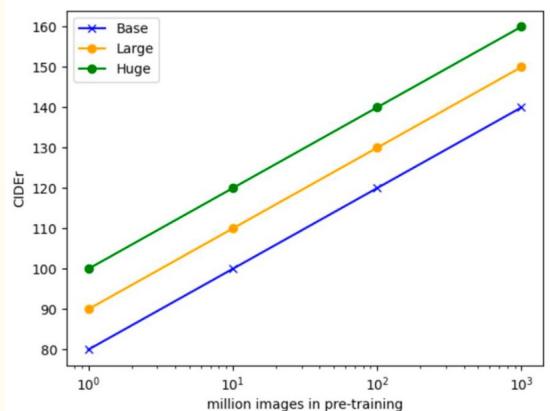
Sec. 4.6 Coding Capability with Vision

Prompt:

Generate python code to draw similar curves.

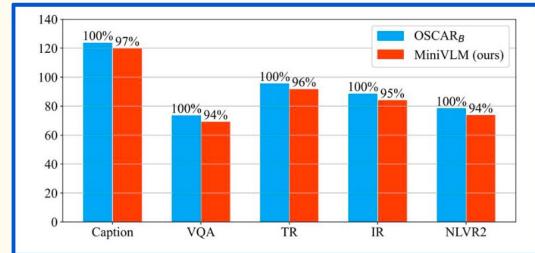


GPT-4V:

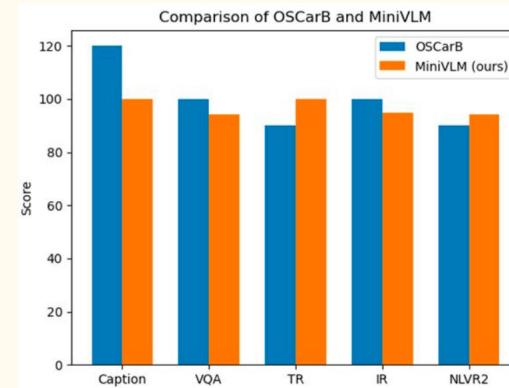


Prompt:

Write python codes to generate similar figures.



GPT-4V:



Prompt:

Write svg codes to generate following image.



Grounded Description (overlaid visual pointers)

Prompt:

Describe the pointed region in the image.

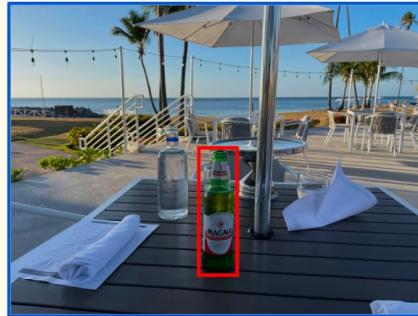


GPT-4V:

The pointed region in the image is a green glass bottle of Magna beer, with a red label and a gold cap. It is sitting on a black table, surrounded by other items such as a water bottle, napkins, and umbrellas.

Prompt:

Describe the pointed region in the image.



GPT-4V:

The pointed region in the image is a green beer bottle with a red label that reads "MAGNA". The bottle is standing upright on a black table with a white napkin next to it.

Prompt:

What is the icon used for?

GPT-4V:

The icon circled in the image is a microphone icon. It is used to activate the voice search feature on Bing. Users can click on this icon and speak into their device's microphone to ask a question or search for something instead of typing it in. This can be helpful for hands-free searching or for users who may have difficulty typing.

Prompt:

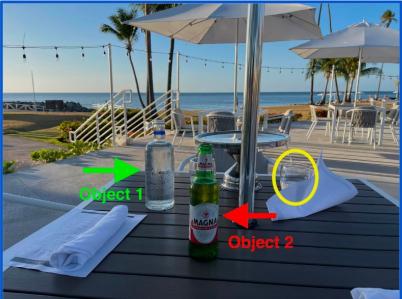
What does the pointed news talk about?

GPT-4V:

The pointed news is talking about the new feature of ChatGPT, which allows users to create personalized instructions. This feature is said to help users better manage their conversations and improve their experience with the chatbot.

(1) Prompt:

What is in the circled glass? Object 1 or object 2?
Describe what is Object 1 and object 2. Then check what is in the circled glass.

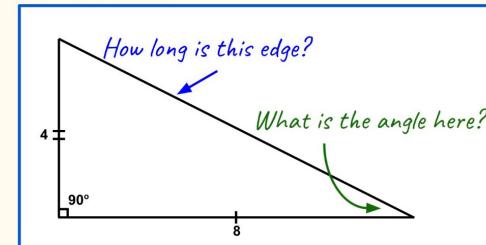


GPT-4V:

Object 1 is a glass bottle of water. It is clear and has a blue label on it.
Object 2 is a glass bottle of Magna beer. It is green and has a red and white label on it.
In the circled glass, it appears to be a clear liquid, likely water. So, it is likely that the contents of the circled glass are from Object 1, the glass bottle of water.

(2) Prompt:

Answer the question in the following figure. Provide intermediate steps.



GPT-4V:

To find the length of the blue arrowed edge, we can use the Pythagorean theorem since we have a right triangle. The theorem states that $a^2 + b^2 = c^2$, where a and b are the two legs of the right triangle and c is the hypotenuse. Plugging in the values we have, we get:
 $4^2 + 8^2 = c^2 \quad 16 + 64 = c^2 \quad 80 = c^2 \quad \sqrt{80} = c \quad 8.94 \approx c$
Therefore, the length of the blue arrowed edge is approximately 8.94.
To find the angle at the green arrow, we can use the tangent function since we have the opposite and adjacent sides of the angle. The tangent function states that $\tan\theta = \text{opposite}/\text{adjacent}$. Plugging in the values we have, we get:
 $\tan\theta = 4/8 \quad \tan\theta = 1/2 \quad \theta = \tan^{-1}(1/2) \quad \theta \approx 26.57^\circ$
Therefore, the angle at the green arrow is approximately 26.57°.

(3) Prompt:

Which circled line shows a different comparison trend?
Provide intermediate steps.

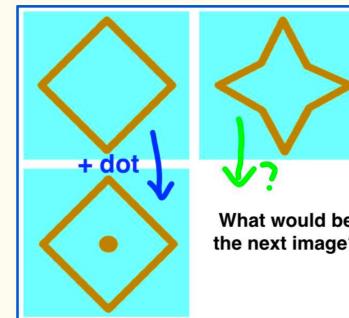
Exam	GPT-4	GPT-4 (no vision)
Uniform Bar Exam (MBE+MEE+MPT) LSAT	298 / 400 (~90th) 163 (~88th)	298 / 400 (~90th) 161 (~83rd)
SAT Evidence-Based Reading & Writing SAT Math	710 / 800 (~93rd)	710 / 800 (~93rd)
Graduate Record Examination (GRE) Quantitative	700 / 800 (~89th)	690 / 800 (~89th)
Graduate Record Examination (GRE) Verbal	163 / 170 (~80th)	157 / 170 (~62nd)
Graduate Record Examination (GRE) Writing	169 / 170 (~99th)	165 / 170 (~96th)
USABO Semifinal Exam 2020 USNCO Local Section Exam 2022	4 / 6 (~54th)	4 / 6 (~54th)
Medical Knowledge Self-Assessment Program Codeforces Rating	87 / 150 (99th - 100th) 36 / 60 75 % 392 (below 5th)	87 / 150 (99th - 100th) 38 / 60 75 % 392 (below 5th)

GPT-4V:

The circled line that shows a different comparison trend is the USNCO Local Section Exam 2022. Intermediate steps:
1. Analyze the data in the table and look for any discrepancies in the comparison trends. 2. Notice that in all the rows except for the USNCO Local Section Exam 2022, the GPT-4 (no vision) column has a lower score than the GPT-4 column. 3. This means that in all the other rows, the GPT-4 model with vision performed better than the GPT-4 model without vision. 4. However, in the USNCO Local Section Exam 2022 row, the GPT-4 (no vision) column has a higher score than the GPT-4 column. 5. This is a different comparison trend than the rest of the rows, indicating that the GPT-4 model without vision performed better than the GPT-4 model with vision in this particular exam.

(4) Prompt:

Answer questions in the following figure. Provide intermediate steps.



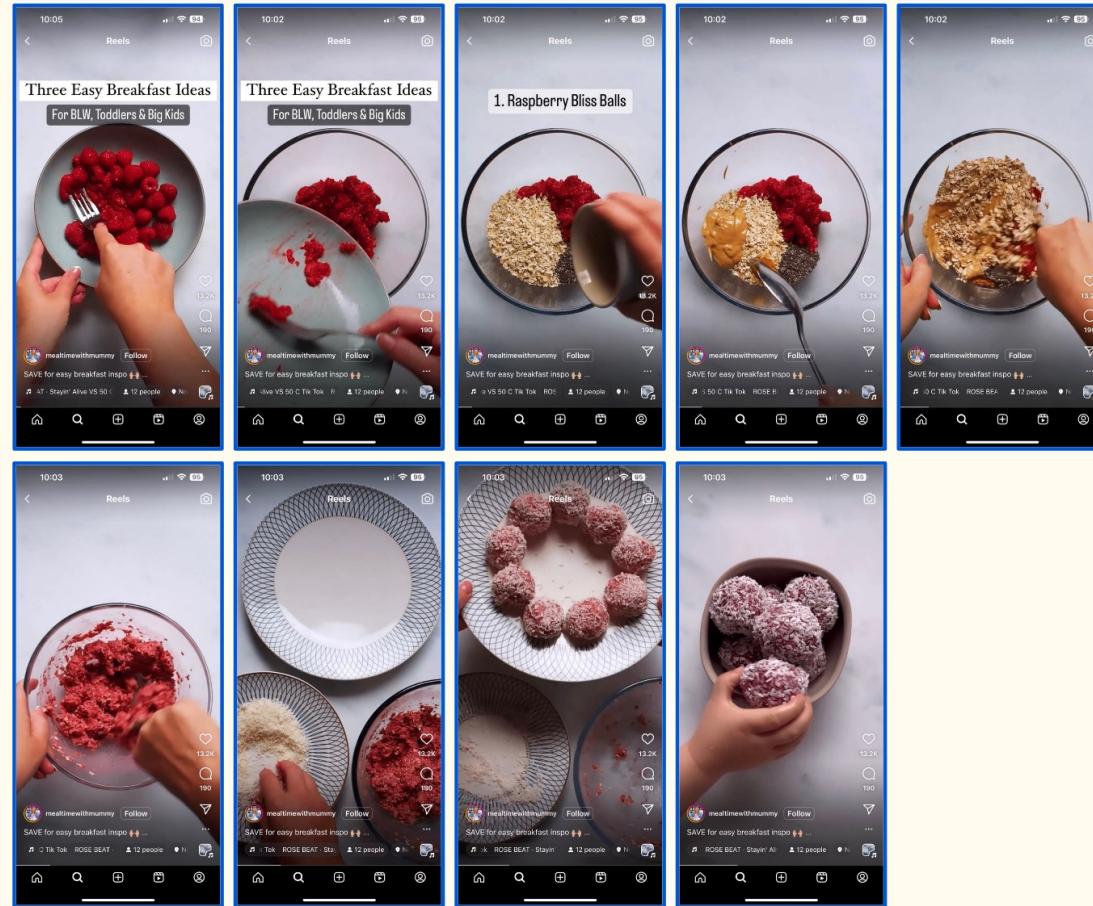
GPT-4V:

The next image in the sequence would be a star with a dot in the center. This is because the sequence appears to be following a pattern where the shape alternates between a diamond and a star, and a dot is added to the center of the shape every other image. Intermediate steps:
- First, we can observe that the sequence alternates between a diamond and a star shape.
- Next, we can see that a dot is added to the center of the shape every other image.
- Based on this pattern, we can predict that the next image will be a star with a dot in the center.

Watching Videos

Prompt:

Tell me about the video content, given the frames below. Please include as many details as possible.



GPT-4V:

The video content is a tutorial on how to make Raspberry Bliss Balls, a healthy breakfast option for toddlers and big kids. The video is posted by the Instagram account "mealtime with mummy" and has over 13k likes.

The video starts with a bowl of fresh raspberries and a fork, which is then mashed into a puree. The next frame shows a larger mixing bowl with oats, chia seeds, and peanut butter being added to the raspberry puree. The mixture is then stirred together until well combined.

The following frame shows a plate with shredded coconut on it, and the raspberry mixture being rolled into small balls and coated in the coconut. The final frames show the finished Raspberry Bliss Balls being placed in a bowl and ready to eat.

Throughout the video, there are text overlays with instructions and tips, making it easy for viewers to follow along and make the recipe themselves.

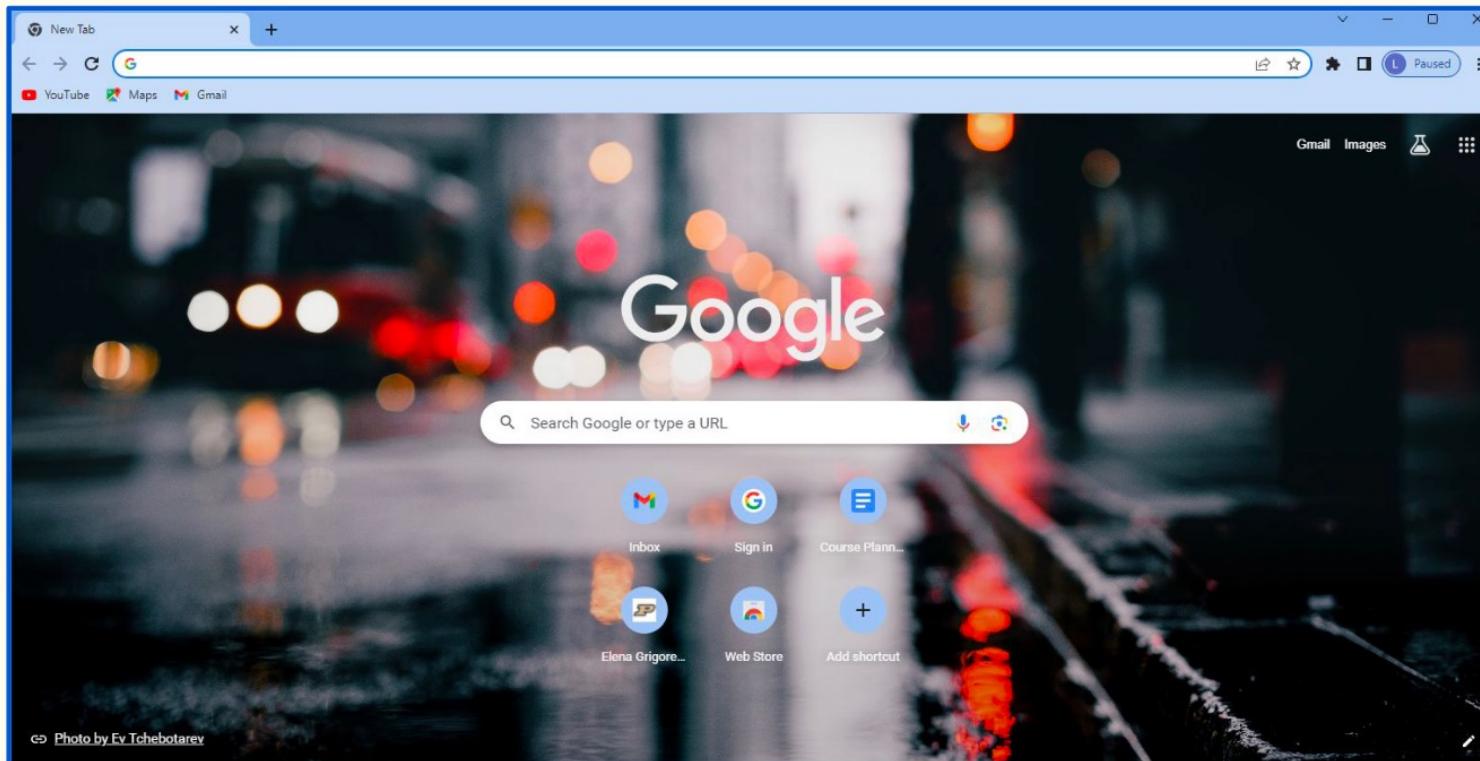
Prompt:

Imagine that you are a robot operating a computer. Like how humans operate the computer, you can move the mouse, click an icon with the mouse, or type some texts with the keyboard.

You are asked to find a detailed recipe of Mapo Tofu.

Previous action: "I would move the mouse to the Google Chrome icon located on the left side of the screen, approximately 1/3 of the way down from the top. I would then click on the icon to open the web browser."

You have now opened the web browser. Below is what you see on the computer screen. What will you do next? If the action involves moving the mouse, describe the location as detailed as possible.



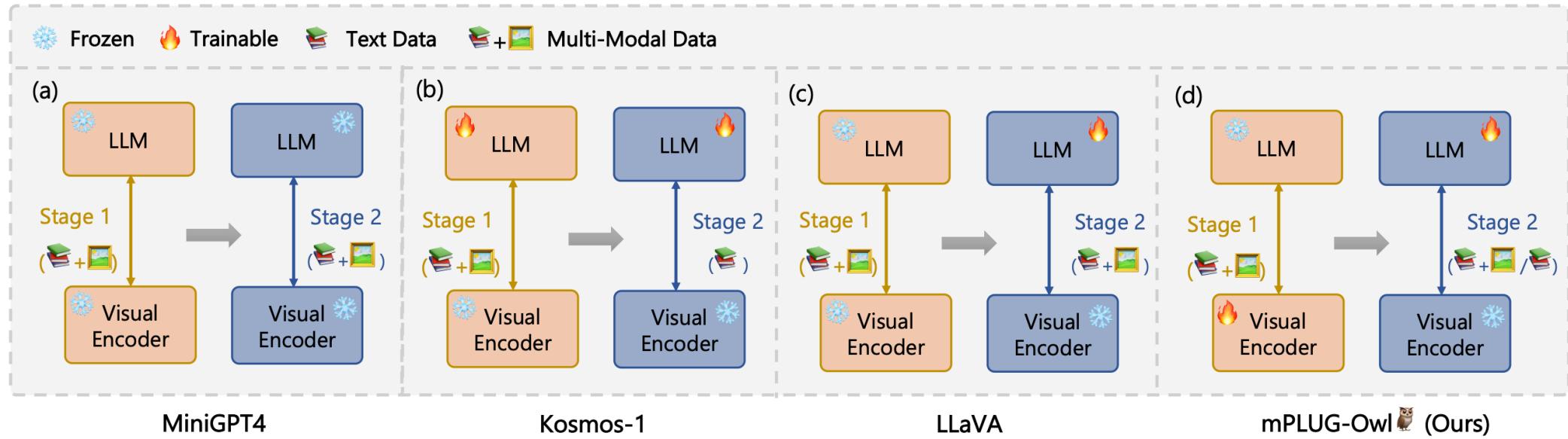
GPT-4V:

Next, I would move the mouse to the search bar located in the center of the screen, approximately 1/2 of the way down from the top. I would then click on the search bar and type in "Mapo Tofu recipe" using the keyboard. After typing in the search query, I would press the "Enter" key on the keyboard to initiate the search.

现状与挑战

- GPT4-V架构方案未知
- GPT4-V的训练数据未知
- GPT4-V过于强大，强到你不知道从何下手

开源工作



难点与挑战

- 如何高效灵活广泛的收集数据集？

 图文align是一件大工程， 数据集很noisy， 想做强数据就不得不做广

- 如何构造更多样的instruction？

 基于Benchmark 标注来构造提问回答很受限

- 专业领域的图如何处理？

 例如图表， 介于文字和图的中间状态如何处理

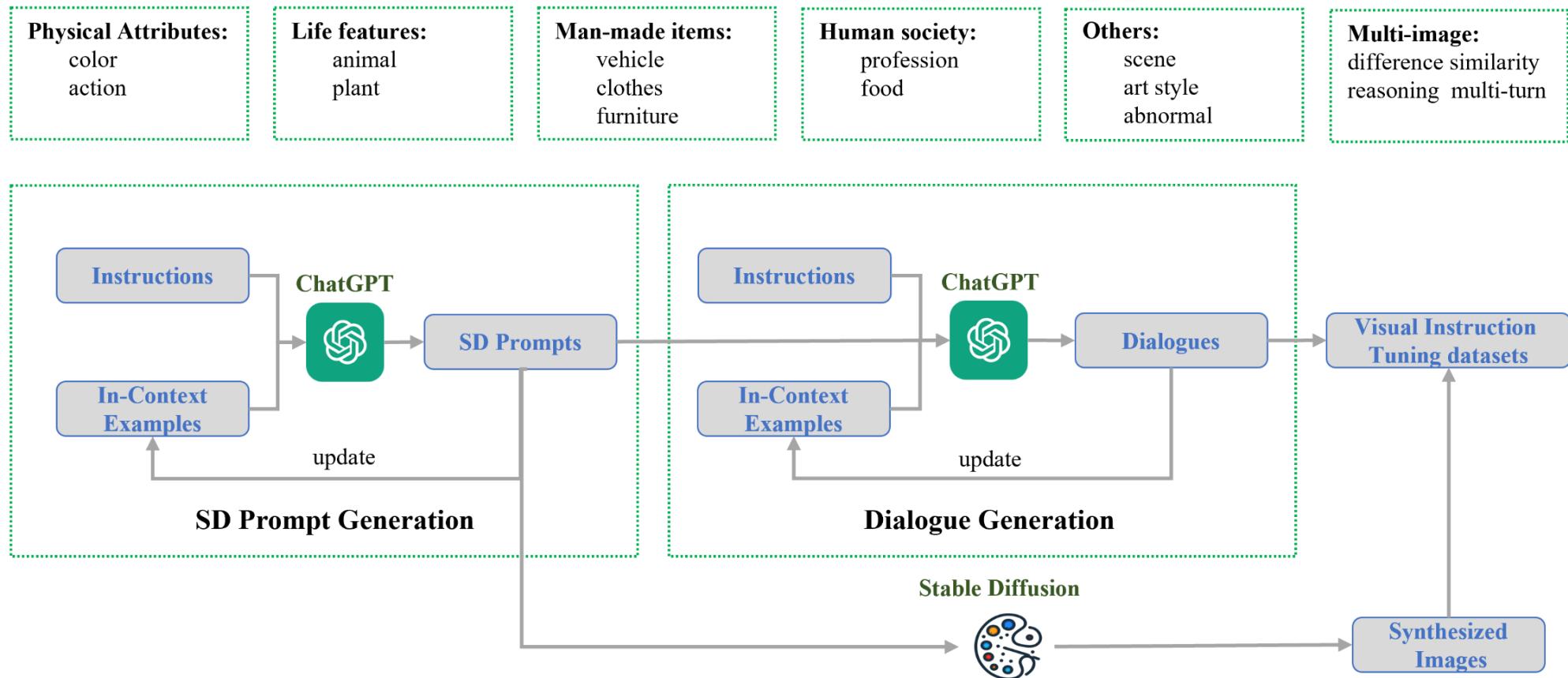
Motivation1 :

- 更灵活高效的构造数据，增强一个特别需要的能力（例如风格图像下的内容理解）
- 构造一些不太常见的问题（比如GPT4展示的玩梗图像理解）
- 构造更贴近现实场景的提问形式（多图，多问混合）

StableLLaVA

- 同时生成图像和instruction对话
- 由于图像是自己生成的， 内容已知且可控， 可以构造对话
- 主题多样可控， 形式丰富 （多图）
- 数据集规模可控

StableLLaVA





■ Q: "Can you teach me how to prepare a classic spaghetti dish? This is a picture of the ingredients I have."

■ A: "Great! First, boil a pot of water and add spaghetti to cook."



■ Q: "Oaky, the water is boiling and I added the spaghetti. What should I do next?"

■ A: "Add the chopped tomatoes to the pan until the tomatoes have softened and created a sauce. Then add salt and pepper."



■ Q: "The tomatoes have softened and I added salt and pepper. What is next?"

■ A: "Once the spaghetti is cooked, drain it and add it to the pan with tomato sauce. And let it cook for a few more minutes."



■ Q: "Okay, I've cooked it, here's the spaghetti."

■ A: "Great! You did a good job."

■ Q: "What stands out as unusual in this image?"



■ A: "Elephants, being wild animals, don't have the need or capability to wear graduation gowns and hold diplomas, making this image both surreal and abnormal."

■ Q: "What profession does the person in the image represent?"



■ A: "Carpenter. As he is in a woodshop surrounded by woodworking tools and wood shavings."



■ Q: "Is there any inference to be made concerning the link between two images?"

■ A: "Watering the flowers with the garden's watering can makes the flowers bloom beautifully."








...

Figure 1. **Examples of synthesized visual instruction data.** We use ChatGPT and text-to-image generation models to synthesize various forms of visual instruction tuning data, such as multi-round dialogue data, multi-image reasoning data, and anomaly detection data. These data are used to train the multimodal large language models.

StableLLaVA

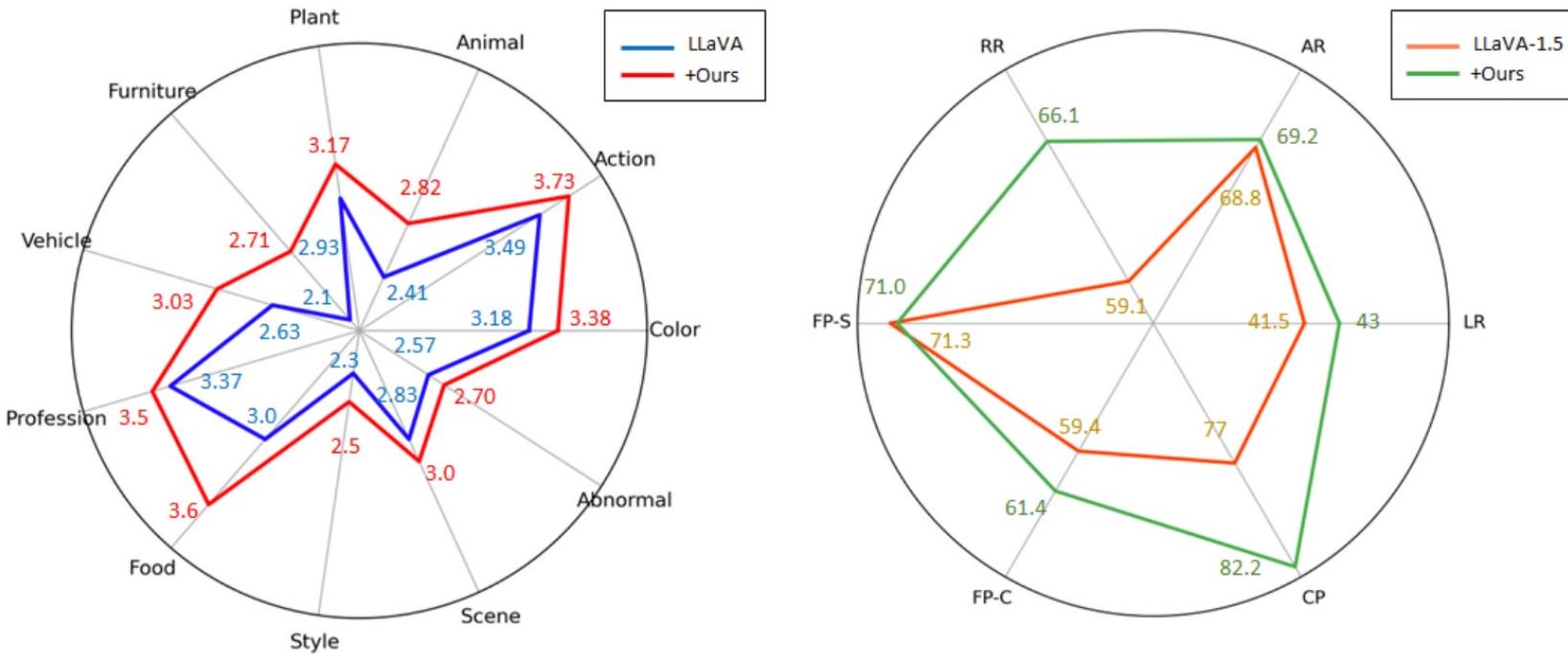


Figure 5. Left: Results on evaluation benchmarks for various abilities. Right: Comparison of various subcategories on MMBench [22] with the baseline. Our model outperforms the baselines on both benchmark datasets.

Table 1. Quantitative comparison with other state-of-the-arts methods on multiple multimodal benchmarks. We achieve state-of-the-art performance on four benchmarks.

Method	VisWiz	MM-Vet	MME	MMB
BLIP2 [16]	19.6	22.4	1293.8	-
InstructBLIP [7]	33.4	25.6	1212.8	-
IDEFICS-9B [12]	35.5	-	-	48.2
IDEFICS-80B	36.0	-	-	54.5
Qwen-VL [2]	35.2	-	-	38.2
Qwen-VL-Chat	38.9	-	1487.5	60.6
LLaVA-1.5 [20]	53.6	35.4	1531.3	67.7
Ours	58.4	36.1	1532.3	69.1

Table 2. Quantitative comparison on the multi-image benchmark. After the addition of multi-image data, various multi-image capabilities have significantly improved.

Method	Difference	Similarity	Logical relations	Average
LLaVA [21]	2.7	2.2	3.1	2.67
Ours	3.6	2.8	3.7	3.37

未来的可能性

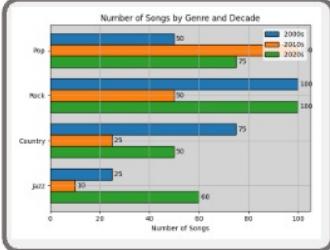
- 融入更准确的生图算法（例如DALLE3），构造更准确灵活的指令
- 融入更高级的生图算法，例如能指定空间位置的生图算法，构造更多样的指令
- 任何文本（代码，自然语言）产生图像（自然图像，特殊图像）都可以反过来构造图文instruction的数据集

Motivation2:

- 更贴近生产力场景的图文理解，关注特殊形式的图
- 目前GPT可以读PDF中的文字，但是无法连图表一起图与理解
- 图表的理解，代码生成，内容编辑

Chart LLama

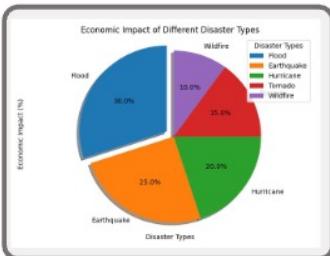
Q&A



Q: What is the number of Pop songs in the 2000s according to the chart?

A: 50

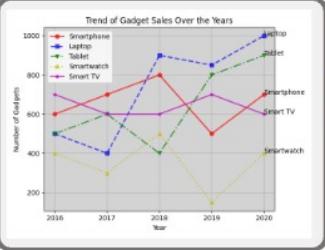
Chart extraction



Q: Convert the figure into table.

Disaster Type	Economic Impact (%)
Flood	30.0%
Earthquake	15.0%
Hurricane	20.0%
Tornado	15.0%
Wildfire	10.0%

Chart description



Q: Please summarize the chart with text.

A: The chart presents data on the popularity of various gadgets from 2016 to...

Chart editing

Q: Change the background of the figure to white and remove the grid lines.



Q: Change the color of the image by setting a different color for each category of bars.

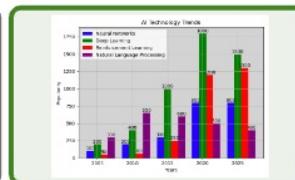
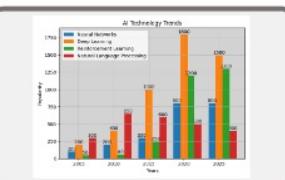
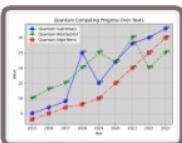


Chart-to-chart

Q: Derender the given image to Python code using matplotlib.

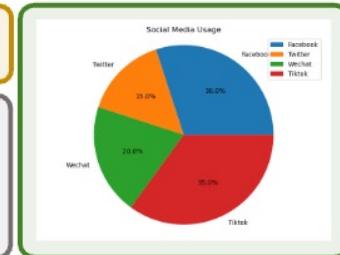


A:
import ...
ax=plt.subplots(figsize=(10, 6))
ax.plot(x, y,
label=column, linewidth=2,
marker='o', markersize=10,
alpha=0.7)
for i,txt in enumerate(df['2020']):...

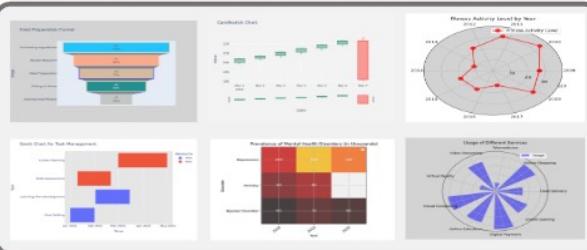
Text-to-chart

Q: Convert the figure into pie table.

Facebook	30.0%
Twitter	15.0%
Wechat	20.0%
Tiktok	35.0%



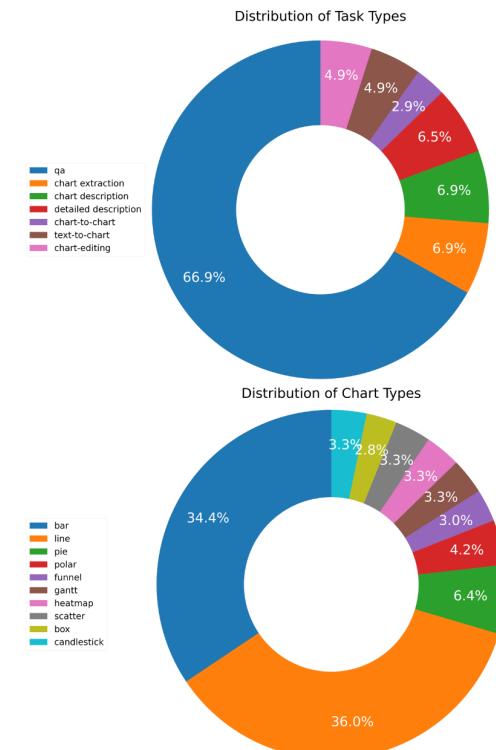
More chart types and abilities...



数据集统计信息

Datasets	#Chart type	#Chart figure	#Instruction tuning data	#Task type
Chartqa	3	21.9K	32.7K	1
PlotQA	3	224K	28M	1
Chart-to-text	6	44K	44K	1
Unichart	3	627K	7M	3
StructChart	3	9K	9K	1
ChartLlama	10	11K	160K	7

Table 1. **Dataset statistics.** Thanks to the flexibility of our data construction method, our proposed dataset supports a wider range of chart types and tasks. We can generate more diverse instruction-tuning data based on specific requirements.



Stage 1: Chart Data Generation

Input

Theme: Global average temperature, Daily traffic, ...

Trend: Rapid increase, Slow increase, ...

...

Output

Detailed descriptions about data: the chart presents the variation in forest cover over time, specifically for the Amazon Rainforest and the Siberian Taiga. ... showcases the irregular fluctuations and sudden drops in forest coverage for ...

Raw Data:

Year	Amazon	Siberian
2010	500	200
2011	600	300
...

Stage 2: Chart Figure Generation

Input

In context examples:

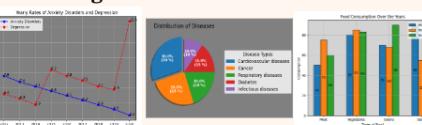
Raw data: tabular data from Stage 1.

...

Output

Detailed descriptions about charts: ...the plot has labels for x and y axis as 'Year' and 'Area (Square Kilometers)', respectively, and the title of the plot is 'Comparison of Amazon Rainforest and Siberian Taiga Area'. A legend is placed at the upper right corner...

Generated figures:



Stage 3: Instruction Data Generation

Input

The descriptions: The chart presents the variation in...

The raw data: Year, Amazon, Siberian\n2010, 500...

...

Instruction tuning data

Q1: What is the number of Pop songs in the 2000s according to the chart? **A1:** 50

Q2: From the chart, can we infer any potential reasons for the more significant reduction in forest coverage? **A2:** It could...

Q3: Extract the raw data from the given chart. **A3:** ...

Q4: Redraw the given chart figure. **A4:** ...

Q5: Draw a funnel chart based on given raw data. **A5:** ...

Q6: Remove the grids in the given chart figure. **A6:** ...



Abilities: Q&A, Chart Descriptions, ...

Figure 3. Pipeline of our data generation method. The innovative data generation process we proposed consists of three important steps relying on GPT-4. The dataset generated using this process exhibits significant advantages compared to previous datasets in terms of data diversity, quality, the number of chart types, and the variety of tasks. ChartLlama, which is trained on this dataset, has the ability to perform various tasks based on the design of the instruction-tuning data.

实验结果

- 更好的结果，更少的数据，数据量可以随意扩展

Method	Chartqa			Chart-to-text		Chart extraction (human)		Chart extraction (augmented)	
	Human	Augmented	Average	Pew	Statista	Precision	F1	precision	F1
Pix2struct [8]	30.50	81.60	56.00	10.30	38.00	—	—	—	—
Matcha [14]	38.20	90.20	64.20	12.20	39.40	—	—	—	—
DePlot [13]	—	—	—	—	—	81.32	81.15	93.42	93.29
Unichart [19]	43.92	88.56	66.24	12.48	38.21	61.51	35.20	79.59	70.21
Baseline* [16]	37.68	72.96	55.32	7.16	24.65	53.48	48.39	55.17	49.50
ChartLlama	48.96	90.36	69.66	14.23	40.71	84.92	84.89	94.94	94.78

Table 2. **Results on traditional tasks.** We compare our work with the previous three open-source models and also compare it with Baseline* trained on the training split of respective benchmarks.

Method	Detailed Description	Chart-to-chart		Text-to-chart		Chart-editing		Chart-to-text	
		GPT Score	Success Rate (%)	GPT Score	Success Rate (%)	GPT Score	Success Rate (%)	Pew	Statista
LLaVA-1.5 [16]	67.2	64.8	46	62.2	77	51.6	38	65.8	73.4
ChartLlama	74.2	74.4	73	81.6	81	75.6	71	81.0	92.6

Table 3. **Results on new tasks.** We primarily compared our work with the baseline model LLaVA-1.5. For the proposed new task, we used GPT for evaluation and validated the effectiveness of our proposed dataset. Evaluation of Chart-to-text using ChatGPT is also listed.

3D + LLM

■ Q: "I am interested in the object in this image <Image>, and I speculate that it is <a modern looking chair with black seat and back rest with grey metal arms and 3 legs>. Please generate a model so I can see what the back looks like.?"



■ A: <Shape>

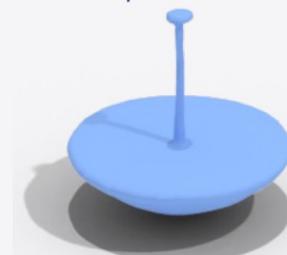


ShapeGPT

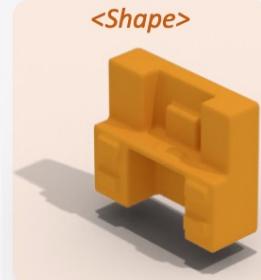
■ Q: "Wow, this image <Image> looks amazing! I wonder how it would look in 3D space."



■ A: <Shape>

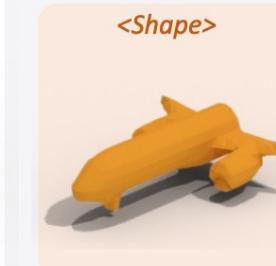


■ Q: "My friend sent me a 3D model <Shape>, can you help me identify what it is."

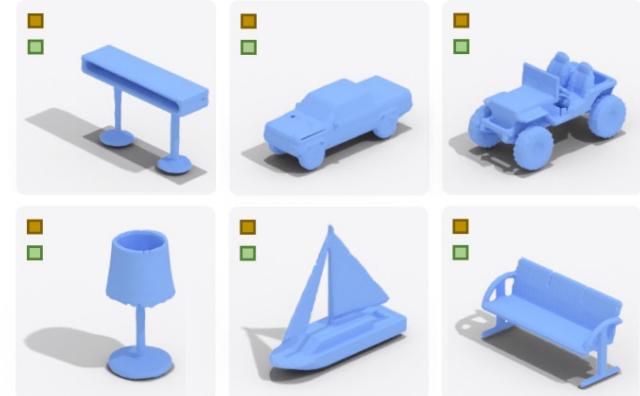


■ A: A brown wooden desk with a small shelf on the top and a small drawer on the bottom.

■ Q: "What a pity! This shape <Shape> is half damaged. Can you help me complete it?"



■ A: <Shape>



ShapeGPT: 3D Shape Generation with A Unified Multi-modal Language Model

3D + LLM

3D Dense Caption

What is this object? [Click]

The couch is located at <obj>229, 142, 30, 38, 63, 53</obj>.

This is a light brown couch, it is situated between two gray ottomans. It is under a big painting. The painting is gray, yellow and brown.

3D Dialogue

Can you please move the picture on the wall?

I have arrived at the picture on the wall near the couch. What would you like me to do with it?

Can you move it to the other wall, near the table?

Sure! I have moved the picture to the other wall near the table. Is there anything else you need me to do?

3D QA

What is behind the ottoman, that is next to the love seat?

Trash Can

What is behind the ottoman, that is next to the love seat? [Click]

CLICK

ottoman

Table is located behind ottoman that is next to love seat

Scene Description

Describe this scene.

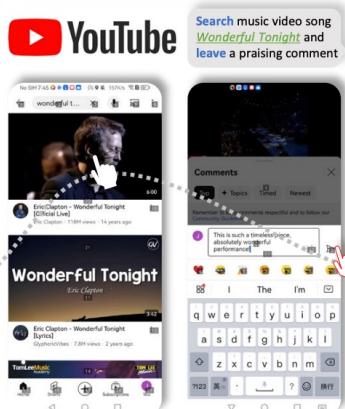
The room contains various couches and armchairs, providing comfortable seating options. Additionally, there are plants placed in different areas of the room, adding a touch of greenery. Lastly, a picture is hung on one of the walls.

LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning

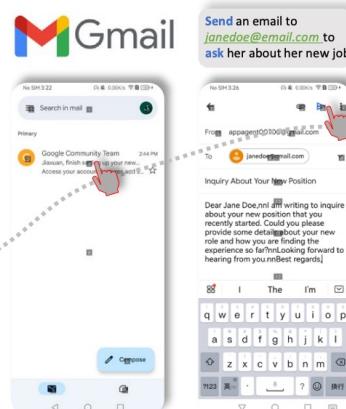
图文大模型的应用AppAgent



Observation: The screenshot shows the Telegram app, a conversation from "Agent Liu" with a message...
Thought: To complete this task, I should input the text with this answer into the input field labeled with the tag "1".
Action: text("The Toronto Raptors won the 2019 NBA....")



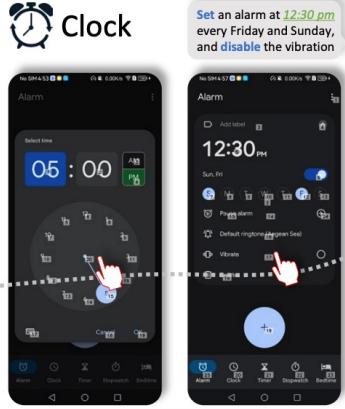
Observation: The screenshot shows a YouTube search result for "Wonderful Tonight" ...
Thought: To complete this task, I should insert a praising comment into the text input field labeled '2'.
Action: text("This is such a timeless piece...")



Observation: The screenshot shows a smartphone's email application ...
Thought: To complete this task, I should type the content of the email asking Jane Doe about her new job.
Action: text("Dear Jane Doe,\n\nI am writing to ...")



Observation: The screenshot shows a photo editing app on a smartphone with an image displayed on the screen...
Thought: To complete this task, I should adjust the noise reduction level using the slider.
Action: Action: swipe(23, "right", "medium")



Observation: The image shows a screenshot of an alarm application on a smartphone...
Thought: To complete this task, I should disable the vibration feature by tapping on the UI element.
Action: tap(17)



Observation: The screenshot displays a shopping app with a variety of headsets listed...
Thought: To complete this task, I should add the selected headset into my shopping cart.
Action: tap(14)

Deployment Phase

Doc Source: Human Demonstration

APP: Gmail

**Task: Send an email to johndoe@email.com
to ask him about his recent trip to Las Vegas**

Deployment Phase

Doc Source: Human Demonstration

APP: Adobe Lightroom

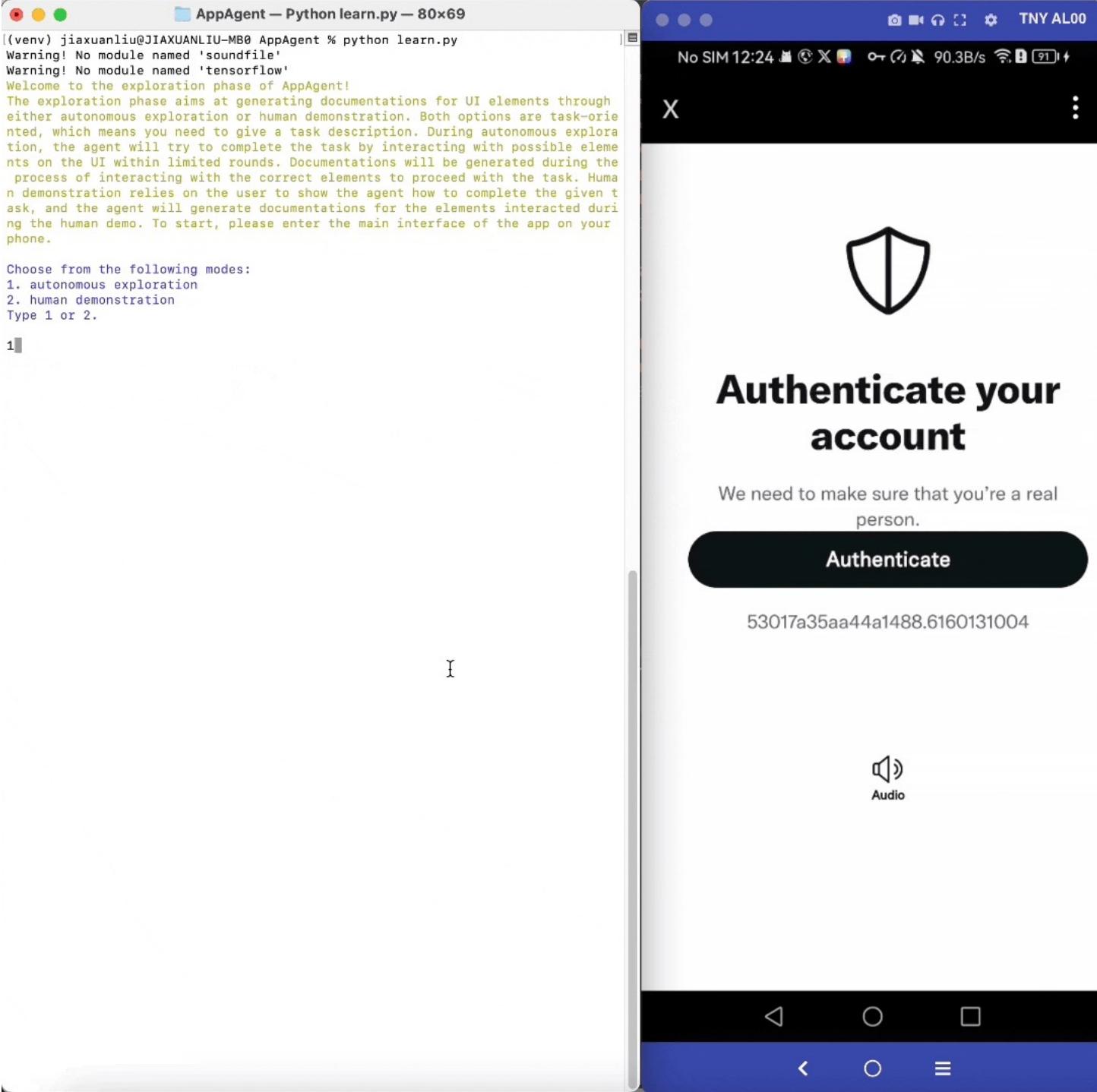
Task: Improve this image until it looks good to you

Deployment Phase

Doc Source: Autonomous Exploration

APP: X

Task: Search for the user Bill Gates and follow him



挑战1：如何高效的输出控制？

- 极端的情况可以类比成机器人的控制
- 手机的控制也不是很容易（参数复杂）

挑战2-真的有可能做到泛化吗？

- 人也无法做到泛化
- 人对App有充分的先验
- 人虽然不能泛化但是可以快速学习使用一个App

挑战3：什么样的学习是好的学习？

- 有先验下的简单学习（逻辑类似，界面不复杂的App）
- 无先验下的学习，复杂学习（专业软件）

目标

- 操作App，完成复杂的，high-level的任务（不需接入App后台）
- 泛化性，能够拓展到不同的App
- 高效的训练与部署范式

Search in mail 1

J 3

Primary



Google Community Team

Dec 12

Jiaxuan, finish setting up your new...

Access your account features and s... ☆

实验环境

- 安卓平台
- Command line interface (CLI) 控制手机
- XML解析出所有可交互的元素，并为每个元素分配一个独立的ID
- 在UI上为每个交互元素动态标记一个数字，覆盖在截屏上

8

Compose 6



9



7

<

O

□

高效简洁的控制方式-简化的action space

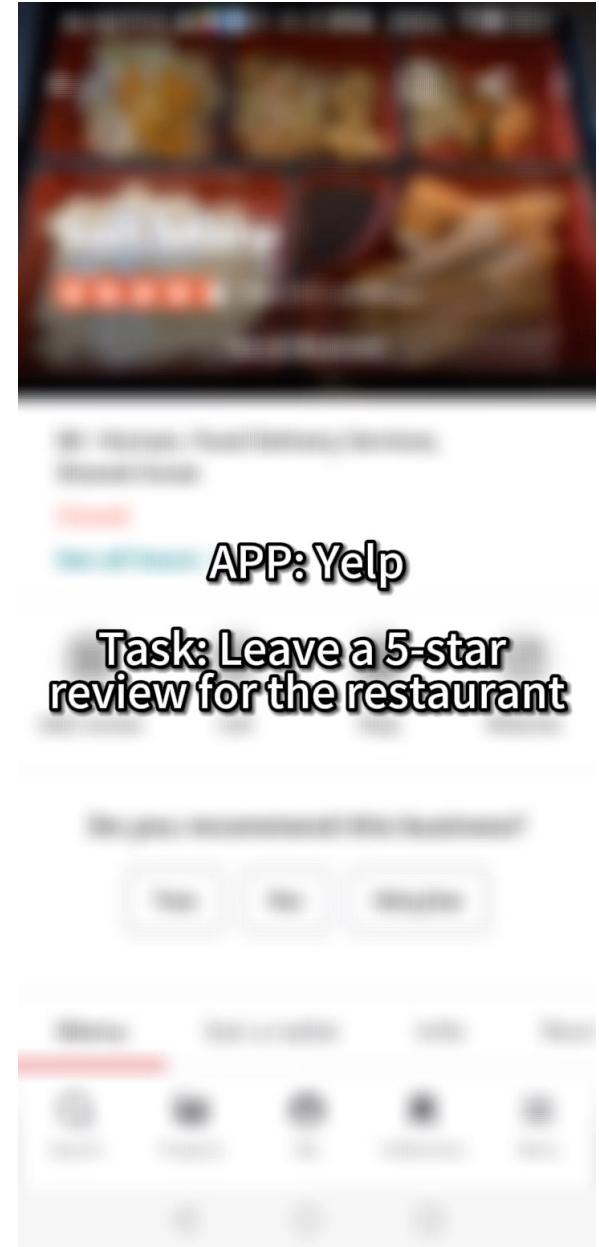
- **Tap(element : int)** : This function simulates a tap on the UI element numbered on the screen. For example, tap(5) would tap the element labeled ‘5’ .
- **Long_press(element : int)** : This function emulates a long press (for 1 second) on a UI element.
- **Swipe (element : int, direction : str, dist : str)**: It allows the agent to swipe on an element in a specified direction (up, down, left, right) and distance (short, medium, long). For instance, swipe(21, “up” , “medium”) would swipe up on element ‘21’ for a medium distance.
- **Text(text : str)** : To bypass inefficient virtual keyboard typing, this function inputs text directly into an input field when a virtual keyboard is visible. For example, text(“Hello, world!”) inputs the string “Hello, world!”.
- **Back()** : A system-level function that helps the agent return to the previous UI page, especially useful for exiting irrelevant pages.
- **Exit()** : A specialized function is employed to conclude processes, typically invoked upon successful task completion.

输出控制

- 原本的CLI控制需要精确的坐标
- GPT4并不擅长预测坐标
- 原手势操作参数过于复杂

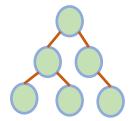
Update after Arxiv submission

- 操作element之外的区域
- 增加“唤出网格”函数，通过选择网格来实现其他区域的控制

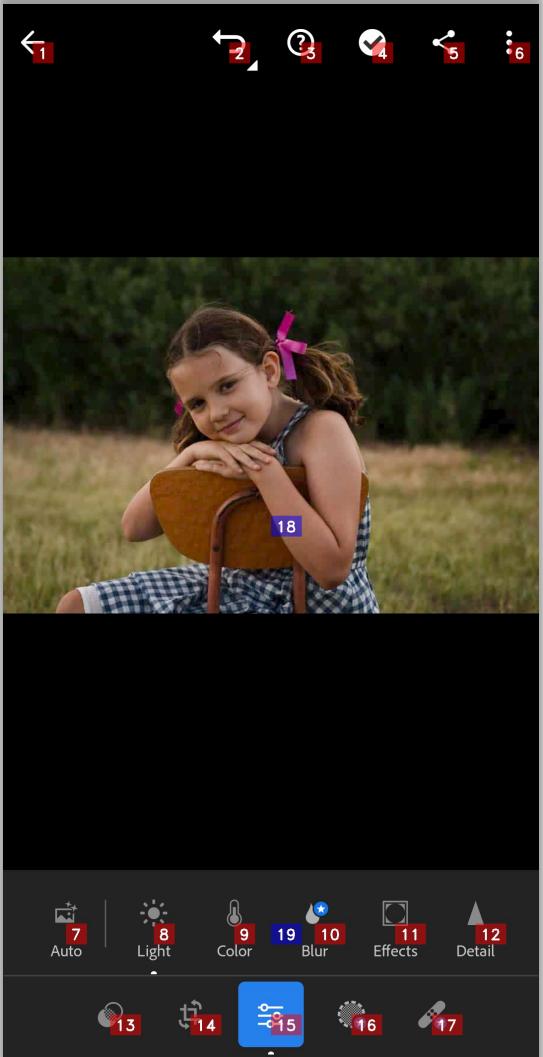


总体思路：探索阶段->部署阶段

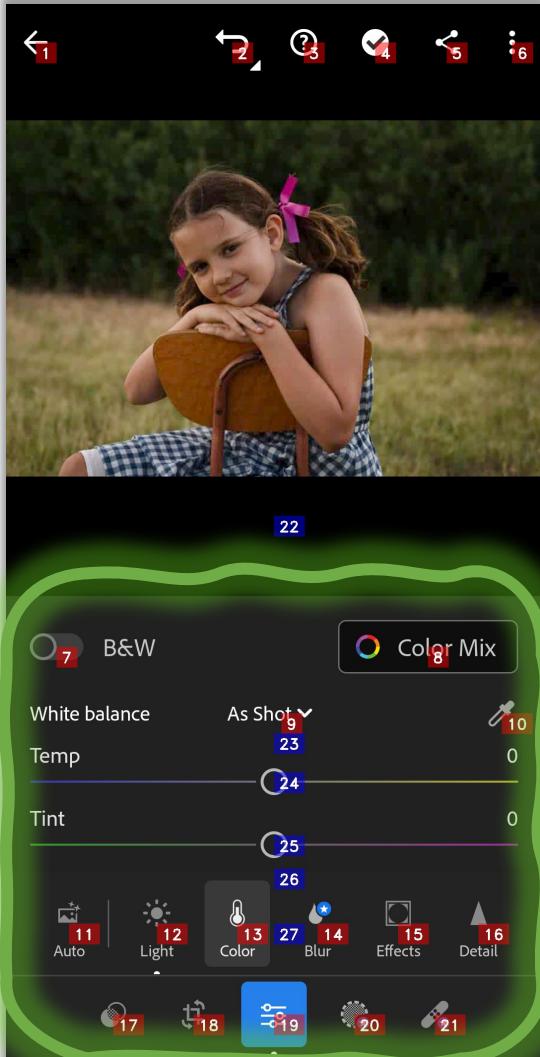
- 探索by自助探索
- 探索by观察人类演示
- 探索完生成一份app文档
- 部署的时候动态调用每个页面的文档



Exploration Phase

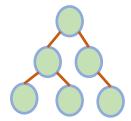


Tap

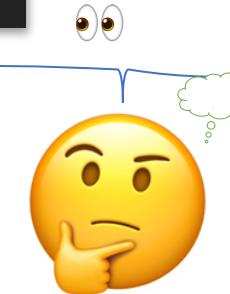
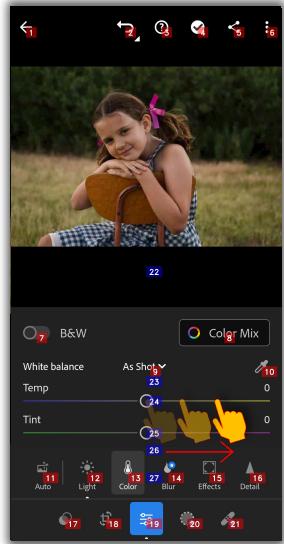


"Tapping this UI element opens the color settings to adjust the color of the image."

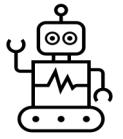




Exploration Phase



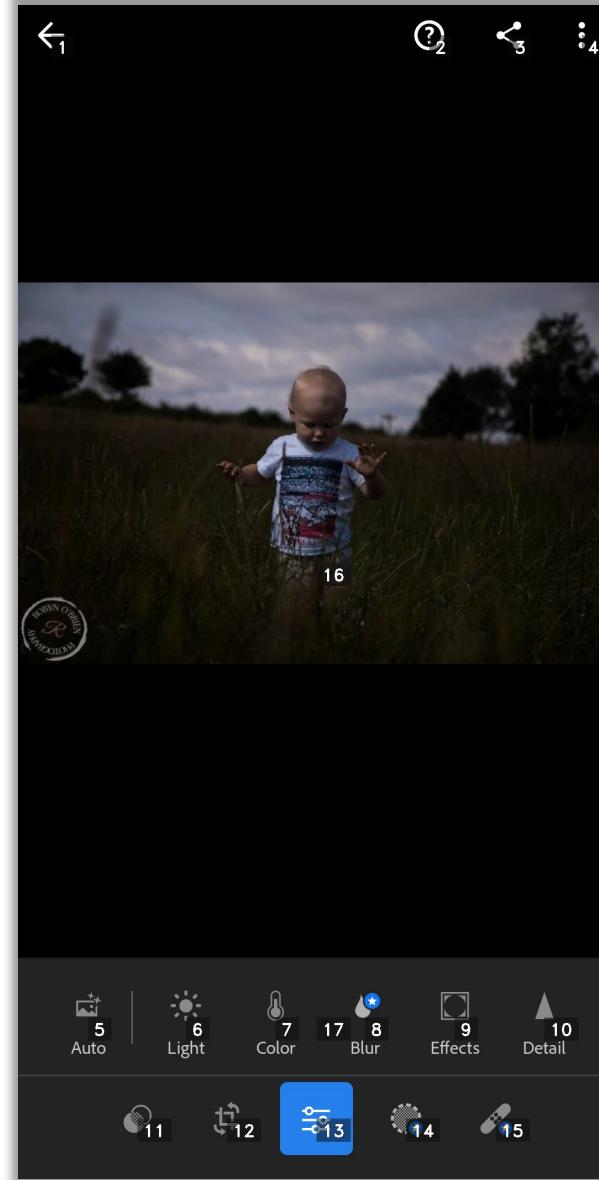
Doc for App

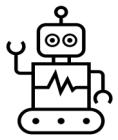


Deployment Phase

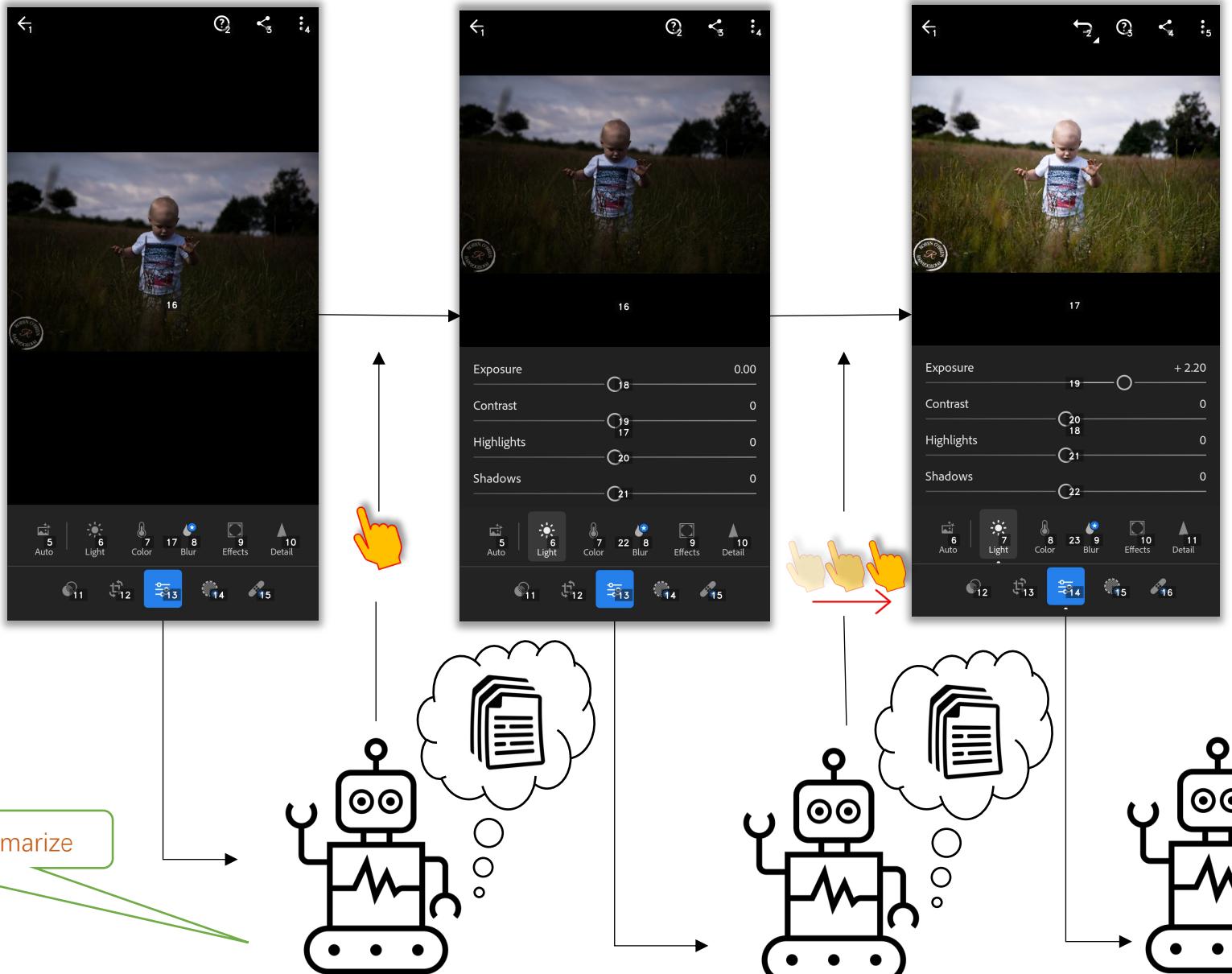
Task:

“beautify this photo for me....”



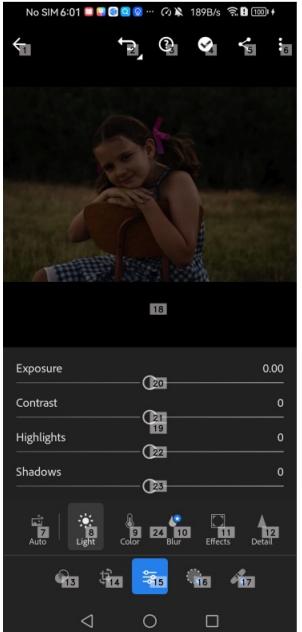


Deployment Phase



Observe, Think, Act, and Summarize

Lightroom: improve the quality of this image until it looks good to you



Observation

The screenshot shows a photo editing application interface with various adjustment settings for **exposure, contrast, highlights, and shadows**. These adjustments have **numeric tags 20, 21, 22, and 23** respectively. The settings are currently at their default values, and the image appears **underexposed**, needing improvement in lighting.

Thought

To improve the image quality and **correct the underexposure**, I need to **increase the exposure**. The slider for the exposure setting is labeled with **numeric tag 20**, which I can interact with to adjust the exposure level. I will **swipe right on this slider to increase the exposure**.

Action

```
swipe(20, "right", "medium")
```



Observation

The **exposure level has been increased**, as indicated by the slider tagged with number 20 being to the right. The **contrast, highlights, and shadows are yet to be adjusted**, with their **sliders at the neutral position**. There are also other adjustment options available, such as color and detail, which could further improve the image quality if needed.

Thought

The next step to improve the quality of the image could be to **adjust the contrast**, as increasing contrast can often **make an image appear more vivid** and can complement the increased exposure. Based on the documentation, **interacting with UI element labeled 21 will adjust the contrast level of the image**.

Action

```
swipe(21, "right", "short")
```



Observation

The exposure and contrast sliders are visible, with the **exposure increased to +2.20** and the **contrast increased to +24**. The **highlights and shadows sliders are at 0**, indicating **no adjustment has been made to them**.

Thought

Since I have previously increased the exposure and adjusted the contrast slightly, the next step could be to **adjust the highlights or shadows** to further improve the image quality. **Adjusting the highlights** might help in **bringing out more detail in the brighter parts** of the image without affecting the overall exposure too much.

Action

```
tap(22)
```

实验

- 50个任务，
10个App

Method	Document	Action Space	SR ↑	Reward ↑	Avg. Steps
GPT4 (Baseline)	None	Raw	2.2%	0.6	4.0
	None	Ours	48.9%	3.5	6.9
AppAgent	Auto. Exploration	Ours	73.3%	5.1	4.4
	Watching Demos	Ours	84.4%	4.7	5.1
	Manually Crafted	Ours	95.6%	5.5	5.5

本地模型能力适配

- 多图，长context，理解能力
- OCR能力（多语言）
- Visual pointing的能力（通过网格，数字指代图像元素）
- 调用函数、工具的能力
- In-context能力（格式跟随）
- App/UI的先验
- 特殊图像内容判断能力（P图）

应用探索

- 软件开发测试
- 车载操作系统
- App内部的helper
- 专业软件的操作

Q & A

- 个人主页 : <https://icoz69.github.io/>
- 邮箱 : dr.zhang.chi@outlook.com