



2024 AI+研发数字峰会

AI+ Development Digital summit

AI驱动研发迈进数智化时代

中国·上海 05/17-18

基于多模态大语言模型的 GUI智能体

张驰 博士 腾讯QQ光影实验室/西湖大学 AGI 实验室

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **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**  **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



张弛

入选美国斯坦福大学发布的《2023全球前2%顶尖科学家榜单》

张弛博士，现就职于腾讯公司担任研究科学家。将在 2024 年秋季加入西湖大学担任博导助理教授，并创建 AGI 实验室。2021年博士毕业于新加坡南洋理工大学计算机学院。他的博士研究课题专注于高效低成本的机器学习和计算机视觉算法，目前的研究方向是大模型时代下的多模态模型与 AIGC 模型。曾在 CVPR、ICCV、NeurIPS、TPAMI 等顶级国际会议期刊上发表多篇学术论文。

目录

CONTENTS

1. 背景：大语言模型与 Agent
2. 难点与挑战
3. 用户界面与人机交互
4. 基于 GUI 的智能体 AppAgent

▶ Large language models as agent

Prompt:

Imagine that you are a home robot, and is asked to go to the kitchen to **fetch something from the fridge**. The image below shows your current position. Please **plan your next action**.



GPT-4V:

My next action would **be turn right and move forward towards the hallway**, as I can see that the kitchen is likely located in that direction. I would then continue to navigate through the hallway until I reach the kitchen and locate the fridge.

▶ Large language models as agent

Prompt:

Now you are at the kitchen, this is what you can see from there. Can you see the fridge from here? What is your next action?



GPT-4V:

From this position, I can see the fridge on the right side of the kitchen. My next action would be **to move forward and turn slightly to the right to approach the fridge**

▶ Large language models as agent

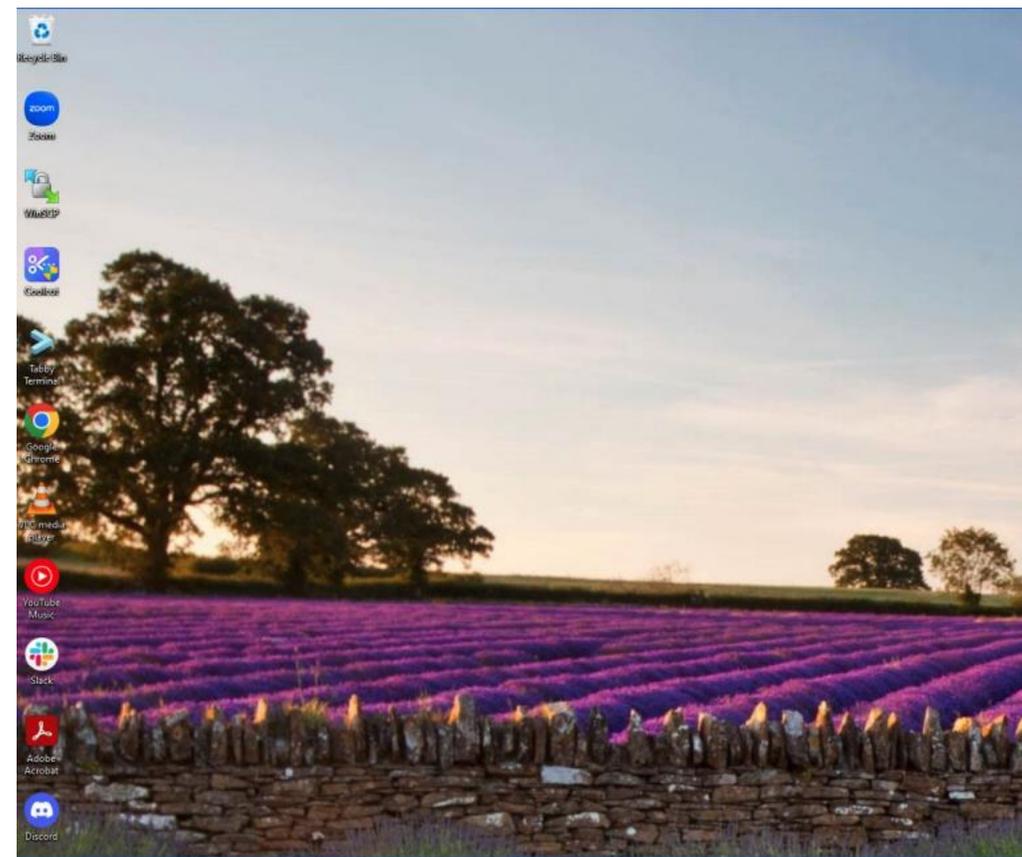
Prompt:

Imagine that you are a robot operating a computer... You are asked to **find a detailed recipe of Mapo Tofu**.

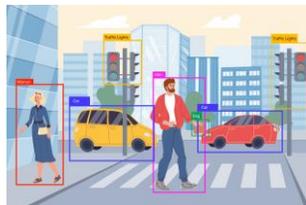
Below is what you see on the computer screen, **predict your next move...**

GPT-4V:

Next move: 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.



▶ A Simple Agent Prototype



Challenges

- Limitation of **text-only** input in current agent design
- Control alone is a big discipline
- Insufficient knowledge in LLM & high learning cost

▶ Significance of Interface



▶ Agent on Interface

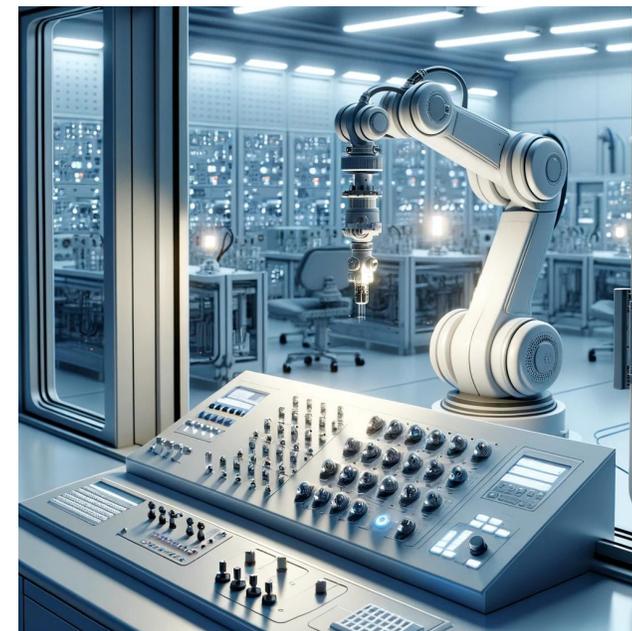
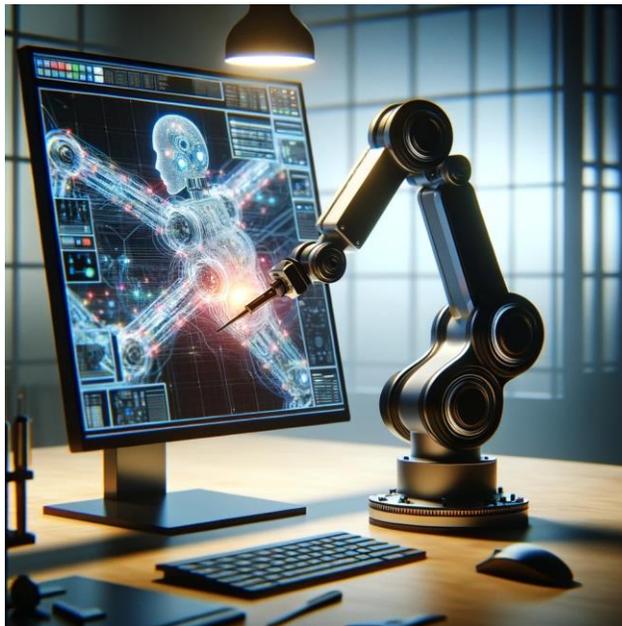
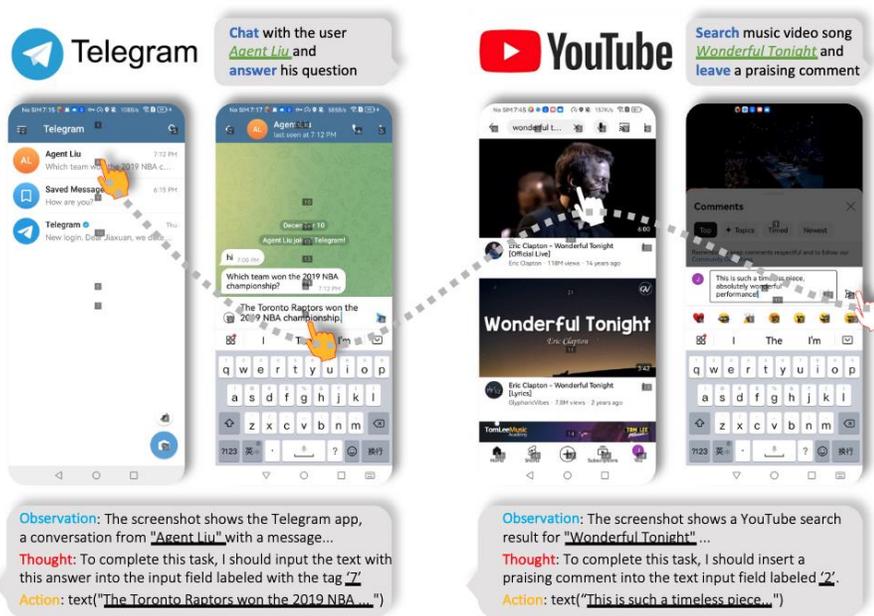
Advantages

- Actions **have clear semantic meaning**, providing a user-friendly interaction method
- Large language model **focus on decision making**
- Relatively easy perception of environment and state



▶ Goal

- Develop UI-based agent frameworks with strong generalization capabilities
 - Develop an multimodal LLM that can **accept multimodal inputs**
 - Developing an embodied agent that can interact with interfaces, such as mobile phones



▶ AppAgent

Telegram Chat with the user *Agent Liu* and answer his question

YouTube Search music video song *Wonderful Tonight* and leave a praising comment

Gmail Send an email to *janedoe@email.com* to ask her about her new job

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 '7'

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...")

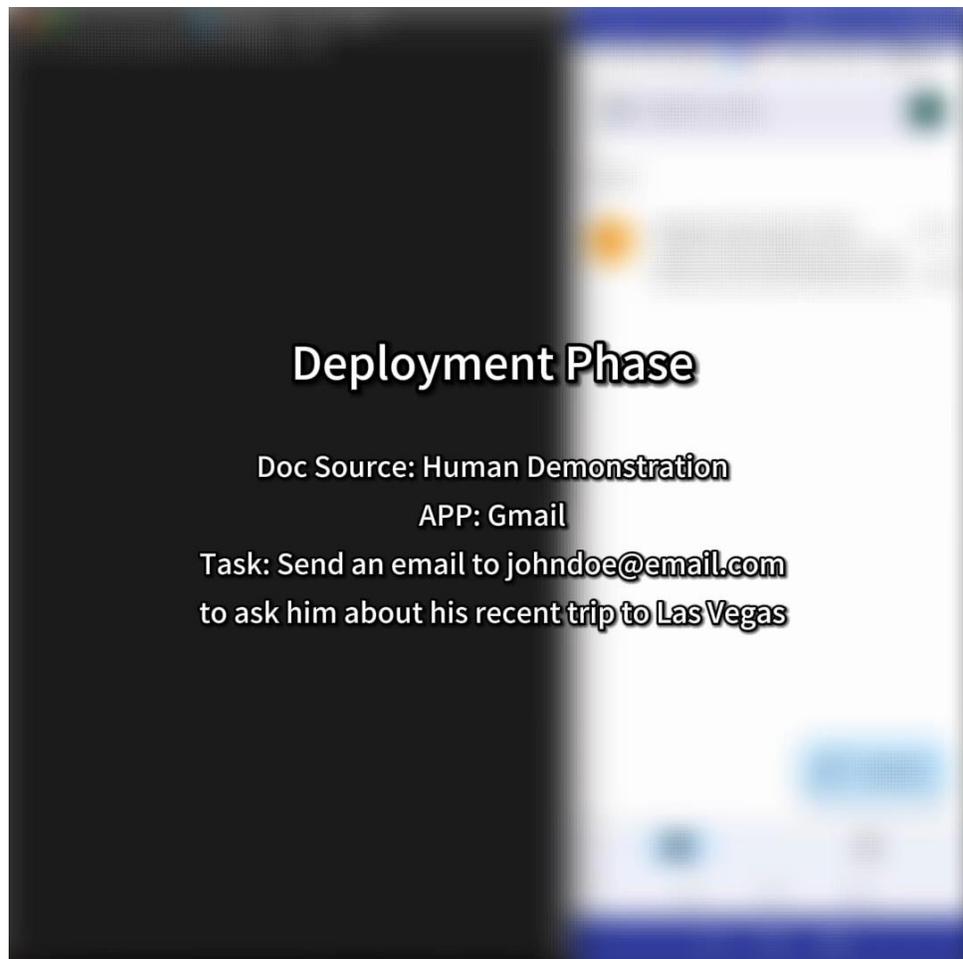
Observation: The screenshot shows a smartphone's email application ...

Thought: To complete this task, I should type the content of the email asking *Jane Doe* about her new job.

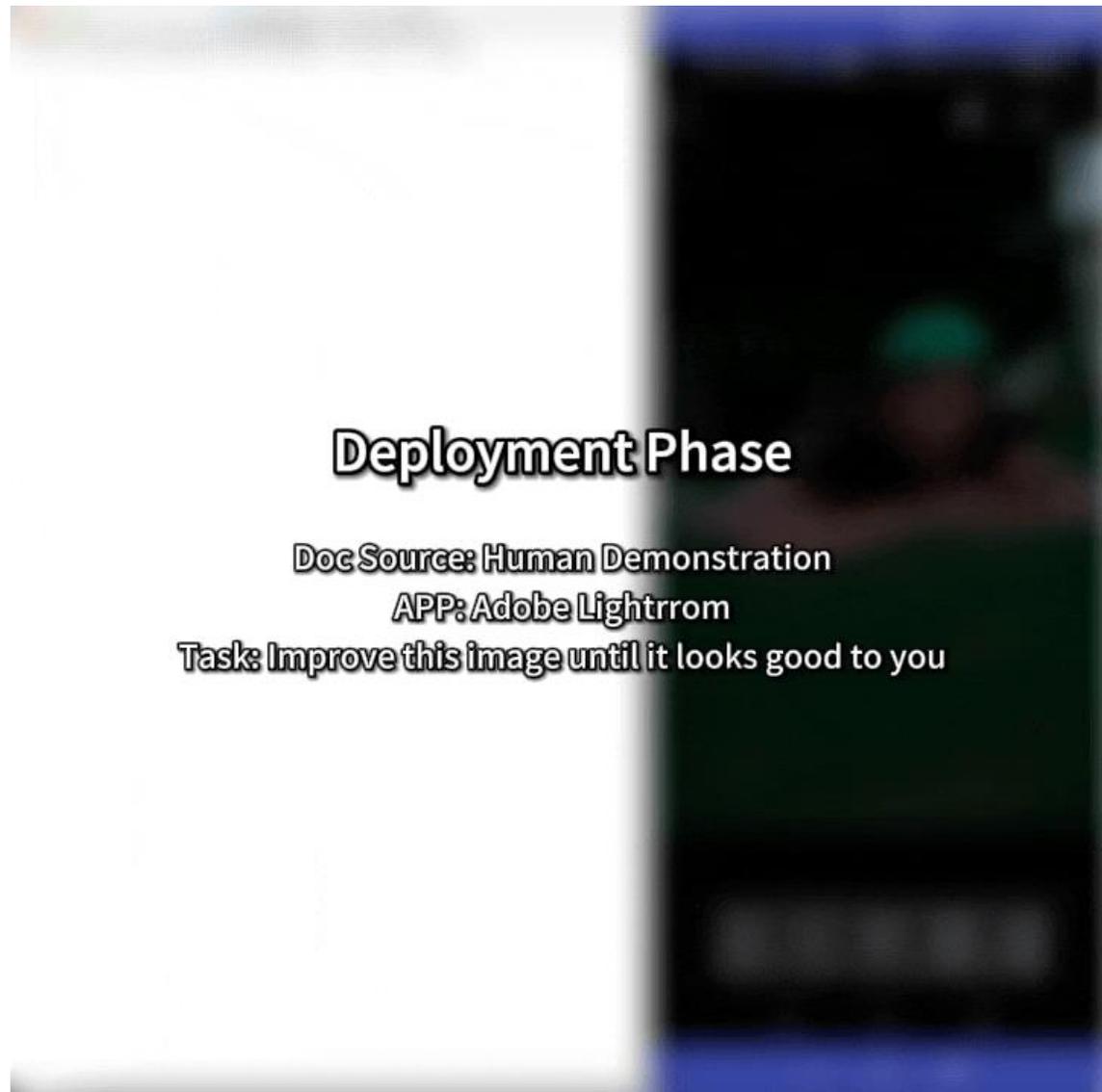
Action: text("Dear Jane Doe,\n\nI am writing to ...")

AppAgent: Multimodal Agents as Smartphone Users

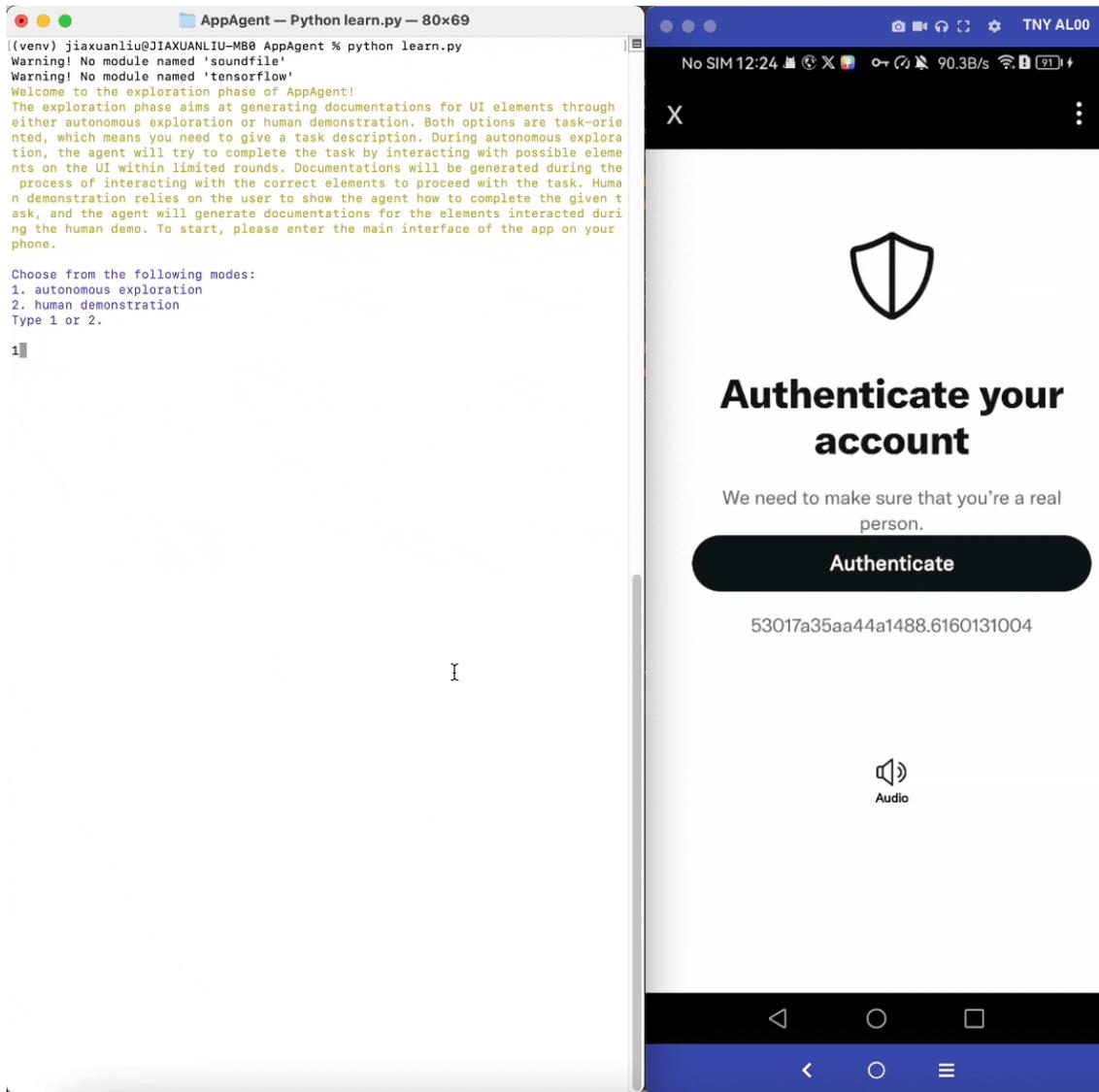
▶ AppAgent



▶ AppAgent

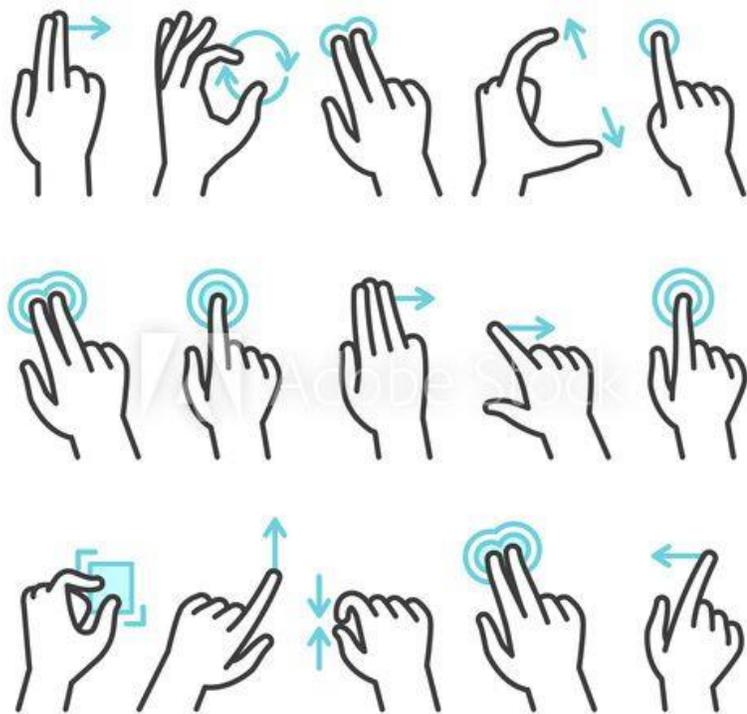


▶ AppAgent

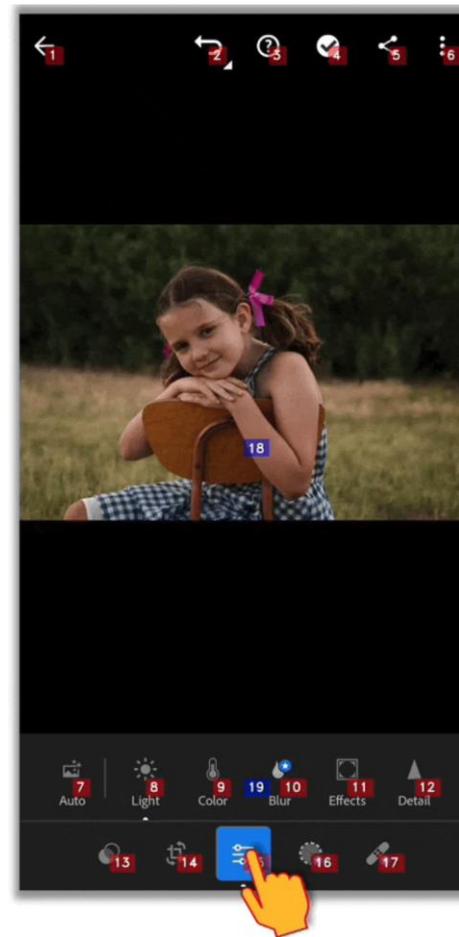


▶ AppAgent

Challenge 1 : How to output control?



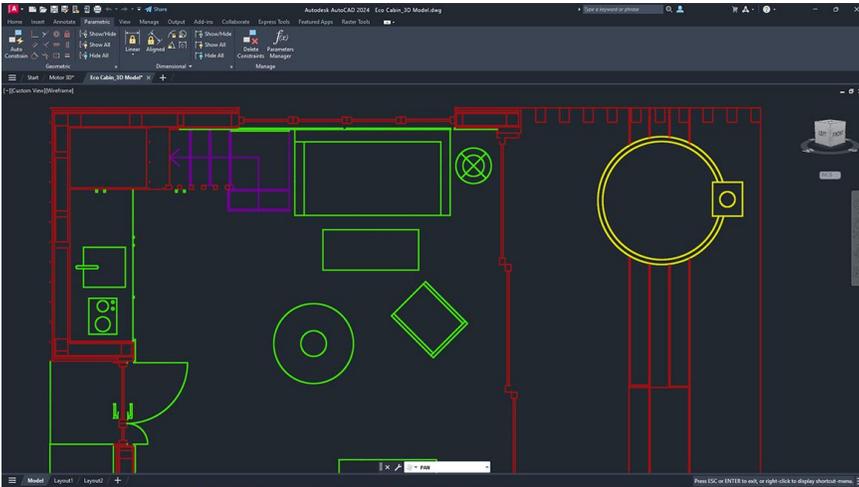
#272300566



▶ AppAgent

Challenge 2 : How to learn?

- Complex software
- Easy software

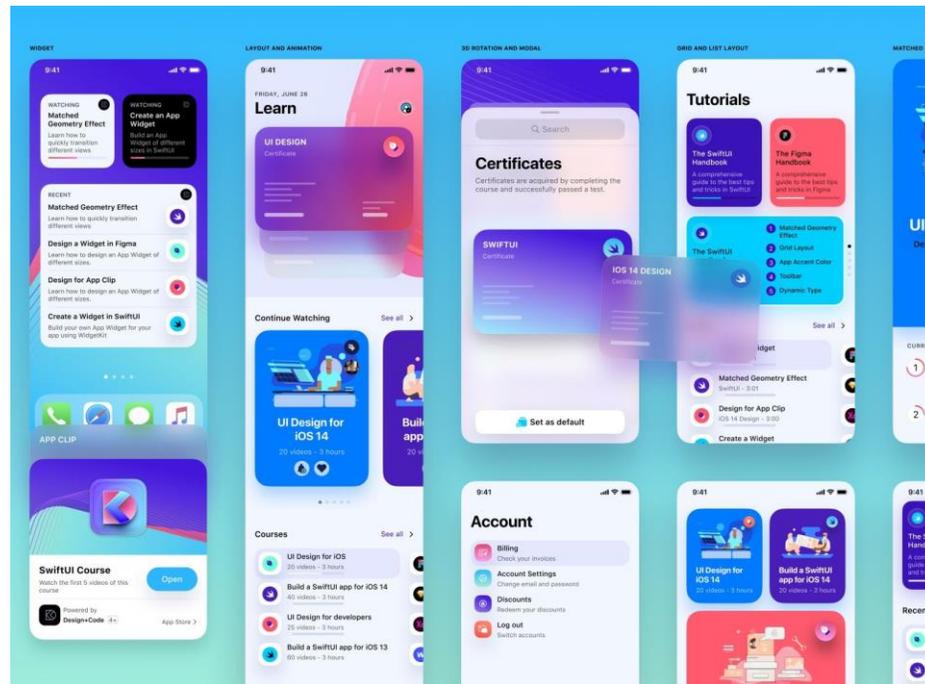
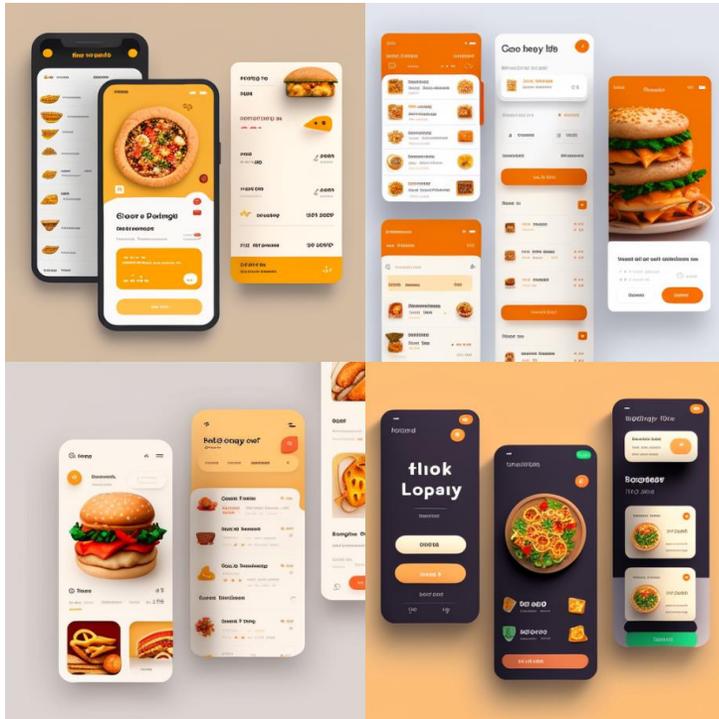


VS



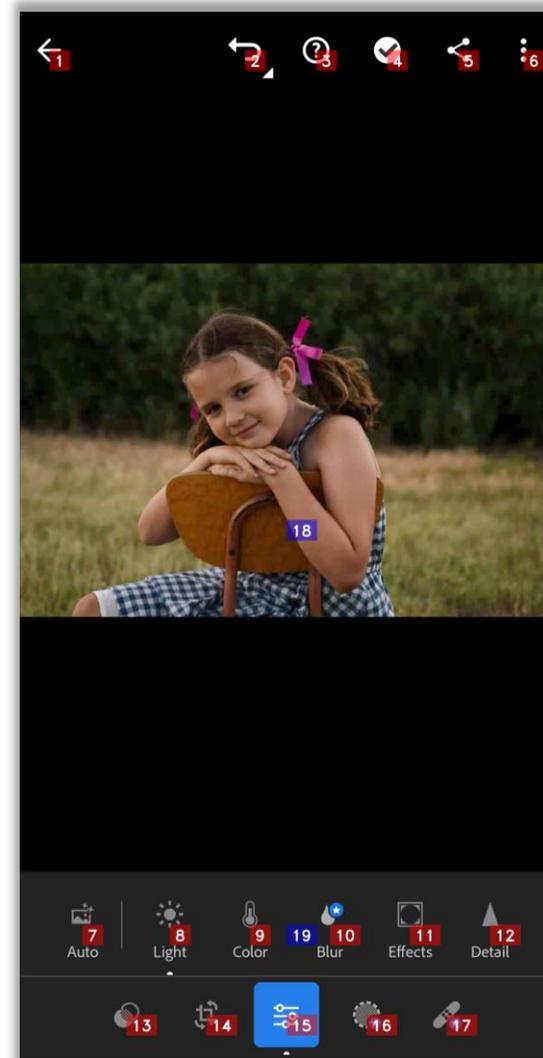
▶ AppAgent

Challenge 3 : How to generalize?



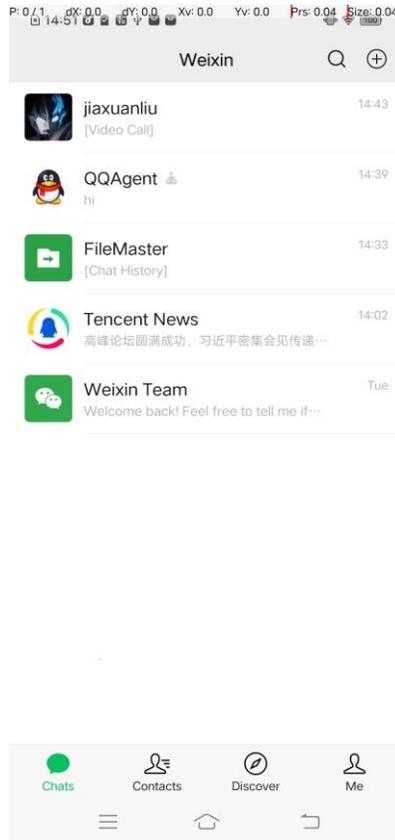
▶ Action Space

- Tap(element : int)
- Long_press(element : int)
- Swipe (element : int, direction : str, dist : str)
- Text(text : str)
- Back()
- Exit()



▶ Human-Like system design

Supervised finetuning of LLM — Are you sure?



“Briefly describe the functions of the four buttons in the bottom bar of this screenshot.”

“Can you provide the bounding box coordinates for the Profile Photo button on the screen?”

▶ Human-Like system design

How do we human learn?

- Quick Learning process
 - Self-exploration
 - Watching demos
- Good generalization capability

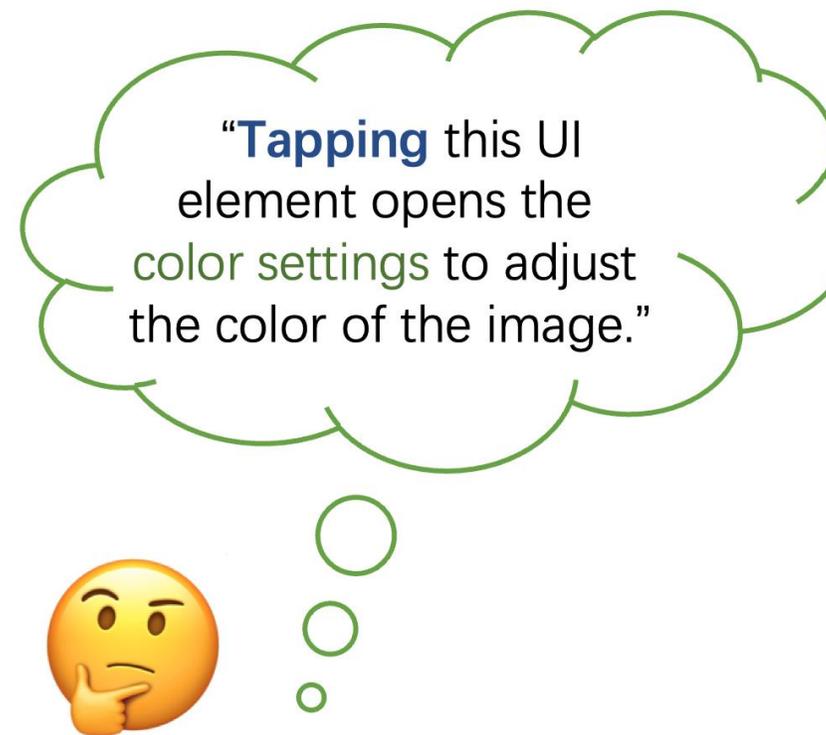
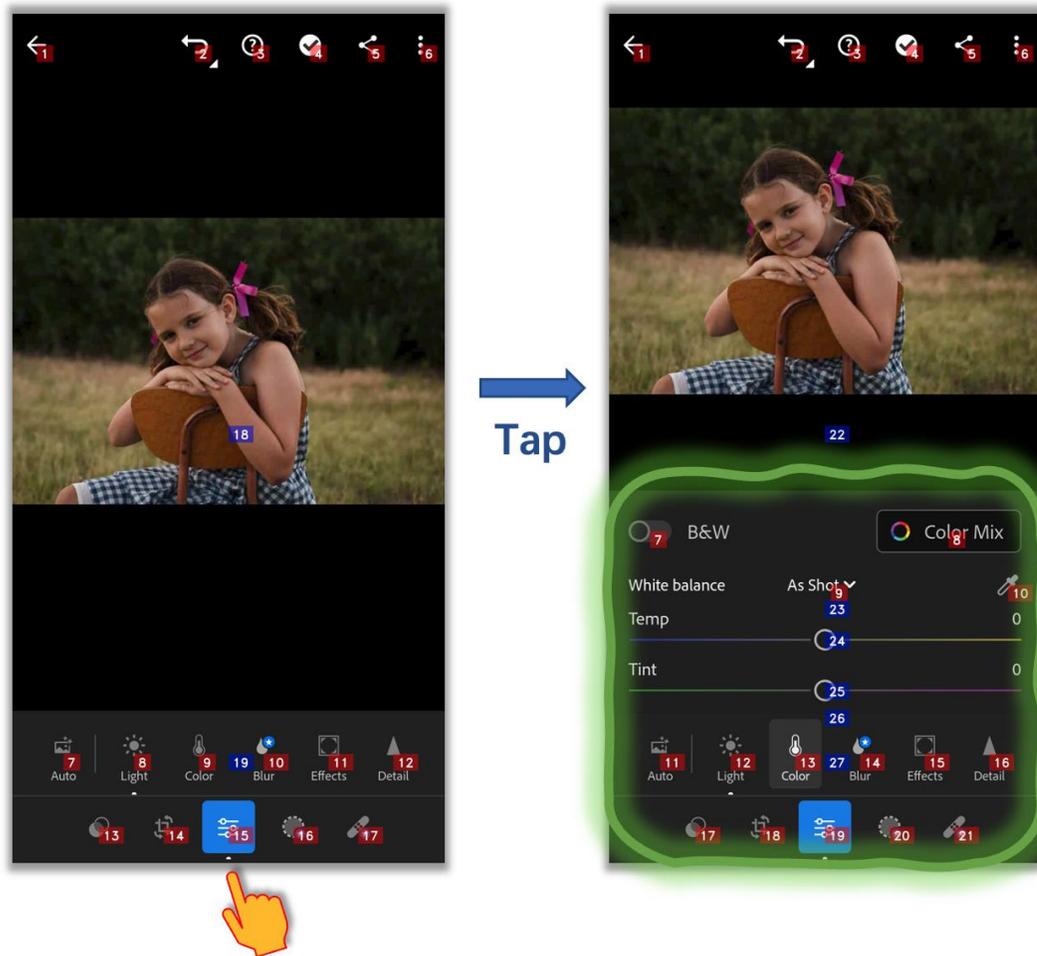


▶ Human-Like system design

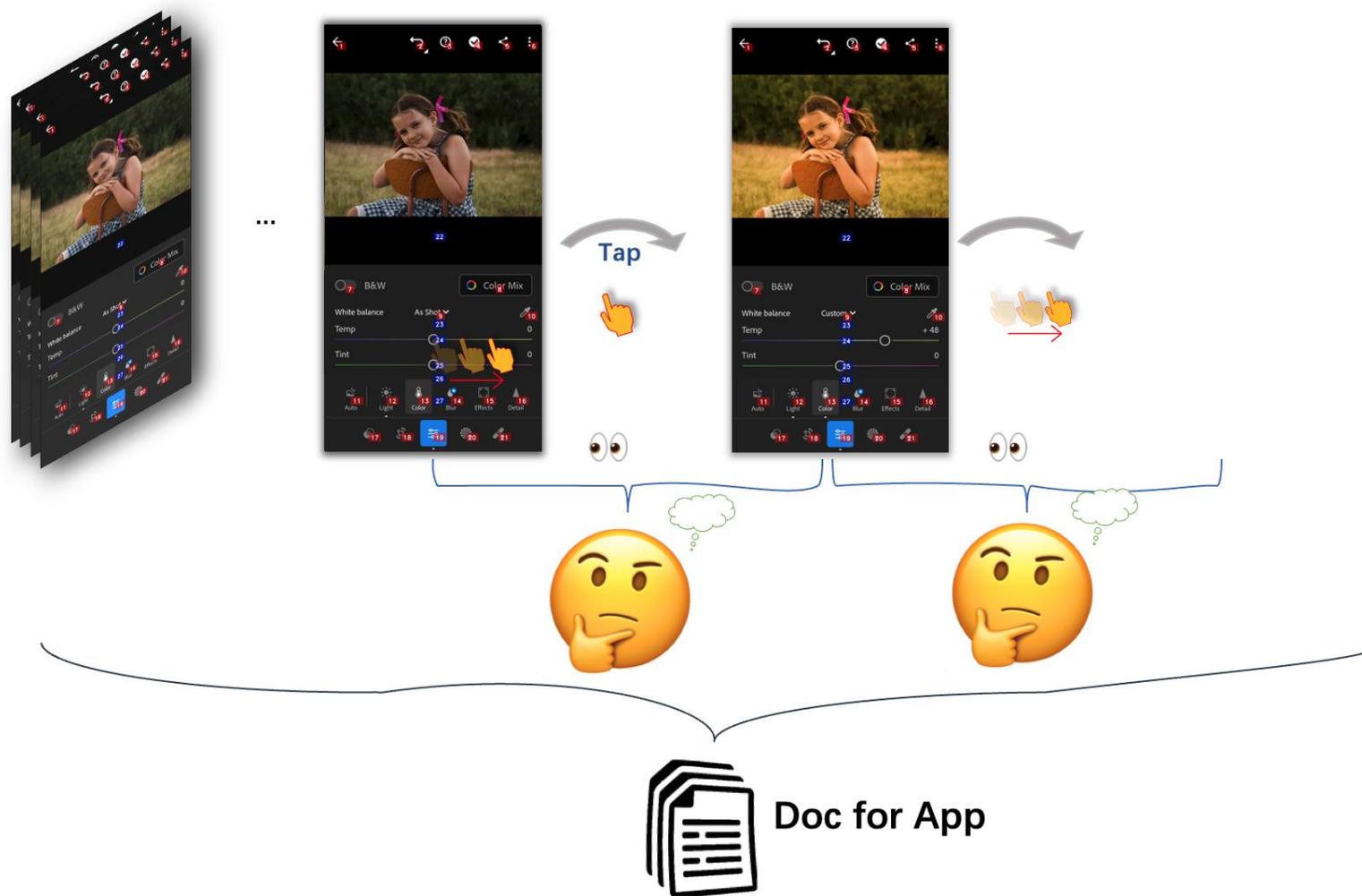
Overall Idea: Exploration->Deployment

- **Explore** how to use an App (self-exploration or watching demos)
- Generate a **document** after exploration, used as an external database
- Refer to the Doc for better decision making

▶ Exploration Phase



▶ Exploration Phase

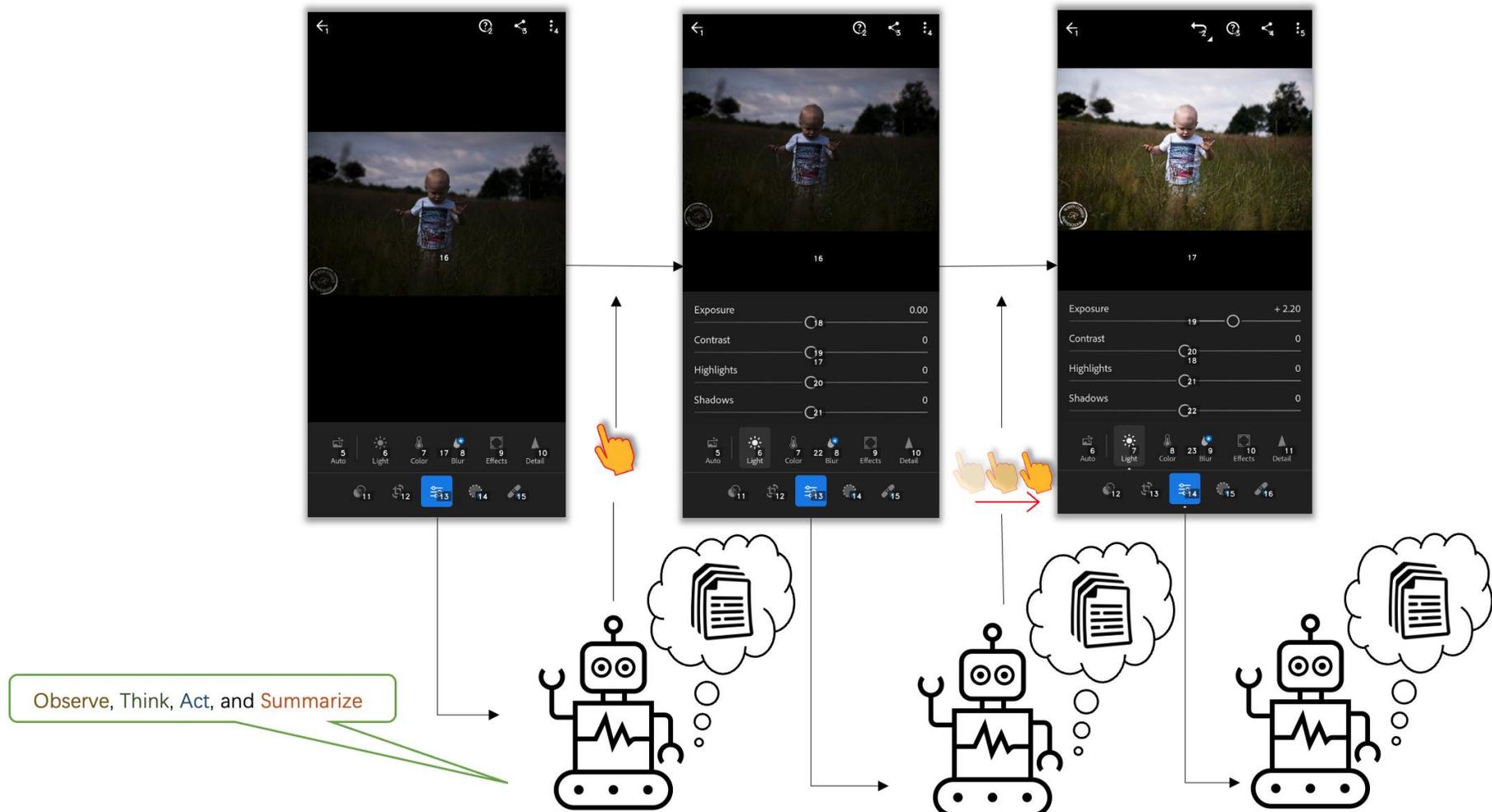


▶ Deployment Phase

Task:
“beautify this photo for me...”

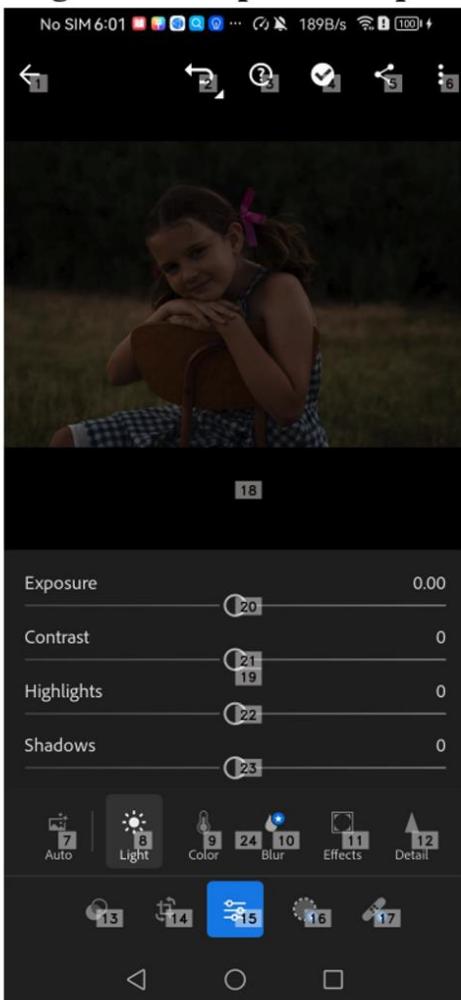


Deployment Phase



Deployment Phase

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



Observation

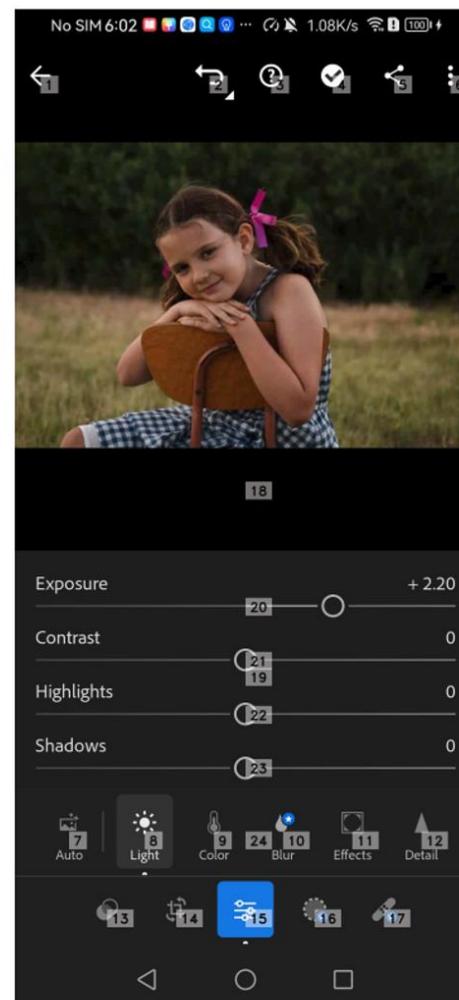
The screenshot shows a photo editing application interface with various adjustment settings for **exposure, contrast, highlights, and shadows**. These adjustments have **numeric tags 20, 21, 22, and 23** respectively. The settings are currently at their default values, and the image appears **underexposed**,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



▶ Experiment

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

▶ Significance and Impact

- Revolutionizing Human-Computer Interaction

CMD



GUI

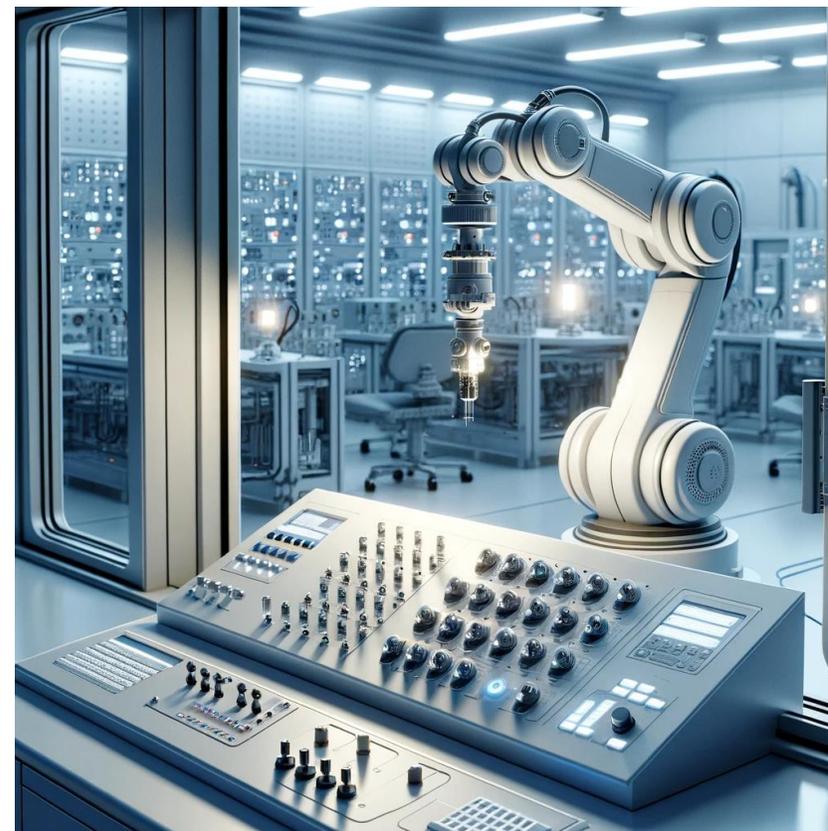
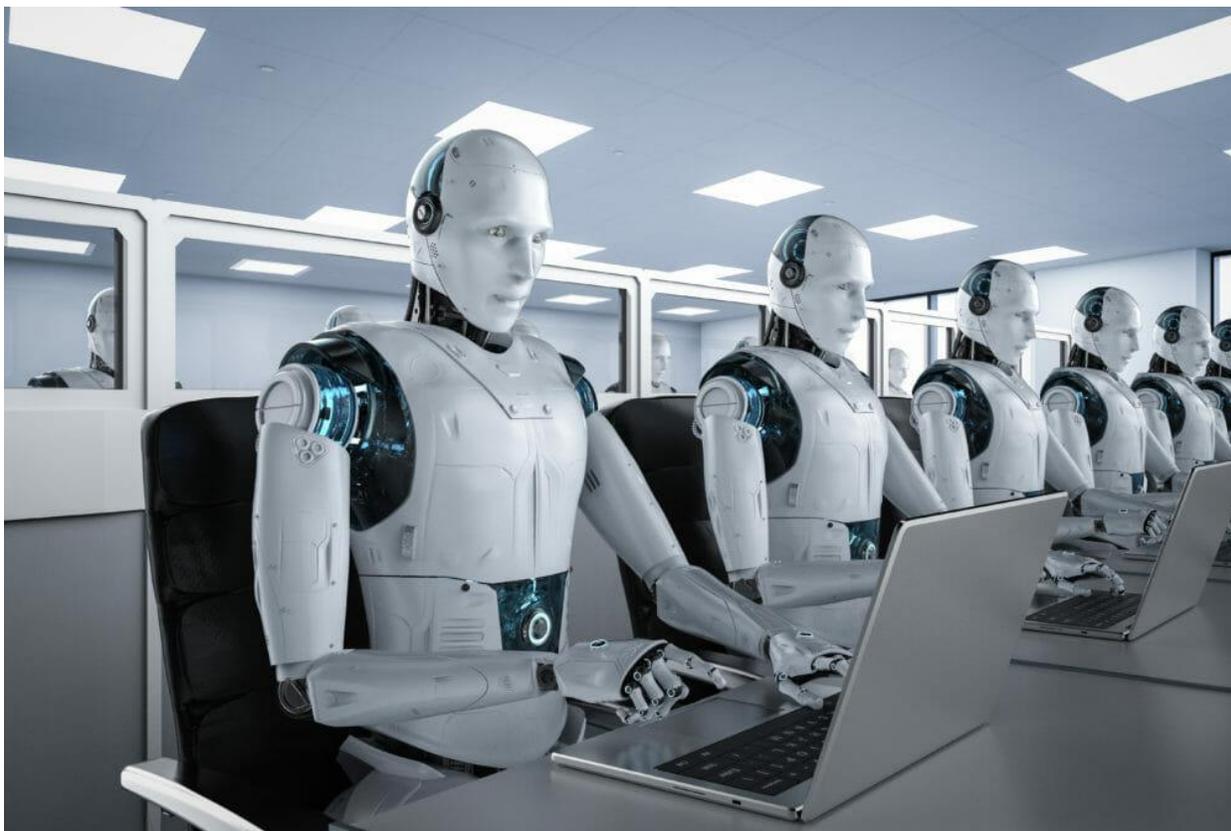


LUI



▶ Significance and Impact

- Digital employee and super assistants



▶ Significance and Impact

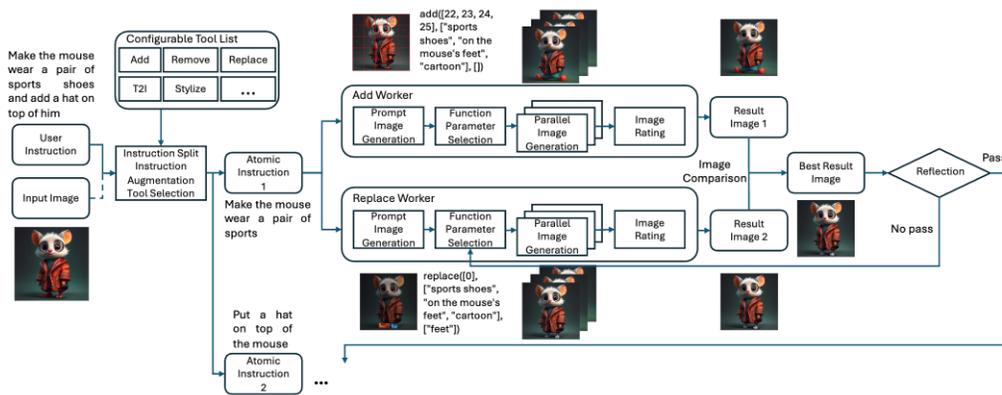
- Software testing and user study



More Research on Agents

Application

- The AppAgent framework and ideas have been validated in the company's key products such as **image editing and 3D room layout generation**, and will continue to be explored and developed in the future.

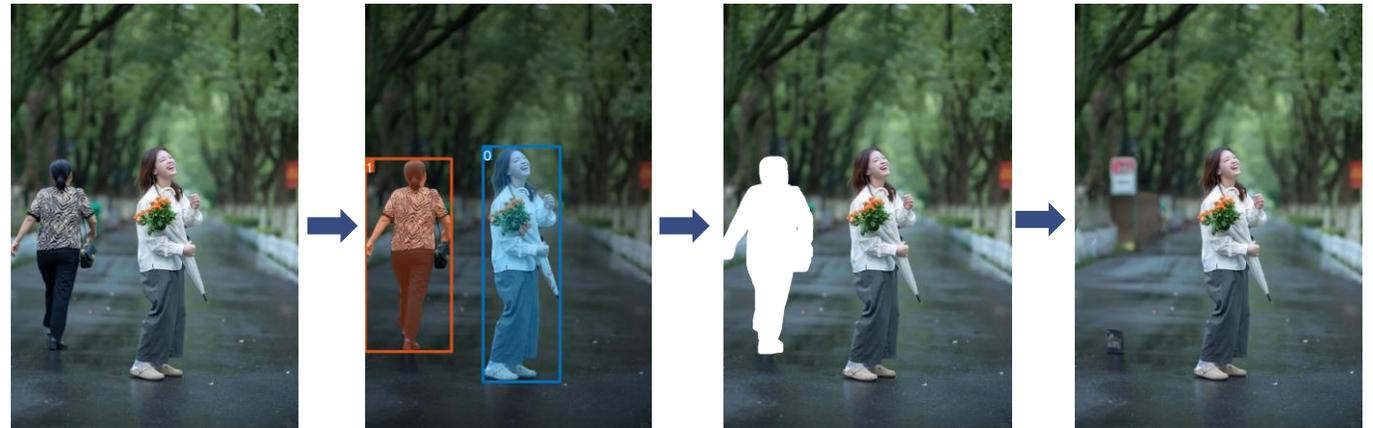


"remove the left person"

segments the image first

Erase the person

Inpaint the region



▶ More Research on Agents

Application

- We also design agent, MovieLLM, to create movies for training long-video understanding models.



Frame 386 Frame 387

Line:
I wonder what mysteries I will find in these old pages.



Frame 388 Frame 389

Line:
Could this... could this be for me?



Frame 390 Frame 391

Line:
Words penned with such passion... Could it be Thomas?



Frame 392 Frame 393

Line:
How did he know of my feelings for him?



Frame 394 Frame 395

Line:
I must let Thomas know, he must know how I feel.

Question: Why is Rosamund so surprised and emotional after reading the letter?

Answer: Because she believes that the love letter reflects the same feelings she has for Thomas.

▶ More Research on Agents

AppAgent的应用

- AppAgent框架与思路在公司的图像编辑，3D小屋等重点业务能上得到验证，未来会被继续探索与发展

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
201	202	203	204	205	206	207	208	209	210	211	212	213	214	215	216	217	218	219	220
221	222	223	224	225	226	227	228	229	230	231	232	233	234	235	236	237	238	239	240
241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259	260
261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279	280
281	282	283	284	285	286	287	288	289	290	291	292	293	294	295	296	297	298	299	300
301	302	303	304	305	306	307	308	309	310	311	312	313	314	315	316	317	318	319	320
321	322	323	324	325	326	327	328	329	330	331	332	333	334	335	336	337	338	339	340
341	342	343	344	345	346	347	348	349	350	351	352	353	354	355	356	357	358	359	360
361	362	363	364	365	366	367	368	369	370	371	372	373	374	375	376	377	378	379	380
381	382	383	384	385	386	387	388	389	390	391	392	393	394	395	396	397	398	399	400



▶ LLM for Agents

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的先验增强
- 特殊图像内容判断能力 (P图)
- 提高 LLM 预测速度

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **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**  **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



THANKS

