



2025 AI+ Development
Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

05/23-24 | 上海站





2025 AI+研发数字峰会

拥抱AI 重塑研发 AI+ Development Digital Summit

下一站预告

08/08-09 | 北京站

11/14-15 | 深圳站



查看会议详情

北京站论坛设置

大模型和 AI 应用评测

智能存储与检索技术

下一代知识工程

AI+ 金融业务创新

智能需求工程

智能体与研发效率工具

AI 产品运营与出海策略

大模型安全与对齐

大模型应用开发框架与实践

智能体经济 (Agentic Economy)

智能测试工具的开发与应用

具身智能与机器人

代码生成及其改进

AI+ 新能源汽车

AI 前沿技术探索与实践

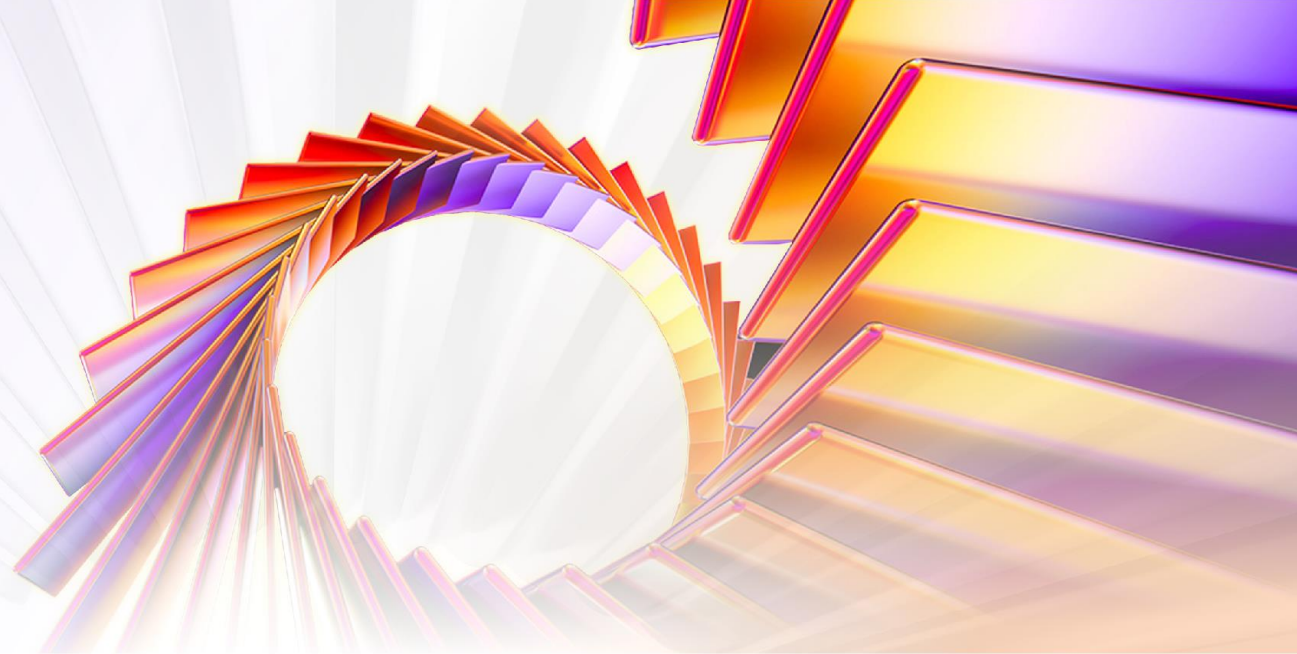


| 05/23-24 | 上海站

2025 AI+ Development
Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发



AppAgentX可进化GUI智能体框架

张驰 | 西湖大学AGI实验室 助理教授



张弛 博士

连续两年入选美国斯坦福大学发布的《全球前2%顶尖科学家榜单》

2021年博士毕业于新加坡南洋理工大学计算机学院，博士毕业后加入腾讯公司担任研究科学家。2024年秋加入西湖大学担任博导，并创建 AGI 实验室。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

- I. 背景: 大语言模型与 Agent
- II. 基于 GUI 的Agent: AppAgent
- III. 可进化的GUI Agent: AppAgentX

► 大语言模型作为智能体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.



► 大语言模型作为智能体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**



▶ 大语言模型作为智能体Agent

Prompt:

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

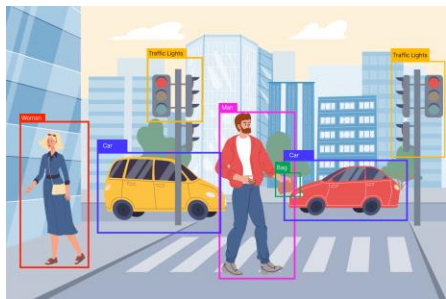
Below is what you see on the computer screen, **predict your next move...**

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



感知

决策

执行

挑战

- 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 作用在用户界面上 (UI)

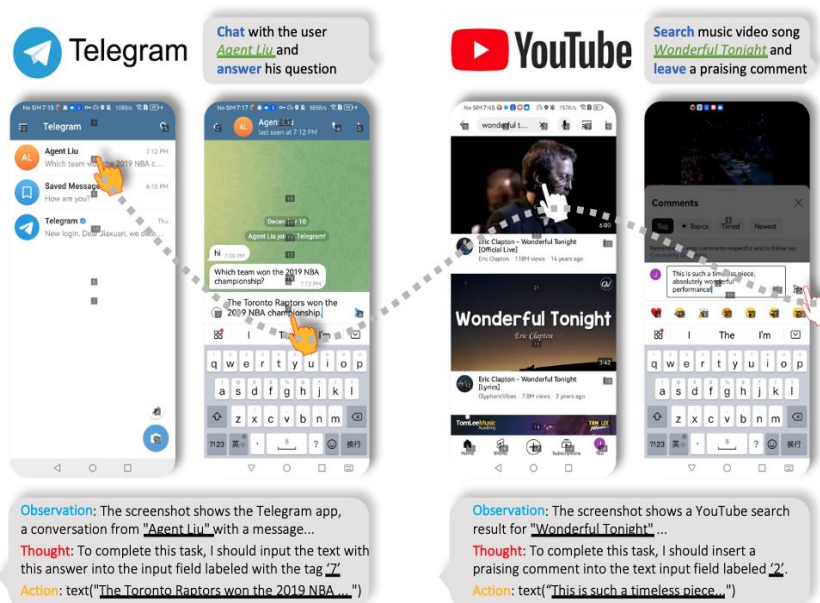
Advantages

- 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 an multimodal Agent that can **accept multimodal inputs**
- Developing an embodied agent that can interact with interfaces, such as mobile phones





Chat with the user Agent Liu and answer his question



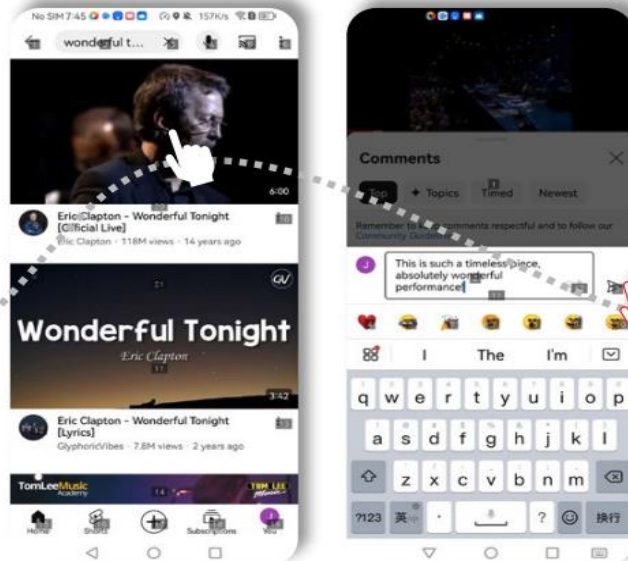
Observation: The screenshot shows the Telegram app, a conversation from "Agent Liu" with a message...

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 ...")



Search music video song Wonderful Tonight and leave a praising comment



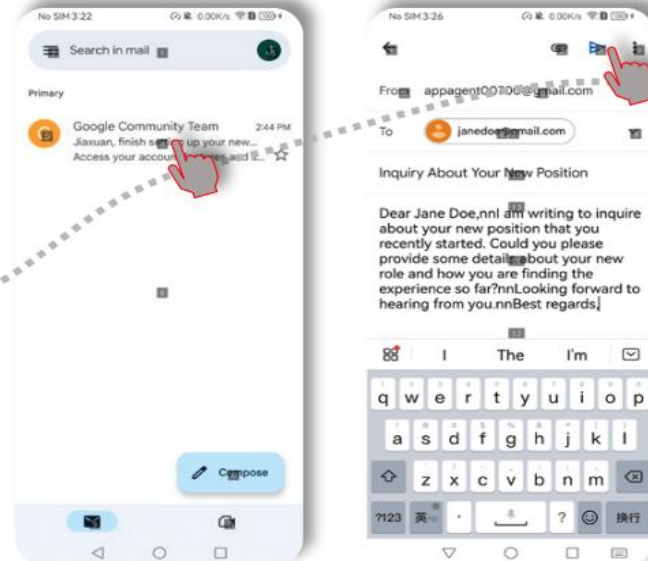
Observation: The screenshot shows a YouTube search result for "Wonderful Tonight" ...

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...")



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



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 (CHI 2025)

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

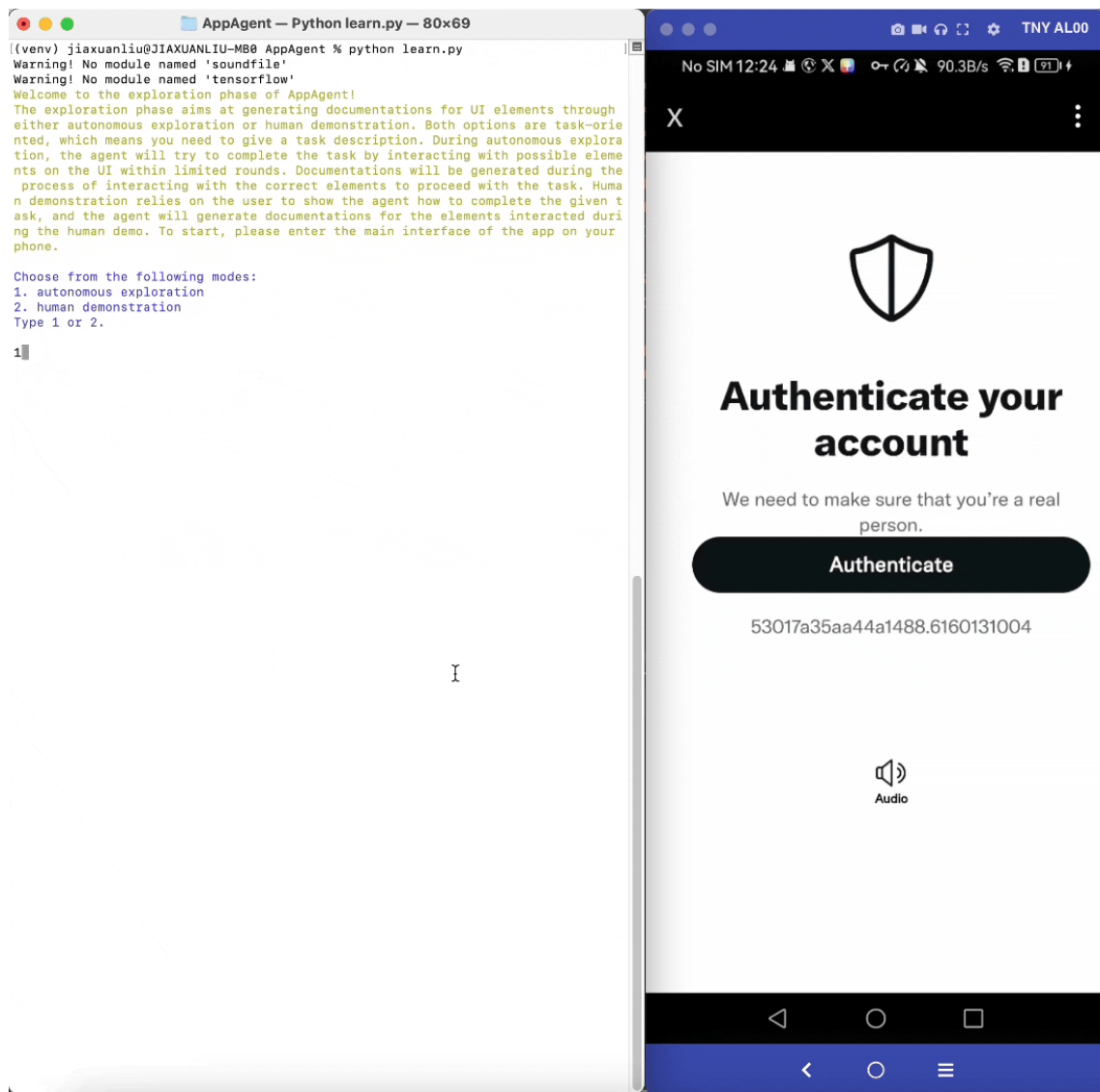


Deployment Phase

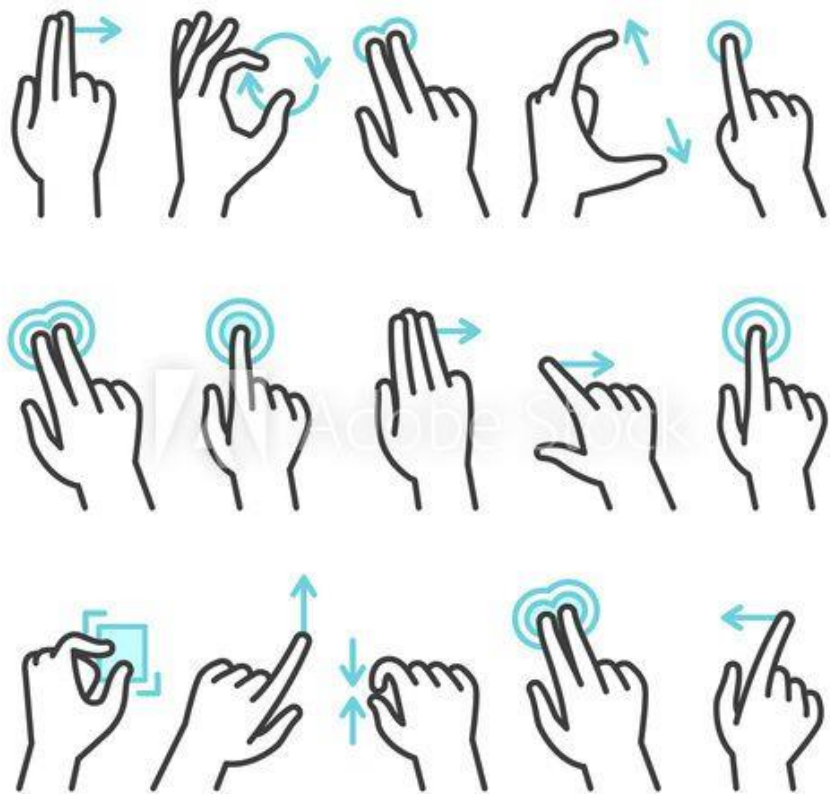
Doc Source: Human Demonstration

APP: Adobe Lightroom

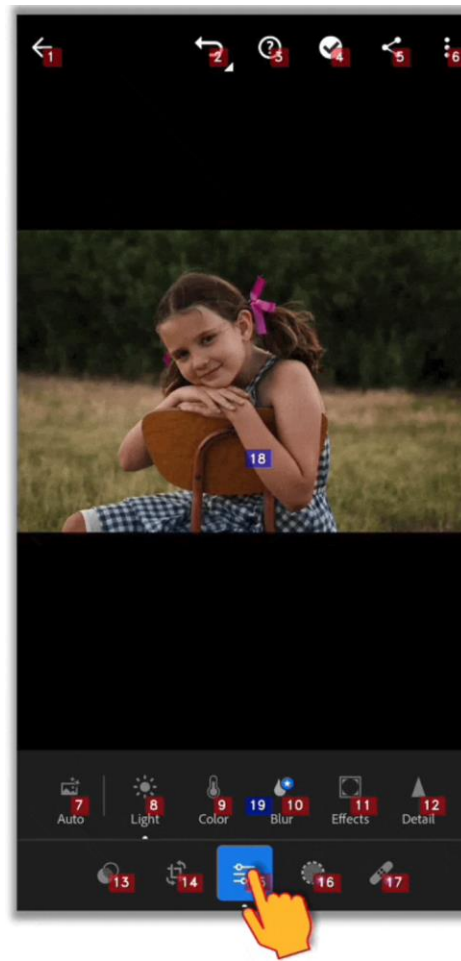
Task: Improve this image until it looks good to you



Challenge 1 : How to output control?



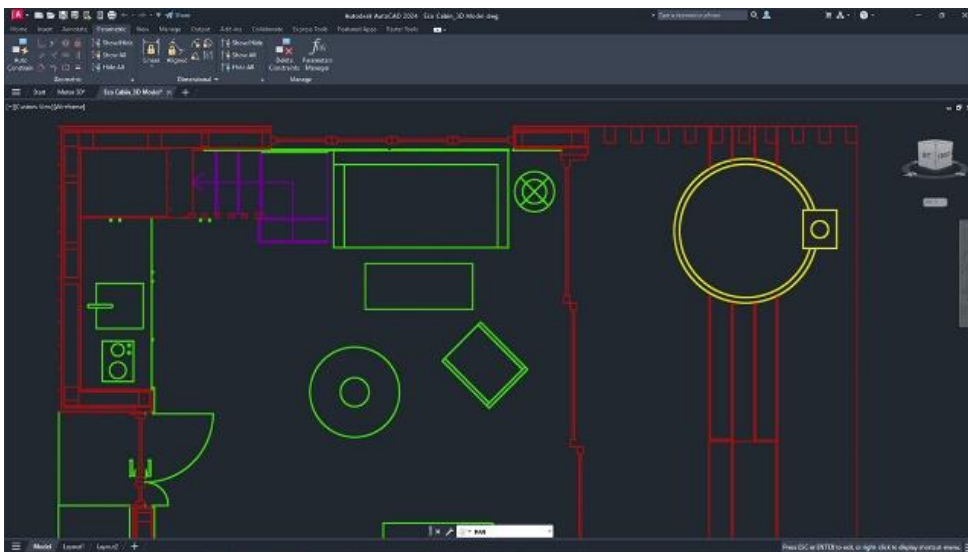
#272300566



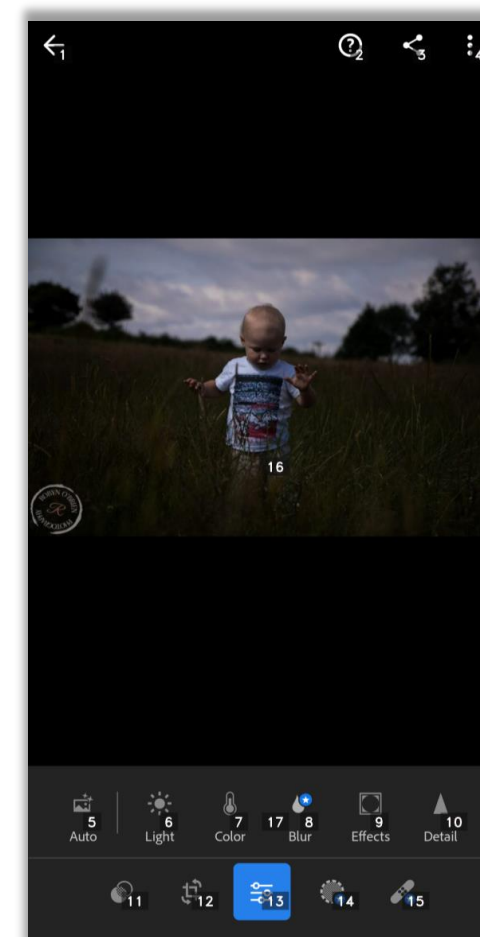
▶ AppAgent

Challenge 2 : How to learn?

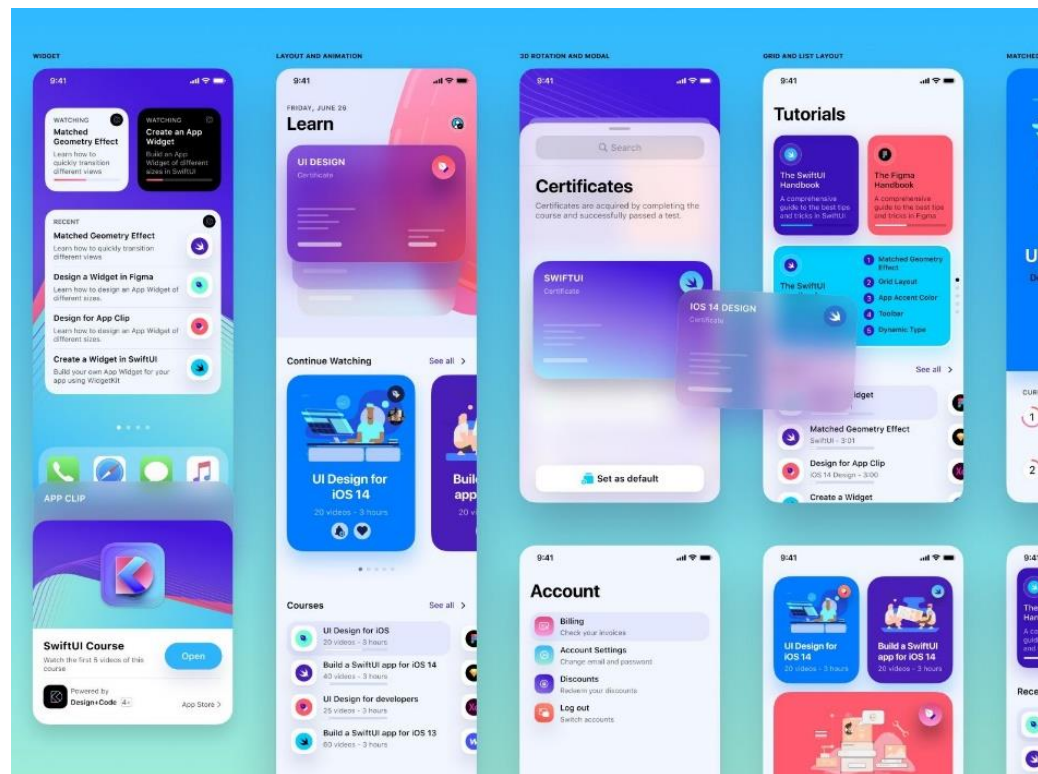
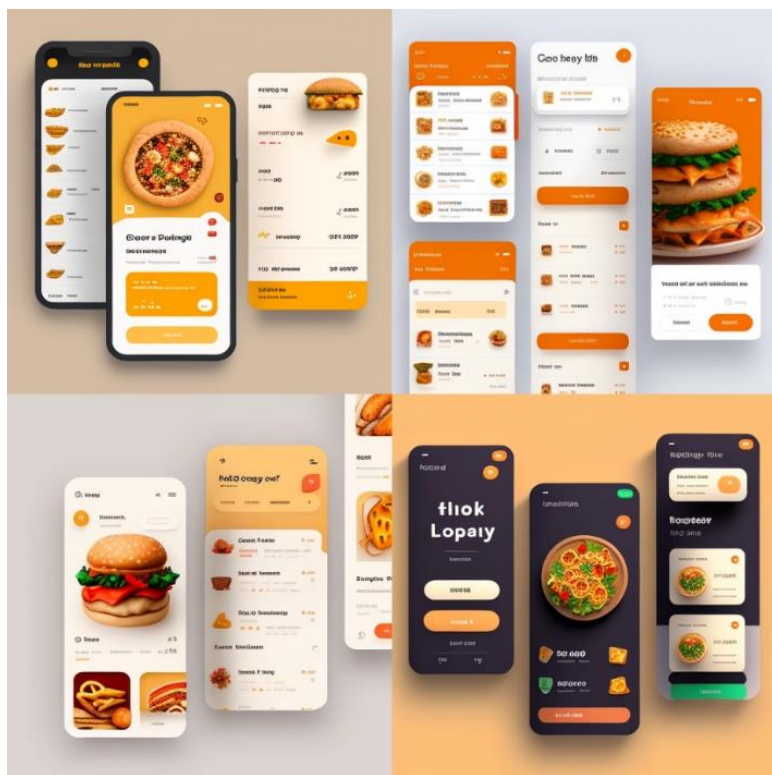
- Complex software
- Easy software



VS

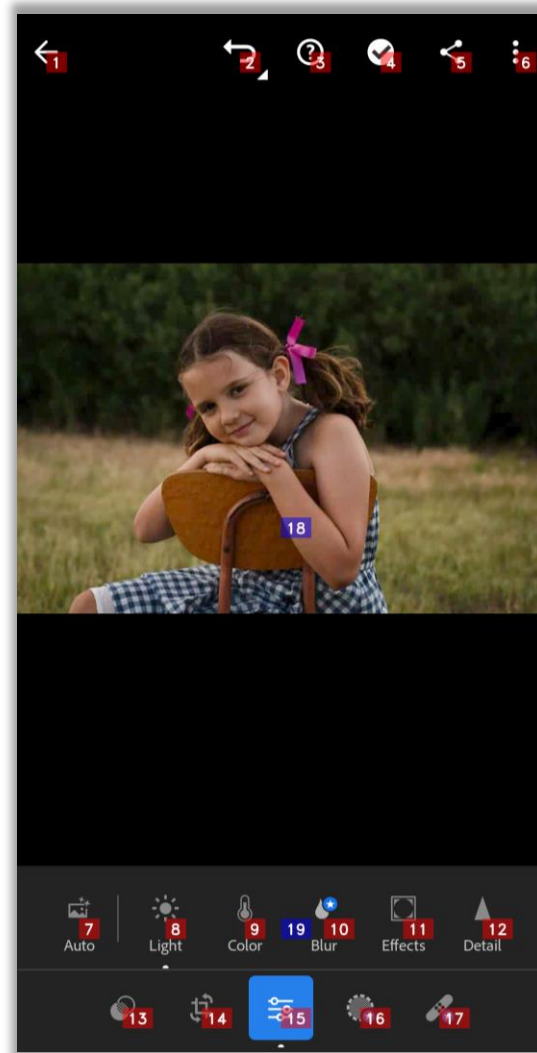


Challenge 3 : How to generalize?



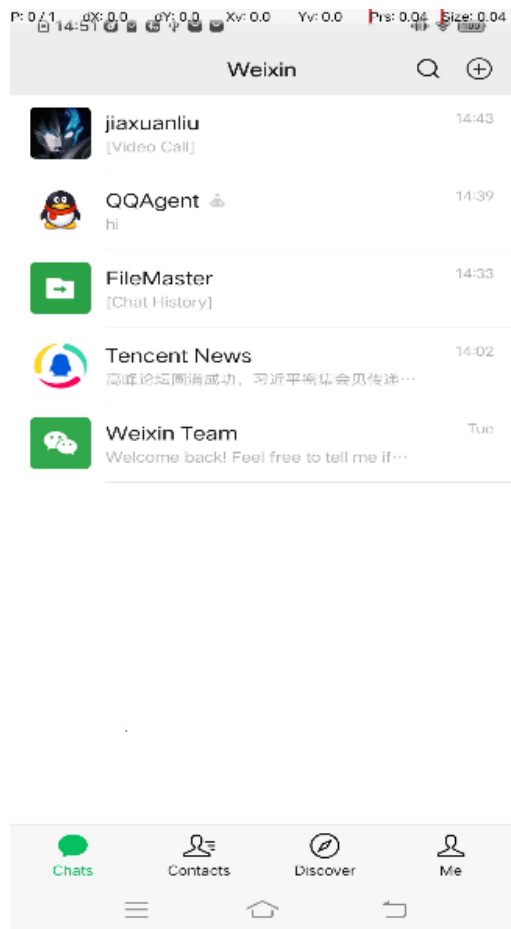
▶ 执行空间

- 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?”

Zhang, Chi, et al. "Appagent: Multimodal agents as smartphone users. (2023).



▶ Human-like system design

How do we human learn?

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



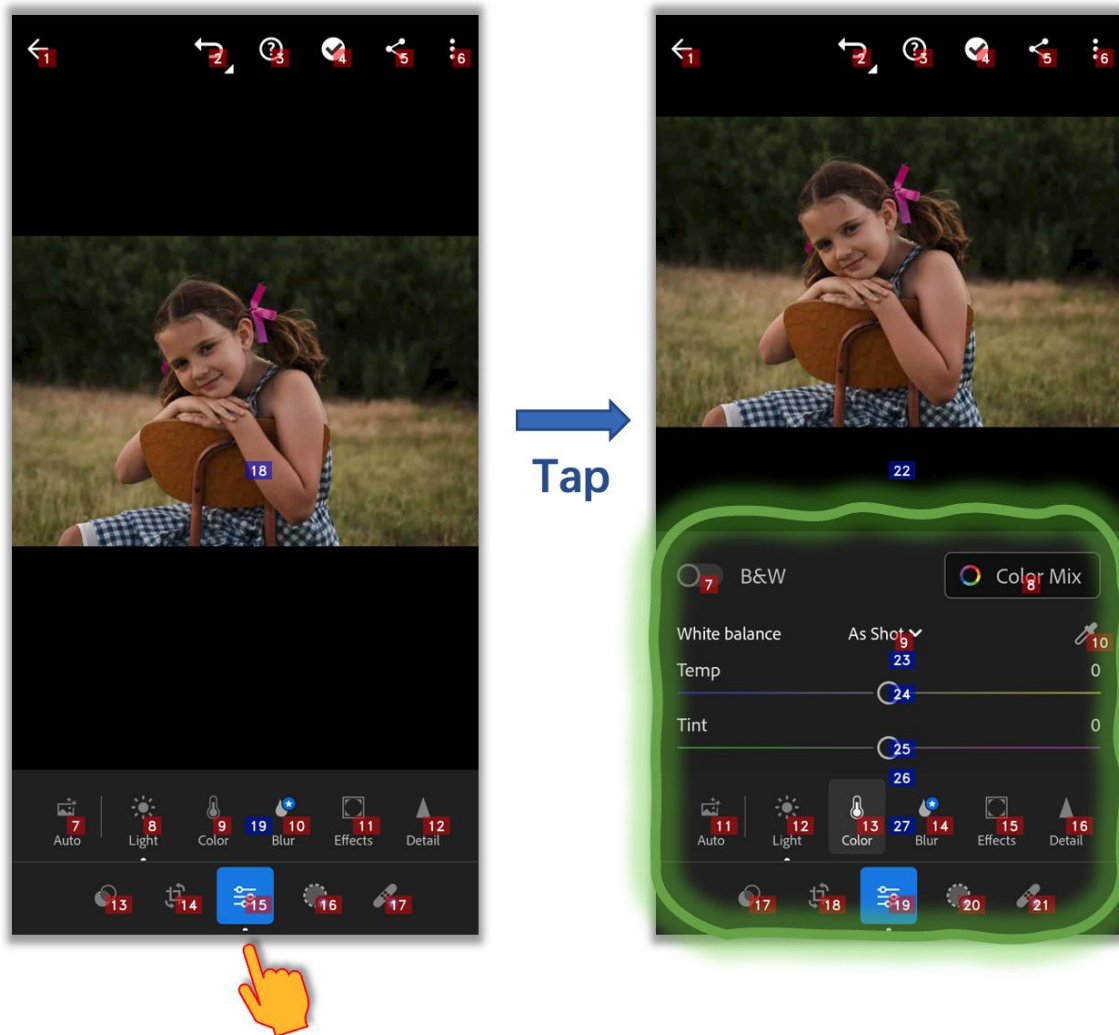
Overall Idea: Exploration>Deployment

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



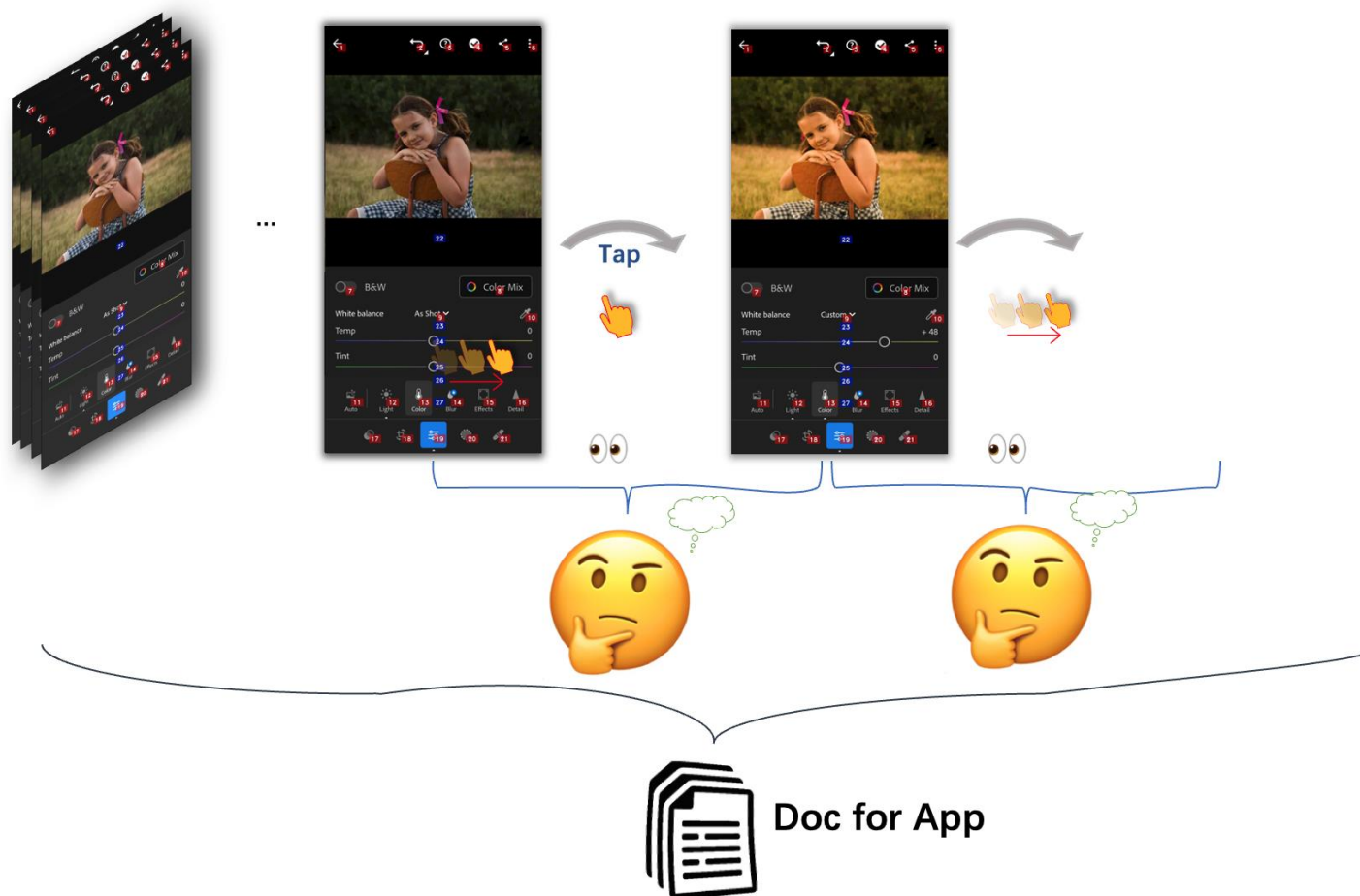
► Exploration Phase



“**Tapping** this UI element opens the color settings to adjust the color of the image.”



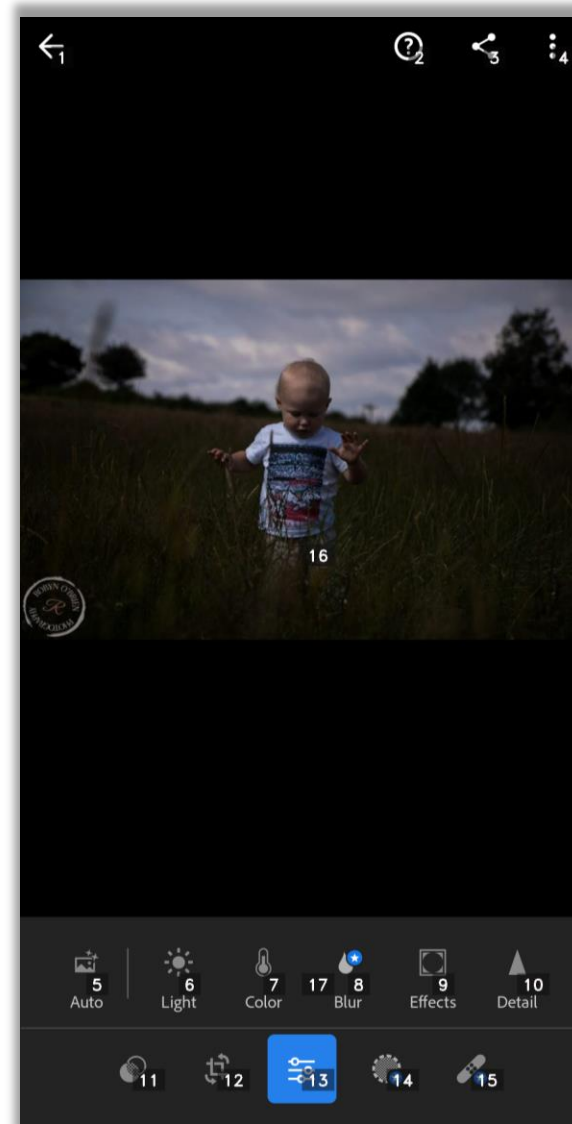
►► Exploration Phase



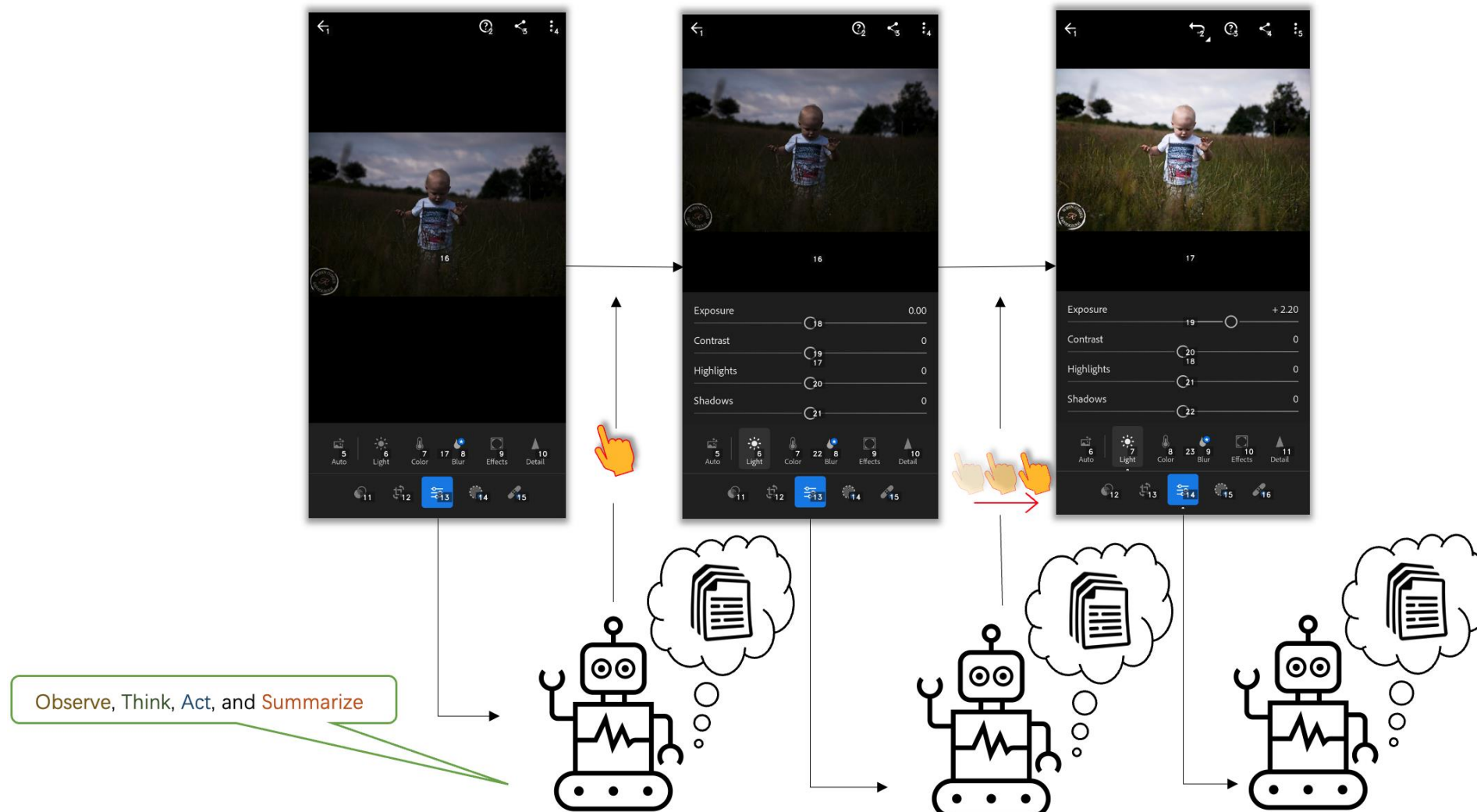
►► Deployment Phase

任务:

“帮我美化一下这张图...”

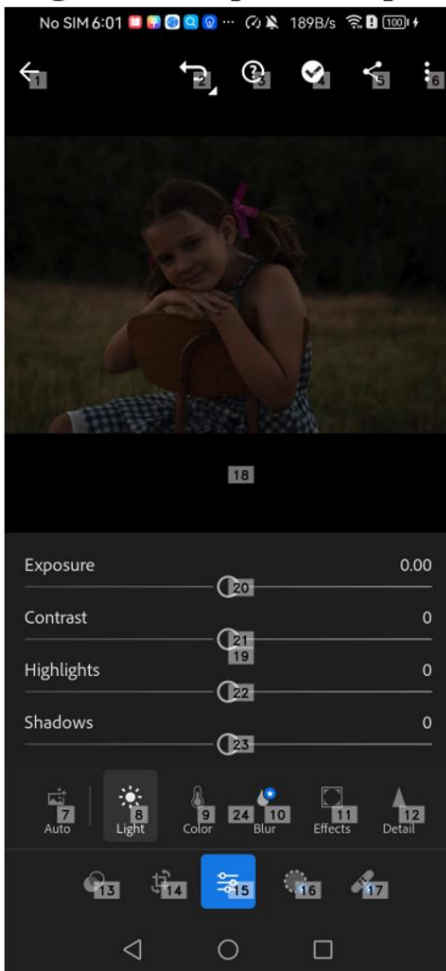


Deployment Phase



► Deployment Phase

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



Observation

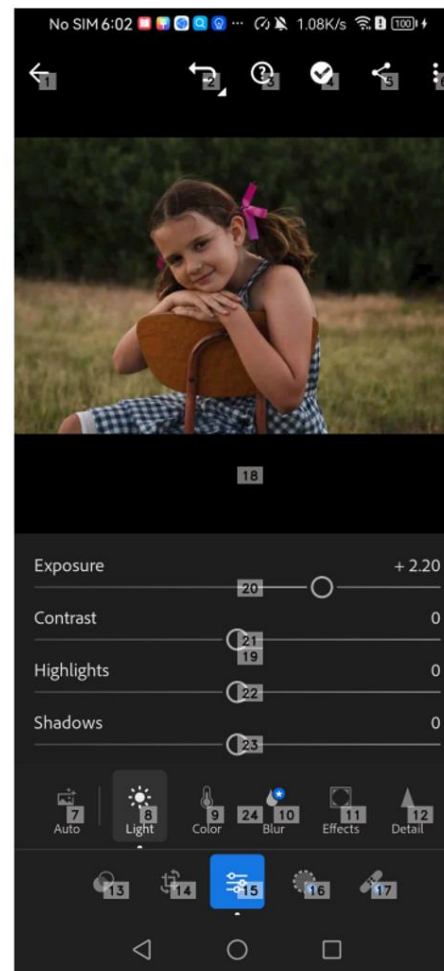
The screenshot shows a photo editing application interface with various adjustment settings for **exposure, contrast, highlights, and shadows**. These adjustments have **numeric tags 20, 21, 22, and 23** respectively. The settings are currently at their default values, and the image appears **underexposed**,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



► 智能化向高效化过渡



每次的推理时间过长，时间效率比不过RPA方法。



没有人类的屏幕操作知识，容易陷入死循环。



对于屏幕动态内容的适配不够。





“First do it right, then do it fast.”

—Jon Bentley (1986年《编程珠玑》作者提出)



AppAgentX的设计

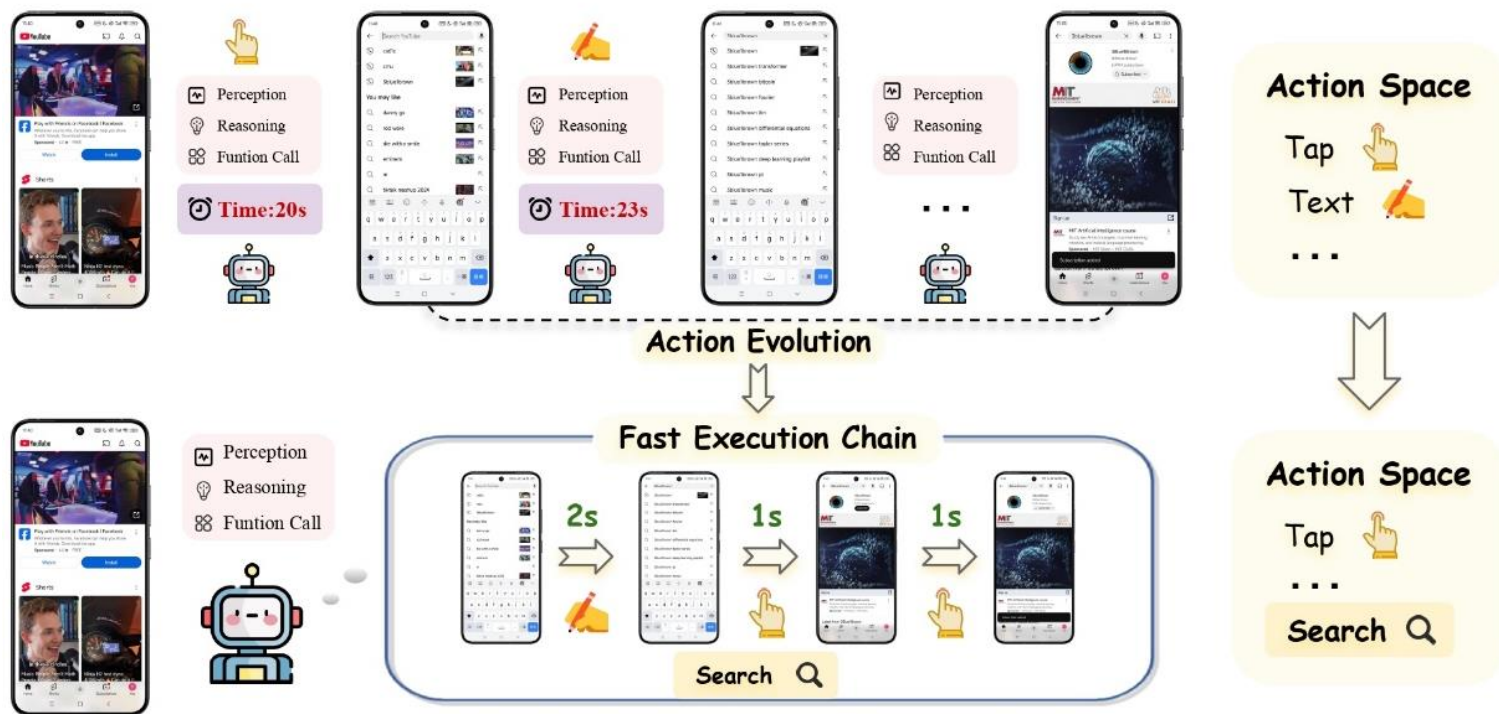
1. 进化的动作空间:

将重复的执行路径提取成“捷径”，在下次执行的过程中只需要利用“捷径”，而不是每一次都进行推理。

2. 动态的匹配策略:

在执行过程中，动态的匹配操作的元素与当前屏幕，抗干扰能力因此增强。

Task: "Subscribe '3Blue1 Brown' on YouTube" 



• 人是如何操作UI的?

➤ 泛化的**学习能力**

- 知道页面的跳转逻辑
- 能够猜测元素作用

➤ 习惯性动作的**肌肉记忆**

- 下拉刷新
- 侧滑菜单



让 LLM 像人类一样产生“肌肉记忆”，需要从用户的历史操作轨迹中进行存储与归纳，从而为智能体构建高级的动作空间。

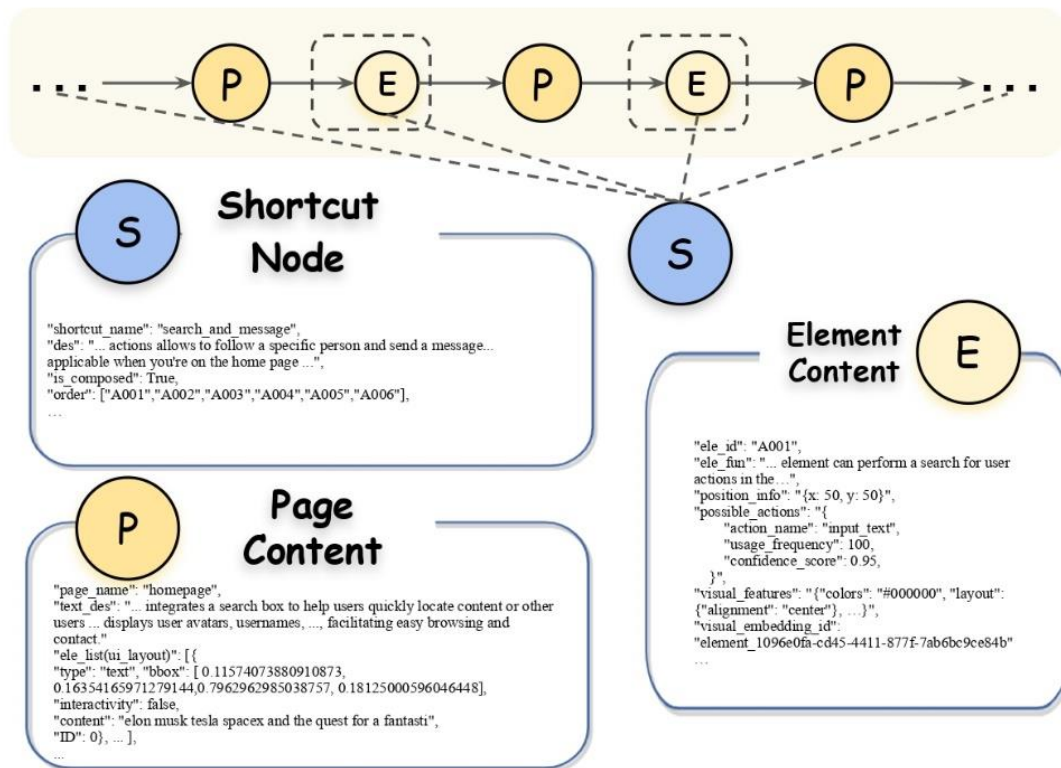


我们在操作UI的时候都发生了什么？



用户通过一系列操作完成屏幕的动态变化



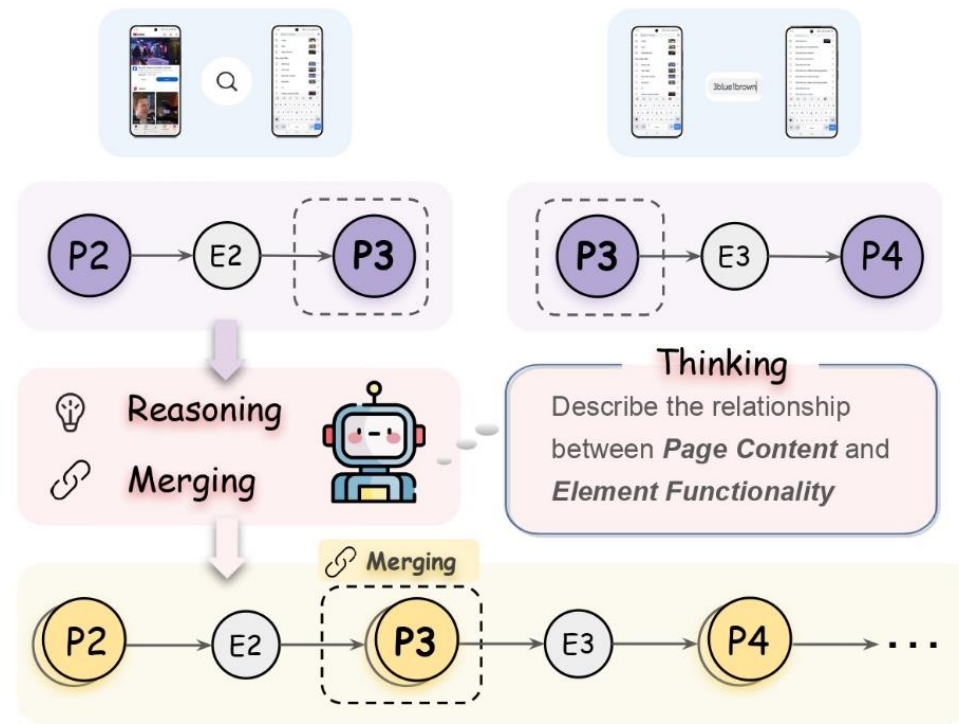


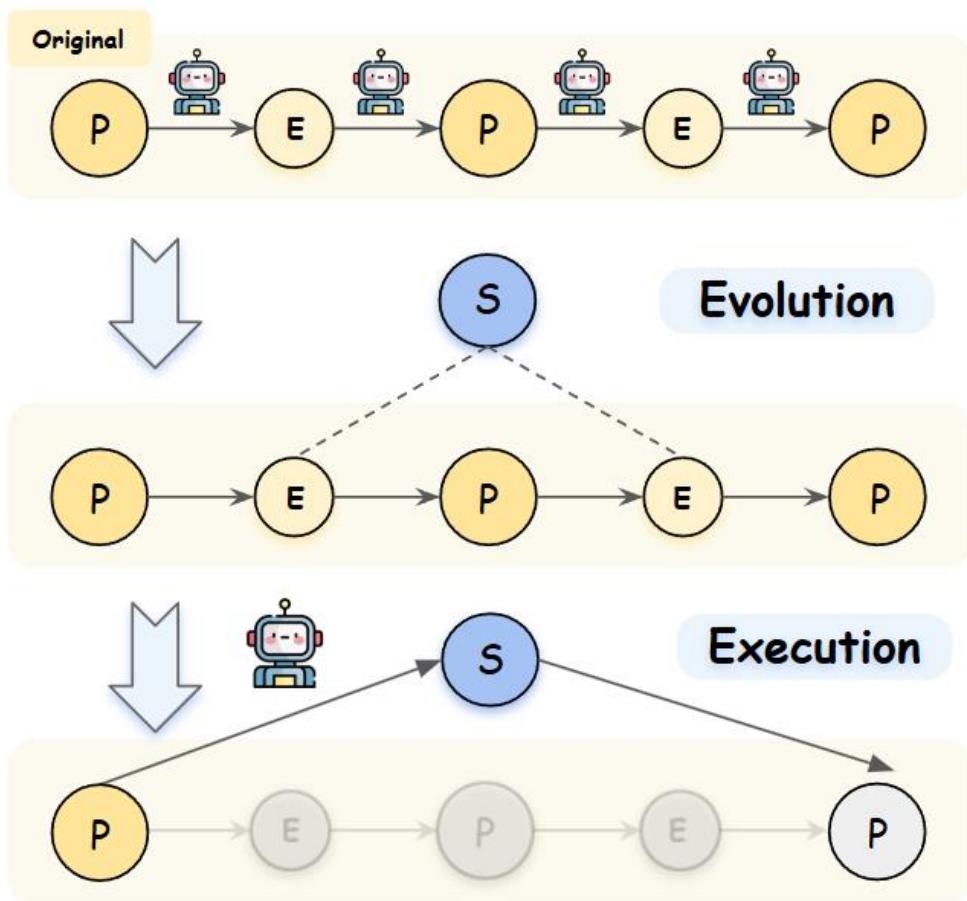
如何存储交互过程?

- 交互过程中必要的元素
 - 页面
 - 交互元素
 - 交互动作
- 交互过程实际上是在页面中对于相应的元素使用特定的交互动作的循环过程。
- 我们将这个过程抽象成一个由页面与页面包含的**元素的链条**，并记录相应的信息。
- 对于页面以及操作的含义使用LLM的视觉能力推理获得。

如何理解页面与元素?

- 对于一个操作链条, 每次采样一个 页面-元素-页面 (PEP) 三元组并使用LLM进行推理。
- 推理的内容:
 - 页面内容 (布局、功能、状态)
 - 动作含义 (跳转、切换、返回)
- 在进行采样推理的过程中, 会进行滑动重叠的方式。这样的设计使得三元组进行推理拥有更大的上下文, 并使得页面的理解更加鲁棒。
- 这样存储下的动作链条, 具有了更强的通用性与可迁移性。

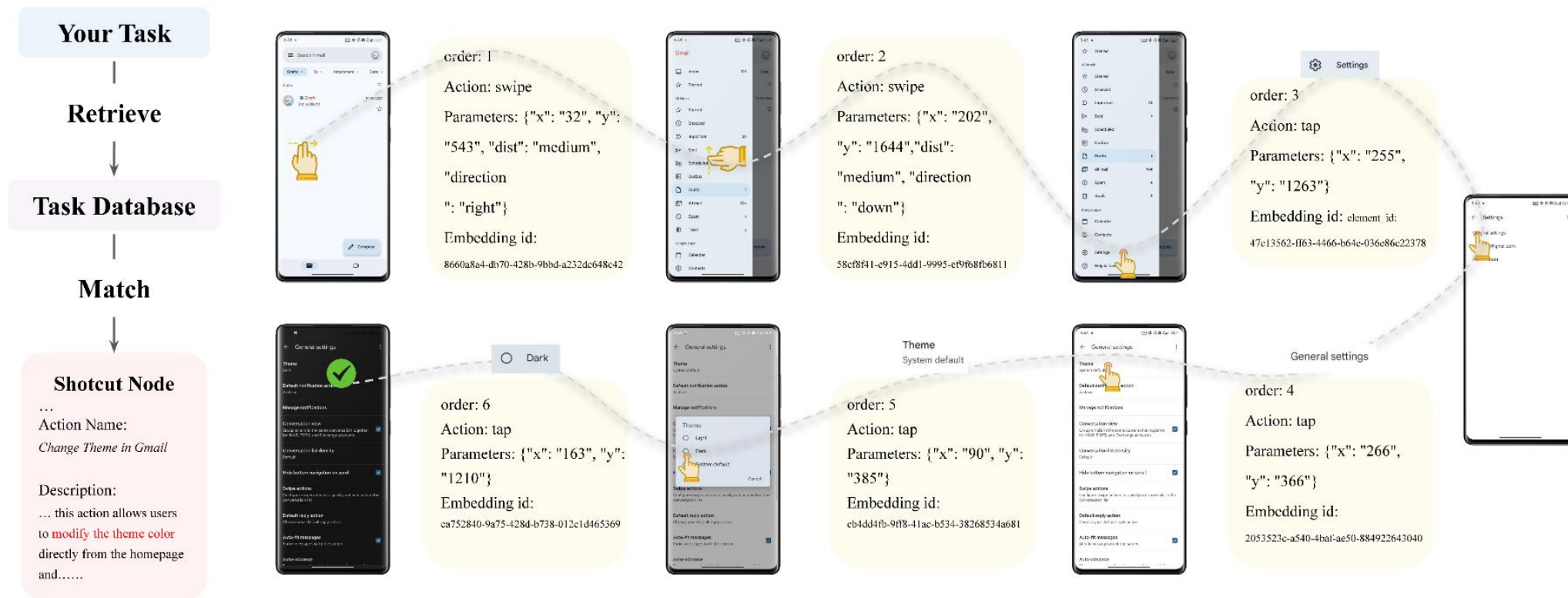




如何完成动作的进化?

- 在记录用户完整操作轨迹的基础上，通过合适的 Shortcut 跳跃连接，我们可以对复杂的操作序列进行抽象，从而形成一个高级动作。
- 一旦建立起这样的高级动作，仅需少量的 reasoning，就可以复用历史轨迹，实现对既有操作的“进化压缩”，大幅减少重复执行的成本。
- 这一过程中，系统逐步构建出更高级、更抽象的动作空间。随着用户轨迹数据的积累，高级动作的提取也变得更加鲁棒，系统对任务的理解和执行能力持续增强。

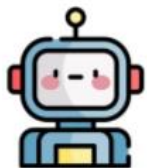
Task: Set Gmail's theme to dark mode

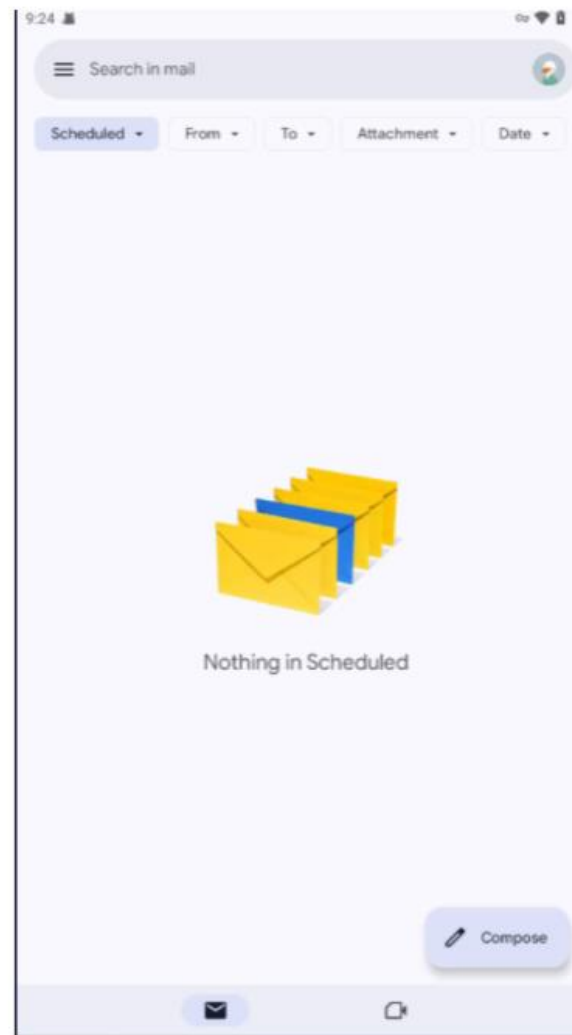


在执行的过程中，利用存储的链式记忆生成可能的执行步骤，减少每次都进行推理的耗时。当执行出现失败时，回退基础动作空间保证任务的成功率。

Task

Navigate on foot from
your current location to
Xihu and start the
navigateion.





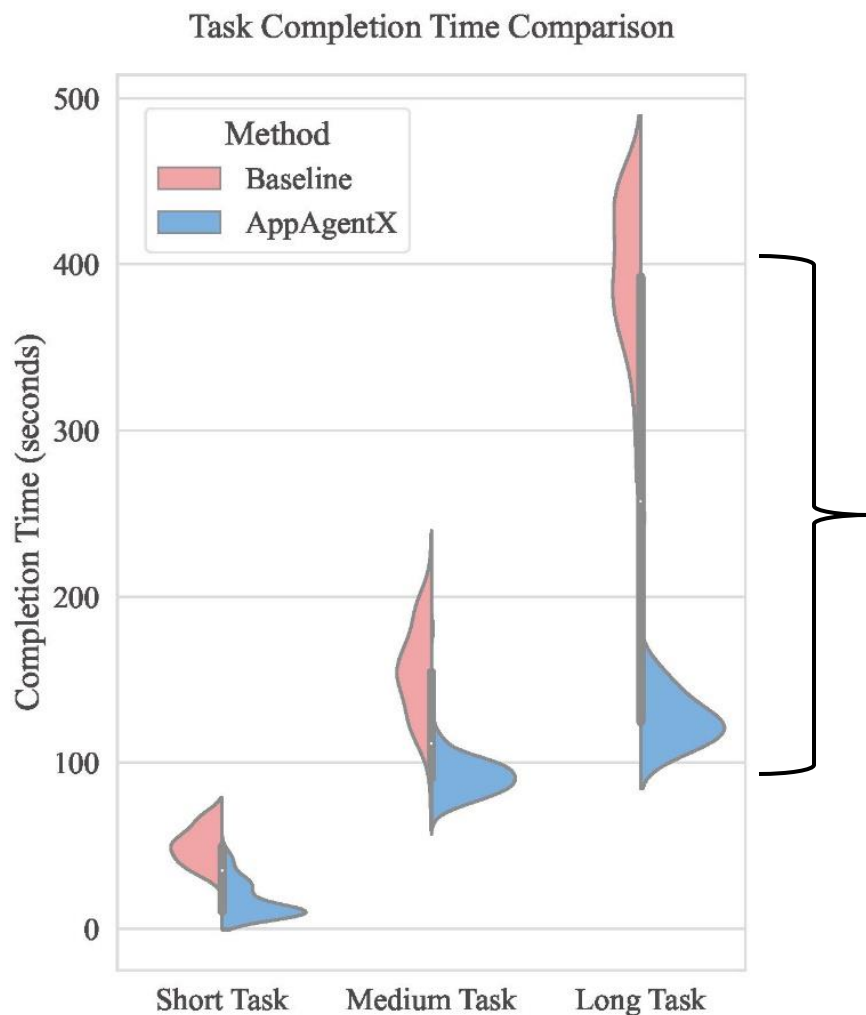
Method	Memory Type	Action Space	Steps↓	Step Time (s)↓	Tokens (k)↓	SR ↑
GPT-4o (Baseline)	None	Basic	10.8	26	6.72	16.9%
	Element	Basic	9.3	24	8.46	69.7%
AppAgentX	Chain	Basic	9.1	23	9.26	70.8%
	Chain	Basic+Evolve	5.7	16	4.94	71.4%

Table 1: **Analysis of Different Components in AppAgentX.** This table compares the performance differences resulting from the different designs with the baseline. Both our memory design and evolution mechanism can improve success rate and efficiency.

AppAgentX显著降低了平均每一步的执行时间，同时大幅降低了Token消耗



► AppAgentX对长任务的高效性

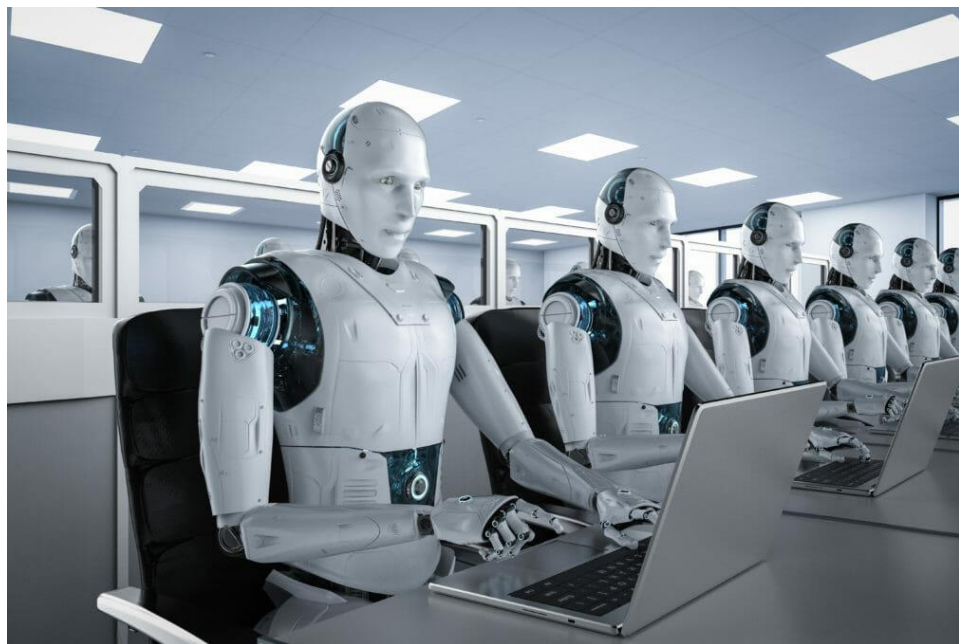


随着任务的长度与复杂度的提升，AppAgentX完成单个任务的总时间消耗明显减少。

这是因为AppAgentX利用进化机制完成了，对于耗时部分（reasoning）的压缩，使得操作的智能体拥有了长短时记忆。



- 智能助手
- 数字员工



- 自动化测试领域
- 人机交互方式的革命



► Significance and Impact

- Revolutionizing Human-Computer Interaction

CMD

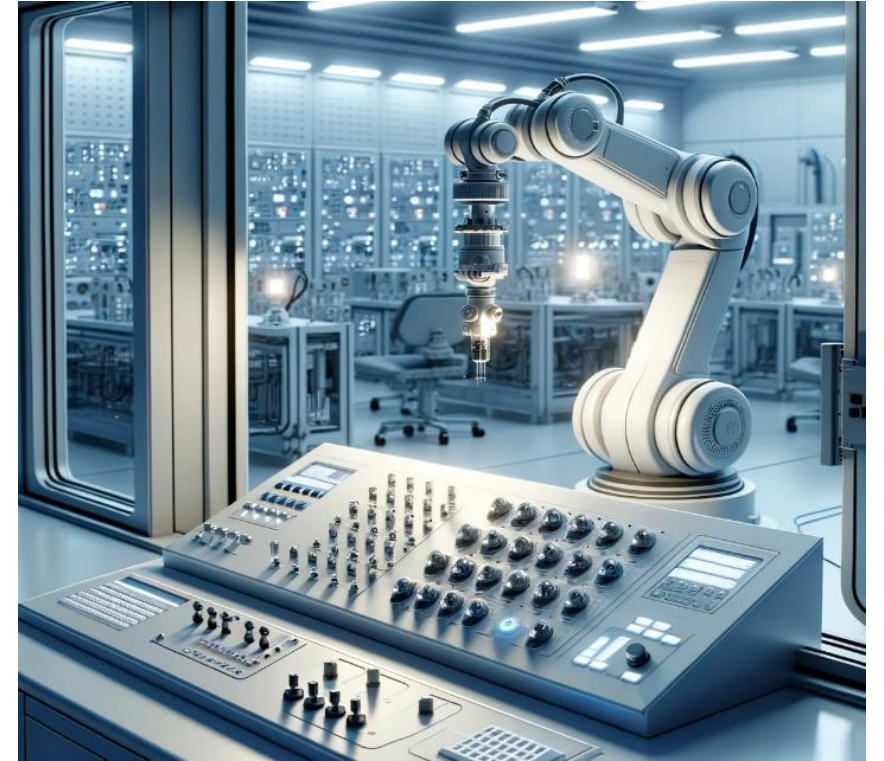
GUI

LUI



► Significance and Impact

- Digital employee and super assistants



►► Significance and Impact

- **Software testing and user study**



参与调研您将优先获得



AiDD定制版
《AI+软件研发精选案例》



专属学习顾问
1对1需求对接

AiDD会后小调研

AiDD峰会致力于协助企业利用AI技术深化计算机对现实世界的理解,推动研发进入智能化和数字化的新时代。作为峰会的重要共建者,您的真知灼见对我们至关重要。衷心感谢您的参与支持!



扫码参与调研

2025 AI+研发数字峰会

拥抱 AI 重塑研发

科技生态圈峰会 + 深度研习

——1000+ 技术团队的共同选择



敦煌站

K+ 思考周®研习社

时间: 2025.08.29-30



上海站

K+ 金融专场

时间: 2025.09.26-27



香港站

K+ 思考周®研习社

时间: 2025.11.17-18



K+峰会详情



上海站

AI+研发数字峰会

时间: 2025.05.23-24



北京站

AI+研发数字峰会

时间: 2025.08.08-09



深圳站

AI+研发数字峰会

时间: 2025.11.14-15



AiDD峰会详情

AiDD 峰会



2025 AI+研发数字峰会

AI+ Development Digital Summit

感谢聆听!

扫码领取会议PPT资料

