CoCo-Agent: A Comprehensive Cognitive MLLM Agent for Smartphone GUI Automation

Xinbei Ma[1,2,3,4], **Zhuosheng Zhang**[1,*], **Hai Zhao**[1,2,3,4*]
[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[3]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
[4]Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3
sjtumaxb@sjtu.edu.cn, zhangzs@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Multimodal large language models (MLLMs) have shown remarkable potential as human-like autonomous language agents to interact with real-world environments, especially for graphical user interface (GUI) automation. However, those GUI agents require comprehensive cognition including exhaustive perception and reliable action response. We propose a Comprehensive Cognitive LLM Agent, CoCo-Agent, with two novel approaches, comprehensive environment perception (CEP) and conditional action prediction (CAP), to systematically improve the GUI automation performance. First, CEP facilitates the GUI perception through different aspects and granularity, including screenshots and complementary detailed layouts for the visual channel and historical actions for the textual channel. Second, CAP decomposes the action prediction into sub-problems: determining the action type and then identifying the action target conditioned on the action type. With our technical design, our agent achieves state-of-the-art performance on AITW and META-GUI benchmarks, showing promising abilities in realistic scenarios. Code is available at https://github.com/xbmxb/CoCo-Agent.

1 Introduction

Graphical user interface (GUI) automation aims to enable human-like operations on operating systems with artificial intelligence instead of human efforts. Large language models (LLMs) have demonstrated commendable performance as human-like agents, showing emergent abilities of perceiving (Yao et al., 2022), reasoning (Li et al., 2023b; Park et al., 2023), and acting (Wang et al., 2023; Richards, 2023). With the multimodal enhancement, MLLM agents become promising autonomous GUI assistants to deal with complex tasks on behalf of human operators. To interact with the GUI environment, those agents require comprehensive cognition, including exhaustive perception and reliable action response.

Current vital challenges for autonomous agents lie in two aspects. One is **(i) the dependence on strong (M)LLMs**, and the other is **(ii) the insufficient GUI environment modeling**.

Although *strong (M)LLMs* like GPT-4V (OpenAI, 2023) and ChatGPT (Ouyang et al., 2022) ignite the development of autonomous agents, they exhibit shortcomings in realistic use. First, the alignment requires a trustworthy domain transfer as there is a large disparity between GUI commands and natural languages. GUI agents are expected to generate accurate and well-formulated responses as executable GUI commands, which is non-trivial for zero/few-shot prompting. For example, given the GUI that accepts commands as "`{action: click, touch_point:[`y_0`, `x_0`], touch_point:[`y_1`, `x_1`], typed_text: ''}`", semantically equivalent generations like "`Open the address book on your phone`" is plausible but unavailable. Second, the black-box APIs are likely to cause unexpected safety issues. Risks to privacy and integrity may arise when granting personal device authority to a black-box API. This significantly reduces realistic usability. Third, agent performance relies on the prompt design. The issues mentioned above leave heavy burdens on the design of prompt lines for those agents. Besides necessary environment descriptions, the prompts (and post-processing) must be sophisticated for domain alignment, instruction following, and security risk mitigation in different circumstances.

Second, GUI agents necessitate a comprehensive multimodal perception for the informative GUI environment modeling. Existing methods for visual language models are mainly endowed with favorable abilities in semantic alignment between the

[*] Corresponding authors. This research was supported by the Joint Research Project of Yangtze River Delta Science and Technology Innovation Community (No. 2022CSJGG1400), the Joint Funds of the National Natural Science Foundation of China (Grant No. U21B2020).

Findings of the Association for Computational Linguistics: ACL 2024, pages 9097–9110
August 11-16, 2024 ©2024 Association for Computational Linguistics

vision and language modalities (Dai et al., 2023; Ye et al., 2023; Zhao et al., 2023). However, GUI contains fine-grained details and intricate semantic connections, presenting a challenge for agents to comprehend (Rawles et al., 2023; Li et al., 2023a). Consider a screenshot that includes a magnifier icon, where the conventionally accepted meaning of "*search*" is conveyed. It implies a potential action through implicit semantics despite its small pixel size. Thus, only leveraging general image semantics like captioning is insufficient for GUI environment modeling. In addition, the GUI environmental perception is limited by the finite input window, where a balance between the visual and textual feature length needs to be struck.

This work proposes CoCo-Agent, a Comprehensive Cognitive MLLM Agent, to address the challenges above for smartphone GUI automation. CoCo-Agent adopts a multimodal backbone of LLaVA (Liu et al., 2023) and further enhances comprehensive cognition, respectively for exhaustive perception and reliable action response. The two proposed approaches are comprehensive environment perception (CEP) and conditional action prediction (CAP). Specifically, CEP integrates GUI perception elements of textual goal, historical action, and high-level and detailed description of the vision channel. CAP decomposes the complex and redundant GUI action commands into sub-problems following a *top-down* order. Our experiments cover diverse tasks in two GUI benchmarks, AITW (Rawles et al., 2023) and META-GUI (Sun et al., 2022a), including application manipulation, web operation, and dialogues. CoCo-Agent achieves SOTA performance with a limited parameter size. Subsequently, we present deep analyses including element ablation, visual module selection, and future action prediction. We show the significant effect of each perception element and the favorable choice of visual module. We also analyze the limitations of existing datasets and illustrate the additional potential of CoCo-Agent for realistic scenarios.

Our contributions are summarized as follows:

○ We propose CoCo-Agent, an autonomous agent with comprehensive cognition for GUI, with novel approaches to enhance the perception and action response, namely comprehensive environment perception (CEP) and conditional action prediction (CAP).

○ CoCo-Agent achieves state-of-the-art performance on representative GUI benchmarks, demonstrating superior performance.

○ Extensive analyses for a systematic study of GUI automation demonstrate our significant effectiveness and realistic potential.

2 Related Work

This section introduces studies on autonomous language agents and multimodal perception of LLMs.

2.1 Autonomous Language Agents

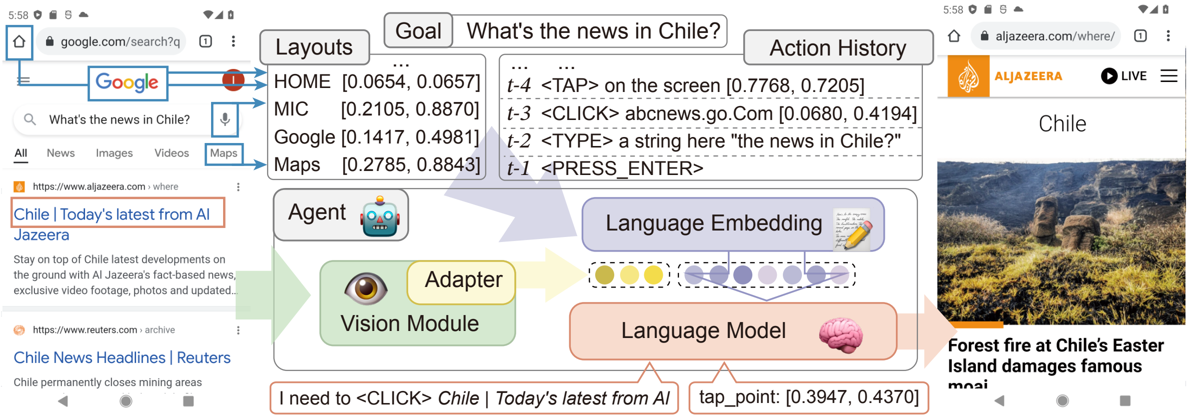
Recent studies (Li et al., 2023b; Richards, 2023) use the term *language agent* to refer to language models that interact with an environment or other agents to solve a problem. This paper investigates the autonomous language agents that perceive the environment and then act on the environment.

One research line relies on the strong fundamental competencies of (M)LLMs. Based on ChatGPT or GPT-4, autonomous agents can be built by only well-written prompts. Existing works have proved the reasoning, planning, and generalizing abilities of GPT-based agents, e.g., AutoGPT (Richards, 2023), BabyAGI (Nakajima, 2023), AgentGPT (Reworkd, 2023), HuggingGPT (Shen et al., 2023), and MM-Navigator (Yan et al., 2023).

However, when we expect practicality and reliability, we pursue the trainable language agent that can be customized and privatized to align with given environments (Shao et al., 2023). Thus, another research line turns to training methods for open-source language models. m-BASH (Sun et al., 2022b) adopted ROI pooling to present GUI icons in a BERT-based multi-task system. The Auto-GUI (Zhang and Zhang, 2023) was trained on a multimodal T5 (Raffel et al., 2020), formulating the GUI interaction to a first-principal form. CogAgent (Hong et al., 2023) integrated an extra attention-based high-resolution visual module with alignment pre-training. This paper follows the second research line to discuss the trainable, open-source language agent.

2.2 Multimodal Perception

Beyond language modeling, recent works have studied the fusion with channels of other modalities. Because of the development of LLMs, mainstream methods usually follow a language-centric framework, i.e., encoding information of other modalities into the language embedding space. These models consist of a pre-trained encoder of other


Goal
What's the news in Chile?
Layouts
Action History
Agent
Adapter
Vision Module
Language Embedding
Language Model
google.com/search?q
Google
What's the news in Chile?
All
News
Images
Videos
Maps
Chile | Today's latest from Al Jazeera
Stay on top of Chile latest developments on the ground with Al Jazeera's fact-based news, exclusive video footage, photos and updated...
Chile News Headlines | Reuters
Chile permanently closes mining areas
aljazeera.com/where/
ALJAZEERA
LIVE
Chile
Forest fire at Chile's Easter Island damages famous moai

HOME [0.0654, 0.0657]
MIC [0.2105, 0.8870]
Google [0.1417, 0.4981]
Maps [0.2785, 0.8843]

t-4 <TAP> on the screen [0.7768, 0.7205]
t-3 <CLICK> abcnews.go.Com [0.0680, 0.4194]
t-2 <TYPE> a string here "the news in Chile?"
t-1 <PRESS_ENTER>

I need to <CLICK> *Chile | Today's latest from Al*

tap_point: [0.3947, 0.4370]

Figure 1: Overview of CoCo-Agent, illustrating the perception and action response on a time step. *CEP* integrates the shown fine-grained elements. The predicted actions are formulated following *CAP*.

modalities, a language model, and an adapter (or a projector) as the bridge. For example, LLaVA (Liu et al., 2023) uses a linear layer to map the vision encoding from CLIP, while BLIP-2 (Li et al., 2023c) adopts a Q-former to learn a query vector to represent the image. This endeavor has given rise to the emergence of various multimodal LLMs, such as Flamingo (Alayrac et al., 2022), mPLUG (Ye et al., 2023), MiniGPT-4&v2 (Zhu et al., 2023; Chen et al., 2023), Video-LLaMA (Zhang et al., 2023b), and SpeechGPT (Zhang et al., 2023a).

However, the multimodal perception is even more challenging for GUI agents. Because GUI contains extensive detailed information with intricate semantic connections, such as very small icons conveying customary meanings (shown in Figure 1). A gap remains between existing visual LLMs and the perception necessitated for GUI agents.

3 Methodology

In this section, we will first formulate the GUI automation task and then propose our CoCo-Agent. Concretely, we will describe our technical designs of cognition, namely, comprehensive environment perception (CEP) and conditional action prediction (CAP), to improve the GUI automation performance systematically. Figure 1 shows an overall illustration.

3.1 Task Formalization

The task of GUI automation is defined as an interactive sequence generation problem. First, the user instructs the agent with a goal g that can be achieved in the GUI environment in several steps. At each step, the agent first perceives the present GUI *state*, s_t, and predicts the next *action* a_t, leading to the next GUI state s_{t+1}. The sequential (s, a) that accomplishes a goal forms an *episode*. An interaction record is formulated as

$$\textsc{Episode} = (g, [(s_t, a_t)]_{t=1}^{n}). \tag{1}$$

The action space is a finite operation command set with limited parameters. Examples are illustrated in Table 1. The state space includes any possible display from the smartphone. As the output recipient of agents is not human but a GUI, accurate actions are expected instead of flexible expressions like natural language.

3.2 Backbone

Our backbone follows LLaVA (Liu et al., 2023), which uncovers the generalization of LLM to vision modality. LLaVA consists of a language model (LM), Llama-2-chat-7B (Touvron et al., 2023), a vision encoder ($\textsc{Encoder}_{image}$), CLIP (Radford et al., 2021), and a one-layer linear projector ($\textsc{Prj}$) to bridge the image features to the space of language embedding ($\textsc{Embed}_{text}$). The input is denoted as X, including text X_{text} and image X_{image}, The output is denoted as Y. The backbone can be formulated as

$$\begin{aligned} H_{text} &= \textsc{Embed}_{text}(\ X_{text} \circ \hat{Y}^{0:t-1}\ ), \\ Z_{image} &= \textsc{Encoder}_{image}(\ X_{image}\ ), \\ H_{image} &= \textsc{Prj}(\ Z_{image}\ ), \\ H_t^{Decoder} &= \textsc{Decoder}(\ H_{image} \circ H_{text}^{t}), \\ P_t &= \text{LM}_{head}(\ H_t^{Decoder}\ ), \\ \mathcal{L} &= \sum_t \text{CE}(\ P_t, Y_t\ ), \end{aligned} \tag{2}$$

where $\circ$ denotes the concatenation operation. The training objective $\mathcal{L}$ is cross entropy (CE).

Action Type	Touch_point	Lift_point	Typed_text	Redefined Actions in CAP
PRESS_HOME	"[-1.0, -1.0]"	"[-1.0, -1.0]"	""	I need to <PRESS_HOME>
PRESS_BACK	"[-1.0, -1.0]"	"[-1.0, -1.0]"	""	I need to <PRESS_BACK>
PRESS_ENTER	"[-1.0, -1.0]"	"[-1.0, -1.0]"	""	I need to <PRESS_ENTER>
STATUS_TASK_COMPLETE	"[-1.0, -1.0]"	"[-1.0, -1.0]"	""	For this goal, no more action is needed, so <STATUS_TASK_COMPLETE>
TYPE	"[-1.0, -1.0]"	"[-1.0, -1.0]"	"{string}"	I need to <TYPE> a string here, "typed_text": "{string}"
DUAL_POINT	"{coordinate}"	"{coordinate}"	"{string}"	I need to <SCROLL> {direction}
DUAL_POINT	"{coordinate}"	"{coordinate}"	"{string}"	I need to <CLICK> {item name}, the location of {item name} is "tap_point": "{coordinate}"
DUAL_POINT	"{coordinate}"	"{coordinate}"	"{string}"	I need to <TAP> on the screen, the location is "tap_point": "{coordinate}"

Table 1: Illustration of JSON-formatted GUI commands in AITW (left) and our definition in CAP style (right). *"[-1.0, -1.0]"* follows the default value in AITW. *String*, *item name*, *coordinate*, and *direction* are required parameters.

3.3 Comprehensive Environment Perception

Environment perception is a crucial prerequisite for action responses. The environment can be simplified to only a GUI screenshot (Zhang and Zhang, 2023), which is highly subject to the upper-bound ability of the vision encoder. However, there is a bottleneck for the vision channel. First, the size of the encoder is restricted to a relatively low resolution, e.g., 224 × 224. Second, the existing pre-training objectives on vision encoders mainly focus on general, high-level semantic modeling, like image captioning (Radford et al., 2021; Li et al., 2023c). Thus, fine-grained information on the screen needs to be enhanced to complement the high-level perception.

Our proposed comprehensive environment perception fully leverages tools like optical character recognition (OCR) and IconNET (Sunkara et al., 2022), which gives fine-grained layouts with readable textual hints, e.g., "*ICON_SETTINGS: [0.1783, 0.8701]*". Besides the global goal, g, the environment state is perceived from three aspects, the present screenshot, X_{image}, the layouts from OCR, L, and the previous actions in the present episode, $a_{t-h:t-1}$. The total input can be noted as

$$X_{text} = \textsc{Prompt}(\ g, L, a_{t-h:t-1}\ ), X_{image}, \quad (3)$$

where Prompt is our prompt template (Appendix A). h denotes the number of action histories involved. The layouts L are listed *(item name, item coordinate)*, where *items* are OCR results or icons.

3.4 Conditional Action Prediction

Regarding action response, we propose to refactor GUI actions following the thinking order. As is shown in the left part of Table 1, existing GUI actions involve redundant parameters of each command, including the action type, the beginning coordinates, the ending coordinates, and the possible input text. However, these parameters are not independent of each other but show significant relations. For example, the coordinates depend on the action type. If the action is to click on an icon, then the touch and lift coordinates are accordingly determined. Predicting such JSON-formatted parameters would cause inconsistency and waste efforts.

Thus, we propose conditional action prediction. The GUI actions are refactored for relation decomposition as illustrated in Table 1. The actions are decomposed into two sub-problems to address, (i) action type prediction and (ii) optional action target prediction conditioned on the action type prediction. Also, we use natural language-like expressions without compromising the accuracy. As illustrated in Table 1, we change the action to a prompt line *step-by-step*, explicitly decomposing and clarifying those actions. Notably, the *dual_point* action is refined into three types: (i) *scroll* action, if the beginning and ending points are farther apart than the threshold (Rawles et al., 2023); (ii) *click* action involving *item name*, if the tap point falls in a bounding box; (iii) *tap* action, if it is not a *scroll* action but matches no bounding box.

In this way, the action prediction follows a *top-down* order. First, the agent decides on action types, conditioned on which the agent further decides on the target item and coordinates.

Normalization. Based on CEP and CAP, the actions are normalized to alleviate noise, which is inevitable in real-world data. Specifically, the target coordinates of *click* actions are normalized to the centroid of the bounding box from OCR. The *scroll* actions are normalized into four-direction swipes (Zhang and Zhang, 2023).

4 Experiments

This section will introduce the experimental settings including the dataset, implementation details, and baselines, followed by our empirical results.

4.1 Dataset

The following benchmarks of GUI automation are considered in the empirical evaluation. The dataset statistics are presented in Table 2.

AITW (Rawles et al., 2023) is a benchmark for smartphone GUI, containing 715K operation episodes under 30K reality intentions. Each entry includes a goal in natural language, screenshots, and actions as Eq. 1. Humans collect the data on various devices and operation systems in various screen resolutions. According to the applications domain, AITW consists of five subsets: General, Install, GoogleApps, Single, and WebShopping.

META-GUI (Sun et al., 2022b) smartphone dataset is originally released for multimodal dialogue generation. Differently, the agent is enabled to communicate with the user to verify the present state or further operation, e.g., "*Is this good?*". Eq. 1 is expanded with optional utterances,

$$(s_t, a_t) \rightarrow (s_t, a_t, u_t^{agent}, u_t^{user}), \qquad (4)$$

where utterances from the agent and the user are denoted as u^{agent} and u^{user}. These utterances cut an episode into several dialogue turns. META-GUI consists of 1k episodes with 18k steps. The data diversity lies in 11 applications of 6 domains including weather, calendar, search, etc.

AITW	Episode	Screen	Goal
General	9,476	85,413	545
Install	25,760	250,058	688
Google Apps	625,542	4,903,601	306
Single	26,303	85,668	15,366
Web Shopping	28,061	365,253	13,473
META-GUI	**Episode**	**Dial. turn**	**Screen**
Train	897	3692	14,539
Dev	112	509	1,875

Table 2: Dataset statistics.

4.2 Implementation

Our implementation adopts LLaVA (Liu et al., 2023) with a LLaMA-2-chat-7B and a vision encoder, CLIP.[1] The maximum input length is 2048 following the vision instruct tuning. For the subsets of AITW, our experiments include two setups, i.e., separate training on each subset and unified training on the whole set. Details are shown in A.

Metrics. Accuracy is computed at each time step of all parameters as our metric. The refactored action is parsed to JSON format and each parameter is compared to the action label following existing work (Rawles et al., 2023). The predicted coordinate is considered correct if it falls in the same element bounding box as the labeled one or falls within a 14% screen distance from the labeled one. A scroll action is correct if its main direction is correct. The accuracy of other parameters are exact match except for *typed_text* or dialogue responses. The typed text of AITW is correct if the label is in the predicted text. For META-GUI, F1 is computed for input text, and BLEU is computed for response generation. One action is regarded as correct if all the JSON fields are correctly predicted.

4.3 Baselines

For AITW, our proposed approach is compared with the following baselines.

- **Uni-modal API-based methods.** Rawles et al. (2023) and Zhang and Zhang (2023) have evaluated 5-shot performance on **PaLM-2** (Anil et al., 2023) and **ChatGPT** (Ouyang et al., 2022). The images are represented by pseudo HTML codes. The action target prediction is the item name or index, without verifying the coordinate numbers.
- **Multimodal API-based methods. MM-Navigator** (Yan et al., 2023) is a GPT-4V-based multimodal agent, achieving few-shot SOTA.
- **Training-based methods.** (i) **Behavioural Cloning** (Rawles et al., 2023) is a Transformer-based specialized agent that models goals, screens, and historical information using a BERT (Devlin et al., 2019). (ii) **LLaMA-2** is shown as a representative trainable uni-modal LLM with pseudo HTML code inputs instead of images. The results are from Zhang and Zhang (2023). (iii) **Auto-GUI** (Zhang and Zhang, 2023) bases on a multimodal encoder-decoder language model with T5 and BLIP. (iv) **CogAgent** (Hong et al., 2023) is a 9B-parameter visual LLM pre-trained for specializing in GUI understanding with a novel high-resolution cross module, which tops on AITW.

For META-GUI, we present baselines following Sun et al. (2022b) including **LayoutLMs** (Xu et al., 2020, 2021), **BERT**, and **m-BASH** (Sun et al., 2022b). All of those need training. m-BASH achieves SOTA which is a multi-task Transformer with Faster R-CNN (Ren et al., 2015) and ROI pooling for vision modeling.

4.4 Main Results

Tables 3 and 4 present the main experimental results. **Our method surpasses the baselines sig-**

[1] openai/clip-vit-large-patch14.

AITW	API	Modality	Unified	Overall	General	Install	GoogleApps	Single	WebShop.
PaLM-2	PaLM-2	*Text*	✓	39.6	–	–	–	–	–
ChatGPT	ChatGPT	*Text*	✓	7.72	5.93	4.38	10.47	9.39	8.42
MM-Navigator	GPT-4V	*Text+Vision*	✓	50.54	41.66	42.64	49.82	72.83	45.73
MM-Navigator$_{\text{w/ text}}$	GPT-4V	*Text+Vision*	✓	51.92	42.44	49.18	48.26	76.34	43.35
MM-Navigator$_{\text{w/ history}}$	GPT-4V	*Text+Vision*	✓	52.96	43.01	46.14	49.18	78.29	48.18
BC	N/A	*Text+Vision*	✗	68.7	–	–	–	–	–
BC $_{\text{w/ history}}$	N/A	*Text+Vision*	✗	73.1	63.7	77.5	75.7	80.3	68.5
LLaMA-2	N/A	*Text*	✗	28.40	28.56	35.18	30.99	27.35	19.92
Auto-GUI$_{\text{separate}}$	N/A	*Text+Vision*	✗	74.22	65.94	77.62	76.45	81.39	69.72
Auto-GUI$_{\text{unified}}$	N/A	*Text+Vision*	✓	74.27	68.24	76.89	71.37	84.58	70.26
CogAgent	N/A	*Text+Vision*	✗	76.88	65.38	78.86	74.95	**93.49**	71.73
LLaVA$_{\text{unified}}$	N/A	*Text+Vision*	✓	70.37	58.93	72.41	70.81	83.73	65.98
CoCo-Agent$_{\text{separate}}$	N/A	*Text+Vision*	✗	77.82	69.92	80.60	75.76	88.81	74.02
CoCo-Agent$_{\text{unified}}$	N/A	*Text+Vision*	✓	**79.05**	**70.96**	**81.46**	**76.45**	91.41	**75.00**

AITW	Model	Action	Act. type	CoT. type	Item	Direction	Input(F1)
General	LLaVA$_{\text{unified}}$	58.93	80.08	N/A	56.76	63.31	93.29
	CoCo-Agent$_{\text{unified}}$	70.96	87.49	76.72	68.91	75.80	97.10
Install	LLaVA$_{\text{unified}}$	72.41	85.11	N/A	72.52	70.20	94.31
	CoCo-Agent$_{\text{unified}}$	81.46	90.82	85.12	81.52	80.49	97.36
GoogleApps	LLaVA$_{\text{unified}}$	70.81	88.49	N/A	65.55	74.95	98.75
	CoCo-Agent$_{\text{unified}}$	75.30	92.10	79.80	70.03	82.03	99.03
Single	LLaVA$_{\text{unified}}$	83.73	88.19	N/A	85.63	83.95	93.83
	CoCo-Agent$_{\text{unified}}$	91.41	95.34	92.49	91.84	92.74	98.15
WebShopping	LLaVA$_{\text{unified}}$	65.98	85.43	N/A	64.81	68.61	92.60
	CoCo-Agent$_{\text{unified}}$	76.10	89.80	80.20	73.88	78.48	96.96

Table 3: Results on AITW. Part 1: Action accuracy, where primary setups are labeled: the reliance on API backends ("API"), the perceptual modalities ("Modality"), and the general ability across subsets ("Unified"). Part 2: Detailed parameter accuracies comparing our unified CoCo-Agent and LLaVA baseline.

META-GUI	API	Modality	Action	Act. type	Item	Direction	Input (F1)	Utter. (BLEU)
LayoutLM	N/A	*Text*	67.76	82.22	71.98	94.87	90.56	50.43
LayoutLMv2	N/A	*Text+Vision*	64.48	85.60	64.38	92.95	70.76	58.20
BERT	N/A	*Text*	78.42	87.52	82.84	93.59	97.24	62.19
m-BASH	N/A	*Text+Vision*	82.74	90.80	85.90	96.42	94.23	63.11
LLaVA	N/A	*Text+Vision*	76.27	87.47	77.49	98.18	96.06	67.24
LLaVA $_{\text{w/ history}}$	N/A	*Text+Vision*	81.08	91.68	81.23	97.62	96.93	66.57
CoCo-Agent	N/A	*Text+Vision*	**88.27**	**92.59**	**91.72**	**98.39**	96.15	65.90

Table 4: Results on META-GUI.

nificantly and achieves overall state-of-the-art performance (except for Single subset). The unified model shows consistent advances compared to the separate training, indicating that the model learns generality across various situations.

The lower part of Table 3 shows the detailed performance on AITW. **CoCo-Agent is enabled to mimic the behavior patterns on GUI, while the limitations lie in predicting target items and scroll directions.** (i) The action type scores achieve around 90%. This high level indicates that the agent can learn the action patterns to operate GUI. The lower CoT type scores indicate that it is harder for the agent to differentiate *dual point* actions. This is reasonable as these action types are more flexible than others whose effects are more definite (like *press back* and *press home*). (ii) Predicting items (the locations to click) and directions are more difficult for agents. Especially, the item accuracy is close to the action accuracy. (iii) Input prediction is relatively easy, and F1 scores are up to 97%.

On META-GUI, we can also observe a significant improvement in action accuracy (12%). The accuracy of the item increases by a very large margin, while the action type and swipe direction accuracy is close to the perfect score. The input and utterance predictions are relatively consistent.

5 Analysis

This section presents further analysis and discussions. Section 5.1 shows the effects of comprehensive cognition elements by ablation and replacement. Section 5.2 discusses the capability of visual modules, followed by future action prediction in Section 5.3. Dataset features are analyzed in Section 5.4, including action type categories and human evaluation for realistic scenarios.

5.1 Effects of Environment Elements

5.1.1 Ablation

Our method combines CEP and CAP to characterize the GUI environment. We ablate each CEP element along with the refactoring method of CAP to observe their significance for the CoCo-Agent. Results are shown in Table 5. The improvement of each element is significant, especially for layouts (+5.82%) and action history (+5.63%).

Besides the coordinates, the layouts provide the item names like icon names and texts shown on the screen. When combined with CAP, they bridge the prediction through rationales, making predictions easier than direct coordinate grounding. Notably, although no historical screens are provided, historical actions with complete parameters lead to better scores than historical action types.

Goal	Img.	CAP	Layout	History	**General**	**META**
✓	✓	✗	✗	0	57.81	73.60
✓	✓	✓	✗	0	58.47	76.90
✓	✓	✓	✓	0	64.29	85.55
✓	✓	✓	✓	8 act. types	67.80	86.99
✓	✓	✓	✓	8 full actions	69.92	88.27

Table 5: Action accuracy for ablation studies on General of AITW and META-GUI ("META"). "Act. types" denotes only to provide history action types, while "full actions" denotes to provide actions with complete parameters.

5.1.2 Replacement

Actions on GUI can be flexible and the dependency on the environment is complex. Thus, we design a related task to probe the environment modeling further. We randomly choose one element from the goal, the screen, the layouts, and the action history and replace it with another from a random data point. With such corrupted input, the agent is trained to select the replaced element and also predict the next action.

Replace	None	Goal	Img	Layout	Hist.	Avg.
Detection	N/A	94.60	93.94	91.02	92.69	93.06
Action	70.96	63.21	59.69	57.55	58.45	59.73

Table 6: Results after replacement on General subset. *Detection* denotes the accuracy of the replaced element selection. *Action* denotes action accuracy.

The results are consistent with ablations and intuitions. (i) The wrong image and goal are more obvious replacements and get higher detection accuracy, while the layouts and action history are more complex to distinguish. (ii) Regarding action prediction, the accuracy decreases more with wrong layouts or wrong action history. Those are hard to memorize and require complex modeling, therefore rely more on correct inputs. (iii) The damage caused by a wrong image is limited due to the complement from layouts. This suggests again that layouts are important fine-grained complements for the screen image, while the image gives an overall impression and the layouts describe detailed information.

Model	Vision Encoder	LM	Acc
Auto-GUI	BILP-2 Encoder	FLAN-Alpaca	65.9
MMICL[2]	Q-Former	Flan-T5-xxl	56.4
mPLUG[2]	Abstractor	Vicuna	53.0
CogAgent	Low & High-Res.	Vicuna	65.4
LLaVA	CLIP	Llama-2-chat	58.9
Ours	CLIP & Layout	Llama-2-chat	71.0

Table 7: GUI agents with different vision encoders. Llama-2-chat and Vicuna are in 7B size. "Low & High-Res." is short for "low and high-resolution encoding".

Train	1-next	3-next	3-next	3-next
Test	1-next	1-next	2-next	3-next
Accuracy	70.96	58.90	43.40	37.74

Table 8: Future action prediction accuracy.

5.2 Visual Capability

The vision encoder and projector largely influence the visual capability of GUI agents. We compare a range of visual LMs with various vision encoders.

The results are shown in Table 7. Our integrated CLIP with projector encodes an image to a 256-length vector with a 4096 hidden size.[1] With fine-grained layouts, CoCo-Agent receives exhaustive visual information. Auto-GUI uses a BLIP-2 with pooling leading to a 1-length image vector, which is integrated into language embedding by an attention-based fusion module. Differently, MMICL adopts Q-Former to learn a 32-length query vector, and mPLUG adopts Abstractor to learn a 64-length vector. [2] CogAgent uses an EVA2-CLIP-E (Sun et al., 2023) as a *low-resolution encoder* and an EVA2-CLIP-L (Sun et al., 2023) with cross-attention layers as a *high-resolution encoder*. In cross-attention

[2] The first 1000 samples in *General*, the most difficult subset of AITW.

<table>
<tr><th>Dataset</th><th>Model</th><th colspan="2">Dual Point</th><th colspan="2">Text</th><th colspan="2">Press Back</th><th colspan="2">Press Home</th><th colspan="2">Press Enter</th><th colspan="2">Complete</th></tr>
<tr><td></td><td></td><td>Prop.</td><td>Acc</td><td>Prop.</td><td>Acc</td><td>Prop.</td><td>Acc</td><td>Prop.</td><td>Acc</td><td>Prop.</td><td>Acc</td><td>Prop.</td><td>Acc</td></tr>
<tr><td>General</td><td>LLaVA</td><td>86.09</td><td>54.16|84.71</td><td>10.90</td><td>86.09</td><td>1.17</td><td>8.16</td><td>5.36</td><td>77.73</td><td>2.61</td><td>41.55</td><td>10.67</td><td>62.53</td></tr>
<tr><td></td><td>CoCo-Agent</td><td></td><td>67.73|91.59</td><td></td><td>84.34</td><td></td><td>15.31</td><td></td><td>89.76</td><td></td><td>57.99</td><td></td><td>78.19</td></tr>
<tr><td>Install</td><td>LLaVA</td><td>69.26</td><td>72.53|90.77</td><td>11.77</td><td>92.19</td><td>1.96</td><td>14.98</td><td>5.79</td><td>67.38</td><td>0.81</td><td>12.69</td><td>10.38</td><td>67.66</td></tr>
<tr><td></td><td>CoCo-Agent</td><td></td><td>80.46|93.84</td><td></td><td>93.20</td><td></td><td>41.35</td><td></td><td>82.19</td><td></td><td>68.53</td><td></td><td>83.15</td></tr>
<tr><td>GoogleApps</td><td>LLaVA</td><td>78.42</td><td>71.41|93.95</td><td>1.54</td><td>75.73</td><td>1.34</td><td>11.13</td><td>6.00</td><td>74.14</td><td>0.07</td><td>18.18</td><td>12.63</td><td>71.52</td></tr>
<tr><td></td><td>CoCo-Agent</td><td></td><td>75.41|95.04</td><td></td><td>80.11</td><td></td><td>23.93</td><td></td><td>86.60</td><td></td><td>39.39</td><td></td><td>83.40</td></tr>
<tr><td>Single</td><td>LLaVA</td><td>49.28</td><td>79.04|88.09</td><td>14.06</td><td>89.74</td><td>0.17</td><td>57.14</td><td>0.23</td><td>89.47</td><td>4.52</td><td>72.43</td><td>31.74</td><td>90.06</td></tr>
<tr><td></td><td>CoCo-Agent</td><td></td><td>88.94|96.92</td><td></td><td>93.22</td><td></td><td>64.29</td><td></td><td>100.00</td><td></td><td>75.95</td><td></td><td>96.73</td></tr>
<tr><td>WebShop.</td><td>LLaVA</td><td>72.64</td><td>64.78|91.55</td><td>11.96</td><td>86.51</td><td>0.56</td><td>14.00</td><td>3.82</td><td>75.67</td><td>3.29</td><td>36.32</td><td>7.73</td><td>57.08</td></tr>
<tr><td></td><td>CoCo-Agent</td><td></td><td>72.52|94.09</td><td></td><td>88.72</td><td></td><td>28.00</td><td></td><td>89.15</td><td></td><td>73.90</td><td></td><td>74.04</td></tr>
</table>

Table 9: Accuracy of different types of actions. The agents are in the unified training setting. *Prop.* is type proportion in datasets. *Acc* is action accuracy|type accuracy for *Dual Point* and action accuracy for others.

layers, the high-resolution image features interact with each layer of the language model.

The models with only an image encoder outperform those with learnable queries. Learning queries from *image-text attention* can be unsuitable for GUI tasks, as input texts are complex and different from generic captions. Besides the overall semantic impression of images, straightforward textual layouts can work as an even better high-resolution module for image detail enhancement.

5.3 Future Actions

This section considers a more challenging setting of $n-$next actions prediction. The task is much less trivial as the agent only receives the environment state s_t, without the perception of future states $s_{t+1:t+n-1}$. Thus, predicting future actions $a_{t:t+n-1}$ involves harder reasoning, planning, and environment simulation. Table 8 shows the results with $n = 3$. Although the next action can be predicted with 70.96% accuracy, predicting the following actions without environmental feedback remains to be improved.

5.4 Dataset Features

5.4.1 Action Type Categories

Each action type has very different proportions in AITW, which is decided by the unbalanced distribution in natural operations. For example, the *click* action is the most frequent but the *complete* action appears at most once in each episode. Thus, we divide datasets into categories according to the ground truth action type. Table 9 shows the proportion and action accuracy. (i) *Dual point* action (including click, tap, and scroll) accounts for 69.26% - 86.09% in long-episode tasks and accounts for around half even in the Single subset. Other types, especially *press back* and *press enter* consistently account for low proportions. Such an unbalance can limit the performance of less frequent actions. (ii) For *dual point* type, the data is sufficient and type accuracy scores are up to above 90%. The action accuracy scores are limited by the difficulties in predicting target items and directions as shown in Table 3. (iii) Single subset shows better performance on all action types and more significant on less frequent actions. This is because the samples in Single are segmented sub-goals that give clear instructions and require fewer steps.

5.4.2 Potential for Realistic Scenarios

There is a disparity between the evaluation and realistic scenarios. The benchmarks show randomness because the actions can be stochastically chosen with different paths for the same goal. However, when the predicted action indicates an alternative path, it is reasonable but fails to match the label. This leads to an underestimation. Thus, the agent has extra potential in practice.

We study the first 500 samples of General dataset to see this phenomenon. In the first 500 actions, there are 118 (23.6%) actions whose predictions are different from the labels. Our human evaluation considers two criteria. We check the predicted action with the goal, the last action, the present screen, and the next screen to see if (i) the predicted action can result in a situation that is similar to the next screen, or (ii) the predicted action is consistent with the goal. We observed that only 54 (10.8%) samples strictly contradict the episode, while other 64 predictions do not betray the goal. Among the 64 samples, there are 25 (5%) predicted actions that can lead to a similar next state. For example, after typing a query in the search bar of a search engine, many actions can lead to the search results including *press enter*, *click the magnifier icon*, or *click a proper query suggestion* (Figure 3 and 4).

6 Conclusion

This paper proposes CoCo-Agent, an MLLM-based autonomous agent for smartphone GUI. Our method facilitates comprehensive cognition with exhaustive perception and reliable action response. CoCo-Agent is enhanced with two approaches. Comprehensive environment perception (CEP) integrates GUI perception, and conditional action prediction (CAP) enables a decomposition of complex commands. CoCo-Agent achieves SOTA performance in experiments on two GUI datasets. Further analysis shows that the agent learns the behavior patterns in smartphone GUI. The significance of the proposed enhancements is also verified. We discuss the unbalanced category distributions and the underestimation of agent performance. Our work reveals the promising capabilities of MLLMs as autonomous agents, especially for complex environment perception and action response.

Limitations

We acknowledge the limitations of this work. (i) Resource consumption. The data is of a large amount. The training process costs computational resources compared to the zero-shot methods. Whereas, the training process only needs to be conducted once. And our unified model achieves the SOTA performance. Efficient LLM training or inference methods can improve the balance between resource cost and performance. (ii) More complex settings. As is shown in Section 5.3, the predictions of future action remain to be improved. The ultimate goal is to empower the agent to operate the full episodes in a simulated GUI environment or a realistic device.

Future Work

The following directions for future work are proposed based on our study. First, GUI agents require generalization ability to support new instructions, new applications, or even new operating systems. The effectiveness of CoCo-Agent in task adaptation across domains of AITW has been preliminarily validated. Second, GUI agents require improved multimodal training strategies. Multimodal LLMs can be strengthened by integrating GUI perception into vision-language pre-training or instruction tuning. Third, GUI agents require comprehensive measurements. The measurements can be improved to reflect different paths for the same goal in practical scenarios. Last but not least, there is still room for performance improvement (79.05% for AITW and 88.27% for META-GUI) and future action reasoning and planning for full-episode prediction.

Ethics Statement

This section presents ethics statements in the following aspects. (i) Data Privacy. The research dataset AITW (Rawles et al., 2023) has declared control over Personal Identifiable Information. The instructions contain no toxic, biased, or misleading content. The META-GUI is based on a task-oriented dialogue dataset, SMCalFlow (Andreas et al., 2020), crowdsourced on Amazon Mechanical Turk with qualification requirements. CoCo-Agent does not rely on LLM APIs, preserving privacy data from leakage to LLM companies. (ii) System security. Compared to the command-line interface (CLI), the GUI is more interpretable and controllable. Since GUI actions follow human behavior, security considerations can better align with human-oriented mechanisms, which are already implemented in existing GUIs for operating systems. However, the potential risks of GUI agents have yet to be well explored (Yuan et al., 2024; Yang et al., 2024). (iii) Potential social impacts. On the one hand, GUI automation can facilitate efficiency and save human resources for more high-level work. On the other hand, malicious actors could abuse GUI agents to achieve undesirable purposes. For practical applications of GUI agents, platforms may need to update detection, authorization, and governance mechanisms to control potential risks for social impacts.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, et al. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *ArXiv preprint*, abs/2305.10403.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *ArXiv preprint*, abs/2310.09478.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint*, abs/2305.06500.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023. Cogagent: A visual language model for gui agents. *ArXiv preprint*, abs/2312.08914.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. 2023a. Otterhd: A high-resolution multi-modality model. *ArXiv preprint*, abs/2311.04219.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for" mind" exploration of large scale language model society. *ArXiv preprint*, abs/2303.17760.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint*, abs/2301.12597.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Yohei Nakajima. 2023. Babyagi. https://github.com/yoheinakajima/babyagi.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy P Lillicrap. 2023. Androidinthewild: A large-scale dataset for android device control. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Reworkd. 2023. Agentgpt. https://github.com/reworkd/AgentGPT.

Toran Bruce Richards. 2023. Auto-gpt: An autonomous gpt-4 experiment. https://github.com/Significant-Gravitas/Auto-GPT.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *ArXiv preprint*, abs/2303.17580.

Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. 2022a. META-GUI: Towards multi-modal conversational agents on mobile GUI. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6699–6712, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liangtai Sun, Xingyu Chen, Lu Chen, Tianle Dai, Zichen Zhu, and Kai Yu. 2022b. META-GUI: Towards multi-modal conversational agents on mobile GUI. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 6699–6712, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *ArXiv preprint*, abs/2303.15389.

Srinivas Sunkara, Maria Wang, Lijuan Liu, Gilles Baechler, Yu-Chung Hsiao, Abhanshu Sharma, James Stout, et al. 2022. Towards better semantic understanding of mobile interfaces. *arXiv preprint arXiv:2210.02663*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *ArXiv preprint*, abs/2305.16291.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2579–2591, Online. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM.

An Yan, Zhengyuan Yang, Wanrong Zhu, Kevin Lin, Linjie Li, Jianfeng Wang, Jianwei Yang, Yiwu Zhong, Julian McAuley, Jianfeng Gao, et al. 2023. Gpt-4v in wonderland: Large multimodal models for zero-shot smartphone gui navigation. *ArXiv preprint*, abs/2311.07562.

Wenkai Yang, Xiaohan Bi, Yankai Lin, Sishuo Chen, Jie Zhou, and Xu Sun. 2024. Watch out for your agents! investigating backdoor threats to llm-based agents. *ArXiv*, abs/2402.11208.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. ReAct: Synergizing reasoning and acting in language models. volume abs/2210.03629.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality.

Tongxin Yuan, Zhiwei He, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, and Gongshen Liu. 2024. R-judge: Benchmarking safety risk awareness for llm agents. *arXiv preprint arXiv:2401.10019*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *ArXiv preprint*, abs/2306.02858.

Zhuosheng Zhang and Aston Zhang. 2023. You only look at screens: Multimodal chain-of-action agents. *ArXiv preprint*, abs/2309.11436.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *ArXiv preprint*, abs/2309.07915.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv preprint*, abs/2304.10592.

A Implementation Details

Prompt Template. We show the prompt templates following an example. <image> is the special token for image position in LLaVA.

Prompt template for GUI agent

<image>
{item name}$_0$ location: {item coordinate}$_0$
...
{item name}$_1$ location: {item coordinate}$_1$
...
Previous Actions:
{a_{t-h}}
...
{a_{t-1}}
Goal: {g}
Next action:

Example of input and output

Input:
<image> ICON_HOME location: [0.0654, 0.0657]
ICON_THREE_DOTS location: [0.0649, 0.9213]
google.com/search?q location: [0.0689, 0.4704]
Google location: [0.1417, 0.4981]
ICON_THREE_BARS location: [0.1412, 0.0796]
ICON_MIC location: [0.2105, 0.8870]
ICON_MAGNIFYING_GLASS location: [0.2132, 0.1074]
What’s the news in Chile? location: [0.2136, 0.4648]
Al location: [0.2772, 0.0722]
News location: [0.2768, 0.2343]
Images location: [0.2789, 0.4481]
Videos location: [0.2772, 0.6759]
Maps location: [0.2785, 0.8843]
ICON_THREE_DOTS location: [0.3408, 0.9407]
4 location: [0.3425, 0.0667]
https://www.aljazeera.com> where location: [0.3425, 0.4120]
Chile | Today’s latest from Al location: [0.3947, 0.4370]
Jazeera location: [0.4316, 0.1574]
Stay on top of Chile latest developments on location: [0.4803, 0.4667]
the ground with Al location: [0.5101, 0.2194]
Jazeera’s fact-based location: [0.5079, 0.6083]
news, location: [0.5114, 0.8759]
exclusive video footage, location: [0.5395, 0.2778]
photos location: [0.5395, 0.5907]
and location: [0.5373, 0.7056]
updated... location: [0.5395, 0.8463]
ICON_THREE_DOTS location: [0.6132, 0.9407]
https://www.reuters.com» archive location: [0.6132, 0.3991]
Chile News Headlines |Reuters location: [0.6658, 0.4778]
Chile permanently closes location: [0.7136, 0.2889]
mining areas location: [0.7140, 0.6713]
Chile files location: [0.7421, 0.7019]
connected to giant sinkhole location: [0.7430, 0.3139]
charges against mining company for giant... location: [0.7724, 0.4694]
ICON_THREE_DOTS location: [0.8452, 0.9407]
https://www.independent.co.uk> topic location: [0.8461, 0.4324]
Chile location: [0.8978, 0.1167]
latest news, location: [0.8991, 0.4111]
breaking location: [0.9009, 0.7157]
- location: [0.8991, 0.2167]
stories and comMent - The location: [0.9338, 0.4241]
ICON_NAV_BAR_CIRCLE location: [0.9693, 0.4963]
ICON_NAV_BAR_RECT location: [0.9693, 0.7463]
ICON_V_BACKWARD location: [0.9697, 0.2454]
Previous Actions:I need to <PRESS_HOME>
I need to <TAP> on the screen, the location is "tap_point": "[0.7768, 0.7205]"
I need to <CLICK> abcnews.go.Com, the location of abcnews.go.Com on the screen is "tap_point": "[0.0680, 0.4194]"
I need to <TYPE> a string here, "typed_text": "Whats the news in Chile?"
I need to <TYPE> a string here, "typed_text": ""
I need to <PRESS_ENTER>
Goal: What’s the news in Chile? Next action:
Output:
I need to <CLICK> Chile | Today’s latest from Al, the location of Chile | Today’s latest from Al on the screen is "tap_point": "[0.3947, 0.4370]"

Experiment Details. For AITW, we follow the training and test set splits on *General*, *Install*, *Google Apps*, and *Web shopping* of Rawles et al. (2023), and use the splits on *Single* of (Zhang and Zhang, 2023), which is more available. Each subset is split for training, validation, and test sets by 8:1:1. The first 1000 samples of the test set are used as the validation sets. The reported scores are the results of the full test sets. For META-GUI, we follow the original dataset splits. The maximum length of action history is 8-lastest action. The model is trained for {8,10,12} epochs, with a learning rate of 2e-5. The batch size is set to {12,16} per device. All experiments are conducted on 4 Nvidia A800 GPUs.

B Examples

We show examples of a full episode in Figure 2. Examples of the randomness in AITW are illustrated here. Figure 3 is a case where the predicted action leads to the same situation but does not match the label (i, in Section 5.4.2). Figure 4 is a case where the predicted action is different from the label but is still reasonable for the goal (ii, in Section 5.4.2). Blue boxes highlight the target items on the screen for label actions and orange boxes for predicted actions.

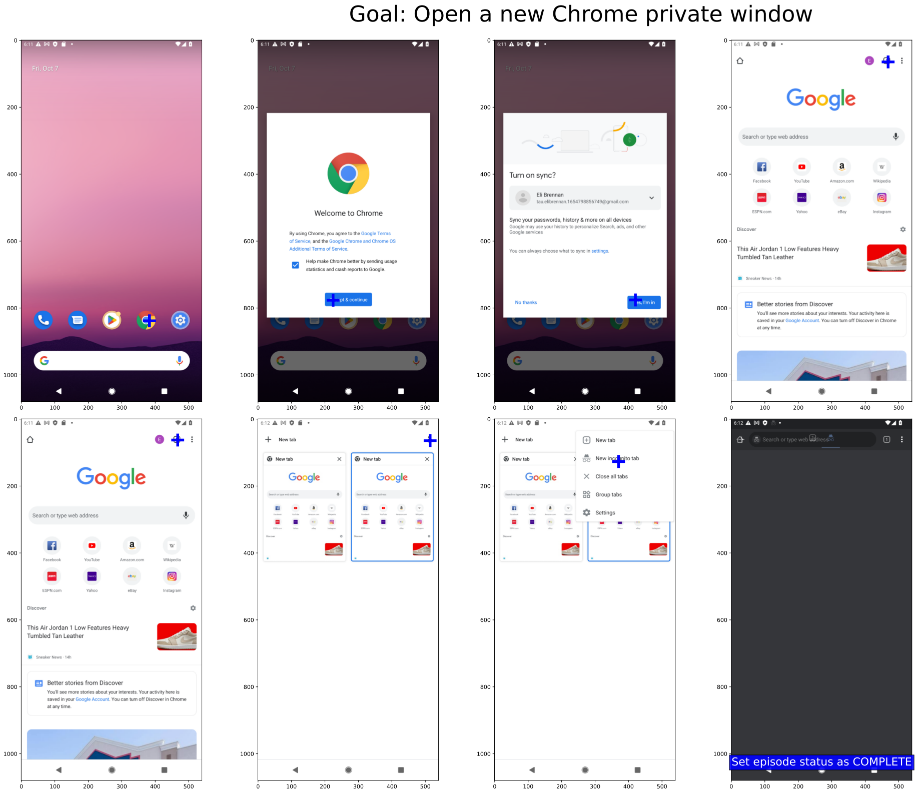
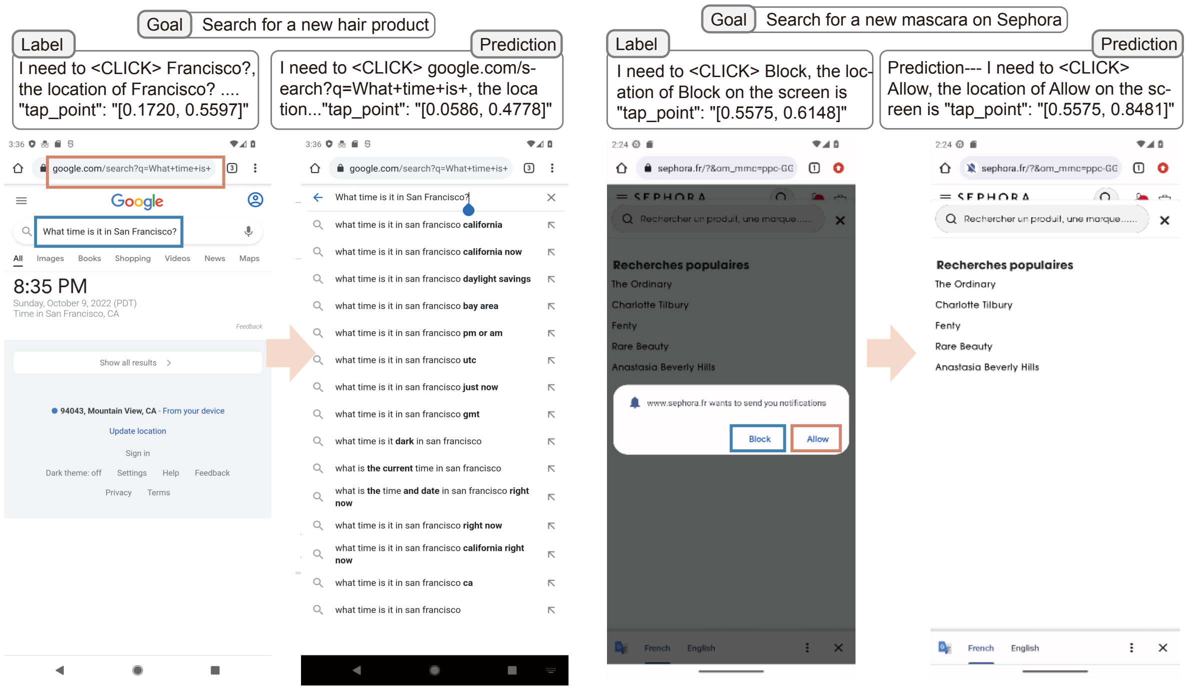


Figure 2: An example of a full episode.


Goal Search for a new hair product

Label
I need to <CLICK> Francisco?, the location of Francisco? .... "tap_point": "[0.1720, 0.5597]"

Prediction
I need to <CLICK> google.com/s-earch?q=What+time+is+, the loca tion..."tap_point": "[0.0586, 0.4778]"

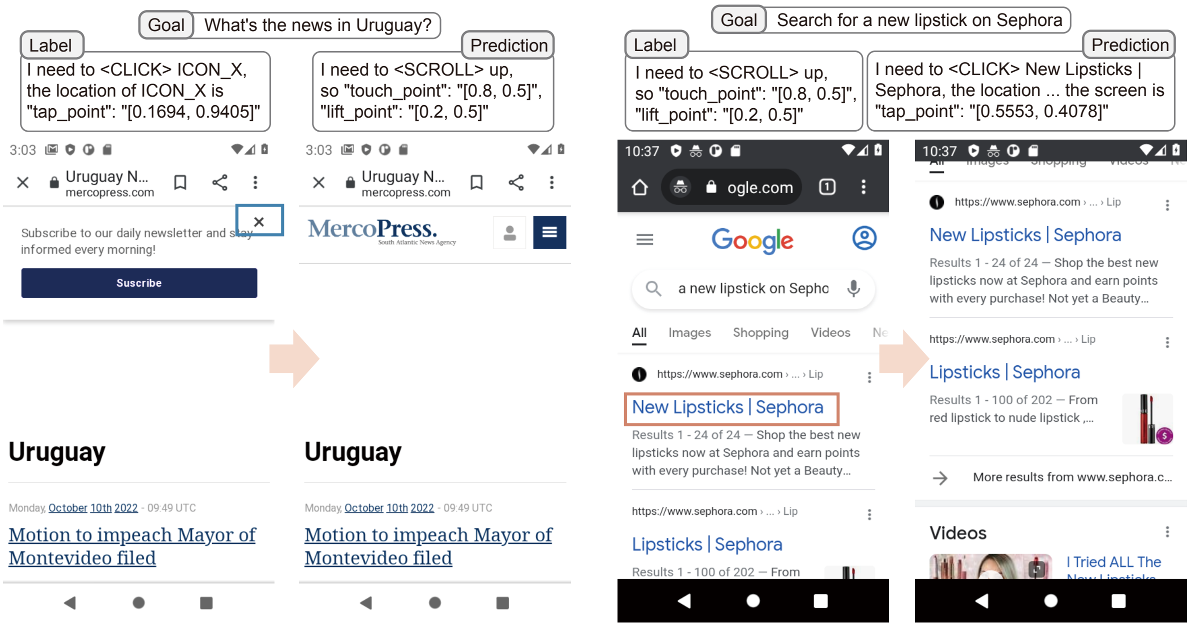
Goal Search for a new mascara on Sephora

Label
I need to <CLICK> Block, the loc-ation of Block on the screen is "tap_point": "[0.5575, 0.6148]"

Prediction
Prediction--- I need to <CLICK> Allow, the location of Allow on the sc-reen is "tap_point": "[0.5575, 0.8481]"

google.com/search?q=What+time+is+
What time is it in San Francisco?
8:35 PM
Sunday, October 9, 2022 (PDT)
Time in San Francisco, CA
Show all results
94043, Mountain View, CA - From your device
Update location
what time is it in san francisco california
what time is it in san francisco california now
what time is it in san francisco daylight savings
what time is it in san francisco bay area
what time is it in san francisco pm or am
what time is it in san francisco utc
what time is it in san francisco just now
what time is it in san francisco gmt
what time is it dark in san francisco
what is the current time in san francisco
what is the time and date in san francisco right now
what time is it in san francisco right now
what time is it in san francisco california right now
what time is it in san francisco ca
what time is it in san francisco
sephora.fr/?&om_mmc=ppc-GG
SEPHORA
Rechercher un produit, une marque......
Recherches populaires
The Ordinary
Charlotte Tilbury
Fenty
Rare Beauty
Anastasia Beverly Hills
www.sephora.fr wants to send you notifications
Block
Allow
French
English

Figure 3: The predicted action leads to the same situation but does not match the label (i, in Section 5.4.2).



Goal What's the news in Uruguay?

Label I need to <CLICK> ICON_X, the location of ICON_X is "tap_point": "[0.1694, 0.9405]"

Prediction I need to <SCROLL> up, so "touch_point": "[0.8, 0.5]", "lift_point": "[0.2, 0.5]"

Goal Search for a new lipstick on Sephora

Label I need to <SCROLL> up, so "touch_point": "[0.8, 0.5]", "lift_point": "[0.2, 0.5]"

Prediction I need to <CLICK> New Lipsticks | Sephora, the location ... the screen is "tap_point": "[0.5553, 0.4078]"

3:03
Uruguay N...
mercopress.com
Subscribe to our daily newsletter and stay informed every morning!
Suscribe
MercoPress.
South Atlantic News Agency
Uruguay
Monday, October 10th 2022 - 09:49 UTC
Motion to impeach Mayor of Montevideo filed
10:37
ogle.com
Google
a new lipstick on Sepho
All
Images
Shopping
Videos
Ne
https://www.sephora.com › ... › Lip
New Lipsticks | Sephora
Results 1 - 24 of 24 — Shop the best new lipsticks now at Sephora and earn points with every purchase! Not yet a Beauty...
https://www.sephora.com › ... › Lip
Lipsticks | Sephora
Results 1 - 100 of 202 — From
Results 1 - 100 of 202 — From red lipstick to nude lipstick ,...
More results from www.sephora.c...
Videos
I Tried ALL The New Lipsticks

Figure 4: The predicted action is different from the label but is still reasonable for the goal (ii, in Section 5.4.2.)