

# CogAgent: A Visual Language Model for GUI Agents

## CogAgent: 用于图形用户界面代理的视觉语言模型

Appendix

附录

### 1. Details of Training Configurations

#### 1. 训练配置详情

We report the detailed training settings of CogAgent in Table 7, and model configurations of CogAgent in Table 8.

我们在表 7 中报告了 CogAgent 的详细训练设置，在表 8 中报告了 CogAgent 的模型配置。

Configurations	Pre-train	Multi-task
Total steps	60,000	10,000
Warmup steps	500	500
Batch size	4,608	1,024
Learning rate	$2 \times 10^{-5}$	
Learning rate decay	Cosine	
Weight decay	0.05	
Dropout ratio	0.1	
Adam $\epsilon$	$1 \times 10^{-5}$	
Adam $\beta$	(0.9, 0.95)	

配置	预训练	多任务
总步数	60,000	10,000
预热步数	500	500
批量大小	4,608	1,024
学习率	$2 \times 10^{-5}$	
学习率衰减	余弦	
权重衰减	0.05	
丢弃率	0.1	
Adam $\epsilon$	$1 \times 10^{-5}$	
Adam $\beta$	(0.9, 0.95)	

Table 7. Training settings of pre-training and multi-task fine-tuning.

表 7. 预训练和多任务微调的训练设置。

VLM decoder	
Architecture	Vicuna-1.5-7B + visual expert
Layers	32
Hidden size	4,096
Attention heads	32
Low-resolution visual encoder	
Architecture	EVA2-CLIP-E
Input resolution	$224 \times 224$
Patch size	$14 \times 14$
High-resolution visual encoder	
Visual encoder	EVA2-CLIP-L
Input resolution	$1120 \times 1120$
Patch size	$14 \times 14$
Cross Attention	
Hidden size	1,024
Attention heads	32

VLM 解码器	
架构	Vicuna-1.5-7B + 视觉专家
层数	32
隐藏层大小	4,096
注意力头数	32
低分辨率视觉编码器	
架构	EVA2-CLIP-E
输入分辨率	$224 \times 224$
图像块大小	$14 \times 14$
高分辨率视觉编码器	
视觉编码器	EVA2-CLIP-L
输入分辨率	$1120 \times 1120$
图像块大小	$14 \times 14$
交叉注意力	
隐藏层大小	1,024
注意力头数	32

Table 8. Model configurations of CogAgent.

表 8. CogAgent 的模型配置。

## 2. Details of Evaluation Datasets

### 2. 评估数据集详情

In this section, we will provide a detailed overview of the datasets used in our evaluations .

本节将详细介绍我们评估中使用的数据集。

## 2.1. General VQA

### 2.1. 通用视觉问答 (VQA)

- VQAv2 [1]. VQAv2 is designed for visual question answering with natural images, covering a wide range of question types including yes/no, numerical counting, and more open-ended inquiries. The dataset comprised of a collection exceeding 200,000 images, paired with more than 1.1 million questions. Each question is paired with 10 answers annotated by different annotators.
  - VQAv2 [1]。VQAv2 旨在针对自然图像进行视觉问答，涵盖包括是/否、数值计数及更开放式问题在内的多种问题类型。该数据集包含超过 20 万张图像，配有超过 110 万条问题。每个问题配 10 个由不同标注者提供的答案。
- OK-VQA [23]. OK-VQA (Outside Knowledge Visual Question Answering) dataset is constructed to evaluate visual question-answering skills that require external knowledge. Models need to combine image content and common sense to answer questions. The dataset includes 14,055 open-ended questions, each accompanied by 5 ground truth answers.
  - OK-VQA [23]。OK-VQA(外部知识视觉问答) 数据集用于评估需要外部知识的视觉问答能力。模型需结合图像内容与常识来回答问题。数据集包含 14,055 个开放式问题，每个问题配 5 个标准答案。
- MM-Vet [41]. MM-Vet is designed to evaluate the overall capability of generalist visual language models in a zero-shot manner. It integrates 6 core VL capabilities to solve complex tasks (including recognition, OCR, knowledge, language generation, spatial awareness, and math), and explores 16 distinct integrations resulting from the combination of these capabilities. As for evaluation metrics, it utilizes a language model-based evaluator tailored for open-ended responses.
  - MM-Vet [41]。MM-Vet 旨在以零样本方式评估通用视觉语言模型的整体能力。它整合了 6 项核心视觉语言能力以解决复杂任务(包括识别、光学字符识别 (OCR)、知识、语言生成、空间感知和数学)，并探索这些能力组合产生的 16 种不同集成方式。评估指标采用基于语言模型的评估器，专为开放式回答设计。
- POPE [19]. POPE (Polling-based Object Probing Evaluation) is a dataset constructed to assess the object hallucination problem in large visual language models. It employs a polling-based object probing method, transforming hallucination assessment into a binary classification challenge. This is achieved by prompting large vision-language models (LVLMs) with straightforward Yes-or-No queries regarding the objects in question (for example, "Is there a car in the image?"). Our evaluation is conducted under the dataset's most challenging setting: the adversarial setting.

- POPE [19]。POPE(基于投票的对象探测评估) 数据集用于评估大型视觉语言模型中的对象幻觉问题。它采用基于投票的对象探测方法，将幻觉评估转化为二分类问题。通过向大型视觉语言模型 (LVLMs) 提出简单的是/否问题(例如，“图中有车吗？” ) 进行检测。我们的评估在数据集最具挑战性的对抗设置下进行。

## 2.2. Text-rich VQA

### 2.2. 富文本视觉问答

- OCR-VQA [27]。OCR-VQA dataset comprises 207,572 images of book covers, paired with more than 1 million question-answer pairs. The questions inquire about book information including title, edition, year, author, and genre of the book, which requires text recognition and comprehension abilities.
- OCR-VQA [27]。OCR-VQA 数据集包含 207,572 张书籍封面图像，配有超过 100 万对问答。问题涉及书籍信息，包括标题、版本、年份、作者和类别，要求具备文本识别和理解能力。
- TextVQA [34]。TextVQA is a benchmark of visual reasoning based on text in images. Models need to
  - TextVQA [34]。TextVQA 是基于图像中文本的视觉推理基准。模型需要

Task	Dataset	Description	Split	Metrics
General VQA	VQAv2	VQA on natural images.	test-dev	VQA Score(↑)
	OK-VQA	VQA on natural images requiring outside knowledge.	val	VQA Score (↑)
	MM-Vet	Conversational style VQA on integrated capabilities.	test	GPT-4 score(↑)
	POPE	VQA for hallucination assessment. The adversarial setting is used.	test	F1 score(↑)
Text-rich VQA	OCR-VQA	VQA on images of book covers.	test	EM (↑)
	TextVQA	VQA on natural images containing text.	test	VQA Score (↑)
	ST-VOA	VQA on natural images requiring textual understanding.	test	ANLS (↑)
	ChartQA	VQA about charts with visual and logical reasoning.	test	VQA Score (↑)
	InfoVQA	VQA on infographics.	test	ANLS (↑)
	DocVQA	VQA on document images.	test	ANLS (↑)
GUI Agent	Mind2Web	Web behavior prediction given snapshots and historical actions.	test	step SR (↑)
	AITW	Android behavior prediction given snapshots and historical actions.	test	Matching Score (↑)

任务	数据集	描述	划分	指标
通用视觉问答	VQAv2	自然图像上的视觉问答 (VQA)。	测试开发集	视觉问答得分 (↑)
	OK-VQA	需要外部知识的自然图像视觉问答。	验证集	视觉问答得分 (↑)
	MM-Vet	基于综合能力的对话式视觉问答。	测试集	GPT-4 得分 (↑)
	POPE	用于幻觉评估的视觉问答。采用对抗设置。	测试集	F1 得分 (↑)
文本丰富的视觉问答	OCR-VQA	书籍封面图像上的视觉问答。	测试集	完全匹配率 (EM) (↑)
	TextVQA	包含文本的自然图像视觉问答。	测试集	视觉问答得分 (↑)
	ST-VOA	需要文本理解的自然图像视觉问答。	测试集	ANLS 指标 (↑)
	ChartQA	关于图表的视觉问答，涉及视觉和逻辑推理。	测试集	视觉问答得分 (↑)
	InfoVQA	信息图上的视觉问答。	测试集	ANLS 指标 (↑)
	DocVQA	文档图像上的视觉问答。	测试集	ANLS 指标 (↑)
图形用户界面代理	Mind2Web	基于快照和历史操作的网页行为预测。	测试集	步骤成功率 (step SR) (↑)
	AITW	基于快照和历史操作的安卓行为预测。	测试集	匹配得分 (↑)

Table 9. Summary of the evaluation benchmarks.

表 9. 评估基准汇总。

incorporate the textual information in the images and reason over it to answer TextVQA questions. It comprises a total of 28,408 images and 45,336 questions.

整合图像中的文本信息并进行推理以回答 TextVQA 问题。该数据集共包含 28,408 张图像和 45,336 个问题。

- ST-VQA [4]. ST-VQA is designed to emphasize the significance of exploiting the semantic information present within images in textual form during the VQA process. It comprises tasks of diverse difficulties, for which recognizing the scene text and performing necessary reasoning is required to generate the answer. The dataset comprises 23,038 images sourced from multiple public datasets and 31,791 question-answer pairs.

• ST-VQA [4]。ST-VQA 旨在强调在视觉问答 (VQA) 过程中利用图像中以文本形式存在的语义信息的重要性。该任务包含不同难度级别，需要识别场景文本并进行必要推理以生成答案。数据集包含 23,038 张来自多个公开数据集的图像和 31,791 个问答对。

- ChartQA [24]. ChartQA is a benchmark of question-answering about logical and visual reasoning on charts. It consists of 20,882 charts curated from four different online sources, 9,608 manual written questions, as well as 23,111 questions automatically generated with T5 according to human-written chart summaries.

• ChartQA [24]。ChartQA 是一个关于图表逻辑和视觉推理的问答基准。它由 20,882 个图表组成，图表来源于四个不同的在线资源，包含 9,608 个手工编写的问题，以及 23,111 个根据人工编写的图表摘要通过 T5 自动生成的问题。

- InfographicVQA(InfoVQA) [26]. The task of In-foVQA is to answer questions centering on a given infographic image. The answers to most questions can be extracted from the given documents, while the answers to a small percentage of questions are not extractive. There are 5K Images collected from the Internet and 30K manually annotated questions in the dataset.

• InfographicVQA(InfoVQA) [26]。InfoVQA 任务是针对给定的信息图像回答问题。大多数问题的答案可以从给定文档中提取，而少部分问题的答案则非提取式。数据集中收集了 5 千张网络图片和 3 万个手工标注的问题。

- DocVQA [25]. DocVQA focuses on question-answering given a document image. The answer for questions is often a span of text from the given documents. There are 12 K images and 50 K manually annotated questions in the datasets.

• DocVQA [25]。DocVQA 聚焦于基于文档图像的问答。问题的答案通常是给定文档中的一段文本。数据集中包含 12 K 张图像和 50 K 个手工标注的问题。

## 2.3.GUI Agent

### 2.3.GUI 代理

CogAgent is evaluated on two GUI agent datasets, Mind2Web and Android in the Wild (AITW), corresponding to computer agent and smartphone agent respectively.

CogAgent 在两个 GUI 代理数据集上进行评估，分别是 Mind2Web 和 Android in the Wild (AITW)，对应计算机代理和智能手机代理。

- Mind2Web [10]. Mind2Web is designed to develop and evaluate web agents capable of executing intricate tasks on various websites based on language directions. While existing datasets for web agents commonly rely on simulated or overly simplified web data, Mind2Web utilizes real-world websites and is annotated by human annotators. It gathers data from 137 websites covering 31 domains, and collects over 2,000 open-ended tasks, each accompanied by a crowd-sourced action sequence.

- Mind2Web [10]。Mind2Web 旨在开发和评估能够根据语言指令在各种网站上执行复杂任务的网页代理。现有的网页代理数据集通常依赖模拟或过于简化的网页数据，而 Mind2Web 使用真实网站并由人工标注。它收集了 137 个网站，涵盖 31 个领域，收集了 2000 多个开放式任务，每个任务配有众包的操作序列。

In mind2web, evaluated agents are asked to accomplish a designated task on a chosen website by performing a sequence of actions. Each instance (i.e. a specific task) in Mind2Web contains a task description, action sequence, and webpage snapshots. Each action in the sequence is a (Target element, Operation) pair, and Operation includes Click, Type (with additional value), and Select (with additional value). Each action is paired with the concurrent webpage snapshots in a variety of formats including raw HTML code, DOM tree, screenshot, etc. As for CogAgent, we choose screenshot images as the input representation of web-sites; as for other language-based agents, HTML is chosen as the input representation.

在 Mind2Web 中，被评估的代理需通过执行一系列操作完成指定网站上的任务。Mind2Web 中的每个实例（即具体任务）包含任务描述、操作序列和网页快照。操作序列中的每个操作是（目标元素，操作）对，操作包括点击、输入（附带具体值）和选择（附带具体值）。每个操作配有很多种格式的网页快照，包括原始 HTML 代码、DOM 树、截图等。对于 CogAgent，我们选择截图图像作为网页的输入表示；对于其他基于语言的代理，则选择 HTML 作为输入表示。

Following Deng et al. [10] and Zeng et al. [42], we formalize the problem as: first choose the target webpage element among top-k ( $k=10$  or  $50$ ) candidates, then predict specific operations. The top-k candidates are provided by the candidate generation model in Deng et al. [10]. Step success rate (step SR) is reported on 3 out-of-domain test sets (cross-website, cross-domain,

遵循 Deng 等人 [10] 和 Zeng 等人 [42] 的方法，我们将问题形式化为：首先从前  $k$  ( $k=10$  或  $50$ ) 候选目标网页元素中选择目标元素，然后预测具体操作。前  $k$  候选由 Deng 等人 [10] 的候选生成模型提供。步骤成功率 (step SR) 在三个域外测试集 (跨网站、跨领域、

cross-task) as metric. Only predictions with the same target element and operation as the ground truth are regarded as correct.

跨任务) 上报告作为评估指标。只有预测的目标元素和操作均与真实值相同的情况才被视为正确。

- Android in the Wild (AITW) [31]. AITW is constructed to develop and evaluate Android device-control

systems that are capable of understanding and acting upon human natural language instructions by directly manipulating the device’s user interface. This dataset significantly surpasses others in its category in terms of size, encompassing 715k episodes across 30k distinct instructions, and covering four Android versions (v10-13). It also includes eight types of devices, ranging from Pixel 2 XL to Pixel 6, each with different screen resolutions. AITW consists of five subsets: GoogleApps, Install, WebShopping, General, and Single. The subsets have distinct tasks, while are in the same data format. Each episode (i.e. action sequence) comprises three components: a goal instruction provided in natural language, a user action sequence, and a corresponding screenshots sequence.

- Android in the Wild (AITW) [31]。AITW 旨在开发和评估能够理解并执行基于自然语言指令的 Android 设备控制系统，通过直接操作设备用户界面实现。该数据集在同类中规模显著领先，包含 715,000 个操作序列，涵盖 30,000 条不同指令，支持四个 Android 版本 (v10-13)。数据集包含八种设备类型，从 Pixel 2 XL 到 Pixel 6，屏幕分辨率各异。AITW 包含五个子集：GoogleApps、Install、WebShopping、General 和 Single。各子集任务不同，但数据格式一致。每个操作序列包含三个部分：以自然语言提供的目标指令、用户操作序列及对应的截图序列。

As for screenshots, AITW only provides screenshot images and does not provide tree-based representations of UI. This is because a large portion of them in Smartphone applications are of low quality or even do not exist, and adopting the tree-based representation would strongly limit agents’ applications. For visual agents, screenshots are provided to the agents in image format; for language-model-based agents evaluated by Zhan and Zhang [43], the textual representations of OCR and icons formatted in HTML syntax are provided. As for actions, AITW considers a variety of action types including tapping, swiping, typing, going home, going back, entering, etc. For each action, models are required to predict the exact action type; for tap, swipe and type, models are further required to predict the position, direction, and content to be typed, respectively. The detailed standard for computing matching scores is provided in Rawles et al. [31].

关于截图，AITW 仅提供截图图像，不提供基于树的 UI 表示。这是因为智能手机应用中大量截图质量较低甚至不存在，采用基于树的表示会极大限制代理的应用。对于视觉代理，截图以图像格式提供；对于由 Zhan 和 Zhang [43] 评估的基于语言模型的代理，提供了以 HTML 语法规式化的 OCR 和图标的文本表示。关于操作，AITW 考虑了多种操作类型，包括点击、滑动、输入、回到主页、返回、进入等。对于每种操作，模型需预测准确的操作类型；对于点击、滑动和输入，模型还需分别预测位置、方向和输入内容。匹配分数的详细计算标准见 Rawles 等人 [31]。

### 3. Derivation of Acceleration for High- Resolution Cross-Module

#### 3. 高分辨率跨模块加速推导

Suppose that  $L_{I_{lo}}$ ,  $L_{I_{hi}}$  and  $L_T$  are the lengths of the low-resolution image, high-resolution image and text sequences. Let  $H_{\text{cross}}$ ,  $H_{\text{dec}}$  be the number of attention heads in cross-attention and self-attention, and  $d_{\text{cross}}$ ,  $d_{\text{dec}}$  be the dimension of each attention head.

假设  $L_{I_{lo}}$ ,  $L_{I_{hi}}$  和  $L_T$  分别为低分辨率图像、高分辨率图像及文本序列的长度。令  $H_{\text{cross}}$ ,  $H_{\text{dec}}$  为交叉注意力和自注意力中的注意力头数， $d_{\text{cross}}$ ,  $d_{\text{dec}}$  为每个注意力头的维度。

If using our high-resolution cross-module, the computational complexity of attention is

若使用我们的高分辨率跨模块，注意力的计算复杂度为

$$\begin{aligned} T_{\text{improved}} = & \mathbf{O}((L_{I_{\text{lo}}} + L_T) L_{I_{\text{hi}}} H_{\text{cross}} d_{\text{cross}} \\ & + (L_{I_{\text{lo}}} + L_T)^2 H_{\text{dec}} d_{\text{dec}}) . \end{aligned} \quad (5)$$

If not utilizing the high-resolution cross-module and directly substituting low-resolution images with high-resolution ones, the computational complexity would be

若不使用高分辨率跨模块，直接用高分辨率图像替代低分辨率图像，计算复杂度将为

$$T_{\text{original}} = \mathbf{O}((L_{I_{\text{hi}}} + L_T)^2 H_{\text{dec}} d_{\text{dec}}) . \quad (6)$$

The reduction factor of the computational complexity in attention,  $T_{\text{original}} / T_{\text{improved}}$  , equals to

注意力计算复杂度的降低因子  $T_{\text{original}} / T_{\text{improved}}$  等于

$$\frac{(L_{I_{\text{hi}}} + L_T)^2 H_{\text{dec}} d_{\text{dec}}}{(L_{I_{\text{lo}}} + L_T) L_{I_{\text{hi}}} H_{\text{cross}} d_{\text{cross}} + (L_{I_{\text{lo}}} + L_T)^2 H_{\text{dec}} d_{\text{dec}}} \quad (7)$$

$$= \frac{L_{I_{\text{hi}}} + L_T}{L_{I_{\text{lo}}} + L_T} \frac{(L_{I_{\text{hi}}} + L_T) H_{\text{dec}} d_{\text{dec}}}{L_{I_{\text{hi}}} H_{\text{cross}} d_{\text{cross}} + (L_{I_{\text{lo}}} + L_T) H_{\text{dec}} d_{\text{dec}}} \quad (8)$$

$$= \frac{L_{I_{\text{hi}}} + L_T}{L_{I_{\text{lo}}} + L_T} \frac{(L_{I_{\text{hi}}} + L_T) \frac{H_{\text{dec}} d_{\text{dec}}}{H_{\text{cross}} d_{\text{cross}}}}{L_{I_{\text{hi}}} + (L_{I_{\text{lo}}} + L_T) \frac{H_{\text{dec}} d_{\text{dec}}}{H_{\text{cross}} d_{\text{cross}}}} \quad (9)$$

- Case 1:  $L_{I_{\text{lo}}}, L_T \ll L_{I_{\text{hi}}}$  .

- 情况 1:  $L_{I_{\text{lo}}}, L_T \ll L_{I_{\text{hi}}}$  。

Given that  $L_{I_{\text{lo}}}$  is much smaller than  $L_{I_{\text{hi}}}$  , when  $L_T$  also satisfies being much smaller than  $L_{I_{\text{hi}}}$  , both  $\frac{L_{I_{\text{lo}}}}{L_{I_{\text{hi}}}}$  and  $\frac{L_T}{L_{I_{\text{hi}}}}$  become first-order small quantities. If conducting a 0-th order approximation for the complexity reduction factor, we obtain:

鉴于  $L_{I_{\text{lo}}}$  远小于  $L_{I_{\text{hi}}}$  , 且当  $L_T$  也远小于  $L_{I_{\text{hi}}}$  时，  $\frac{L_{I_{\text{lo}}}}{L_{I_{\text{hi}}}}$  和  $\frac{L_T}{L_{I_{\text{hi}}}}$  均为一阶小量。对复杂度降低因子进行零阶近似，得到：

$$\frac{T_{\text{original}}}{T_{\text{improved}}} = \frac{L_{I_{\text{hi}}} \left(1 + \frac{L_T}{L_{I_{\text{hi}}}}\right)}{L_{I_{\text{lo}}} + L_T} \frac{L_{I_{\text{hi}}} \left(1 + \frac{L_T}{L_{I_{\text{hi}}}}\right) \frac{H_{\text{dec}}}{H_{\text{cross}}} d_{\text{cross}}}{L_{I_{\text{hi}}} \left(1 + \frac{L_{I_{\text{lo}}} + L_T}{L_{I_{\text{hi}}}} \frac{H_{\text{dec}} d_{\text{dec}}}{H_{\text{cross}}}\right)} \quad (10)$$

$$\begin{aligned} & = \frac{L_{I_{\text{hi}}} \left(1 + \frac{L_T}{L_{I_{\text{hi}}}}\right)}{L_{I_{\text{lo}}} + L_T} \frac{\left(1 + \frac{L_T}{L_{I_{\text{hi}}}}\right) \frac{H_{\text{dec}} d_{\text{dec}}}{H_{\text{cross}} d_{\text{cross}}}}{1 + \frac{L_{I_{\text{lo}}} + L_T}{L_{I_{\text{hi}}}} \frac{H_{\text{dec}} d_{\text{dec}}}{H_{\text{cross}} d_{\text{cross}}}} \end{aligned} \quad (11)$$

$$\approx \frac{L_{I_{\text{hi}}}}{L_{I_{\text{lo}}} + L_T} \quad (12)$$

- Case 2: Our implementation.

- 情况 2: 我们的实现。

In our implementation,  $d_{\text{cross}} = 32$ ,  $H_{\text{cross}} = 32$ , and we inherits  $d_{\text{dec}} = 128$ ,  $H_{\text{dec}} = 32$  from CogVLM-17B. Both high- and low-resolution encoders patchify images with  $14 \times 14$ -pixel patches, thus  $L_{I_{\text{hi}}} = 6400$ ,  $L_{I_{\text{lo}}} = 256$ . Subsute the values to Eq. (9), we obtain

在我们的实现中,  $d_{\text{cross}} = 32$ ,  $H_{\text{cross}} = 32$ , 并且继承自 CogVLM-17B 的  $d_{\text{dec}} = 128$ ,  $H_{\text{dec}} = 32$ 。高低分辨率编码器均以  $14 \times 14$  像素的 patch 划分图像, 因此  $L_{I_{\text{hi}}} = 6400$ ,  $L_{I_{\text{lo}}} = 256$ 。将数值代入公式(9), 得到

$$\frac{T_{\text{original}}}{T_{\text{improved}}} = \frac{6400 + L_T}{256 + L_T} \frac{4(6400 + L_T)}{6400 + 4(256 + L_T)} \quad (13)$$

$$> \frac{6400 + L_T}{256 + L_T} \quad (14)$$

That is to say, our method leads to at least  $\frac{L_{I_{\text{hi}}} + L_T}{L_{I_{\text{lo}}} + L_T} = \frac{6400 + L_T}{256 + L_T} \times$  acceleration which is a stringent lower bound. Taking pre-training as an example, with  $L_T$  typically below 512 in earlier stages, our high-resolution cross-module can get an over  $25 \times$  reduction in computational cost.

也就是说, 我们的方法至少带来  $\frac{L_{I_{\text{hi}}} + L_T}{L_{I_{\text{lo}}} + L_T} = \frac{6400 + L_T}{256 + L_T} \times$  倍加速, 这是一个严格的下界。以预训练为例, 早期阶段  $L_T$  通常低于 512, 我们的高分辨率跨模块可实现超过  $25 \times$  的计算成本降低。

## 4. Performance Analysis on AITW

### 4. AITW 上的性能分析

Android in the Wild (AITW) dataset is a large-scale dataset for Android device agents. Our analysis centers on instances where the CogAgent's responses diverge from the annotated ground truth, focusing on these discrepancies.

Android in the Wild (AITW) 数据集是面向 Android 设备代理的大规模数据集。我们的分析聚焦于 CogAgent 响应与标注真实值不一致的实例, 重点研究这些差异。

In our analysis, we conducted a sampling and manual verification of cases where there was a discrepancy between the model's responses and the labeled answers. We sampled instances and discovered that in 42% of these, the model's response actually represented an alternative correct method, rather than an error. This phenomenon is common due to the multiplicity of ways to accomplish a function on a smart-phone. For example, many phones are equipped with both Google's app and a Google search bar, and searches can be conducted using either. During the use of smartphones, human users may exhibit simultaneous needs for both in-depth exploration and extensive browsing. For example, when browsing news, CogAgent might identify relevant results on the current page and select them immediately, while some users may prefer to scroll through the entire page first. Furthermore, the data collection for annotations might involve redundant or overly detailed steps, whereas CogAgent can directly find the most efficient path. Fig. 4 illustrates these typical examples of multiple correct answers.

在我们的分析中，我们对模型回答与标注答案不一致的案例进行了抽样和人工验证。抽样结果显示，在42%的情况下，模型的回答实际上代表了一种替代的正确方法，而非错误。这种现象普遍存在，原因在于智能手机上完成某一功能的方式多样。例如，许多手机同时配备了谷歌应用和谷歌搜索栏，用户可以通过任一途径进行搜索。在使用智能手机时，用户可能同时存在深入探索和广泛浏览的需求。例如，在浏览新闻时，CogAgent可能会立即识别当前页面上的相关结果并选择它们，而部分用户则倾向于先滚动浏览整页。此外，标注数据的收集可能包含冗余或过于详细的步骤，而CogAgent能够直接找到最高效的路径。图4展示了这些多种正确答案的典型例子。

Given the nature of smartphone usage, wherein accomplishing a task often involves multiple viable paths, it is advisable for future studies to explore a virtual environment with an evaluation of the accuracy of task completion following a sequence of actions, aligning more closely with real-world human usage scenarios.

鉴于智能手机使用的特性，即完成任务通常涉及多条可行路径，未来研究建议探索虚拟环境，并评估在一系列操作后任务完成的准确性，以更贴近真实的人类使用场景。

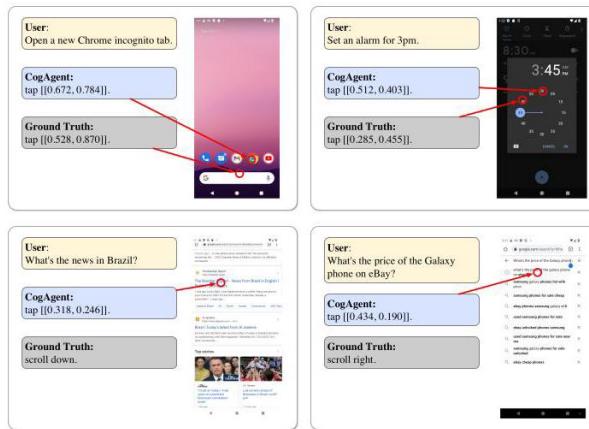


Figure 4. Instances in the AITW dataset where CogAgent provides accurate responses that differ from the annotated ground truth.

图4. AITW 数据集中 CogAgent 提供的与标注标准答案不同但准确的响应实例。

## 5. Samples of Pre-train Data

### 5. 预训练数据样本

Samples of pre-train data are listed in this section, including data for text recognition (Fig. 5), visual grounding (Fig. 6), and webpage-html pairs (Fig. 7).

本节列出了预训练数据样本，包括文本识别(图5)、视觉定位(图6)和网页-HTML对(图7)的数据。

has kept nine players that featured in last year's FIBA Basketball World Cup. While 36-year-old Eduardo Mingas is set to play his sixth consecutive African Championship, Roberto Fortes who has featured for Angola in

consecutive World Championships, will make his AfroBasket debut in Tuni. from 19-30 August. Fortes stood out throughout the training camp, havin helped Angola beat Venezuela in

保留了九名参加去年国际篮球联合会 (FIBA) 篮球世界杯的球员。36 岁的爱德华多·明加斯 (Eduardo Mingas) 将连续第六次参加非洲锦标赛，而曾连续参加世界锦标赛的罗伯托·福尔特斯 (Roberto Fortes) 将在 8 月 19 日至 30 日于突尼斯首次亮相非洲篮球锦标赛。福尔特斯在整个训练营中表现突出，帮助安哥拉击败了委内瑞拉。

being about a half hour from the city's center. I was struck by how lovely it was, until I learned one could n't eat the fish, an it isn't a good idea to do to to much swimming radiation being a prob Halladay no-hitter came in the first possessor start of his care Roy Halladay of the Phillies pitched the second no-hitter in Mi League Baseball postseason history on Wednesday as the Philadelp Phillies beat the Cincinnati Reds 4-0 in their National League

距离市中心约半小时车程。我被这里的美丽所震撼，直到得知鱼不能食用，也不宜过多游泳，因为辐射是个问题。霍拉迪 (Halladay) 的无安打比赛发生在他职业生涯的首次投球局。费城人队的罗伊·霍拉迪 (Roy Halladay) 于周三投出了大联盟季后赛历史上的第二场无安打比赛，费城人以 4-0 击败辛辛那提红人队。



$$g'(y) := \frac{f(x)}{\sqrt{\tilde{d}_{\lambda,h}(x)}} - \frac{f(y)}{\sqrt{\tilde{d}_{\lambda,h}(y)}} \quad L^p \rightarrow L^p_{(2n-1)/p'} \\ \gamma = 6.96 \times 10^8 \text{ m} \quad \kappa^2 = -\frac{1}{2} \nabla^\mu \chi^\nu \nabla_\mu \chi_\nu$$

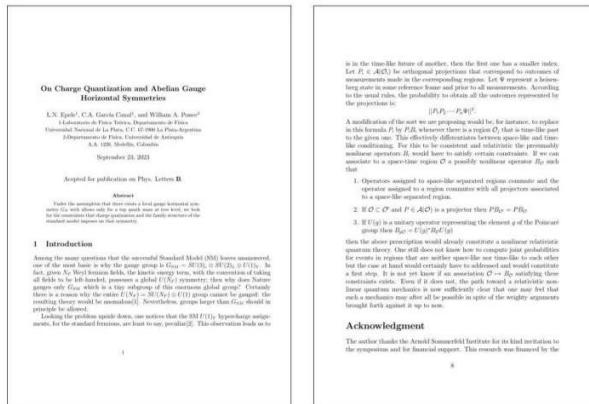


Figure 5. Samples of pre-train data for text recognition.

图 5. 文本识别预训练数据样本。



A bride [[592,179,866,956]] and groom [[717,160,862,882]] standing near the water [[425,426,990,898]] at sunset.

一对新娘 [[592,179,866,956]] 和新郎 [[717,160,862,882]] 站在日落时分的水边 [[425,426,990,898]]。



Chicken and rice dish [[301,160,997,792]] in a pan [[282,107,998,839]], with spoons [[091,476,630,953]] and plate [[002,004,258,374]] next to it.

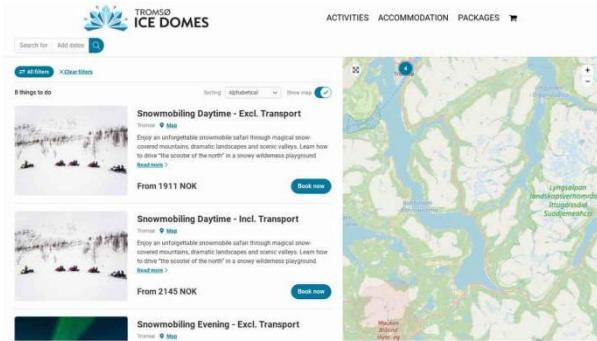
锅中 [[282,107,998,839]] 的鸡肉米饭菜肴 [[301,160,997,792]], 旁边放着勺子 [[091,476,630,953]] 和盘子 [[002,004,258,374]]。



A bluebird [[302,413,640,752]] sitting on a branch coffee mugs [[279,588,677,804]].

Figure 6. Samples of pre-train data for visual grounding. Snowmobiling Daytime - Excl. Transport

图 6. 视觉定位预训练数据样本。白天雪地摩托 - 不含交通



[Box] [[30, 281, 870, 445]]

```
[[30, 281, 870, 445]]
```

[HTML]<div>\n<div>\n<div>\n<h2>\n\n<a>\n<a>\n</a>\n<div>\n<div>\n<!-->\n<span>\nTromsø\n</span>\n<span>\n<a>\n<span>\n</span>\n<b>\nMap\n</b>\n<b>\nMap\n</b>\n</a>\n</span>\n</div>\n</div>\n<a>\n<div>\n<p>\nEnjoy an unforgettable snowmobile safari through magical snow-covered mountains, dramatic landscapes and scenic valleys. Learn how to drive "the scooter of the north" in a snowy wilderness playground.\n</p>\n<div>\n<span>\nRead more\n</span>\n<span>\n<span>\n</span>\n</span>\n</div>\n</div>\n</ad>\n<!-->\n</div>

<div> \n<div> \n<div> \n<h2> \n \n<a> \n<a> \n \n</a> \n<div> \n<div> \n<!—> \n<span> \n 特罗姆瑟 (Tromsø) \n</span> \n<span> \n<a> \n<span> \n</span> \n<b> \n 地图 \n</b> \n<b> \n 地图 \n</b> \n</a> \n</span> \n</div> \n</div> \n</div> \n<div> \n<a> \n<div> \n<p> \n 享受一次难忘的雪地摩托之旅，穿越神奇的雪山、壮丽的风景和秀美的山谷。在雪地荒野游乐场学习驾驶“北方滑雪车”。 \n</p> \n<div> \n<span> \n 阅读更多 \n</span> \n<span> \n<span> \n</span> \n</span> \n</div> \n</div> \n</div> \n</ad> \n<!—> \n</div>



[Box] [[190, 510.203125, 1088.21875, 682.703125]] [HTML] <p>\n WordPress via the Jetpack plugin publish reports on your blogs which are very interesting. If you have a WordPress blog and you have not yet installed

Jetpack then you really should as there are some great features including the 'Publicize' feature where you can auto-update your Facebook, Twitter and Google+ accounts with your posts. My [...] \n</p>

[[190, 510.203125, 1088.21875, 682.703125]] [HTML] <p> \n 通过 Jetpack 插件, WordPress 可以发布关于您的博客的报告, 这些报告非常有趣。如果您有一个 WordPress 博客但还未安装 Jetpack, 那么您真的应该安装, 因为它有一些很棒的功能, 包括“Publicize”功能, 可以自动将您的帖子更新到 Facebook、Twitter 和 Google+ 账户。我的 [...] \n</p>

Figure 7. Samples of webpage-html pairs.

图 7. 网页-HTML 对的样本。

## 6. Details of Fine-Tuning Data

### 6. 微调数据详情

#### 6.1. Human annotation

##### 6.1. 人工标注

To enable CogAgent to function as an agent on various apps and websites, we assembled a team of over ten annotators to collect and label screenshots. To ensure the quality of the annotations and prevent overly simplistic, homogenized content, the annotation process was divided into two phases.

为了使 CogAgent 能够作为各种应用和网站上的代理, 我们组建了一个由十余名标注员组成的团队, 负责收集和标注截图。为确保标注质量并防止内容过于简单和同质化, 标注过程被分为两个阶段。

In the first phase, annotators were required to independently select from a pool of apps and websites, and capture screenshots of pages strongly relevant to the main function of that app/website, with no less than 20 screenshots for each app/website. Subsequently, for each screenshot, the annotators would label them according to the following aspects:

在第一阶段, 标注员需从应用和网站池中独立选择, 并截取与该应用/网站主要功能高度相关的页面截图, 每个应用/网站不少于 20 张截图。随后, 标注员需根据以下方面对每张截图进行标注:

1. Five buttons. List the names of five clickable buttons from the screenshots. The types of buttons should be diverse.

1. 五个按钮。从截图中列出五个可点击按钮的名称, 按钮类型应多样化。

2. Three clickable areas. List three clickable areas from the screenshots, e.g. text boxes, search boxes, click-able images, hyperlinks, etc.

2. 三个可点击区域。从截图中列出三个可点击区域, 例如文本框、搜索框、可点击图片、超链接等。

3. Two questions extracting information from the image. Ask questions involving the textual information in the images. For instance, "In what year did the user in the image register?".

3. 两个从图像中提取信息的问题。提出涉及图像中文本信息的问题。例如，“图中用户在哪一年注册？”

4. One operation requirement. What function of the webpage/app would you use in this interface? For example, adding the comment: "cool!". The requirement shouldn't be vague such as "adding comment".

4. 一个操作需求。在此界面中您会使用网页/应用的什么功能？例如，添加评论：“酷！”操作需求不应模糊，如“添加评论”。

In the second phase, annotators are required to provide grounding annotation for the questions and operational methods proposed in the first phase (the aforementioned parts 3 and 4). For example, for Figure 8, a possible annotation would be:

在第二阶段，标注员需为第一阶段提出的问题和操作方法(上述第3和第4部分)提供定位标注。例如，对于图8，可能的标注为：

1. Buttons: Back; Search; Subscribed; Home; Library.

1. 按钮：返回；搜索；已订阅；首页；库。

2. Clickable areas: Avatar in the middle top; Video preview in the middle bottom; Personal profile.

2. 可点击区域：顶部中间的头像；底部中间的视频预览；个人资料。

3. Question 1: Based on the page, how many followers does this author have?

3. 问题1：根据页面，该作者有多少粉丝？

Answer: According to the personal profile at the top [[013,568,802,188]], this author has 4.97M followers.

答案：根据顶部的个人资料 [[013,568,802,188]]，该作者有 497 万粉丝。

Question 2: Based on the page, how many videos has this author posted?

问题2：根据页面，该作者发布了多少视频？

Answer: According to the personal profile at the top [[013,568,802,188]], this author has posted 502 videos.

答案：根据顶部的个人资料 [[013,568,802,188]]，该作者已发布 502 个视频。

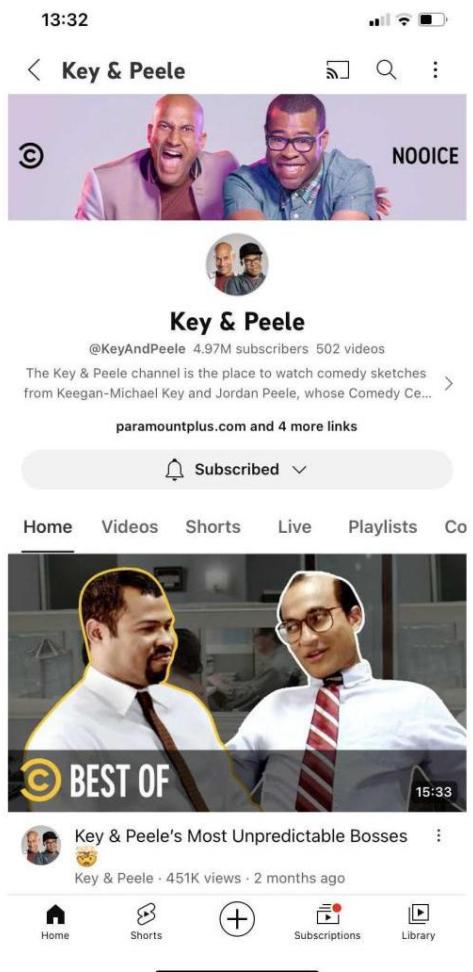


Figure 8. Samples of human-collected screenshot.

图 8. 人工收集的截图样本。

4. Operation requirement: Based on the page, write out the steps to complete the following function: Follow this author.

4. 操作要求: 根据页面, 写出完成以下功能的步骤: 关注该作者。

Answer: Click on Subscribed [[049,826,728,078]] to follow this author.

答案: 点击“已订阅” [[049,826,728,078]] 以关注该作者。

## 6.2. Conversion of Agent Datasets

### 6.2. 代理数据集的转换

To convert Mind2Web to natural language with GPT4, we use the following prompt:

为了使用 GPT4 将 Mind2Web 转换为自然语言，我们使用以下提示：

---

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

想象你是一个操作电脑的机器人。就像人类操作电脑一样，你可以移动鼠标、点击鼠标，或用键盘输入文本。

\*\*Your ultimate task is: "Find the lowest-priced round trip flight with hotel on May 2 from Kathmandu, Nepal KTM to Shanghai, China PVG and return on May 5. Book a double room and check out with the default flights.".\*\*

\*\* 你的最终任务是：“查找 2023 年 5 月 2 日从尼泊尔加德满都 (Kathmandu, Nepal KTM) 到中国上海 (Shanghai, China PVG) 的最低价往返含酒店的机票，并于 5 月 5 日返回。预订一间双人房，并使用默认航班办理退房。” \*\*

---

You are given previous actions: (format: element → operation)

你已执行的操作如下:(格式: 元素 → 操作)

1. [link] Flight + Hotel → CLICK,

1. [链接] 机票 + 酒店 → 点击,

2. [textbox] Where from? → TYPE: KATHMANDU,

2. [文本框] 出发地? → 输入:KATHMANDU,

3. [div] Tribhuvan Intl Airport (KTM), Nepal → CLICK,

3. [div] 特里布万国际机场 (Tribhuvan Intl Airport, KTM), 尼泊尔 → 点击,

4. [textbox] Where to? → TYPE: SHANGHAI,

4. [文本框] 目的地? → 输入:SHANGHAI,

5. [div] Pudong Intl Airport (PVG), China → CLICK,

5. [div] 浦东国际机场 (Pudong Intl Airport, PVG), 中国 → 点击,

6. [span] Sat 8 Apr - Sat 15 Apr → CLICK,

6. [span] 4 月 8 日周六 - 4 月 15 日周六 → 点击,

7. [checkbox] 2 May 2023 → CLICK,

7. [复选框] 2023 年 5 月 2 日 → 点击,

8. [checkbox] 5 May 2023 → CLICK,

8. [checkbox] 2023 年 5 月 5 日 → 点击,

Furthermore, you **secretly** know the following steps of action. **(But you have to pretend this is the result of your derivation and you don't know it before)**

此外, 你 **秘密地** 知道接下来的操作步骤。**(但你必须假装这是你推导的结果, 之前并不知道)**

1. [button] Search → CLICK,

1. [button] 搜索 → 点击,

2. [button] Recommended → CLICK,

2. [button] 推荐 → 点击,

3. [div] Lowest Price → CLICK,

3. [div] 最低价 → 点击,

4. [svg] → CLICK,

4. [svg] → 点击,

5. [button] Choose your room → CLICK,

5. [button] 选择你的房间 → 点击,

6. [button] Book Business Double Room A → CLICK,

6. [button] 预订商务双人房 A → 点击,

7. [svg] → CLICK,

7. [svg] → 点击,

8. [svg] → CLICK,

8. [svg] → 点击,

9. [button] Choose → CLICK

## 9. [button] 选择 → 点击

Please formulate an operational guide for future operations for solving the task. The guide includes:

请制定一个用于未来操作以解决任务的操作指南。该指南包括:

1. Plan: A \*\*multi-step future\*\* plan that does not involve low-level operations \*\*(start from current screen and action, DON'T include previous actions)\*\*; steps indexed by numbers. \*\*Be sure to pretend that you don't know the future interface and actions, and don't know the elements not existing on the current screen.\*\*

1. 计划: 一个 \*\*多步骤的未来\*\* 计划, 不涉及低级操作 \*\*(从当前界面和操作开始, 不包含之前的操作)\*\*; 步骤用数字索引。\*\* 务必假装你不知道未来的界面和操作, 也不知道当前界面不存在的元素。\*\*

2. Action: The specific \*\*immediate\*\* action that needs to be taken (Corresponding to the\*\* very next action\*\* [button] Search → CLICK at the box {"x\_left": 0.876, "y\_left": 0.308, "width": 0.063, "height": 0.034}), which should be detailed enough to directly control the operation, such as specifying the exact location (or box) to click, the exact content of the text to type, the item to select.

2. 操作: 需要执行的具体 \*\*即时\*\* 操作 (对应于 \*\*下一步操作\*\* [button] 搜索 → 点击, 位置在框 {"x\_left": 0.876, "y\_left": 0.308, "width": 0.063, "height": 0.034}), 应详细到可以直接控制操作, 如指定精确点击位置 (或框)、输入的文本内容、选择的项目。

\*\*Output Format:\*\* A JSON dictionary strictly following the format:

\*\*输出格式:\*\* 严格遵循以下格式的 JSON 字典:

{ "plan": "...<Your Plan Here>", "action": "...<Your Action Here>", "operation": "[button] Search → CLICK at the box \"x\_left\": 0.876, \"y\_left\": 0.308, \"width\": 0.063, \"height\": 0.034"} \*\*Please do not output any content other than the JSON format. Don't mention elements that only appear in HTML such as svg, div.\*\*

{ "plan": "...<你的计划>", "action": "...<你的操作>", "operation": "[button] 搜索 → 点击, 位置框 \"x\_left\": 0.876, \"y\_left\": 0.308, \"width\": 0.063, \"height\": 0.034"} \*\*请不要输出除 JSON 格式外的任何内容。不要提及仅出现在 HTML 中的元素如 svg、div。\*\*

The returned answer from GPT-4 would be:

从 GPT-4 返回的答案是:

{"plan": "1. After searching, you'll see a list of flight and hotel packages. 2. Review the recommended options, usually highlighted for convenience. 3. Focus on finding the lowest priced option. You might need to navigate through various sorting or filtering features to do this. 4. Once you've found the lowest priced option, select it to view more details. 5. In the detailed view, you'll have to choose your room type. Look for options that fit your requirements, like a business double room. 6. After selecting the room, proceed to finalize your booking. This might involve confirming your choice and entering any required information. 7. Review all the details carefully before finalizing the booking to ensure everything is correct.", "action": "[button] Search → CLICK at the box \"x\_left\": 0.876, \"y\_left\": 0.308, \"width\": 0.063, \"height\": 0.034", "operation": "[button] 搜索 → 点击, 位置框 \"x\_left\": 0.876, \"y\_left\": 0.308, \"width\": 0.063, \"height\": 0.034"} \*\*请不要输出除 JSON 格式外的任何内容。不要提及仅出现在 HTML 中的元素如 svg、div。\*\*

{"plan": "1. 搜索后，您会看到一系列航班和酒店套餐。2. 查看推荐选项，通常会高亮显示以便于选择。3. 重点寻找价格最低的选项。您可能需要通过各种排序或筛选功能来实现这一点。4. 找到最低价选项后，选择它以查看详细信息。5. 在详细视图中，您需要选择房型。寻找符合您需求的选项，比如商务双人房。6. 选定房间后，继续完成预订。这可能包括确认选择和填写必要信息。7. 在最终确认预订前，仔细核对所有细节，确保信息无误。",

"action": "Click the 'Search' button to proceed with the search based on the entered criteria. This button is located towards the right side of the screen.",

"action": "点击“搜索”按钮，根据输入的条件进行搜索。该按钮位于屏幕右侧。",

"operation": "[button] Search → CLICK at the box {"x\_left": 0.876, "y\_left": 0.308, "width": 0.063, "height": 0.03}" }

"operation": "[按钮] 搜索 → 点击框 {"x\_left": 0.876, "y\_left": 0.308, "width": 0.063, "height": 0.03}" }

## 7. Failure cases

### 7. 失败案例

The failure cases can be summarized into four categories (Fig. 9):

失败案例可归纳为四类 (图 9):

1. Incorrect action prediction;

1. 操作预测错误；

2. Incorrect coordinate prediction;

2. 坐标预测错误；

3. Incorrect GUI observation;

3. 图形用户界面 (GUI) 观察错误；

4. Hallucination.

4. 幻觉。

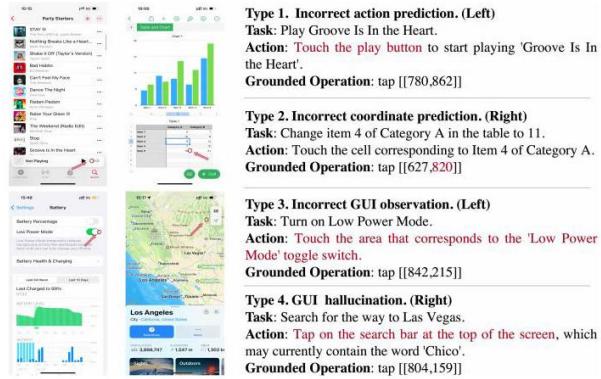


Figure 9. Failure cases of GUI Agent.

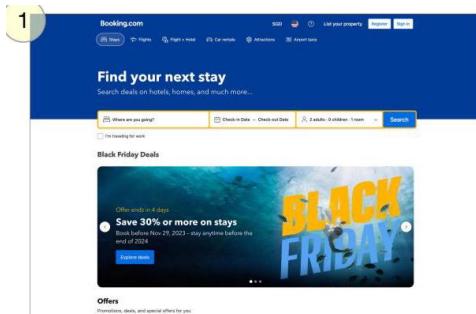
图 9. GUI 代理的失败案例。

## 8. More Generated Samples of CogAgent

### 8. CogAgent 的更多生成样本

To further demonstrate the capabilities of CogAgent, we provide additional generated samples in this section.

为了进一步展示 CogAgent 的能力，我们在本节提供了更多生成的样本。



Plan: 1. Type 'New York' into the search box to specify the destination.

计划:1. 在搜索框中输入“New York”以指定目的地。

2. Select the correct New York location from the auto-suggested list.

2. 从自动建议列表中选择正确的纽约位置。

3. Adjust the check-in and check-out dates to December 1st, respectively.

3. 将入住和退房日期分别调整为 12 月 1 日。

4. Execute the search by clicking the 'Search' button.

4. 点击“搜索”按钮执行搜索。

5. Once the results are displayed, sort them by price, selecting the lowest first.

5. 结果显示后，按价格排序，选择最低价优先。

6. Review the options and select the room offer that is the lowest priced.

6. 审查选项并选择价格最低的房间报价。

7. Proceed to book the selected room following the website's booking procedure.

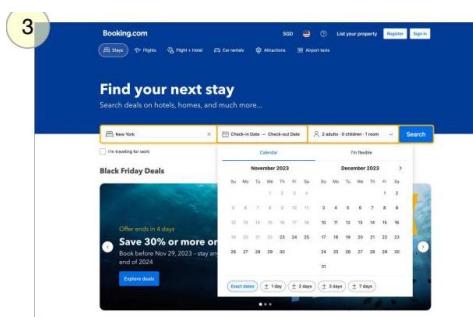
7. 按照网站的预订流程继续预订所选房间。

Next Action: Type 'New York' into the textbox located at the coordinates {'x\_min': 0.175, 'y\_min': 0.309, 'width': 0.28, 'height': 0.047} on the screen.

下一步操作: 在屏幕坐标 {'x\_min': 0.175, 'y\_min': 0.309, 'width': 0.28, 'height': 0.047} 处的文本框中输入“New York”。

Grounded Operation: [textbox] Where are you going? -> TYPE: New York at the box [[175,309,455,356]]

具体操作:[文本框] 你要去哪? -> 在框 [[175,309,455,356]] 中输入:New York



Plan: 1. Confirm the checkout date by clicking on the appropriate date, which in this case is 1 December 2023.

计划:1. 通过点击相应日期确认退房日期，此处为 2023 年 12 月 1 日。

2. Initiate the search for hotels by clicking the 'Search' button to find available rooms.

2. 点击“搜索”按钮开始搜索酒店以查找可用房间。

3. Once the results are displayed, sort the list of hotels by price, selecting the option to order them from the lowest to highest price.

3. 结果显示后，按价格对酒店列表排序，选择从低到高的排序方式。

4. Review the sorted list and select the hotel offering the lowest price.

4. 查看排序后的列表，选择价格最低的酒店。

5. Proceed to book the selected hotel room following the website's booking procedure.

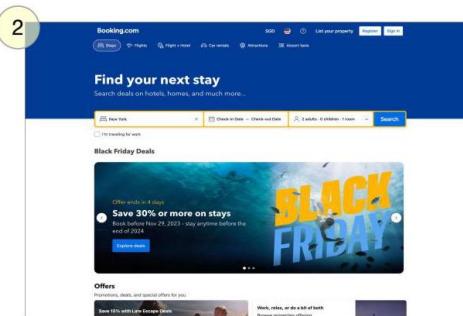
5. 按照网站的预订流程预订所选酒店房间。

Next Action: Click on the calendar box corresponding to the date 1 December 2023 to set it as the checkout date.

下一步操作: 点击对应 2023 年 12 月 1 日的日历框，将其设为退房日期。

Grounded Operation:[gridcell] 1 December 2023 -> CLICK at the box [[739,564,776,611]]

实际操作:[gridcell] 2023 年 12 月 1 日 -> 点击坐标框 [[739,564,776,611]]



Plan: To book the lowest priced room in New York on Dec 1st, the next steps would involve:

计划: 为了预订 12 月 1 日在纽约价格最低的房间，接下来的步骤包括:

1. Adjusting the check-in and check-out dates to December 1st, respectively.

1. 分别将入住和退房日期调整为 12 月 1 日。

2. Executing a search for available rooms.

2. 执行可用房间搜索。

3. Sorting the search results by price, from low to high.

3. 按价格从低到高排序搜索结果。

4. Selecting the room offer with the lowest price.

4. 选择价格最低的房间报价。

5. Proceeding to book the chosen room.

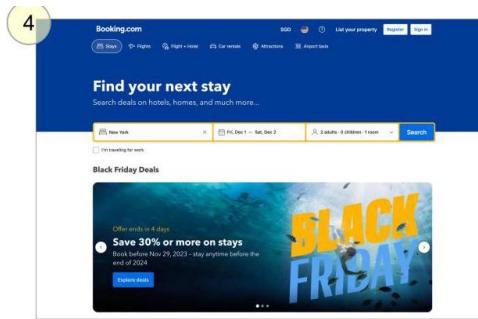
5. 继续预订所选房间。

Next Action: Click on the 'Check-in Date' textbox at the specified coordinates to bring up the date selection interface.

下一步操作: 点击“入住日期”文本框指定坐标, 调出日期选择界面。

Grounded Operation:[textbox] Check-in Date → CLICK at the box [[438,309,549,357]]

实际操作:[textbox] 入住日期 → 点击坐标框 [[438,309,549,357]]



Plan: 1. Click the Search button to initiate the search with the current criteria.

计划:1. 点击搜索按钮, 使用当前条件发起搜索。

2. Once the search results are displayed, sort the listings by price, selecting the option to order them from low to high.

2. 搜索结果显示后, 按价格排序, 选择从低到高的排序方式。

3. Review the sorted list and select the room offer that is the lowest in price.

3. 查看排序后的列表, 选择价格最低的房间报价。

4. Click on the 'Book' button for the chosen offer to proceed with the reservation process.

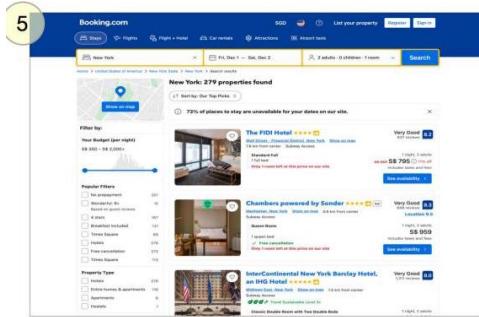
4. 点击所选报价的“预订”按钮, 继续预订流程。

Next Action: Click the Search button located at the bottom right of the search parameters box to execute the search for hotels in New York from December 1st to December 2nd for 2 adults and 0 children in 1 room.

下一步操作: 点击搜索参数框右下角的搜索按钮, 执行针对 12 月 1 日至 12 月 2 日, 2 名成人, 0 名儿童, 1 间房的纽约酒店搜索。

Grounded Operation:[button] Search -> CLICK at the box [[798,367,875,424]]

实际操作:[按钮] 搜索 -> 点击坐标框 [[798,367,875,424]]



Plan: 1. Click on the 'Sort by: Our Top Picks' button to change the sorting criteria.

计划:1. 点击“排序方式: 我们的精选”按钮以更改排序标准。

2. Select 'Price (lowest first)' from the sort options to arrange the listings by price in

2. 从排序选项中选择“价格(从低到高)”以按价格排列列表。

3. Identify the lowest-priced hotel listed after the sort operation.

3. 确认排序后列出的最低价酒店。

4. Click on the 'See availability' button for the lowest-priced hotel to view available rooms.

4. 点击最低价酒店的“查看可用房间”按钮以查看可预订的房型。

5. Choose a room option that fits the requirements, such as a specific price range or room type.

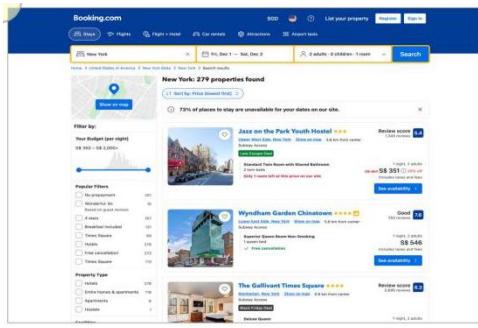
5. 选择符合要求的房型, 如特定价格区间或房间类型。

6. Proceed with booking by clicking 'Reserve' and following the subsequent booking steps. Next Action: Click at the box with coordinates starting approximately at 29.8% from the left of the screen and 27.2% from the top of the screen.

6. 点击“预订”并按照后续步骤完成预订。下一步操作: 点击屏幕左侧约 29.8%、顶部约 27.2% 处的坐标框。

Grounded Operation:[button] Sort by:Our Top Picks -> CLICK at the box [[298,272,420,307]]

实际操作:[按钮] 排序方式: 我们的精选 -> 点击坐标框 [[298,272,420,307]]



Plan: 1. Click on the 'Jazz on the Park Youth Hostel' to view more details about the property and to verify it meets all booking requirements including availability for the specified date, price, and amenities.

计划:1. 点击“Jazz on the Park 青年旅舍”查看详细信息，确认其满足预订要求，包括指定日期的可用性、价格和设施。

2. Review the detailed information provided after the click to ensure it matches the desired criteria.

2. 审核点击后显示的详细信息，确保符合预期条件。

3. If the property is suitable, proceed to reserve a room by following the booking instructions which may include selecting a room type, providing personal details, and making a payment.

3. 若合适，按照预订指引进行房间预订，可能包括选择房型、填写个人信息及付款。

Next Action: Move the cursor over the box that represents the 'Jazz on the Park Youth Hostel' link with coordinates starting at approximately 38.9% from the left of the screen and 38.6% from the top of the screen.

下一步操作: 将光标移至代表“Jazz on the Park 青年旅舍”链接的坐标框，位置约为屏幕左侧 38.9%、顶部 38.6%。

Grounded Operation:[link] Jazz on the Park Youth Hostel Opener West Side, New Yor... -> CLICK at the box [[389,386,810,435]]

实际操作:[链接] Jazz on the Park 青年旅舍 Opener West Side, New Yor... -> 点击坐标框 [[389,386,810,435]]

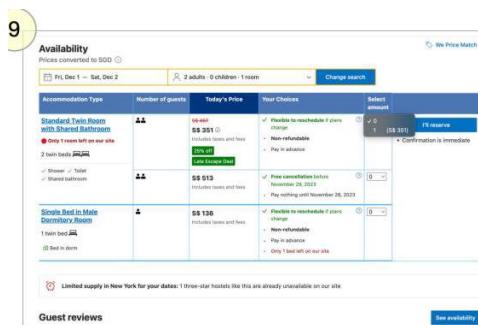
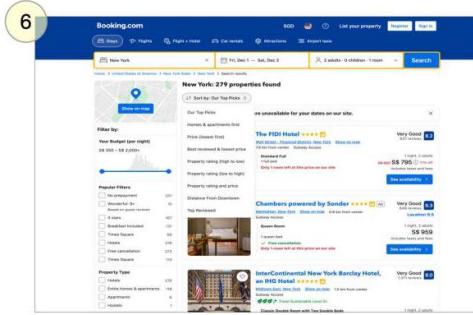


Figure 11. More generated samples of CogAgent.

图 11. CogAgent 生成的更多样本。



Plan: 1. Click on the 'Price (lowest first)' option to sort the properties by price in ascending order.

计划:1. 点击“价格(从低到高)”选项，将房源按价格升序排序。

2. Review the sorted list and identify the property with the lowest price.

2. 查看排序后的列表，找出价格最低的房产。

3. Select the property by clicking on it.

3. 点击该房产进行选择。

4. Once on the property's page, choose a room by clicking on the 'See availability' button for that room.

4. 进入房产页面后，点击该房间的“查看可用性”按钮选择房间。

5. Follow the booking process by selecting the date, room type, and any other necessary options, then proceed to reserve the room by clicking 'Reserve'.

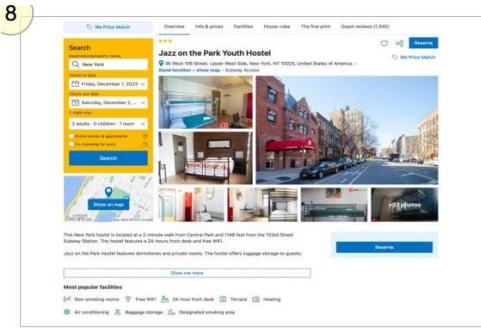
5. 按照预订流程选择日期、房型及其他必要选项，然后点击“预订”按钮完成房间预订。

Next Action: Click on the 'Price (lowest first)' sorting option located towards the top of the webpage, within the specified coordinates on the screen.

下一步操作: 点击网页顶部附近、屏幕指定坐标内的“价格(从低到高)”排序选项。

Grounded Operation:[li] Price (lowest first) -> CLICK at the box [[288,379,494,414]]

实际操作:[li] 价格(从低到高) -> 点击坐标框 [[288,379,494,414]]



Plan: After the current action, the next steps should include confirming the reservation details such as dates, number of guests, and room preferences. Then proceed to enter personal information required for booking. Following that, review all the details for accuracy, choose a payment method, and complete the booking process. Finally, wait for a confirmation message or email from the hotel to ensure the reservation has been successfully made.

计划: 当前操作后, 下一步应确认预订详情, 如日期、入住人数和房间偏好。然后填写预订所需的个人信息。接着核对所有信息的准确性, 选择支付方式并完成预订流程。最后, 等待酒店发送确认信息或邮件以确保预订成功。

Next Action: Click on the 'Reserve' button located towards the right side of the screen in a box with coordinates {'x\_min': 0.781, 'y\_min': 0.724, 'width': 0.106, 'height': 0.038}.

下一步操作: 点击屏幕右侧、坐标框为 {'x\_min': 0.781, 'y\_min': 0.724, 'width': 0.106, 'height': 0.038} 的“预订”按钮。

[[781,724,887,762]]

Plan: 1. Click on the 'Select amount' button for the room with the lowest price listed. 2. Choose the desired payment option, such as 'Pay in advance' if available and offered at a lower rate.

计划: 1. 点击价格最低房间对应的“选择数量”按钮。2. 选择所需的支付方式, 如有提供且价格更低, 可选择“预付”。

3. Confirm the selection and proceed to booking by clicking on 'Continue to book'.

3. 确认选择后, 点击“继续预订”进入下一步。

4. Complete the booking process by entering personal and payment details as required.

4. 按要求填写个人及支付信息, 完成预订流程。

5. Review the reservation summary and finalize the booking by confirming all details are correct and clicking 'Book now'.

5. 审核预订摘要, 确认所有信息无误后点击“立即预订”完成预订。

Next Action: Click on the 'Select amount' button located towards the right side of the screen for the accommodation option with the lowest price displayed, which is the 'Standard'

下一步操作: 点击屏幕右侧、价格最低的住宿选项“标准”对应的“选择数量”按钮。

Grounded Operation: [button] Select amount -> CLICK at the box [[739,272,906,318]]

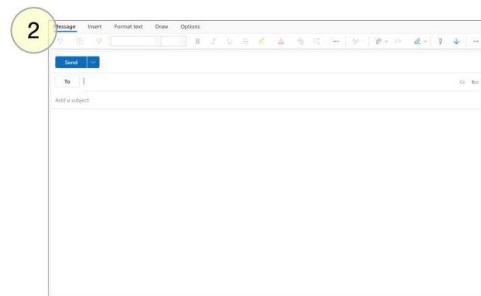
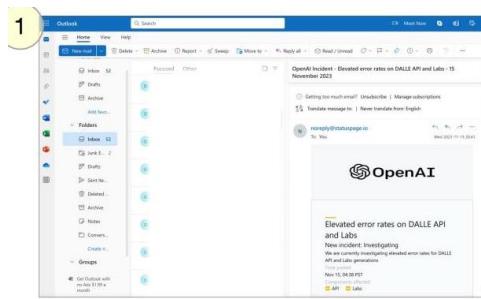
实际操作:[按钮] 选择数量 -> 点击坐标框 [[739,272,906,318]]

User: Could you guide me through the steps to

用户: 你能指导我完成以下步骤吗

”Write an email to cvpr2024@gmail.com, with subject of 'I'm CogAgent', and the main body of 'Here is CogAgent sending the email.' ”?

“写一封邮件给 cvpr2024@gmail.com, 主题为 ‘我是 CogAgent’，正文为 ‘这里是 CogAgent 发送的邮件。’ ”?



**CogAgent:**

**CogAgent:**

CogAgent:

CogAgent:

Plan: 1. Click on the 'New email' button to start a new email. Plan: 1. Type the email address into the 'To' field.

计划:1. 点击“新建邮件”按钮开始撰写新邮件。计划:1. 在“收件人”字段输入电子邮件地址。

2. Enter the recipient's email address in the 'To:' field. 2. Enter the subject line in the 'Subject' field.

2. 在“收件人”字段输入收件人的电子邮件地址。2. 在“主题”字段输入邮件主题。

3. Enter the subject in the 'Subject:' field. 3. Use the keyboard to type the main body of the email.

3. 在“主题”字段输入主题。3. 使用键盘输入邮件正文。

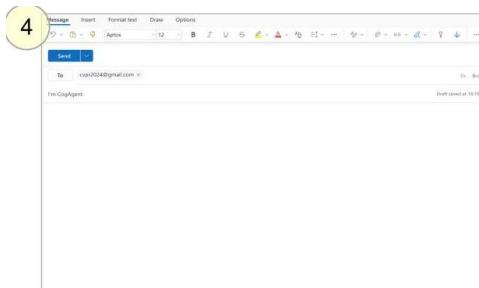
Next Action: Move the cursor to the 'To' textbox, click on it to focus on it, and type 'cvpr2024@gmail.com'.

4. Type the main body of the email in the 'Body' text area.

下一步操作: 将光标移至“收件人”文本框, 点击以聚焦, 然后输入“cvpr2024@gmail.com”。4. 在“正文”文本区域输入邮件正文。

Next Action: Click on the 'New mail' button located at the top left corner of the screen, within the Outlook interface.

下一步操作: 点击位于屏幕左上角 Outlook 界面中的“新建邮件”按钮。



3 的, 因为 8 是  $x = 0$  的最大值的  $\theta = 1$ ,  $\theta = \theta$ ,  $\theta = 1/2$

3 的, 因为 8 是  $x = 0$  的最大值的  $\theta = 1$ ,  $\theta = \theta$ ,  $\theta = 1/2$

Send

发送

To expr2024@gmail.com × It is

收件人 expr2024@gmail.com × 它是

Dreft assemble to 10.16

草稿组装至 10.16

**CogAgent:**

**CogAgent:**

**CogAgent:**

**CogAgent:**

Plan: 1. Click on the text area to bring up the keyboard if it's not already active. Plan: 1. Type the subject line into the 'Subject' field.

计划:1. 点击文本区域调出键盘(如果尚未激活)。计划:1. 在“主题”字段输入邮件主题。

2. Type the main body of the email, 'Here is CogAgent sending the email.'

2. 输入邮件正文，“这里是 CogAgent 发送的邮件。”

2. Enter the main body of the email in the large text area below the subject line.

2. 在主题下方的大文本区域输入邮件正文。

3. Use the 'Send' button to send the email.

3. 使用“发送”按钮发送电子邮件。

Next Action: Click on the 'Subject' textbox, which is currently highlighted and type 'I'm CogAgent'. Then, use the keyboard to type the rest of the message in the large text area below it. 3. Find and click the 'Send' button to send the email.

下一步操作: 点击当前高亮的“主题”文本框，输入“我是 CogAgent”。然后，使用键盘在其下方的大文本区域输入剩余的消息。3. 找到并点击“发送”按钮发送电子邮件。

Next Action: Move the cursor to the text area where the message content needs to be typed and type 'Here is CogAgent sending the email.'

下一步操作: 将光标移至需要输入消息内容的文本区域，输入“这里是 CogAgent 发送的邮件。”

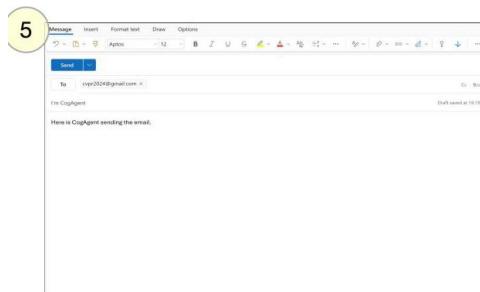


Figure 12. More generated samples of CogAgent.

图 12. 更多 CogAgent 生成的样本。

User: Could you guide me through the steps to "Listen to the music named You Raise Me Up." ?

用户: 你能指导我如何 “听名为 You Raise Me Up 的音乐” 吗?

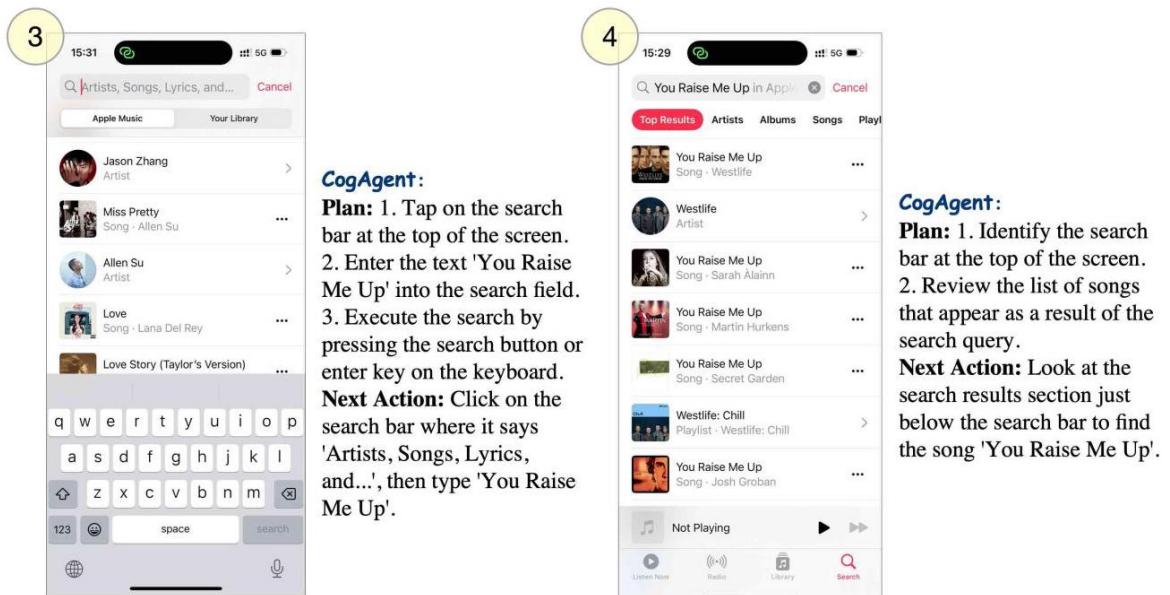
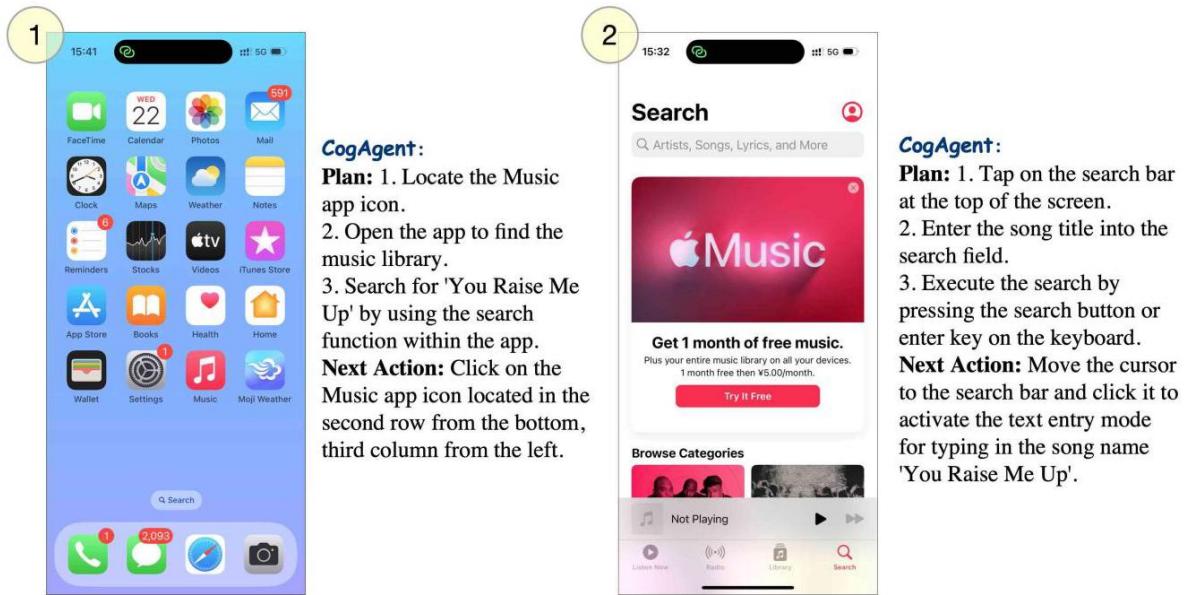


Figure 13. More generated samples of CogAgent.

图 13. 更多 CogAgent 生成的样本。



User: What's shown in this picture? Describe in detail.

用户: 这张图片中显示了什么？请详细描述。

CogAgent: The screenshot shows a character standing in the game Genshin Impact, looking at a teleport waypoint. The HUD displays various elements: on the top left is a mini-map with icons indicating different locations; on the top right are character portraits and health bars for Raiden Shogun, Bennett, and Kaedehara Kazuha, along with an UID number; below these is a quest prompt 'Go to Merusea Village' with a distance indicator of 284 m ; in the center is the glowing Teleport Waypoint; at the bottom are the character's level (Lv. 90), health points (25123/3461), stamina (623), and other game UI elements like food, a key, and a compass.

CogAgent: 截图显示了游戏《原神》(Genshin Impact)中一个角色站在传送点前。界面显示多个元素：左上角是带有不同地点图标的小地图；右上角是雷电将军、班尼特和枫原万叶的角色头像及生命值条，还有一个UID号码；下方是任务提示“前往梅露西亚村”，并显示距离 284 m；中央是发光的传送点；底部显示角色等级(90 级)、生命值(25123/3461)、体力(623)及其他游戏界面元素，如食物、钥匙和指南针。

User: I'm looking for guidance on how to "Go to the teleport waypoint"?

用户: 我想知道如何“前往传送点”？

CogAgent: Plan: 1. Identify the direction to the teleport waypoint from the current location on the map.

CogAgent: 计划:1. 从地图上确定当前位置到传送点的方向。

2. Navigate through the environment or use in-game transportation methods to reach the destination. 3. Confirm arrival at the teleport waypoint and prepare for the next step of the journey within Genshin Impact.

2. 通过环境导航或使用游戏内交通方式到达目的地。3. 确认已到达传送点，准备进行《原神》中的下一步旅程。

Next Action: Move the cursor over the 'Teleport Waypoint' icon located towards the center right of the screen, just above the character's feet, and left-click to activate it.

下一步操作: 将光标移至屏幕中偏右、角色脚上方的“传送点”图标, 左键点击以激活。

Grounded Operation: [generic] Teleport Waypoint -> CLICK at the box [[506,497,686,547]]

具体操作:[通用] 传送点 -> 点击区域 [[506,497,686,547]]

User: Who are the teammates on the current mission?

用户: 当前任务的队友是谁?

CogAgent: The teammates on the current mission are Raiden Shogun, Bennett, and Kaedebara Kazuha.

CogAgent: 当前任务的队友是雷电将军、班尼特和枫原万叶。

User: I'm looking for guidance on how to "Find out how to view a detailed map"?

用户: 我想知道如何“查看详细地图”?

CogAgent: Plan: 1. Identify the map icon on the screen.

CogAgent: 计划:1. 识别屏幕上的地图图标。

2. Click on the map icon to open the detailed map view.

2. 点击地图图标以打开详细地图视图。

Next Action: Move the cursor over the top left corner of the screen and click on the compass-like icon with 'N' inside it.

下一步操作: 将光标移至屏幕左上角, 点击带有“N”字样的指南针图标。



User: When did the Smelly Cat video post?

用户:Smelly Cat 视频是什么时候发布的?

CogAgent: The Smelly Cat video post was made 2 years ago.

CogAgent: Smelly Cat 视频发布于 2 年前。

User: How long is the Smelly Cat video?

用户:Smelly Cat 视频多长时间?

CogAgent: The Smelly Cat video is 3 minutes and 23 seconds long.

CogAgent: Smelly Cat 视频时长为 3 分 23 秒。

User: What should I do if I want to watch the Smelly Cat video (in the middle of the screen) and share it with my friend Rena?

用户: 如果我想观看屏幕中间的 Smelly Cat 视频并分享给我的朋友 Rena, 我该怎么办?

CogAgent: Plan: 1. Play the Smelly Cat video by Taylor Swift & Lisa Kudrow to ensure it is fully loaded and ready for viewing.

CogAgent: 计划:1. 播放 Taylor Swift 和 Lisa Kudrow 的 Smelly Cat 视频, 确保视频已完全加载并准备观看。

2. Locate the share icon on the screen, which is typically represented by a symbol such as a paper airplane or an arrow pointing outwards from a box.

2. 找到屏幕上的分享图标，通常是纸飞机或从盒子向外指的箭头符号。

3. Tap the share icon to open sharing options.

3. 点击分享图标以打开分享选项。

4. Find and select the option to send via messaging app, like WhatsApp, to share the video with friends.

4. 找到并选择通过消息应用 (如 WhatsApp) 发送的选项，以便与朋友分享视频。

5. In the messaging app, locate the contact named Rena.

5. 在消息应用中找到名为 Rena 的联系人。

6. Send the video link to Rena through the messaging app.

6. 通过消息应用将视频链接发送给 Rena。

Next Action: Touch the area in the middle of the screen where the 'Smelly Cat' thumbnail is displayed, corresponding to [Live on The 1989 World Tour] by Taylor Swift & Lisa Kudrow - Smelly Cat. This action is intended to start playing the video.

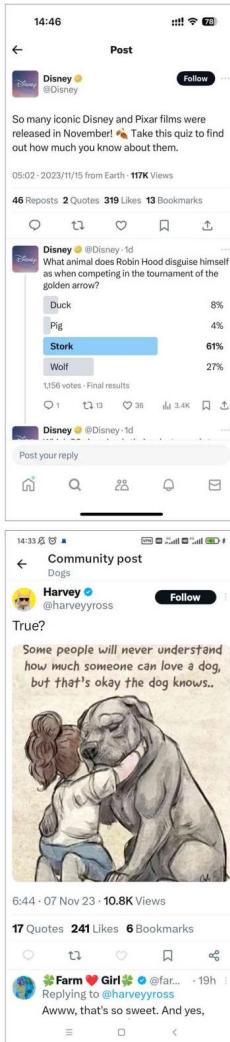
下一步操作: 触摸屏幕中间显示“Smelly Cat”缩略图的区域，该视频为 Taylor Swift 和 Lisa Kudrow 演唱的 [Live on The 1989 World Tour] - Smelly Cat。此操作旨在开始播放视频。

Grounded Operation: tap [[701,498]]

实际操作: 点击 [[701,498]]

Figure 14. More generated samples of CogAgent.

图 14. CogAgent 生成的更多样本。



User: Describe the image in detail.

用户: 请详细描述这张图片。

CogAgent: The image is a screenshot of Twitter posts from the official Disney account. In the first post, there's an emoji-filled statement that reads 'So many iconic Disney and Pixar films were released in November! Take this quiz to find out how much you know about them.' This tweet was posted on 2023/11/15 at 05:02 and has received 46 reposts, 2 quotes, 319 likes, and 13 bookmarks. Below it, another tweet states 'What animal does Robin Hood disguise himself as when competing in the tournament of the golden arrow?' with multiple choice options like Duck, Pig, Stork, Wolf.

CogAgent: 图片是官方迪士尼账号在Twitter上的帖子截图。第一条帖子中有一段充满表情符号的文字：“那么多标志性的迪士尼和皮克斯电影都在十一月上映！来做个测验，看看你了解多少。”该推文发布于2023年11月15日05:02，获得了46次转发、2次引用、319个点赞和13次收藏。下面另一条推文写道：“罗宾汉在黄金箭比赛中伪装成什么动物？”并附有多项选择：鸭子、猪、鹤、狼。

User: Which option has the highest number of votes?

用户: 哪个选项获得的票数最高?

CogAgent: The 'Stork' option has the highest number of votes with 61%.

CogAgent: “鹳” 选项以 61% 的票数领先。

User: How can I "Follow Disney's Twitter account"?

用户: 我怎样才能 “关注迪士尼的 Twitter 账号”?

CogAgent: Plan: 1. Locate the 'Follow' button on the screen. 2. Click the 'Follow' button to follow Disney's Twitter account.

CogAgent: 计划:1. 找到屏幕上的“关注”按钮。2. 点击“关注”按钮以关注迪士尼的 Twitter 账号。

Next Action: Move the cursor over the 'Follow' button next to the Disney logo and click it.

下一步操作: 将光标移至迪士尼标志旁的“关注”按钮并点击。

User: Describe the image in detail.

用户: 请详细描述这张图片。

CogAgent: The image showcases a tweet from an account named 'Harvey' with the handle '@harveyyross'. The tweet contains an illustration of a young girl hugging a large dog. Above this, there is another text that reads: 'Some people will never understand how much someone can love a dog, but that's okay the dog knows...'. Below the illustration, there are statistics showing the tweet has been viewed 10.8K times, liked by 241 users, retweeted 6 times, and bookmarked by 6 users. Additionally, there is a reply to this tweet from an account named 'Farm Girl' with the handle '@far...', which says: 'Awww, that's so sweet. And yes,' indicating agreement or appreciation for the original message.

CogAgent: 图片展示了一条来自名为“Harvey”、账号为“@harveyyross”的推文。推文中有一幅插图，描绘一个小孩拥抱一只大狗。插图上方有文字：“有些人永远无法理解一个人能多么爱一只狗，但没关系，狗知道……”。插图下方显示该推文被浏览了 1.08 万次，获得 241 个点赞，转发 6 次，收藏 6 次。此外，有一条来自名为“Farm Girl”、账号为“@far...”的回复，内容是：“哇，真甜蜜。是的，” 表达了对原推文的认同或赞赏。

User: Does the tweet have any reply? What is it?

用户: 这条推文有回复吗? 回复内容是什么?

CogAgent: Yes, there is a reply to the tweet from an account named 'Farm Girl' with the handle '@far...'. The reply reads: 'Awww, that's so sweet. And yes,' indicating agreement or appreciation for the original message.

CogAgent: 有，推文收到了来自名为“Farm Girl”、账号为“@far...”的回复。回复内容是：“哇，真甜蜜。是的，” 表达了对原推文的认同或赞赏。

Figure 15. More generated samples of CogAgent.

图 15. CogAgent 生成的更多样本。



User: Describe the screen in detail.

用户: 请详细描述屏幕内容。

CogAgent: The picture shows a mobile phone screen with various icons arranged in a grid. There are 24 app icons on the screen, including three cute cartoon bears at the bottom of the screen. These bears have different expressions and colors. In addition to the app icons, there is also a digital clock displaying "11:15" at the top left corner of the screen.

CogAgent: 图片显示一部手机屏幕，图标以网格形式排列。屏幕上 有 24 个应用图标，底部有三只可爱的卡通熊，表情和颜色各异。除了应用图标外，屏幕左上角还有一个显示“11:15”的数字时钟。

User: What apps do I have?

用户: 我有哪些应用?

CogAgent: The apps on the phone are divided into two categories: system apps and third-party apps. System apps include "FaceTime", "Calendar", "Photos", "Mail", "Clock", "Maps", "Weather", "Notes", "Reminders", "Stocks", "Videos", "iTunes Store", "App Store", "Books", "Health", "Home", "Music", and "Moji Weather". Third-party apps include "Q Search" and other unidentified apps, which cannot be identified from their icons alone.

CogAgent: 手机上的应用分为两类: 系统应用和第三方应用。系统应用包括“FaceTime”、“日历”、“照片”、“邮件”、“时钟”、“地图”、“天气”、“备忘录”、“提醒事项”、“股票”、“视频”、“iTunes Store”、“App Store”、“图书”、“健康”、“家庭”、“音乐”和“Moji Weather”。第三方应用包括“Q Search”和其他无法仅凭图标识别的应用。

User: What should I do if I want to see photos taken 2 days ago?

用户: 如果我想查看两天前拍的照片, 该怎么办?

CogAgent: You should tap on the "Photos" app, and it will display all your photos. If you want to see photos taken 2 days ago, you can select a date from the options at the top of the screen or use the search function in the Photos app to find specific photos.

CogAgent: 你应该点击“照片”应用, 它会显示你所有的照片。如果想查看两天前拍的照片, 可以从屏幕顶部的日期选项中选择, 或者使用照片应用中的搜索功能查找特定照片。

User: If I want to take a photo, what should I do?

用户: 如果我想拍照, 该怎么办?

CogAgent: You should tap on the "Camera" app to take a photo.

CogAgent: 你应该点击“相机”应用来拍照。

User: What should I do if I want to call Susan?

用户: 如果我想给 Susan 打电话, 该怎么办?

CogAgent: You should tap on the "Phone" app to call Susan.

CogAgent: 你应该点击“电话”应用来给 Susan 打电话。

User: How can I know Susan's number?

用户: 我怎么知道 Susan 的号码?

CogAgent: You can call Susan by searching for her number in the phone's contact list or on your phone's address book. If you have saved her number, you can directly dial it from there.

CogAgent: 你可以在手机的联系人列表或通讯录中搜索 Susan 的号码。如果你已经保存了她的号码，可以直接从那里拨打。

User: What should I do if I want to download youtube?

用户: 如果我想下载 YouTube, 该怎么办?

CogAgent: You should tap on the "App Store" app, and then search for YouTube in it. Once you find the official YouTube app, you can download and install it to watch videos.

CogAgent: 你应该点击“App Store”应用，然后在其中搜索 YouTube。找到官方的 YouTube 应用后，可以下载并安装以观看视频。

User: Is there a preferred method for "Join another zoom meeting with id 13572468"?

用户: 关于“加入另一个 ID 为 13572468 的 Zoom 会议”，有没有推荐的方法?

CogAgent: Plan: 1. Identify and select the 'Join' option for an existing or new meeting ID.

CogAgent: 计划:1. 识别并选择“加入”选项，输入现有或新的会议 ID。

2. Enter the specific Meeting ID when prompted to join a particular Zoom meeting.

2. 在提示时输入特定的会议 ID 以加入指定的 Zoom 会议。

3. Confirm any additional required details such as passcodes if necessary, and then join the meeting successfully.

3. 如有需要，确认任何额外的必填信息(如密码)，然后成功加入会议。

Next Action: Tap on the 'Join' button located in the middle of the screen, which corresponds to the coordinates [x, y]=[0.456,0.18]. This action is intended to initiate the process of joining a Zoom meeting.

下一步操作: 点击屏幕中间位置的“加入”按钮，坐标为 [x, y]=[0.456,0.18]。此操作旨在启动加入 Zoom 会议的流程。

Grounded Operation: tap [[456,180]]

接地操作: 点击 [[456,180]]

User: In how many minutes will the next meeting start? Please think step by step.

用户: 下一场会议将在多少分钟后开始？请一步步思考。

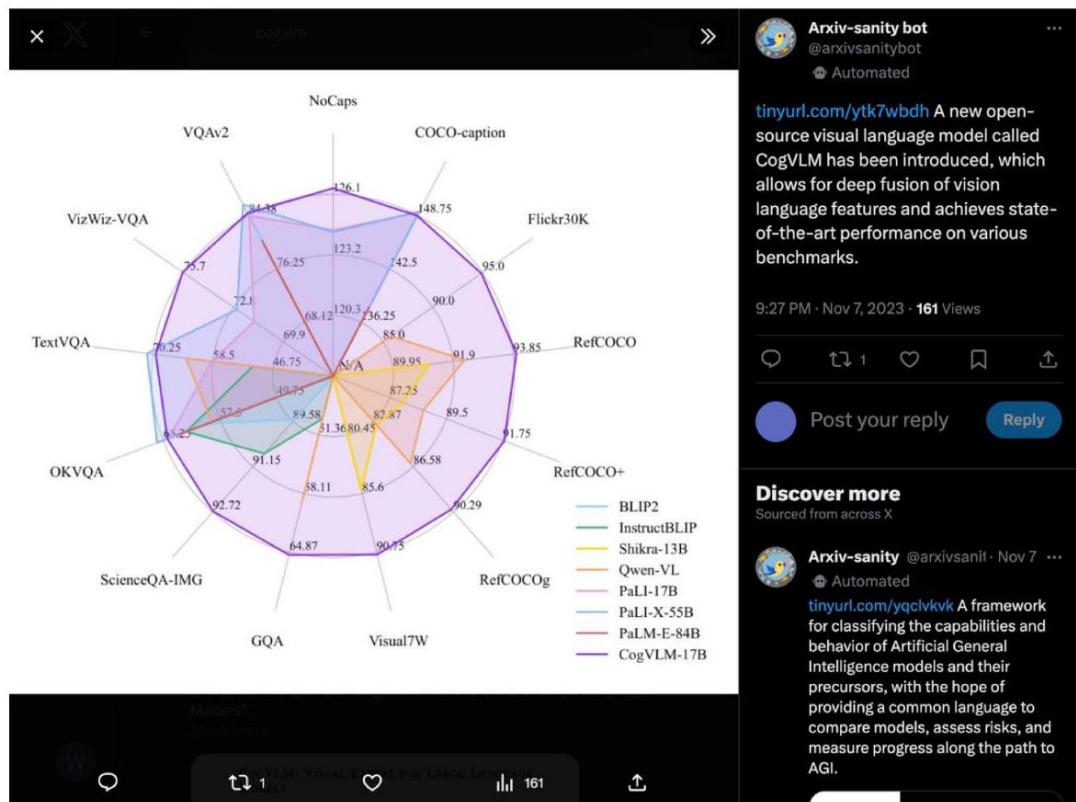
CogAgent: The next meeting will start in 23 minutes. This is determined by subtracting the current time (22:35) from the scheduled start time of the meeting (23:00).

CogAgent: 下一场会议将在 23 分钟后开始。这个时间是通过用会议的预定开始时间 (23:00) 减去当前时间 (22:35) 得出的。



Figure 16. More generated samples of CogAgent.

图 16. 更多 CogAgent 生成的样本。



User: What's the performance on CogVLM-17B on Visual7W?

用户: CogVLM-17B 在 Visual7W 上的表现如何?

CogAgent: The performance on CogVLM-17B on Visual7W is 90.75.

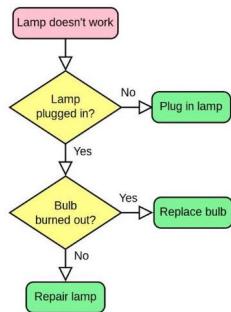
CogAgent: CogVLM-17B 在 Visual7W 上的表现是 90.75。

User: What's Arxiv-sanity's comment?

用户: Arxiv-sanity 的评论是什么?

CogAgent: Arxiv-sanity's comment is: 'A new open-source visual language model called CogVLM has been introduced, which allows for deep fusion of vision language features and achieves state-of-the-art performance on various benchmarks.'

CogAgent: Arxiv-sanity 的评论是: “引入了一种名为 CogVLM 的新开源视觉语言模型，能够实现视觉语言特征的深度融合，并在多个基准测试中达到最先进的性能。”



User: What is the solution if the lamp is plugged in and the bulb has burned out? CogAgent: If the lamp is plugged in and the bulb has burned out, the solution is to 'Replace bulb'.

用户: 如果灯插上电了但灯泡烧坏了, 怎么办? CogAgent: 如果灯插上电且灯泡烧坏, 解决方案是“更换灯泡”。

User: I am getting one funghi pizza and one prosciutto pizza. How much should I pay in total?

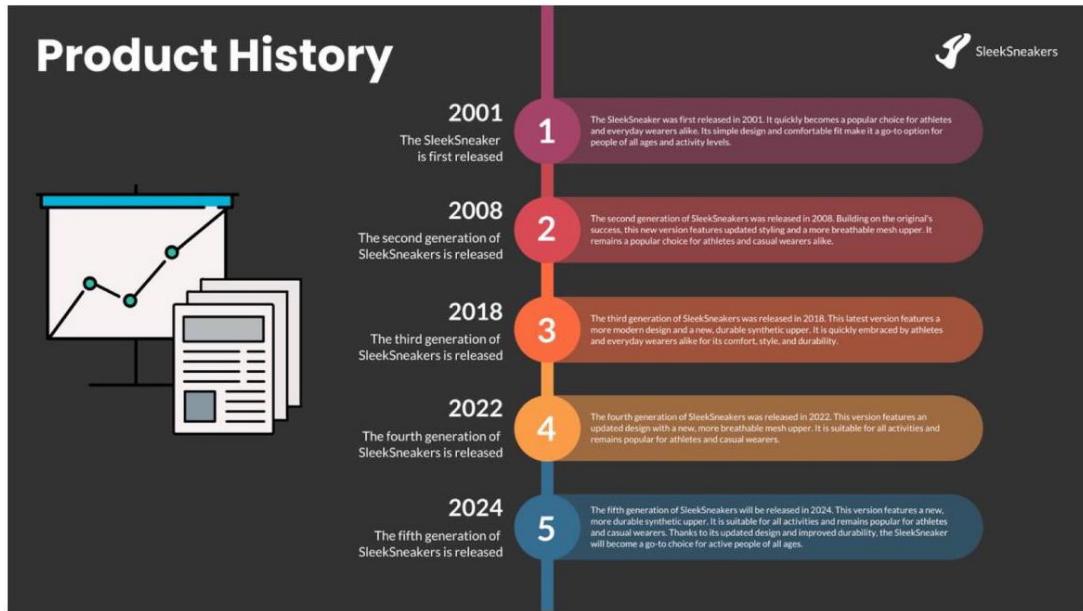
用户: 我要一份蘑菇披萨和一份火腿披萨。总共需要付多少钱?

CogAgent: The funghi pizza costs 6,90 \u20ac and the prosciutto pizza costs 7,90 \u20ac. Therefore, the total cost for both pizzas would be 14,80 \u20ac.

CogAgent: 蘑菇披萨价格为 6.90 欧元，火腿披萨价格为 7.90 欧元。因此，两份披萨的总价为 14.80 欧元。

Figure 17. More generated samples of CogAgent.

图 17. 更多 CogAgent 生成的样本。



User: Describe this image.

用户: 描述这张图片。

CogAgent: The image is a timeline showcasing the product history of 'SleekSneakers'. It starts from 2001, where 'The SleekSneaker' was first released, and ends in 2024, where the fifth generation of SleekSneakers was released. The timeline highlights key milestones, such as the introduction of the second generation of SleekSneakers in 2008, the fourth generation in 2022, and the fifth generation in 2024. Each milestone is accompanied by a brief description.

CogAgent: 这张图片是一条展示“SleekSneakers”产品历史的时间线。起始于 2001 年，当时发布了第一代“SleekSneaker”，终止于 2024 年，发布了第五代 SleekSneakers。时间线突出显示了关键里程碑，如 2008 年推出第二代，2022 年推出第四代，以及 2024 年推出第五代。每个里程碑都有简要描述。

User: Read the content in 2022.

用户: 请读出 2022 年的内容。

CogAgent: In 2022, the fourth generation of SleekSneakers was released. This version features an updated design with a new, more breathable mesh upper. It is suitable for all activities and remains popular for athletes and casual wearers.

CogAgent:2022 年，发布了第四代 SleekSneakers。该版本采用了更新的设计，配备了更透气的网面鞋面。适合各种活动，深受运动员和休闲用户欢迎。



User: In which country was this photo taken? CogAgent: The photo was taken in Australia, as the Australian Open is one of the major tennis tournaments held there.

用户: 这张照片是在什么国家拍的? CogAgent: 这张照片是在澳大利亚拍摄的, 因为澳大利亚网球公开赛是该地举办的主要网球赛事之一。

User: Who is the person on the left?

用户: 左边的人是谁?

CogAgent: The person on the left is Keira

CogAgent: 左边的人是 Keira。

Knightley.

奈特利。

Figure 18. More generated samples of CogAgent.

图 18. CogAgent 的更多生成样本。

method	top-1 err.	top-5 err.
VGG [41] (ILSVRC'14)	-	8.43†
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	19.38	4.49

方法	top-1 错误率	top-5 错误率
VGG [41] (ILSVRC'14)	-	8.43 <sup>†</sup>
GoogLeNet [44] (ILSVRC'14)	-	7.89
VGG [41] (v5)	24.4	7.1
PReLU-net [13]	21.59	5.71
BN-inception [16]	21.99	5.81
ResNet-34 B	21.84	5.71
ResNet-34 C	21.53	5.60
ResNet-50	20.74	5.25
ResNet-101	19.87	4.60
ResNet-152	19.38	4.49

Table 4. Error rates (%) of single-model results on the ImageNet validation set (except <sup>†</sup> reported on the test set).

表 4. 单模型在 ImageNet 验证集上的错误率 (%)(除 <sup>†</sup> 外，均报告于测试集)。

User: Convert it to the markdown format.

用户: 请将其转换为 Markdown 格式。

CogAgent: | method | top-1 err. | top-5 err. |

CogAgent:| 方法 | top-1 错误率 | top-5 错误率 |



Dataset 1

数据集 1

Department	Name	Salary
Administration	Emily	\$2,000
Finance	Nick	\$1,800
Marketing	John	\$2,600
Marketing	Gloria	\$1,800
Administration	Jane	\$2,500
Finance	Max	\$2,000
Administration	Tyler	\$3,000
Finance	Wilkins	\$2,500
Marketing	Miles	\$2,800
Administration	Sabine	\$1,900

部门	姓名	薪资
行政管理	艾米丽	\$2,000
财务	尼克	\$1,800
市场营销	约翰	\$2,600
市场营销	格洛丽亚	\$1,800
行政管理	简	\$2,500
财务	马克斯	\$2,000
行政管理	泰勒	\$3,000
财务	威尔金斯	\$2,500
市场营销	迈尔斯	\$2,800
行政管理	萨宾	\$1,900

User: Based on the table, how much is Wilkins earning?

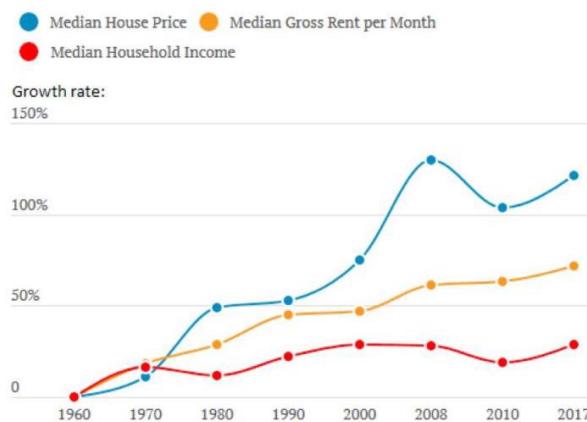
用户: 根据表格, Wilkins 的收入是多少?

CogAgent: Wilkins is earning \$2,500.

CogAgent: Wilkins 的收入是 2500 美元。

US Nationwide:

美国全国范围:



User: Which year has the highest growth rate of median house price?

用户: 哪一年中位房价的增长率最高?

CogAgent: The year 2008 has the highest growth rate of median house price, which is 150%.

CogAgent: 2008 年中位房价的增长率最高, 达到 150%。