CogAgent: A Visual Language Model for GUI Agents

CogAgent: 面向 GUI 代理的视觉语言模型

Wenyi Hong [1*] Weihan Wang [1*] Qingsong Lv [2] Jiazheng Xu [1*] Wenmeng Yu [2] Junhui Ji[2] Yan Wang [2] Zihan Wang [1*] Yuxiao Dong [1] Ming Ding [2†] Jie Tang [1† 1] Tsinghua University [2] Zhipu AI

洪文怡 [1*] 王伟涵 [1*] 吕庆松 [2] 徐家正 [1*] 余文萌 [2] 君辉 Ji[2] 王岩 [2] 王子涵 [1*] 董宇霄 [1] 丁明 [2†] 唐杰 [1† 1] 清华大学 [2] 智谱 AI

{hwy22@mails, jietang@}.tsinghua.edu.cn, ming.ding@zhipuai.cn

{hwy22@mails, jietang@}.tsinghua.edu.cn, ming.ding@zhipuai.cn

Abstract

摘要

People are spending an enormous amount of time on digital devices through graphical user interfaces (GUIs), e.g., computer or smartphone screens. Large language models (LLMs) such as ChatGPT can assist people in tasks like writing emails, but struggle to understand and interact with GUIs, thus limiting their potential to increase automation levels. In this paper, we introduce CogAgent, an 18- billion-parameter visual language model (VLM) specializing in GUI understanding and navigation. By utilizing both low-resolution and high-resolution image encoders, CogA-gent supports input at a resolution of 1120×1120 , enabling it to recognize tiny page elements and text. As a generalist visual language model, CogAgent achieves the state of the art on five text-rich and four general VQA benchmarks, including VQAv2, OK-VQA, Text-VQA, ST-VQA, ChartQA, infoVQA, DocVQA, MM-Vet, and POPE. CogAgent, using only screenshots as input, outperforms LLM-based methods that consume extracted HTML text on both PC and Android GUI navigation tasks-Mind2Web and AITW, advancing the state of the art. The model and codes are available at https://github.com/THUDM/CogVLM.

人们通过图形用户界面 (GUI), 如电脑或智能手机屏幕, 在数字设备上花费了大量时间。大型语言模型 (LLMs), 例如 ChatGPT, 可以协助人们完成写邮件等任务, 但在理解和交互 GUI 方面存在困难, 限制了其提升自动化水平的潜力。本文介绍了 CogAgent, 一种拥有 180 亿参数的视觉语言模型 (VLM), 专注于 GUI 的理解与导航。通过结合低分辨率和高分辨率图像编码器，CogAgent 支持 1120×1120 分辨率的输入，能够识别微小的页面元素和文本。作为一款通用视觉语言模型，CogAgent 在五个文本丰富和四个通用视觉问答 (VQA) 基准上达到了最新水平，包括 VQAv2、OK-VQA、Text-VQA、ST-VQA、ChartQA、infoVQA、DocVQA、MM-Vet 和 POPE。仅使用截图作为输入，CogAgent 在 PC 和安卓 GUI 导航任务——Mind2Web 和 AITW 上，均超越了基于提取 HTML 文本的 LLM 方法，推动了该领域的最新进展。模型和代码可在 https://github.com/THUDM/CogVLM 获取。

1. Introduction

1. 引言

Autonomous agents in the digital world are ideal assistants that many modern people dream of. Picture this scenario: You type in a task description, then relax and enjoy a cup of coffee while watching tasks like booking tickets online, conducting web searches, managing files, and creating Pow-erPoint presentations get completed automatically.

数字世界中的自主代理是许多现代人梦寐以求的理想助手。设想这样一个场景: 你输入任务描述, 然后放松地喝杯咖啡，观看诸如在线订票、网页搜索、文件管理和制作 PowerPoint 演示等任务自动完成。

Recently, the emergence of agents based on large language models (LLMs) is bringing us closer to this dream. For example, AutoGPT [33], a 150,000-star open-source project, leverages ChatGPT [29] to integrate language understanding with pre-defined actions like Google searches

近年来，基于大型语言模型 (LLMs) 的代理的出现正使我们更接近这一梦想。例如，AutoGPT [33]，一个拥有 15 万星的开源项目，利用 ChatGPT [29] 将语言理解与预定义操作如谷歌搜索结合起来。

and local file operations. Researchers are also starting to develop agent-oriented LLMs [7, 42]. However, the potential of purely language-based agents is quite limited in real-world scenarios, as most applications interact with humans through Graphical User Interfaces (GUIs), which are characterized by the following perspectives:

以及本地文件操作。研究人员也开始开发面向代理的 LLM [7, 42]。然而，纯语言代理在现实场景中的潜力相当有限，因为大多数应用通过图形用户界面 (GUI) 与人交互，GUI 具有以下特点:

- Standard APIs for interaction are often lacking.
    - 通常缺乏标准的交互 API。
- Important information including icons, images, diagrams, and spatial relations are difficult to directly convey in words.
    - 重要信息包括图标、图像、图表和空间关系难以用文字直接传达。
- Even in text-rendered GUIs like web pages, elements like canvas and iframe cannot be parsed to grasp their functionality via HTML.
    - 即使在文本渲染的 GUI 如网页中, 画布 (canvas) 和内嵌框架 (iframe) 等元素也无法通过 HTML 解析其功能。

Agents based on visual language models (VLMs) have the potential to overcome these limitations. Instead of relying exclusively on textual inputs such as HTML [28] or OCR results [31], VLM-based agents directly perceive visual GUI signals. Since GUIs are designed for human users, VLM-based agents can perform as effectively as humans, as long as the VLMs match human-level vision understanding. In addition, VLMs are also capable of skills such as extremely fast reading and programming that are usually beyond the reach of most human users,

extending the potential of VLM-based agents. A few prior studies utilized visual features merely as auxiliaries in specific scenarios. e.g. WebShop [39] which employs visual features primarily for object recognition purposes. With the rapid development of VLM, can we naturally achieve universality on GUIs by relying solely on visual inputs?

基于视觉语言模型 (VLM) 的代理有潜力克服这些限制。VLM 代理不依赖于 HTML [28] 或 OCR 结果 [31] 等纯文本输入，而是直接感知视觉 GUI 信号。由于 GUI 设计面向人类用户，只要 VLM 达到人类视觉理解水平，VLM 代理就能像人类一样高效。此外，VLM 还具备极快阅读和编程等通常超出大多数人类用户能力的技能，拓展了 VLM 代理的潜力。此前少数研究仅在特定场景中将视觉特征作为辅助，例如 WebShop [39] 主要利用视觉特征进行物体识别。随着 VLM 的快速发展，我们能否仅依靠视觉输入自然实现 GUI 的通用性?

In this work, we present CogAgent, a visual language foundation model specializing in GUI understanding and planning while maintaining a strong ability for general cross-modality tasks. By building upon CogVLM [38]—a recent open-source VLM, CogAgent tackles the following challenges for building GUI agents:

本文提出了 CogAgent，一种专注于 GUI 理解与规划的视觉语言基础模型，同时保持强大的通用跨模态任务能力。基于近期开源 VLM CogVLM [38]，CogAgent 解决了构建 GUI 代理的以下挑战:

- Training Data. Most current VLMs are pre-trained on datasets like LAION [32], consisting of natural images on the Web. However, we notice that the GUI

- 训练数据。目前大多数 VLM 预训练于如 LAION [32] 这类包含网络自然图像的数据集。然而，我们注意到 GUI

---

*Work was done when interned at Zhipu AI.

* 工作完成于在智谱 AI 实习期间。

† Corresponding authors

† 通讯作者

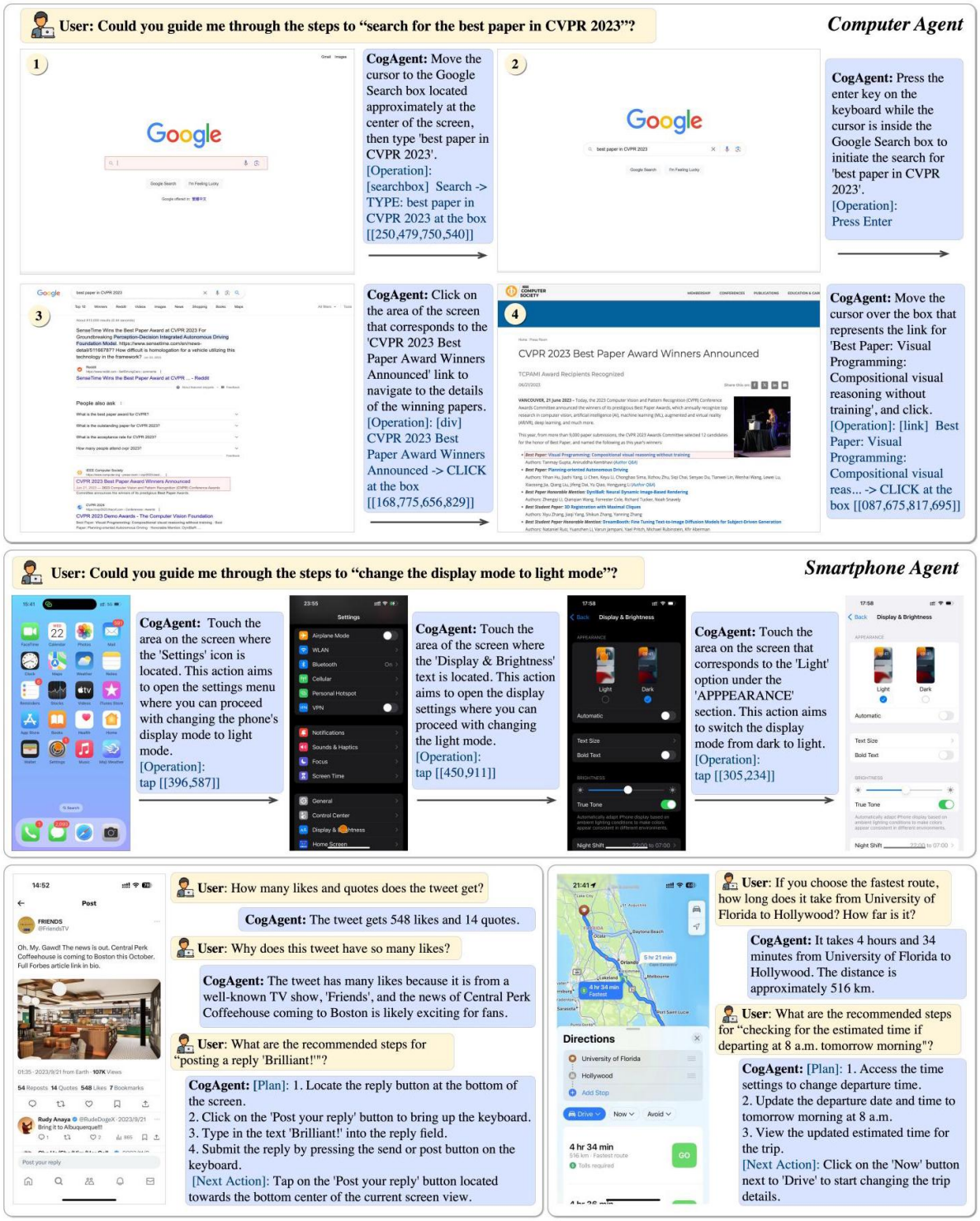

Figure 1. Samples of visual agents generated by CogAgent. More samples are demonstrated in the Appendix.

图 1. 由 CogAgent 生成的视觉代理样本。更多样本展示见附录。

images share a different distribution from natural images. We thus construct a large-scale annotated dataset about GUIs and OCR for continual pre-training.

图像与自然图像分布不同。因此，我们构建了一个关于图形用户界面 (GUI) 和光学字符识别 (OCR) 的规模化标注数据集，用于持续预训练。

- High-Resolution vs. Compute. In GUIs, tiny icons and text are ubiquitous, and it is hard to recognize them in commonly-used 224×224 resolution. However, increasing the resolution of input images results in significantly long sequence length in language models. For example, a 1120×1120 image corresponds to a sequence of 6400 tokens if the patch size is 14, demanding excessive training and inference compute. To address this, we design a cross-attention branch that allows for a trade-off between the resolution and the hidden size within a proper computation budget. Specifically, we propose to combine the original large ViT [12] (4.4B parameters) used in CogVLM [38] and a new small high-resolution cross-module (with image encoder of 0.30B parameters) to jointly model visual features.

  - 高分辨率与计算量。在 GUI 中，微小的图标和文本无处不在，常用 224×224 分辨率下难以识别。然而，提升输入图像分辨率会导致语言模型序列长度显著增加。例如，若 patch 大小为 14，1120×1120 的图像对应 6400 个 token，训练和推理计算需求极高。为此，我们设计了一个跨注意力分支，在合理计算预算内实现分辨率与隐藏层大小的权衡。具体而言，我们提出结合 CogVLM [38] 中使用的原始大型 ViT [12](44 亿参数) 与一个新的小型高分辨率跨模块 (图像编码器参数为 3 亿)，共同建模视觉特征。

Our experiments show that:

我们的实验表明:

- CogAgent tops popular GUI understanding and decision-making benchmarks, including AITW [31] and Mind2Web [10]. To the best of our knowledge, this is the first time that a generalist VLM can outperform LLM-based methods with extracted structured text.

  - CogAgent 在包括 AITW [31] 和 Mind2Web [10] 的流行 GUI 理解与决策基准测试中表现领先。据我们所知，这是首个通用视觉语言模型 (VLM) 能够超越基于大语言模型 (LLM) 提取结构化文本的方法。

- Though CogAgent focuses on GUIs, it achieves state-of-the-art generalist performance on nine visual question-answering benchmarks including VQAv2 [1], OK-VQA [23], TextVQA [34], ST-VQA [4], ChartQA [24], infoVQA [26], DocVQA [25], MM-Vet [41], and POPE [19].

  - 尽管 CogAgent 专注于 GUI，但在九个视觉问答基准上实现了最先进的通用性能，包括 VQAv2 [1]、OK-VQA [23]、TextVQA [34]、ST-VQA [4]、ChartQA [24]、infoVQA [26]、DocVQA [25]、MM-Vet [41] 和 POPE [19]。

- The separated design of high- and low-resolution branches in CogAgent significantly lows the compute cost for consuming high-resolution images, e.g., the number of the floating-point operations (FLOPs) for CogAgent-18B with 1120×1120 inputs is less than half that of CogVLM-17B with its default 490×490 inputs.

- CogAgent 中高低分辨率分支的分离设计显著降低了处理高分辨率图像的计算成本，例如，CogAgent-18B 在 1120×1120 输入下的浮点运算次数 (FLOPs) 不到默认 490 × 490 输入的 CogVLM-17B 的一半。

CogAgent is open-sourced at https://github.com/THUDM/CogVLM.It represents an effort to promote the future research and application of AI agents, facilitated by advanced VLMs.

CogAgent 已开源于 https://github.com/THUDM/CogVLM。它代表了利用先进视觉语言模型推动未来 AI 代理研究与应用的努力。

2. Method

2. 方法

In this section, we will first introduce the architecture of Co-gAgent, especially the novel high-resolution cross-module, and then illustrate the process of pre-training and alignment in detail.

本节首先介绍 CogAgent 的架构，特别是新颖的高分辨率跨模块，然后详细说明预训练与对齐过程。

2.1. Architecture

2.1. 架构

The architecture of CogAgent is depicted in Fig. 2. We build our model based on a pre-trained VLM (on the right side of the image), and propose to add a cross-attention module to process high-resolution input (on the left side of the image). As our base VLM, We select CogVLM- 17B [38], an open-sourced and state-of-the-art large vison-language model. Specifically, We employ EVA2-CLIP-E [35] as the encoder for low-resolution images (224×224 pixels), complemented by an MLP adapter that maps its output into the feature space of the visual-language decoder. The decoder, a pre-trained language model, is enhanced with a visual expert module introduced by Wang et al. [38] to facilitate a deep fusion of visual and language features. The decoder processes a combined input of the low-resolution image feature sequence and text feature sequence, and autoregressively outputs the target text.

CogAgent 的架构如图 2 所示。我们基于预训练视觉语言模型 (图右侧) 构建模型，并提出添加跨注意力模块以处理高分辨率输入 (图左侧)。作为基础 VLM，我们选择了开源且最先进的大型视觉语言模型 CogVLM-17B [38]。具体而言，我们采用 EVA2-CLIP-E [35] 作为低分辨率图像 (224×224 像素) 编码器，配合一个 MLP 适配器将其输出映射到视觉语言解码器的特征空间。解码器为预训练语言模型，结合 Wang 等人 [38] 引入的视觉专家模块，实现视觉与语言特征的深度融合。解码器处理低分辨率图像特征序列与文本特征序列的组合输入，自回归生成目标文本。

Similar to most VLMs, the original CogVLM can only accommodate images of relatively low resolution (224 or 490), which hardly meets the demands of GUI where the screen resolution of computers or smartphones is typically 720p (1280 × 720 pixels) or higher. It is a common problem among VLMs, e.g. LLaVA [21] and PALI-X [8] are pre-trained at a low resolution of 224 × 224 on the general domain. The primary reason is that

high-resolution image brings prohibitive time and memory overhead: VLMs usually concatenate text and image feature sequence as input to the decoder, thus the overhead of self-attention module is quadratic to the number of visual tokens (patches), which is quadratic to the image's side length. There are some initial attempts to reduce costs for high-resolution images. For instance, Qwen-VL [2] proposes a position-aware vision-language adapter to compress image features, but only reduces sequence length by four and has a maximum resolution of 448×448 . Kosmos-2.5 [22] adopts a Perceiver Resampler module to reduce the length of the image sequence. However, the resampled sequence is still long for self-attention in the large visual-language decoder (2,048 tokens), and can only be applied to restricted text recognition tasks.

与大多数视觉语言模型 (VLMs) 类似，原始的 CogVLM 只能处理相对较低分辨率的图像 (224 或 490)，这难以满足图形用户界面 (GUI) 对屏幕分辨率的需求，通常计算机或智能手机的屏幕分辨率为 720p(1280 × 720 像素) 或更高。这是 VLMs 中的一个普遍问题，例如 LLaVA [21] 和 PALI-X [8] 在通用领域的预训练分辨率较低，为 224×224 。主要原因是高分辨率图像带来了极高的时间和内存开销:VLMs 通常将文本和图像特征序列拼接作为解码器输入，因此自注意力模块的开销与视觉标记 (图像块) 数量的平方成正比，而视觉标记数量又与图像边长的平方成正比。已有一些初步尝试来降低高分辨率图像的成本。例如，Qwen-VL [2] 提出了一种位置感知视觉语言适配器来压缩图像特征，但仅将序列长度减少了四倍，最大分辨率为 448×448 。Kosmos-2.5 [22] 采用 Perceiver Resampler 模块来缩短图像序列长度，但重采样后的序列对于大型视觉语言解码器 (2048 个标记) 中的自注意力仍然较长，且仅能应用于受限的文本识别任务。

Therefore, we propose a novel high-resolution cross-module as a potent complement to the existing structure for enhancing understanding at high resolutions, which not only maintains efficiency confronting high-resolution images, but also offers flexible adaptability to a variety of visual-language model architectures.

因此，我们提出了一种新颖的高分辨率交叉模块，作为现有结构的有力补充，用于提升高分辨率下的理解能力，该模块不仅在面对高分辨率图像时保持高效，还能灵活适配多种视觉语言模型架构。

2.2. High-Resolution Cross-Module

2.2. 高分辨率交叉模块

The structural design of high-resolution cross-module is mainly based on the following observations:

高分辨率交叉模块的结构设计主要基于以下观察:

1. At a modest resolution such as 224×224 , images can depict most objects and layouts effectively, yet

1. 在如 224×224 这样适中的分辨率下，图像能够有效描绘大多数物体和布局，但

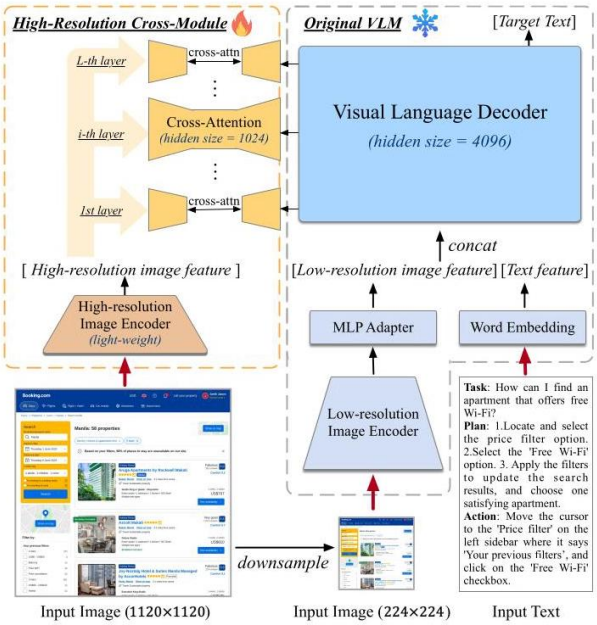

Figure 2. Model architecture of CogAgent. We adopt CogVLM as the original VLM.

图 2. CogAgent 模型架构。我们采用 CogVLM 作为原始视觉语言模型。

the resolution falls short in rendering text with clarity. Hence, our new high-resolution module should emphasize text-related features, which are vital for understanding GUIs.

分辨率不足以清晰呈现文本。因此，我们的新高分辨率模块应重点关注与文本相关的特征，这对理解 GUI 至关重要。

2. While pre-trained VLMs in general domain often need large hidden sizes (e.g. 4,096 in PALI-X and CogVLM, 5,120 in LLaVA), VLMs tailored for text-centered tasks like document OCR require smaller hidden sizes to achieve satisfying performance (e.g. 1,536 in Kosmos-2.5 and Pix2Struct [16]). This suggests that text-related features can be effectively captured using smaller hidden sizes.

2. 通用领域的预训练 VLM 通常需要较大的隐藏层尺寸 (例如 PALI-X 和 CogVLM 为 4096，LLaVA 为 5120)，而针对文本中心任务如文档 OCR 的 VLM 则需要较小的隐藏层尺寸以达到满意的性能 (例如 Kosmos-2.5 和 Pix2Struct [16] 为 1536)。这表明文本相关特征可以通过较小的隐藏层尺寸有效捕捉。

As shown in Fig. 2, the high-resolution cross-module acts as a new branch for higher-resolution input, which accepts images of size 1120×1120 pixels in our implementation. Different from the original low-resolution input branch, the high-resolution cross-module adopts a much smaller pre-trained vision encoder (visual encoder of EVA2-CLIP-L [35] in our implementation, 0.30B parameters), and uses cross-attention of a small hidden size to fuse high-resolution image features with every layer of VLLM decoder, thus reducing the computational cost. To be concrete, for an input image, it is resized to 1120×1120 and 224×224 and fed into the high-resolution cross-module and the low-resolution branch respectively, then encoded into image feature sequences X_{hi} and X_{lo} with two distinct-sized image encoders in parallel. The visual language decoder retains its original computations, while the only change is to

如图 2 所示,高分辨率交叉模块作为高分辨率输入的新分支,在我们的实现中接受尺寸为 1120×1120 像素的图像。与原始的低分辨率输入分支不同，高分辨率交叉模块采用了更小的预训练视觉编码器 (我们实现中使用 EVA2-CLIP-L [35] 的视觉编码器，参数量为 0.30B)，并利用小隐藏层尺寸的交叉注意力将高分辨率图像特征与视觉语言大模型 (VLLM) 解码器的每一层融合，从而降低计算成本。具体来说，对于输入图像，分别调整为 1120 × 1120 和 224 × 224 尺寸，分别输入高分辨率交叉模块和低分辨率分支，然后通过两个不同尺寸的图像编码器并行编码成图像特征序列 X_{hi} 和 X_{lo} 。视觉语言解码器保持其原有计算，唯一的变化是

integrate a cross-attention between X_{hi} and hidden states in every decoder layer.

在每个解码器层中集成 X_{hi} 与隐藏状态之间的交叉注意力。

Formally, suppose that the input hidden states of the i-th attention layer in the decoder are $X_{\text{in}_i} \in \mathbb{R}^{B\times\left(L_{I_{\text{lo}}}+L_T\right)\times D_{\text{dec}}}$, and the output hidden states of cross-module's image encoder are $X_{\text{hi}} \in \mathbb{R}^{B\times\left(L_{I_{\text{hi}}}\right)\times D_{\text{hi}}}$, where B is the batch size, $L_{I_{\text{lo}}}, L_{I_{\text{hi}}}$ and L_T are the lengths of the low-resolution image, high-resolution image and text sequences, D_{dec} and D_{hi} is the hidden size of the decoder and high-resolution encoder's output respectively. Each layer's attention procedure can be formulated as

形式上，假设解码器中第 i 个注意力层的输入隐藏状态为 $X_{\text{in}_i} \in \mathbb{R}^{B\times\left(L_{I_{\text{lo}}}+L_T\right)\times D_{\text{dec}}}$ ，跨模块图像编码器的输出隐藏状态为 $X_{\text{hi}} \in \mathbb{R}^{B\times\left(L_{I_{\text{hi}}}\right)\times D_{\text{hi}}}$ ，其中 B 为批量大小，$L_{I_{\text{lo}}}, L_{I_{\text{hi}}}$ 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，D_{dec} 和 D_{hi} 分别为解码器和高分辨率编码器输出的隐藏层维度。每一层的注意力过程可以表示为

$$X_i' = \text{MSA}\left(\text{layernorm}\left(X_{\text{in}_i}\right)\right) + X_{\text{in}_i}, \tag{1}$$

$$X_{\text{out}_i} = \text{MCA}\left(\text{layernorm}\left(X_i'\right), X_{\text{hi}}\right) + X_i', \tag{2}$$

where MSA and MCA represent multi-head self-attention with visual expert and multi-head cross-attention, while X_i' and X_{out_i} represent their respective output features with the residual connection. To implement cross-attention between them, we add learnable transformation matrices $W^i_{K_{\text{cross}}}, W^i_{V_{\text{cross}}} \in \mathbb{R}^{D_{\text{hi}}\times D_{\text{cross}}}$ to get $K^i_{\text{cross}} = X_{\text{hi}}W^i_{K_{\text{cross}}}$, $V^i_{\text{cross}} = X_{\text{hi}}W^i_{V_{\text{cross}}} \in \mathbb{R}^{L_{I_{\text{hi}}}\times D_{\text{cross}}}$, and $W^i_{Q_{\text{cross}}} \in \mathbb{R}^{D_{\text{dec}}\times D_{\text{cross}}}$ to get $Q^i_{\text{cross}} = X_i'W^i_{Q_{\text{cross}}} \in \mathbb{R}^{\left(L_{I_{\text{lo}}}+L_T\right)\times D_{\text{cross}}}$ in every decoder layer. With the residual connection in Eq. 2, the cross-attention with high-resolution images can be perceived as a complement to the features of low-resolution images, thereby effectively utilizing the previous pre-trained model in low resolution.

其中 MSA 和 MCA 分别表示带视觉专家的多头自注意力 (multi-head self-attention) 和多头交叉注意力 (multi-head cross-attention)，而 X_i' 和 X_{out_i} 表示它们各自带残差连接的输出特征。为了实现它们之间的交叉注意力，我们在每个解码器层中添加可学习的变换矩阵 $W^i_{K_{\text{cross}}}, W^i_{V_{\text{cross}}} \in \mathbb{R}^{D_{\text{hi}}\times D_{\text{cross}}}$ 得到 $K^i_{\text{cross}} = X_{\text{hi}}W^i_{K_{\text{cross}}}$ ， $V^i_{\text{cross}} = X_{\text{hi}}W^i_{V_{\text{cross}}} \in \mathbb{R}^{L_{I_{\text{hi}}}\times D_{\text{cross}}}$ ，以及 $W^i_{Q_{\text{cross}}} \in \mathbb{R}^{D_{\text{dec}}\times D_{\text{cross}}}$ 得到 $Q^i_{\text{cross}} = X_i'W^i_{Q_{\text{cross}}} \in \mathbb{R}^{\left(L_{I_{\text{lo}}}+L_T\right)\times D_{\text{cross}}}$ 。结合公式 2 中的残差连接，带高分辨率图像的交叉注意力可以视为对低分辨率图像特征的补充，从而有效利用先前预训练的低分辨率模型。

Computational complexity. Let the number of attention head be H_{cross} and H_{dec} in cross-attention and self-attention, and the dimension of each head be $d_{\text{cross}} = D_{\text{cross}}/H_{\text{cross}}$ and $d_{\text{dec}} = D_{\text{dec}}/H_{\text{dec}}$. If using our high-resolution cross-module, the computational complexity of attention is

计算复杂度。设交叉注意力和自注意力中的注意力头数分别为 H_{cross} 和 H_{dec} ，每个头的维度分别为 $d_{\text{cross}} = D_{\text{cross}}/H_{\text{cross}}$ 和 $d_{\text{dec}} = D_{\text{dec}}/H_{\text{dec}}$ 。若使用我们的高分辨率交叉模块，注意力的计算复杂度为

$$\begin{aligned} \mathrm{T}_{\text{improved}} = \mathbf{O}\big(&(L_{I_{\text{lo}}} + L_T)\, L_{I_{\text{hi}}} H_{\text{cross}}\, d_{\text{cross}} \\ &+ (L_{I_{\text{lo}}} + L_T)^2 H_{\text{dec}} d_{\text{dec}}\Big) . \end{aligned} \tag{3}$$

Note that d_{cross} and H_{cross} can be flexibly adjusted according to computational budget and model performance. If not utilizing the high-resolution cross-module and directly substituting low-resolution images with high-resolution ones, the computational complexity would be

注意，d_{cross} 和 H_{cross} 可根据计算预算和模型性能灵活调整。如果不使用高分辨率交叉模块，直接用高分辨率图像替代低分辨率图像，计算复杂度将为

$$\mathrm{T}_{\text{original}} = \mathbf{O}\left((L_{I_{\text{hi}}} + L_T)^2 H_{\text{dec}} d_{\text{dec}}\right) . \tag{4}$$

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 14×14 -pixel patches, thus $L_{I_{\text{hi}}} = 6400, L_{I_{\text{lo}}} = 256$. 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 (refer to Appendix for detailed derivation), and reduces memory overhead at the same time.

在我们的实现中，$d_{\text{cross}} = 32, H_{\text{cross}} = 32$ ，我们继承了 CogVLM-17B 的 $d_{\text{dec}} = 128, H_{\text{dec}} = 32$ 。高分辨率和低分辨率编码器均采用 14×14 像素的图像分块，因此 $L_{I_{\text{hi}}} = 6400, L_{I_{\text{lo}}} = 256$ 。我们的方法至少带来了 $\frac{L_{I_{\text{hi}}} + L_T}{L_{I_{\text{lo}}} + L_T} = \frac{6400 + L_T}{256 + L_T} \times$ 的加速，这是一个严格的下界 (详见附录推导)，同时减少了内存开销。

2.3. Pre-training

2.3. 预训练

To enhance the model's ability to comprehend high-resolution images and adapt it for GUI application scenarios, we focus our pre-training efforts on the following aspects: the capability to recognize texts of various sizes, orientations, and fonts in high-resolution images, the grounding ability of text and objects in the image, and a specialized understanding capability for GUI imagery such as web page. We divide our pre-train data into three parts based on the aforementioned aspects, with samples in the Appendix. All the pre-training data are derived from publicly available datasets. The construction methods are detailed below.

为了增强模型理解高分辨率图像的能力并使其适应 GUI 应用场景，我们的预训练重点包括: 识别高分辨率图像中不同大小、方向和字体的文本能力，图像中文本与对象的定位能力，以及对网页等 GUI 图像的专门理解能力。我们根据上述方面将预训练数据分为三部分，附录中有样本。所有预训练数据均来自公开数据集，构建方法详述如下。

Text recognition. Our data includes (1) Synthetic renderings with text from language pre-training dataset (80M). This is similar to the Synthetic Document Generator in Kim et al. [15], with text of varying font, size, color

and orientation, and diverse image background from LAION-2B [32]. (2) Optical Character Recognition (OCR) of natural images (18M). We collect natural images from COYO [6] and LAION-2B [32] and employ Paddle-OCR [13] to extract the texts and their bounding boxes, and filter out images with no text boxes. (3) Academic documents (9M). We follow Nougat [5] to construct image-text pairs including text, formula and tables from the source code (LaTeX) release on arXiv. For (1)(3), we apply the same data augmentation as Nougat. For (2), we additionally employed more aggressive rotation and flipping data augmentation techniques.

文本识别。我们的数据包括:(1) 来自语言预训练数据集 (8000 万) 的合成渲染文本。类似于 Kim 等 [15] 中的合成文档生成器，文本字体、大小、颜色和方向多样，背景图像来自 LAION-2B[32]。(2) 自然图像的光学字符识别 (OCR)(1800 万)。我们从 COYO[6] 和 LAION-2B[32] 收集自然图像，使用 Paddle-OCR[13] 提取文本及其边界框，过滤掉无文本框的图像。(3) 学术文档 (900 万)。我们遵循 Nougat[5] 构建图文对，包括来自 arXiv 源码 (LaTeX) 的文本、公式和表格。对 (1)(3) 采用与 Nougat 相同的数据增强，对 (2) 则额外采用更激进的旋转和翻转增强技术。

Visual grounding. It is imperative for GUI agents to possess the capability to accurately comprehend and locate diverse elements within images. We follow CogVLM [38] to use a constructed visual grounding dataset of 40M images with image-caption pairs sampled from LAION-115M [18], which associate entities in the caption with bounding boxes to indicate their positions. The format of the bounding box is $[[x_0, y_0, x_1, y_1]]$, where (x_0, y_0) and (x_1, y_1) represent the coordinates of upper-left and lower-right corners which are normalized to $[000, 999]$. If multiple objects are indicated by a single noun phrase, their boxes are separated by semicolons in double square brackets.

视觉定位。GUI 代理必须具备准确理解和定位图像中多样元素的能力。我们遵循 CogVLM[38]，使用构建的视觉定位数据集，包含从 LAION-115M[18] 采样的 40M 张图像及其图文对，图文对将标题中的实体与边界框关联以指示其位置。边界框格式为 $[[x_0, y_0, x_1, y_1]]$ ，其中 (x_0, y_0) 和 (x_1, y_1) 表示左上角和右下角坐标，均归一化至 $[000, 999]$ 。若单一名词短语指代多个对象，其边界框用双中括号内的分号分隔。

GUI imagery. Our approach innovatively addresses the scarcity and limited relevance of GUI images in datasets like LAION and COYO, which predominantly feature natural images. GUI images, with their distinct elements such as input fields, hyperlinks, icons, and unique layout characteristics, require specialized handling. To boost the model's capability in interpreting GUI imagery, we have conceptualized two pioneering GUI grounding tasks: (1) GUI Referring Expression Generation (REG) - where the model is tasked with generating HTML code for DOM (Document Object Model) elements based on a specified area in a screenshot, and (2) GUI Referring Expression Comprehension (REC) - which involves creating bounding boxes for given DOM elements. To facilitate robust training in GUI

GUI 图像。我们的方法创新性地解决了 LAION 和 COYO 等数据集中 GUI 图像稀缺且相关性有限的问题，这些数据集主要包含自然图像。GUI 图像具有输入框、超链接、图标及独特布局等特征，需专门处理。为提升模型对 GUI 图像的理解能力，我们设计了两项开创性的 GUI 定位任务:(1) GUI 指代表达生成 (REG)——模型根据截图中指定区域生成 DOM(文档对象模型) 元素的 HTML 代码；(2) GUI 指代表达理解 (REC)——为给定 DOM 元素创建边界框。为促进 GUI 的稳健训练

grounding, we have constructed the CCS400K (Common Crawl Screenshot 400K) dataset. This extensive dataset is formed by extracting URLs from the latest Common Crawl data, followed by capturing 400,000 web page screenshots. Alongside these screenshots, we compile all visible DOM elements and their corresponding rendered

boxes using Playwright [1] , supplementing the dataset with 140 million REC and REG question-answer pairs. This rich dataset ensures comprehensive training and understanding of GUI elements. To mitigate the risk of overfitting, we employ a diverse range of screen resolutions for rendering, selected randomly from a list of commonly used resolutions across various devices. Additionally, to prevent the HTML code from becoming overly extensive and unwieldy, we perform necessary data cleaning by omitting redundant attributes in the DOM elements, following the method outlined in [16].

为实现界面元素定位，我们构建了 CCS400K(Common Crawl Screenshot 400K) 数据集。该大型数据集通过提取最新 Common Crawl 数据中的 URL，随后截取 40 万个网页截图而成。除截图外，我们还利用 Playwright [1] 收集所有可见的 DOM 元素及其对应的渲染框，并补充了 1.4 亿对 REC 和 REG 问答对。该丰富数据集确保了对 GUI 元素的全面训练与理解。为降低过拟合风险，我们采用多种屏幕分辨率进行渲染，分辨率从各类设备常用列表中随机选取。此外，为防止 HTML 代码过于庞大难以处理，我们按照文献 [16] 的方法对 DOM 元素中冗余属性进行了必要的数据清理。

We also incorporate publicly available text-image datasets including LAION-2B and COYO-700M (after removing the broken URLs, NSFW images, and images with noisy captions and political bias) during pre-training.

我们还在预训练阶段整合了公开可用的文本-图像数据集，包括 LAION-2B 和 COYO-700M(剔除失效 URL、不适宜内容 (NSFW) 图片、带有噪声标题及政治偏见的图片后)。

We pre-train our CogAgent model for a total of 60,000 iterations with a batch size of 4,608 and a learning rate of 2e-5. We freeze all parameters except the newly added high-resolution cross-module for the first 20,000 steps, resulting in a total number of $646\text{M}\,(3.5\%)$ trainable parameters, then additionally unfreeze the visual expert in CogVLM for the next 40,000 steps. We warm up with curriculum learning by first training on easier text recognition (synthetic renderings and OCR on natural images) and image captioning, then sequentially incorporating harder text recognition (academic document), grounding data and web page data, as we observed that it leads to faster convergence and more stable training in our preliminary experiments.

我们对 CogAgent 模型进行了总计 60,000 次迭代的预训练，批量大小为 4,608，学习率为 2e-5。前 20,000 步冻结除新加入的高分辨率跨模块外的所有参数，训练参数总数为 $646\text{M}\,(3.5\%)$ ，随后在接下来的 40,000 步中解冻 CogVLM 中的视觉专家模块。我们采用课程学习进行预热，先训练较简单的文本识别 (合成渲染和自然图像 OCR) 及图像描述任务，再依次加入更复杂的文本识别 (学术文档)、定位数据和网页数据。初步实验表明，此策略有助于加快收敛速度并提升训练稳定性。

2.4. Multi-task Fine-tuning and Alignment

2.4. 多任务微调与对齐

To enhance our model's performance for diverse tasks and ensure it aligns with free-form human instructions in the GUI setting, we further fine-tune our model on a broad range of tasks. We manually collected over two thousand screenshots from mobile phones and computers, each annotated with screen elements, potential tasks, and methods of operation in the question-answering format by human annotators (details illustrated in the Appendix). We also utilize Mind2Web [10] and AITW [31], datasets focusing on web and Android behaviors which comprise tasks, sequences of actions and corresponding screenshots, and convert them into a natural language question-and-answer format using GPT-4. Besides, we incorporate multiple publicly available visual question-answering (VQA) datasets

encompassing a variety of tasks into our alignment dataset. We unfreeze all model parameters during this stage and train for 10k iterations with a batch size of 1024 and a learning rate of 2e-5.

为提升模型在多样任务上的表现并确保其在 GUI 环境中能遵循自由形式的人类指令，我们进一步对模型进行了广泛任务的微调。我们人工收集了两千多张手机和电脑截图，每张截图均由人工标注屏幕元素、潜在任务及操作方法，采用问答格式 (详见附录)。此外，我们利用了专注于网页和安卓行为的 Mind2Web [10] 和 AITW [31] 数据集，这些数据集包含任务、操作序列及对应截图，并通过 GPT-4 转换为自然语言问答格式。我们还整合了多个公开的视觉问答 (VQA) 数据集，涵盖多种任务，构建对齐数据集。此阶段解冻所有模型参数，训练迭代次数为 10k，批量大小为 1024，学习率为 2e-5。

---

[1] https://playwright.dev

[1] https://playwright.dev

Method	General VQA		Text-rich VQA					
	VQAv2	OKVQA	OCRVQA	TextVQA	STVQA	ChartQA	InfoVQA	DocVQA
task-specific fine-tuning models								
Pix2Struct [16]	-	-	-	-	-	58.6	40.0	76.6
BLIP-2 [18]	82.2	59.3	72.7	-	-	-	-	-
PALI-X-55B [8]	86.0	66.1	75.0	71.4	79.9	70.9	49.2	80.0
CogVLMusk-specific [38]	84.7	64.7	74.5	69.7	-	-	-	-
generalist models								
UReader [40]	-	57.6	-	-	-	59.3	42.2	65.4
Qwen-VL [2]	79.5	58.6	75.7	63.8	-	65.7	-	65.1
Qwen-VL-chat [2]	78.2	56.6	70.5	61.5	-	66.3	-	62.6
Llava-1.5 [20]	80.0	-	-	61.5	-	-	-	-
Fuyu-8B [3]	74.2	60.6	-	-	-	-	-	-
CogVLM $_{\text{generalist}}$ [38]	83.4	58.9	74.1	68.1	-	-	-	-
CogAgent (Ours)	83.7	61.2	75.0	76.1	80.5	68.4	44.5	81.6

方法	通用视觉问答		富文本视觉问答					
	VQAv2	OKVQA	OCR 视觉问答	TextVQA	STVQA	图表问答	信息视觉问答	文档视觉问答
特定任务微调模型								
Pix2Struct [16]	-	-	-	-	-	58.6	40.0	76.6
BLIP-2 [18]	82.2	59.3	72.7	-	-	-	-	-
PALI-X-55B [8]	86.0	66.1	75.0	71.4	79.9	70.9	49.2	80.0
CogVLMusk-specific [38]	84.7	64.7	74.5	69.7	-	-	-	-
通用模型								
UReader [40]	-	57.6	-	-	-	59.3	42.2	65.4
Qwen-VL [2]	79.5	58.6	75.7	63.8	-	65.7	-	65.1
Qwen-VL-chat [2]	78.2	56.6	70.5	61.5	-	66.3	-	62.6
Llava-1.5 [20]	80.0	-	-	61.5	-	-	-	-
Fuyu-8B [3]	74.2	60.6	-	-	-	-	-	-
CogVLM $_{\text{generalist}}$ [38]	83.4	58.9	74.1	68.1	-	-	-	-
CogAgent(本工作)	83.7	61.2	75.0	76.1	80.5	68.4	44.5	81.6

Table 1. Performance on Visual Question Answering benchmarks. Bold text indicates the best score among the generalist category, and underlined text represents the best score across both generalist and task-specific categories.

表 1. 视觉问答基准测试的性能表现。加粗文本表示通用类别中的最佳得分，下划线文本表示通用类别和特定任务类别中的最佳得分。

3. Experiments

3. 实验

To evaluate the foundational capabilities and GUI-related performance of our model, we conduct extensive experiments on a broad range of datasets. First, we conduct evaluations on eight VQA benchmarks, as well as MM-Vet [41] and POPE [19], which validate our model's enhanced ability in visual understanding, especially on those that are reliant on text recognition. Then we evaluate our model on Mind2Web and AITW datasets, as the representative of two major GUI scenarios - computers and smartphones.

为了评估我们模型的基础能力及图形用户界面 (GUI) 相关性能，我们在广泛的数据集上进行了大量实验。首先，我们在八个视觉问答 (VQA) 基准测试集以及 MM-Vet [41] 和 POPE [19] 上进行了评估，这些验证了我们模型在视觉理解方面的增强能力，尤其是在依赖文本识别的任务上。随后，我们在 Mind2Web 和 AITW 数据集上对模型进行了评估，这两个数据集分别代表了计算机和智能手机这两大 GUI 场景。

3.1. Foundational Visual Understanding

3.1. 基础视觉理解

We first extensively evaluate CogAgent's foundational visual understanding capability across eight VQA benchmarks, covering a wide range of visual scenes. The benchmarks can be divided into two categories: general VQA, including VQAv2 [1] and OK-VQA [23], and text-rich VQA, including TextVQA [34], OCR-VQA [27], ST-VQA [4], DocVQA [25], InfoVQA [26] and ChartQA [24]. The latter category emphasizes the understanding of visually-situated text, including documents, charts, photographs containing text, etc. To demonstrate the model's versatility and robustness across tasks, our model is fine-tuned collectively on all datasets simultaneously, yielding a single generalist model which is then evaluated across all datasets.

我们首先在八个视觉问答 (VQA) 基准上广泛评估 CogAgent 的基础视觉理解能力，涵盖了各种视觉场景。基准测试可分为两类: 通用视觉问答，包括 VQAv2 [1] 和 OK-VQA [23]，以及富文本视觉问答，包括 TextVQA [34]、OCR-VQA [27]、ST-VQA [4]、DocVQA [25]、InfoVQA [26] 和 ChartQA [24]。后一类侧重于对视觉中嵌入文本的理解，包括文档、图表、含文本的照片等。为了展示模型在各任务中的多功能性和鲁棒性，我们的模型在所有数据集上同时进行微调，生成一个通用模型，随后在所有数据集上进行评估。

The results are presented in Tab. 1. For general VQA, CogAgent achieves state-of-the-art generalist results on both datasets. For text-rich VQA, CogAgent achieves state-of-the-art results on 5 out of 6 benchmarks, significantly surpassing generalist competitors (TextVQA+8.0, ChartQA+2.1, InfoVQA+2.3, DocVQA+16.2), even

outperforming the task-specific state-of-the-art models on TextVQA(+4.7), STVQA(+0.6) and DocVQA(+1.6). Notably, compared to the generalist results of CogVLM which

结果如表 1 所示。对于通用视觉问答 (VQA)，CogAgent 在两个数据集上均实现了最先进的通用模型表现。对于文本丰富的视觉问答，CogAgent 在 6 个基准测试中取得了 5 项最先进的成绩，显著超越了通用模型竞争者 (TextVQA 提升 8.0，ChartQA 提升 2.1，InfoVQA 提升 2.3，DocVQA 提升 16.2)，甚至在 TextVQA(提升 4.7)、STVQA(提升 0.6) 和 DocVQA(提升 1.6) 上超越了针对特定任务的最先进模型。值得注意的是，相较于 CogVLM 的通用模型结果，

CogAgent is initially based on, CogAgent demonstrates certain improvements on both general and Text-rich VQA tasks, suggesting the efficacy of our proposed model architecture and training methods.

CogAgent 最初基于，CogAgent 在通用和文本丰富的视觉问答 (VQA) 任务上均表现出一定的提升，表明我们所提出的模型架构和训练方法的有效性。

Furthermore, we conducted zero-shot tests of our model on the challenging MM-Vet [41] and POPE [19] datasets, both of which are instrumental in gauging the multi-modal capabilities and the generalization performance in complex tasks including conversation question-answering, detailed descriptions, complex reasoning tasks. MM-Vet is designed with six core tasks to assess multi-modal models' proficiency in handling intricate assignments, and POPE-adversarial models on their susceptibility to hallucinations. Our experimental results, as detailed in Table 2, showcase that our model significantly outperforms other existing models in both datasets. Notably, on the MM-Vet dataset, our model achieved a remarkable score of 52.8, surpassing the closest competitor, LLaVA-1.5, by a substantial margin (+16.5). On the POPE-adversarial evaluation, our model attained a score of 85.9, demonstrating superior handling of hallucinations compared to other models.

此外，我们在具有挑战性的 MM-Vet [41] 和 POPE [19] 数据集上对模型进行了零样本测试，这两个数据集对于评估多模态能力及在包括对话问答、详细描述和复杂推理任务等复杂任务中的泛化性能具有重要作用。MM-Vet 设计了六个核心任务，以评估多模态模型处理复杂任务的能力，POPE-对抗性则用于测试模型对幻觉现象的敏感性。我们的实验结果详见表 2，显示我们的模型在两个数据集上均显著优于现有其他模型。值得注意的是，在 MM-Vet 数据集上，我们的模型取得了 52.8 的优异成绩，较最接近的竞争者 LLaVA-1.5 高出 16.5 分。在 POPE-对抗性评估中，我们的模型获得了 85.9 的分数，表现出较其他模型更优的幻觉处理能力。

Method	LLM	MM-Vet	POPEadv
BLIP-2 [18]	Vicuna-13B	22.4	-
Otter [17]	MPT-7B	24.7	-
MiniGPT4 [44]	Vicuna-13B	24.4	70.4
InstructBLIP [9]	Vicuna-13B	25.6	77.3
LLaVA [21]	LLaMA2-7B	28.1	66.3
LLaMA-Adapter v2 [14]	LLaMA-7B	31.4	-
DreamLLM [11]	Vicuna-7B	35.9	76.5
LLaVA-1.5 [20]	Vicuna-13B	36.3	84.5
Emu [36]	LLaMA-13B	36.3	-
CogAgent (Ours)	Vicuna-7B	52.8	85.9

方法	大型语言模型 (LLM)	MM-Vet	POPEadv
BLIP-2 [18]	Vicuna-13B	22.4	-
Otter [17]	MPT-7B	24.7	-
MiniGPT4 [44]	Vicuna-13B	24.4	70.4
InstructBLIP [9]	Vicuna-13B	25.6	77.3
LLaVA [21]	LLaMA2-7B	28.1	66.3
LLaMA-Adapter v2 [14]	LLaMA-7B	31.4	-
DreamLLM [11]	Vicuna-7B	35.9	76.5
LLaVA-1.5 [20]	Vicuna-13B	36.3	84.5
Emu [36]	LLaMA-13B	36.3	-
CogAgent(本工作)	Vicuna-7B	52.8	85.9

Table 2. Evaluation of CogAgent on conversational style QA and hallucination assessment. Regarding the POPE dataset, we use its adversarial subset for this evaluation.

表 2. CogAgent 在对话风格问答和幻觉评估中的表现。关于 POPE 数据集，我们使用其对抗子集进行评估。

Method			cross-task cross-website cross-domain overall	
Representations of screen inputs: HTML				
GPT-3.5[29](htw-shot)	18.6	17.4	16.2	17.4
GPT-4[30] † (few-shot)	36.2	30.1	26.4	30.9
Flan-T5xL [10]	52.0	38.9	39.6	43.5
LLaMA2-7B[37]	52.7	47.1	50.3	50.1
LLaMA2-70B[37]	55.8	51.6	55.7	54.4
Representations of screen inputs: Image				
Qwen-VL[2]	12.6	10.1	8.0	10.2
CogVLM[38]	37.1	23.4	26.3	23.9
CogAgent (Ours)	62.3	54.0	59.4	58.2

方法			跨任务跨网站跨领域综合	
屏幕输入的表示:HTML				
GPT-3.5[29](少样本学习)	18.6	17.4	16.2	17.4
GPT-4[30] † (少样本学习)	36.2	30.1	26.4	30.9
Flan-T5xL [10]	52.0	38.9	39.6	43.5
LLaMA2-7B[37]	52.7	47.1	50.3	50.1
LLaMA2-70B[37]	55.8	51.6	55.7	54.4
屏幕输入的表示: 图像				
Qwen-VL[2]	12.6	10.1	8.0	10.2
CogVLM[38]	37.1	23.4	26.3	23.9
CogAgent(本方法)	62.3	54.0	59.4	58.2

Table 3. Performance on Mind2Web. † denotes element selection from top-10 element candidates, others from top-50, following Deng et al. [10]. Results for GPT-3.5 and GPT-4 are from Deng et al. [10].

表 3. Mind2Web 上的性能表现。† 表示从前 10 个元素候选中选择元素，其他则从前 50 个中选择，遵循 Deng 等人 [10] 的方法。GPT-3.5 和 GPT-4 的结果来自 Deng 等人 [10]。

3.2.GUI Agent: Computer Interface

3.2.GUI 代理: 计算机界面

We evaluate CogAgent on Mind2Web, a dataset for web agents that includes over 2,000 open-ended tasks collected from 137 real-world websites across 31 domains. Given the task description, current webpage snapshot and previous actions as inputs, agents are expected to predict the subsequent action. We follow the setting of Deng et al. [10] in our experiments, and report step success rate (step SR) metric.

我们在 Mind2Web 上评估 CogAgent，该数据集针对网页代理，包含来自 31 个领域 137 个真实网站的 2000 多个开放式任务。给定任务描述、当前网页快照和先前操作作为输入，代理需预测下一步操作。我们的实验遵循 Deng 等人 [10] 的设置，并报告步骤成功率 (step SR) 指标。

Several language models were evaluated on this benchmark. For instance, AgentTuning [42] and MindAct [10] evaluated Llama2-70B and Flan-T5-XL in a fine-tuned setting, and GPT-3.5 and GPT-4 in a in-context learning setting. However, limited by the input modality of language models, these models could only use heavily cleansed HTML as the representation of screen inputs. To the best of our knowledge, no visually-based web agents have been experimented with on this benchmark.

多个语言模型在该基准上进行了评估。例如，AgentTuning [42] 和 MindAct [10] 在微调设置下评估了 Llama2-70B 和 Flan-T5-XL，在上下文学习设置下评估了 GPT-3.5 和 GPT-4。然而，受限于语言模型的输入模态，这些模型只能使用经过严格清洗的 HTML 作为屏幕输入的表示。据我们所知，尚无基于视觉的网页代理在该基准上进行过实验。

We fine-tune our model on the train set and evaluate on three out-of-domain subsets, i.e. cross-website, cross-domain, and cross-task. We additionally fine-tune LLaMA2-7B and LLaMA2-70B as the baseline of fine-tuned LLMs, and adopt the same HTML cleansing process as Deng et al. [10] to construct HTML input. The results are presented in Sec. 3.2. Compared to other methods, our approach achieved significant performance improvements across all three subsets, surpassing LLaMA2-70B, which is nearly $4\times$ the scale of CogAgent, by 11.6%,4.7%, and 6.6%, respectively. This reflects not only the capability of our model but also the advantages of employing a visual agent in computer GUI scenarios.

我们在训练集上微调模型，并在三个域外子集上进行评估，即跨网站、跨领域和跨任务。我们还微调了 LLaMA2-7B 和 LLaMA2-70B 作为微调大语言模型的基线，并采用与 Deng 等人 [10] 相同的 HTML 清洗流程构建 HTML 输入。结果见第 3.2 节。与其他方法相比，我们的方法在所有三个子集上均取得显著性能提升，分别超越了规模约为 CogAgent 近 $4\times$ 一半的 LLaMA2-70B 11.6%、4.7% 和 6.6%。这不仅体现了我们模型的能力，也反映了在计算机 GUI 场景中采用视觉代理的优势。

3.3.GUI Agent: Smartphone Interface

3.3.GUI 代理: 智能手机界面

To evaluate our model on diverse smartphone interfaces and tasks, we utilize Android in the Wild (AITW) dataset [31], a large-scale dataset for Android device agents. It comprises 715k operation episodes covering varying Android versions

为了评估我们模型在多样化智能手机界面和任务上的表现，我们使用了 Android in the Wild (AITW) 数据集 [31]，这是一个针对 Android 设备代理的大规模数据集，包含覆盖不同 Android 版本的 715k 操作序列。

Method	GoogleApp	Install	WebShop	General	Single	Overall
Representations of screen inputs: textual description (OCR+icon)						
GPT-3.5[29](few-shot)	10.47	4.38	8.42	5.93	9.39	7.72
LLaMA2-7B[37]†	30.99	35.18	19.92	28.56	27.35	28.40
Representations of screen inputs: image						
Auto-UIunified[43]	71.37	76.89	70.26	68.24	84.58	74.27
CogAgent (Ours)	74.95	78.86	71.73	65.38	93.49	76.88

方法	Google	应用安装	网店	通用	单一	总体
屏幕输入的表示: 文本描述 (OCR+ 图标)						
GPT-3.5[29](少样本学习)	10.47	4.38	8.42	5.93	9.39	7.72
LLaMA2-7B[37]†	30.99	35.18	19.92	28.56	27.35	28.40
屏幕输入的表示: 图像						
Auto-UIunified[43]	71.37	76.89	70.26	68.24	84.58	74.27
CogAgent(本方法)	74.95	78.86	71.73	65.38	93.49	76.88

Table 4. Performance on Android in the Wild (AITW) dataset. † represents models individually fine-tuned on each subset, while others are unified models across all subsets. The results of LLaMA2 and GPT-3.5 are from Zhan and Zhang [43].

表 4. 在 Android in the Wild (AITW) 数据集上的性能表现。† 表示在每个子集上单独微调的模型，其他则为跨所有子集的统一模型。LLaMA2 和 GPT-3.5 的结果来自 Zhan 和 Zhang [43]。

and device types. Each episode in the dataset consists of a goal described in natural language, followed by a sequence of actions and corresponding screenshots. The training target is to predict the next action based on the given goal, historical actions, and the screenshot. 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.

和设备类型。数据集中的每个情节包含一个用自然语言描述的目标，随后是一系列动作及对应的截图。训练目标是基于给定的目标、历史动作和截图预测下一步动作。对于每个动作，模型需预测准确的动作类型；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。

We conduct comparisons with two kinds of baselines: language models using the textual description of UI elements provided by the original dataset (text OCR and icon) as the representations of screen inputs [2] , and visual-language models using images as the screen inputs. We simultaneously fine-tuned on all the subsets, yielding a unified model which is then evaluated on all test sets. As the GoogleApps subset is 10-100 times larger than other subsets, we down-sample it to 10% to avoid data imbalance.

我们与两类基线方法进行了比较: 使用原始数据集提供的 UI 元素文本描述 (文本 OCR 和图标) 作为屏幕输入表示的语言模型 [2] ，以及使用图像作为屏幕输入的视觉语言模型。我们同时在所有子集上进行微调，得到一个统一模型，并在所有测试集上进行评估。由于 GoogleApps 子集比其他子集大 10 到 100 倍，我们将其下采样至 10% 以避免数据不平衡。

Results are shown in Tab. 4. CogAgent achieves state-of-the-art performance compared to all previous methods. In comparison to language-based methods, our model surpasses both baselines by a large margin. In comparison to the visual-language baseline, Auto-UI, our model achieves +2.61 improvements in the overall performance. In instances of inaccuracies, we randomly sample hundreds of cases, and upon reassessment, more than 40% are determined to be correct (refer to the appendix for details). This diversity arises from the multiple valid pathways inherent in mobile interactions, resulting in a range of responses.

结果如表 4 所示。与所有先前方法相比，CogAgent 实现了最先进的性能。与基于语言的方法相比，我们的模型大幅超越了两个基线。与视觉语言基线 Auto-UI 相比，我们的模型整体性能提升了 2.61。在不准确的实例中，我们随机抽取数百个案例，重新评估后发现超过 40% 被判定为正确 (详见附录)。这种多样性源于移动交互中存在多条有效路径，导致响应结果多样。

4. Ablation Study

4. 消融研究

To thoroughly comprehend the impact of various components in the methodology, we conduct ablation studies on two aspects, model architecture and training data. The evaluation is conducted on diverse datasets, including multiple VQA datasets (STVQA, OCRVQA, DocVQA) and a web agent dataset (Mind2Web). For VQA datasets, we fine-

为了深入理解方法中各组件的影响，我们在模型架构和训练数据两个方面进行了消融研究。评估涵盖多个数据集，包括多个视觉问答 (VQA) 数据集 (STVQA、OCRVQA、DocVQA) 和一个网页代理数据集 (Mind2Web)。对于 VQA 数据集，我们进行微调——

---

[2] Some Android applications may have View Hierarchy which is more friendly to language-based agents, but most of them tend to be poor quality or missing altogether. Therefore, as a large-scale, general-purpose dataset, AITW retained the results of OCR detection and icon detection as textual representations of screenshots.

[2] 一些 Android 应用可能具有对基于语言的代理更友好的视图层级结构，但大多数质量较差或完全缺失。因此，作为一个大规模通用数据集，AITW 保留了 OCR 检测和图标检测的结果，作为截图的文本表示。

tune the model on four datasets together for 3,000 iters with a batch size of 1,280, and report the generalist score; for Mind2Web, models are fine-tuned for 2,400 iters with a batch size of 128 and use top-10 setting. Training iterations are fewer than those in the main experiment, aiming to control variables within the constraints of a limited budget.

我们在四个数据集上共同微调模型 3,000 次迭代，批量大小为 1,280，并报告通用性能分数；对于 Mind2Web，模型微调 2,400 次迭代，批量大小为 128，采用 top-10 设置。训练迭代次数少于主实验，旨在有限预算内控制变量。

4.1. Model Architecture

4.1. 模型架构

To ascertain the efficacy of the high-resolution cross-module, we compare it with directly increasing the resolution using the original model architecture of CogVLM, and ablate on two perspectives: computational efficiency and model performance.

为验证高分辨率交叉模块的有效性，我们将其与直接使用 CogVLM 原始模型架构提升分辨率的方法进行比较，并从计算效率和模型性能两个角度进行消融。

To measure computational overhead, we use floating point operations (FLOPs) as the metric, and conduct experiments on multiple resolutions including 224, 490, 756, and 1120. From Fig. 3 we can see that, as the image resolution increases, models that use a high-resolution cross-module experience only a modest rise in computational overhead, demonstrating an almost linear relationship with the number of image patches. In contrast, using the original model structure, i.e. CogVLM, leads to a significant increase in the number of FLOPs at higher resolutions. Its FLOPs can even be more than 10 times higher compared to employing a cross-module at a resolution of 1120, which is the resolution utilized by CogAgent.

为衡量计算开销，我们以浮点运算次数 (FLOPs) 为指标，在 224、490、756 和 1120 多个分辨率上进行实验。从图 3 可见，随着图像分辨率提升，采用高分辨率交叉模块的模型计算开销仅略有增加，且与图像块数量几乎呈线性关系。相比之下，使用原始模型结构 (即 CogVLM) 在高分辨率下 FLOPs 显著增加。在 1120 分辨率下，其 FLOPs 甚至比采用交叉模块高出 10 倍以上，而 1120 正是 CogAgent 所用的分辨率。

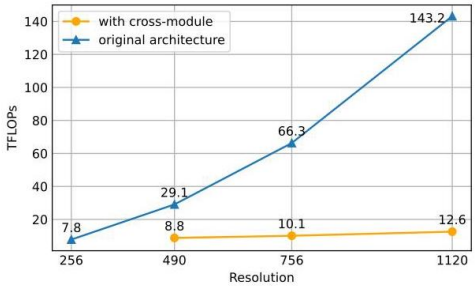

Figure 3. Comparison of FLOPs during forward propagation for different model architectures and resolutions.

图 3. 不同模型架构和分辨率下前向传播的 FLOPs 比较。

We further compare the model performance in Tab. 5, which indicates that models with high-resolution cross-module at the resolution of 756 require only 1/2 of the computational resources used by the original structure at the resolution of 490, while delivering significantly better performance. Additionally, the high-resolution cross-module allows for further increasing models' acceptable resolution within a limited computational budget, thereby yielding additional performance improvements.

我们在表 5 中进一步比较模型性能，结果显示，分辨率为 756 的高分辨率交叉模块模型所需计算资源仅为分辨率 490 的原始结构的一半，同时性能显著提升。此外，高分辨率交叉模块允许在有限计算预算内进一步提升模型可接受的分辨率，从而带来额外性能提升。

4.2. Pre-train Data

4.2. 预训练数据

We further conduct an ablation study on pre-training data, which is an integral part of training visual agents. Building upon the image-caption data commonly used in visual-

我们进一步对预训练数据进行了消融研究，预训练数据是训练视觉代理的重要组成部分。基于视觉领域常用的图像-字幕数据，

high-res module	base							
res	cross							
res	STVQA	OCRVQA	DocVQA	Mind2Web	training time/it (s)	TFLOPs		
✘	224	-	48.0	70.2	28.6	34.6	2.36	7.77
✘	490	-	68.1	74.5	57.6	40.7	6.43	29.14
✓	224	756	73.6	74.2	62.3	40.7	3.57	10.08
✓	224	1120	78.2	75.9	74.1	41.4	5.17	12.56

高分辨率模块	基础							
分辨率	交叉							
分辨率	STVQA	OCRVQA	DocVQA	Mind2Web	训练时间/次 (秒)	万亿次浮点运算 (TFLOPs)		
✘	224	-	48.0	70.2	28.6	34.6	2.36	7.77
✘	490	-	68.1	74.5	57.6	40.7	6.43	29.14
✓	224	756	73.6	74.2	62.3	40.7	3.57	10.08
✓	224	1120	78.2	75.9	74.1	41.4	5.17	12.56

Table 5. Ablation study on model architecture. Training time is evaluated on A800 with the batch size of 8 . Models are pre-trained with Caption+OCR data.

表 5. 模型架构消融研究。训练时间在 A800 上以批量大小为 8 进行评估。模型使用 Caption+OCR 数据进行预训练。

pre-train data base res cross res STVQA OCRVQ DocVQA Mind2Web						
Cap	490	-	68.1	74.5	57.6	38.6
Cap+OCR	490	-	72.5	75.0	59.8	40.7
Cap+OCR	224	1120	78.2	75.9	74.1	41.4
All	224	1120	79.4	75.6	76.4	54.2

预训练数据库 res cross res STVQA OCRVQ DocVQA Mind2Web						
标题	490	-	68.1	74.5	57.6	38.6
标题 + 光学字符识别 (OCR)	490	-	72.5	75.0	59.8	40.7
标题 + 光学字符识别 (OCR)	224	1120	78.2	75.9	74.1	41.4
全部	224	1120	79.4	75.6	76.4	54.2

Table 6. Ablation study on pre-train data with sequentially added image captioning, OCR and other pre-train data.

表 6. 预训练数据消融研究，依次添加图像描述、OCR 及其他预训练数据。

language training, we sequentially add OCR data (denoted as Cap+OCR), as well as GUI and grounding data (denoted as All). The results in Tab. 6 indicate that each part of data broadly contributes to enhanced performance. Notably, web and grounding data have a significant impact on the Mind2Web dataset, underscoring the importance of constructing domain-specific pre-train data in the training of GUI agents.

语言训练中，我们依次添加 OCR 数据 (记为 Cap+OCR)，以及 GUI 和定位数据 (记为 All)。表 6 的结果表明，各部分数据均对性能提升有广泛贡献。值得注意的是，网页和定位数据对 Mind2Web 数据集影响显著，强调了在 GUI 代理训练中构建领域特定预训练数据的重要性。

5. Conclusion

5. 结论

We introduce CogAgent, a VLM-based GUI agent with enhanced pre-train data construction and efficient architecture for high-resolution input. CogAgent achieves state-of-the-art performance on a wide range of VQA and GUI benchmarks, and will be open-sourced. CogAgent is an initial exploration of VLM-based GUI agent, and still has some shortcomings, e.g. imprecise output coordinates and incapability of processing multiple images, necessitating further research.

我们提出了 CogAgent，一种基于视觉语言模型 (VLM) 的 GUI 代理，具备增强的预训练数据构建和高效的高分辨率输入架构。CogAgent 在多种视觉问答 (VQA) 和 GUI 基准测试中实现了最先进的性能，并将开源。CogAgent 是基于 VLM 的 GUI 代理的初步探索，仍存在一些不足，如输出坐标不精确和无法处理多图像，需进一步研究。

Acknowledgments

致谢

This work is supported by Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grant 2022ZD0118600, Natural Science Foundation of China (NSFC) 62277033 and the New Cornerstone Science Foundation through the XPLORER PRIZE. It also got partial support from the National Engineering Laboratory for Cyberlearning and Intelligent Technology, Beijing Key Lab of Networked Multimedia, Daimler Greater China Ltd. -Tsinghua University Joint Institute for Sustainable Mobility, Tsinghua University(Department of Computer Science and Technology)-Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things (JCIIOT) and a research fund from Zhipu AI.

本工作得到中国科技部科技创新重大专项 (项目号 2022ZD0118600)、国家自然科学基金 (NSFC)62277033 及 XPLORER 奖项下新基石科学基金的支持。部分支持来自国家网络学习与智能技术工程实验室、北京网络多媒体重点实验室、戴姆勒大中华区-清华大学可持续出行联合研究院、清华大学 (计算机科学与技术系)-西门子有限公司、中国工业智能与物联网联合研究中心 (JCIIOT) 及智谱 AI 的研究基金。

References

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425- 2433, 2015. 3, 6, 11

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425- 2433, 2015. 3, 6, 11

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023. 3, 6, 7

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023. 3, 6, 7

[3] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Tagirlar. Introducing our multimodal models, 2023. 6

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sagnak Tagirlar. Introducing our multimodal models, 2023. 6

[4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar,

and Dimos-thenis Karatzas. Scene text visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4291-4301, 2019. 3, 6, 12

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimos-thenis Karatzas. Scene text visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4291-4301, 2019. 3, 6, 12

[5] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418, 2023. 5

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. arXiv preprint arXiv:2308.13418, 2023. 5

[6] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/ kakaobrain/coyo-dataset, 2022. 5

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/ kakaobrain/coyo-dataset, 2022. 5

[7] Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. arXiv preprint arXiv:2310.05915, 2023. 1

Baian Chen, Chang Shu, Ehsan Shareghi, Nigel Collier, Karthik Narasimhan, and Shunyu Yao. Fireact: Toward language agent fine-tuning. arXiv preprint arXiv:2310.05915, 2023. 1

[8] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565, 2023. 3, 6

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. arXiv preprint arXiv:2305.18565, 2023. 3, 6

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 6

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 6

[10] Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. arXiv preprint arXiv:2306.06070, 2023. 3, 5, 7, 12

邓翔，顾宇，郑博远，陈世杰，塞缪尔·史蒂文斯，王博时，孙欢，苏宇。Mind2web: 迈向通用网络智能体。arXiv 预印本 arXiv:2306.06070，2023。3, 5, 7, 12

[11] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. arXiv preprint arXiv:2309.11499, 2023. 6

董润培，韩春蕊，彭元，齐泽坤，葛铮，杨金荣，赵亮，孙剑剑，周宏宇，魏浩然，等。Dreamllm: 协同多模态理解与创作。arXiv 预印本 arXiv:2309.11499，2023。6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-vain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 3

阿列克谢·多索维茨基，卢卡斯·拜耶，亚历山大·科列斯尼科夫，迪尔克·魏森博恩，翟晓华，托马斯·安特廷纳，莫斯塔法·德赫加尼，马蒂亚斯·明德勒，乔治·海戈尔德，西尔万·杰利，等。一张图像胜过 16x16 个词: 大规模图像识别的 Transformer 模型。arXiv 预印本 arXiv:2010.11929，2020。3

[13] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. arXiv preprint arXiv:2009.09941, 2020. 5

杜宇宁，李晨霞，郭若瑜，尹晓婷，刘薇薇，周军，白一凡，余子林，杨业华，党青青，等。Pp-ocr: 实用超轻量级 OCR 系统。arXiv 预印本 arXiv:2009.09941，2020。5

[14] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xi-angyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010, 2023. 6

高鹏，韩嘉明，张仁睿，林子怡，耿世杰，周奥军，张伟，卢攀，何聪辉，岳翔宇，等。Llama-adapter v2: 参数高效的视觉指令模型。arXiv 预印本 arXiv:2304.15010，2023。6

[15] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sang-doo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In European Conference on Computer Vision, pages 498-517. Springer, 2022. 5

金基旭，洪泰圭，任文彬，南正妍，朴珍英，任珍英，黄元硕，尹相斗，韩东允，朴承贤。无 OCR 文档理解 Transformer。欧洲计算机视觉会议论文集，页 498-517。施普林格，2022。5

[16] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandel-wal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In International Conference on Machine Learning, pages 18893-18912. PMLR, 2023. 4, 5, 6

肯顿·李，曼达尔·乔希，尤利娅·拉卢卡·图尔克，胡鹤翔，刘方宇，朱利安·马丁·艾森施洛斯，乌尔瓦希·坎德尔瓦尔，彼得·肖，张明伟，克里斯蒂娜·图塔诺娃。Pix2struct: 截图解析作为视觉语言理解的预训练。国际机器学习大会论文集，页 18893-18912。PMLR，2023。4, 5, 6

[17] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. arXiv preprint arXiv:2306.05425, 2023. 6

李博，张元涵，陈良宇，王景浩，蒲凡一，杨景康，李春元，刘子威。Mimic-it: 多模态上下文指令微调。arXiv 预印本 arXiv:2306.05425，2023。6

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023. 5, 6

李俊南，李东旭，西尔维奥·萨瓦雷塞，霍伊·史蒂文。Blip-2: 利用冻结图像编码器和大型语言模型引导语言-图像预训练。arXiv 预印本 arXiv:2301.12597，2023。5, 6

[19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355, 2023. 3, 6, 11

李一凡，杜一凡，周坤，王金鹏，赵新，文继荣。评估大型视觉语言模型中的对象幻觉。arXiv 预印本 arXiv:2305.10355，2023。3, 6, 11

[20] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023. 6

刘昊天，李春元，李宇恒，李永宰。通过视觉指令微调改进基线。arXiv 预印本 arXiv:2310.03744，2023。6

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023. 3, 6

刘昊天，李春元，吴庆阳，李永宰。视觉指令微调。arXiv 预印本 arXiv:2304.08485，2023。3, 6

[22] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shum-ing Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. Kosmos-2.5: A multimodal literate model. arXiv preprint arXiv:2309.11419, 2023. 3

吕腾超，黄玉潘，陈景业，崔磊，马书铭，常瑶瑶，黄少涵，王文辉，董力，罗伟尧，等。Kosmos-2.5: 多模态通识模型。arXiv 预印本 arXiv:2309.11419，2023。3

[23] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195-3204, 2019. 3, 6, 11

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, 和 Roozbeh Mottaghi. Ok-vqa: 一个需要外部知识的视觉问答基准。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，页码 3195-3204，2019 年。3, 6, 11

[24] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244, 2022. 3, 6, 12

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, 和 Enamul Hoque. Chartqa: 一个关于图表的视觉和逻辑推理问答基准。arXiv 预印本 arXiv:2203.10244，2022 年。3, 6, 12

[25] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2200-2209, 2021. 3, 6, 12

Minesh Mathew, Dimosthenis Karatzas, 和 CV Jawahar. Docvqa: 一个针对文档图像的视觉问答数据集。发表于 IEEE/CVF 冬季计算机视觉应用会议论文集，页码 2200-2209，2021 年。3, 6, 12

[26] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, 和 CV Jawahar. Infographicvqa。

In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697-1706, 2022. 3, 6, 12

发表于 IEEE/CVF 冬季计算机视觉应用会议论文集，页码 1697-1706，2022 年。3, 6, 12

[27] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In 2019 international conference on document analysis and recognition (ICDAR), pages 947- 952. IEEE, 2019. 6, 11

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, 和 Anirban Chakraborty. OCR-VQA: 通过读取图像中的文本进行视觉问答。发表于 2019 年国际文档分析与识别会议 (ICDAR)，页码 947-952。IEEE，2019 年。6, 11

[28] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. arXiv preprint arXiv:2112.09332, 2021. 1

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders 等. WebGPT: 基于浏览器辅助的人类反馈问答系统。arXiv 预印本 arXiv:2112.09332，2021 年。1

[29] OpenAI. Introducing chatgpt. 2022. 1, 7

OpenAI. ChatGPT 介绍。2022 年。1, 7

[30] OpenAI. Gpt-4 technical report, 2023. 7

OpenAI. GPT-4 技术报告，2023 年。7

[31] Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Android in the wild: A large-scale dataset for android device control. arXiv preprint arXiv:2307.10088, 2023. 1, 3, 5, 7, 13

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, 和 Timothy Lillicrap. Android in the wild: 一个用于安卓设备控制的大规模数据集。arXiv 预印本 arXiv:2307.10088，2023 年。1, 3, 5, 7, 13

[32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Worts-man, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems, 35:25278-25294, 2022. 1, 5

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman 等. LAION-5B: 一个用于训练下一代图文模型的开放大规模数据集。神经信息处理系统进展，35 卷:25278-25294，2022 年。1, 5

[33] Significant-Gravitas. Autogpt. https://github.com/ Significant-Gravitas/AutoGPT, 2023. 1

Significant-Gravitas. AutoGPT. https://github.com/Significant-Gravitas/AutoGPT，2023 年。1

[34] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317-8326, 2019. 3, 6, 11

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, 和 Marcus Rohrbach. 面向能阅读的视觉问答模型。发表于 IEEE/CVF 计算机视觉与模式识别会议论文集，页码 8317-8326，2019 年。3, 6, 11

[35] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023. 3, 4

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, 和 Yue Cao. EVA-CLIP: 大规模 CLIP 训练的改进技术。arXiv 预印本 arXiv:2303.15389，2023 年。3, 4

[36] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. arXiv preprint arXiv:2307.05222, 2023. 6

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, 和 Xinlong Wang. 多模态生成预训练。arXiv 预印本 arXiv:2307.05222，2023 年。6

[37] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 7

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale 等. LLaMA 2: 开放基础模型与微调聊天模型。arXiv 预印本 arXiv:2307.09288，2023 年。7

[38] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079, 2023. 1, 3, 5, 6, 7

王伟汉, 吕庆松, 于文萌, 洪文怡, 齐吉, 王岩, 纪俊辉, 杨卓毅, 赵磊, 宋希轩, 等. Cogvlm: 预训练语言模型的视觉专家. arXiv 预印本 arXiv:2311.03079, 2023. 1, 3, 5, 6, 7

[39] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems, 35:20744-20757, 2022. 1

姚顺宇, 陈霍华德, 杨约翰, 和卡尔蒂克·纳拉西姆汉. Webshop: 面向可扩展的真实世界网络交互的基于语言的智能体. 神经信息处理系统进展, 35:20744-20757, 2022. 1

[40] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang,

叶嘉博, 胡安文, 徐海洋, 叶庆浩, 闫明, 徐国海, 李晨亮, 田俊峰, 钱琦, 张吉,

et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. arXiv preprint arXiv:2310.05126, 2023. 6

等. Ureader: 通用无 OCR 视觉语境语言理解的多模态大语言模型. arXiv 预印本 arXiv:2310.05126, 2023. 6

[41] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. arXiv preprint arXiv:2308.02490, 2023. 3, 6, 11

余伟豪, 杨正元, 李林杰, 王建峰, 林凯文, 刘子成, 王新超, 和王丽娟. Mm-vet: 大型多模态模型综合能力评估. arXiv 预印本 arXiv:2308.02490, 2023. 3, 6, 11

[42] Aohan Zeng, Mingdao Liu, Rui Lu, Bowen Wang, Xiao Liu, Yuxiao Dong, and Jie Tang. Agenttuning: Enabling generalized agent abilities for llms. abs/2310.12823, 2023. 1, 7, 12

曾奥涵, 刘明道, 卢锐, 王博文, 刘晓, 董宇霄, 和唐杰. Agenttuning: 赋能大语言模型的通用智能体能力. abs/2310.12823, 2023. 1, 7, 12

[43] Zhuosheng Zhan and Aston Zhang. You only look at screens: Multimodal chain-of-action agents. abs/2309.11436, 2023. 7, 13

詹卓生和张阿斯顿. 你只需看屏幕: 多模态动作链智能体. abs/2309.11436, 2023. 7, 13

[44] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592, 2023. 6

朱德尧, 陈军, 沈晓倩, 李翔, 和穆罕默德·埃尔霍赛尼. Minigpt-4: 利用先进大语言模型增强视觉语言理解. arXiv 预印本 arXiv:2304.10592, 2023. 6