

OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

Tianbao Xie

The University of Hong Kong

10/11/2024



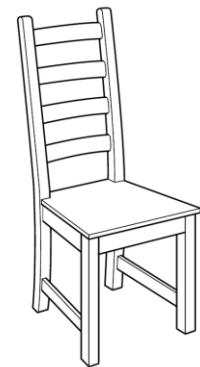
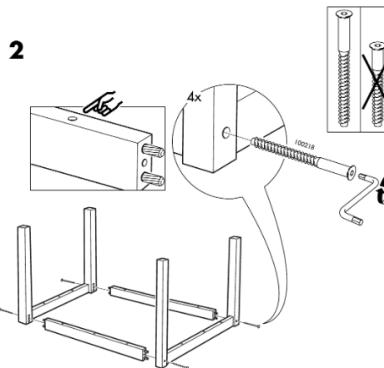
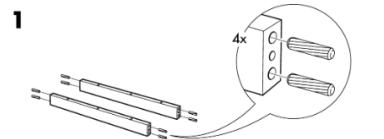
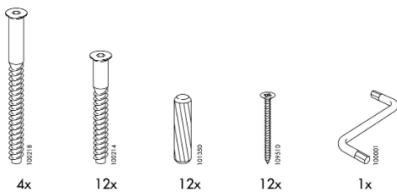
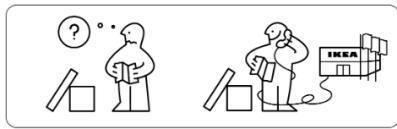
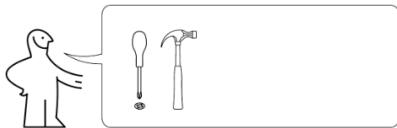
<https://xlang.ai>



香港大學自然語言處理實驗室

Natural Language Processing Group, The University of Hong Kong

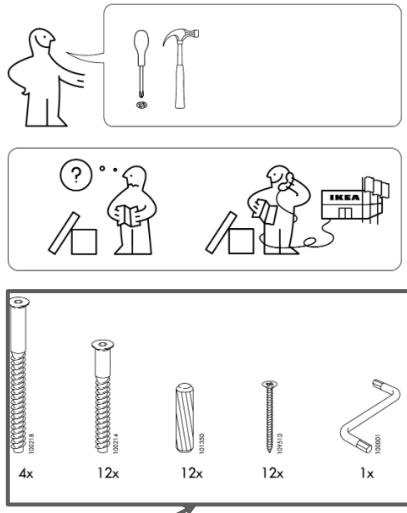
IKEA furniture assembly



Assembled chair

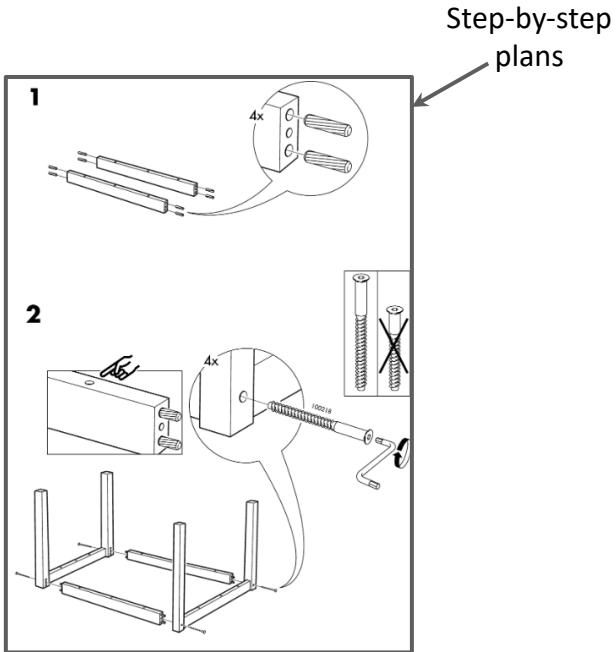
Assembly instructions

Planning with tools

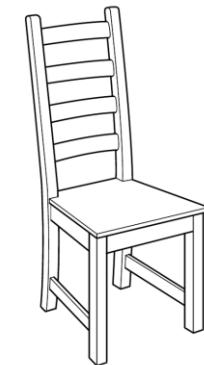


Tool set

Assembly instructions

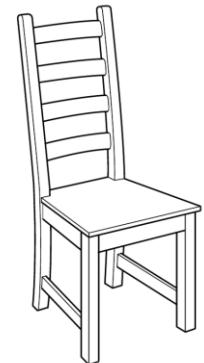
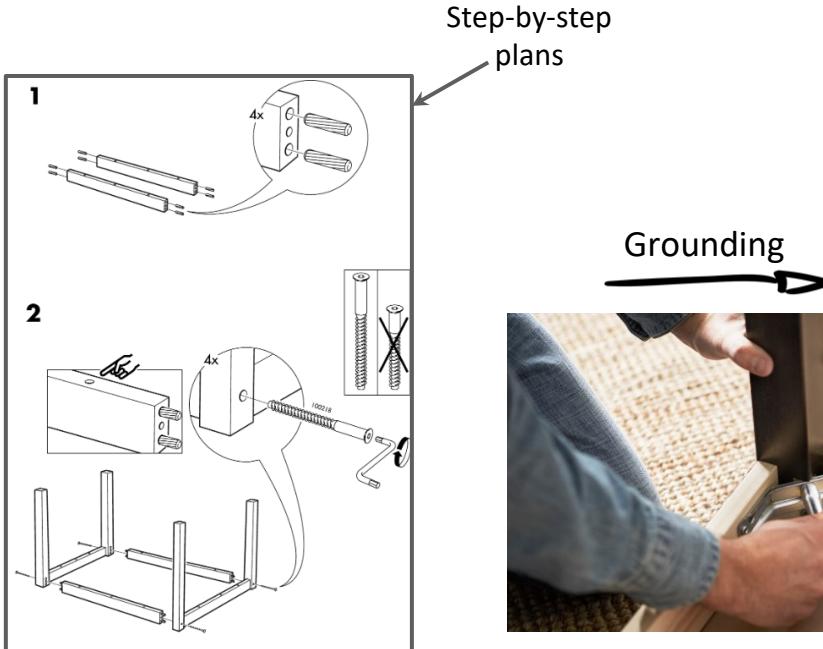
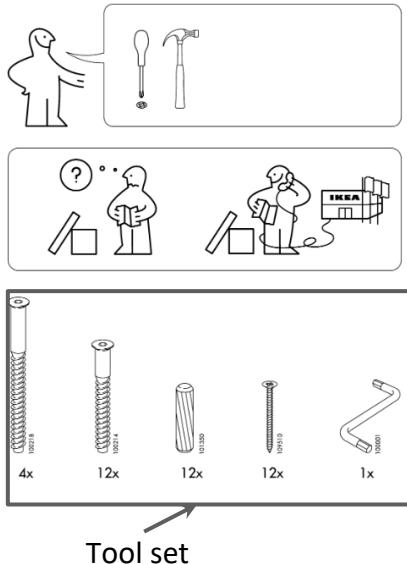


Step-by-step
plans



Assembled chair

Grounding plans into actions in physical world



Computer tasks in digital world

Task instruction: how do I change my Mac desktop background?



macOS environment

Customise the wallpaper on your Mac

You can change the picture that's displayed on your desktop. Choose from a variety of pictures or colours provided by Apple, or use your own images.

1. On your Mac, choose Apple menu > System Settings, then click Wallpaper in the sidebar. (You may need to scroll down.)
2. Select a wallpaper from one of the available categories:
 - *Add Photo / Add Folder or Album:* Controls to choose your own images.
 - *Dynamic Wallpapers:* These images brighten and darken, based on the time of day for your current location.
 - *Landscape, Cityscape, Underwater and Earth aerials:* These still images show dramatic views.



macOS w. new wallpaper!

• • •

Control instructions

Grounding plans into actions in digital world

Task instruction: how do I change my Mac desktop background?



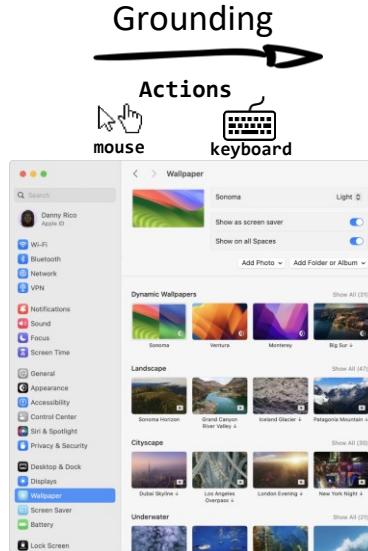
Customise the wallpaper on your Mac

You can change the picture that's displayed on your desktop. Choose from a variety of pictures or colours provided by Apple, or use your own images.

1. On your Mac, choose Apple menu > System Settings, then click Wallpaper in the sidebar. (You may need to scroll down.)
2. Select a wallpaper from one of the available categories:
 - Add Photo / Add Folder or Album: Controls to choose your own images.
 - Dynamic Wallpapers: These images brighten and darken, based on the time of day for your current location.
 - Landscape, Cityscape, Underwater and Earth aerials: These still images show dramatic views.

• • •

Control instructions



Can LLMs/VLMs be used for these tasks?

Yes and No!

You
how do I change my Mac desktop background?

ChatGPT
To change your Mac desktop background, follow these steps:

1. Click the Apple menu (apple) in the top-left corner of your screen and select "System Preferences."
2. Click "Desktop & Screen Saver."
3. In the "Desktop" tab, you'll see a list of folders on the left side. You can browse these folders to find images stored on your Mac.
4. Click on an image to preview it as your desktop background. You can also add your own photo by dragging an image from a Finder window directly into the preview area or by clicking the "+" button below the folders list to add a new folder.
5. Once you select an image, it will automatically set as your desktop background.

If you have multiple displays, you can select a different image for each screen by dragging the preferences window to the display you want to change and then selecting a new image.



You
can you help me assemble my IKEA chair?

ChatGPT
Certainly! While I can provide some general advice for assembling IKEA chairs, the specific steps can vary greatly depending on the model you have. Here's a general approach to assembling most IKEA chairs:

1.
Unpack and Organize: Open the package and carefully lay out all the parts on a clean, flat surface, such as the floor. This helps prevent pieces from getting lost and makes it easier to identify them as you follow the instructions.

2.

Check the Instructions: Locate the assembly instructions within the package. Take a moment to review these instructions thoroughly before starting, to familiarize yourself with the process and the parts involved.

3.

Can LLMs/VLMs be used for these tasks?

Yes and No!

ChatGPT is able to generate step-by-step plans

You
how do I change my Mac desktop background?

ChatGPT
To change your Mac desktop background, follow these steps:

1. Click the Apple menu (apple) in the top-left corner of your screen and select "System Preferences."
2. Click "Desktop & Screen Saver."
3. In the "Desktop" tab, you'll see a list of folders on the left side. You can browse these folders to find images stored on your Mac.
4. Click on an image to preview it as your desktop background. You can also add your own photo by dragging an image from a Finder window directly into the preview area or by clicking the "+" button below the folders list to add a new folder.
5. Once you select an image, it will automatically set as your desktop background.

If you have multiple displays, you can select a different image for each screen by dragging the preferences window to the display you want to change and then selecting a new image.



Task instruction:
how do I change my Mac desktop background?



ChatGPT

Customise the wallpaper on your Mac

You can change the picture that's displayed on your desktop. Choose from a variety of pictures or colours provided by Apple, or use your own images.

1. On your Mac, choose Apple menu (apple) > System Settings, then click Wallpaper (blue square) in the sidebar. (You may need to scroll down.)
2. Select a wallpaper from one of the available categories:
 - *Add Photo / Add Folder or Album:* Controls to choose your own images.
 - *Dynamic Wallpapers:* These images brighten and darken, based on the time of day for your current location.
 - *Landscape, Cityscape, Underwater and Earth aerials:* These still images show dramatic views.

...

Control instructions

Can LLMs/VLMs be used for these tasks?

Yes and No!

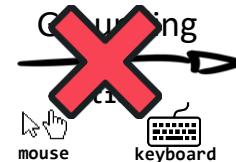
However, ChatGPT **cannot** execute tasks on your Mac by grounding plans into actions.

You
how do I change my Mac desktop background?

ChatGPT
To change your Mac desktop background, follow these steps:

1. Click the Apple menu (apple) in the top-left corner of your screen and select "System Preferences."
2. Click "Desktop & Screen Saver."
3. In the "Desktop" tab, you'll see a list of folders on the left side. You can browse these folders to find images stored on your Mac.
4. Click on an image to preview it as your desktop background. You can also add your own photo by dragging an image from a Finder window directly into the preview area or by clicking the "+" button below the folders list to add a new folder.
5. Once you select an image, it will automatically set as your desktop background.

If you have multiple displays, you can select a different image for each screen by dragging the preferences window to the display you want to change and then selecting a new image.



macOS w. new wallpaper!

Can LLMs/VLMs be used for these tasks?

Yes and No!

At times, ChatGPT also **cannot** generate step-by-step plans w/o interacting in the environment

You
can you help me assemble my IKEA chair?

ChatGPT
Certainly! While I can provide some general advice for assembling IKEA chairs, the specific steps can vary greatly depending on the model you have. Here's a general approach to assembling most IKEA chairs:

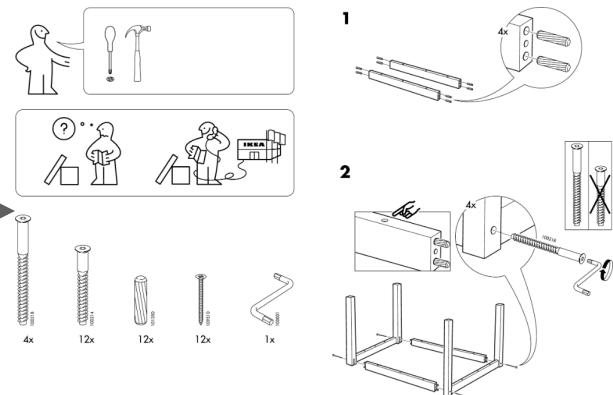
1.

Unpack and Organize: Open the package and carefully lay out all the parts on a clean, flat surface, such as the floor. This helps prevent pieces from getting lost and makes it easier to identify them as you follow the instructions.

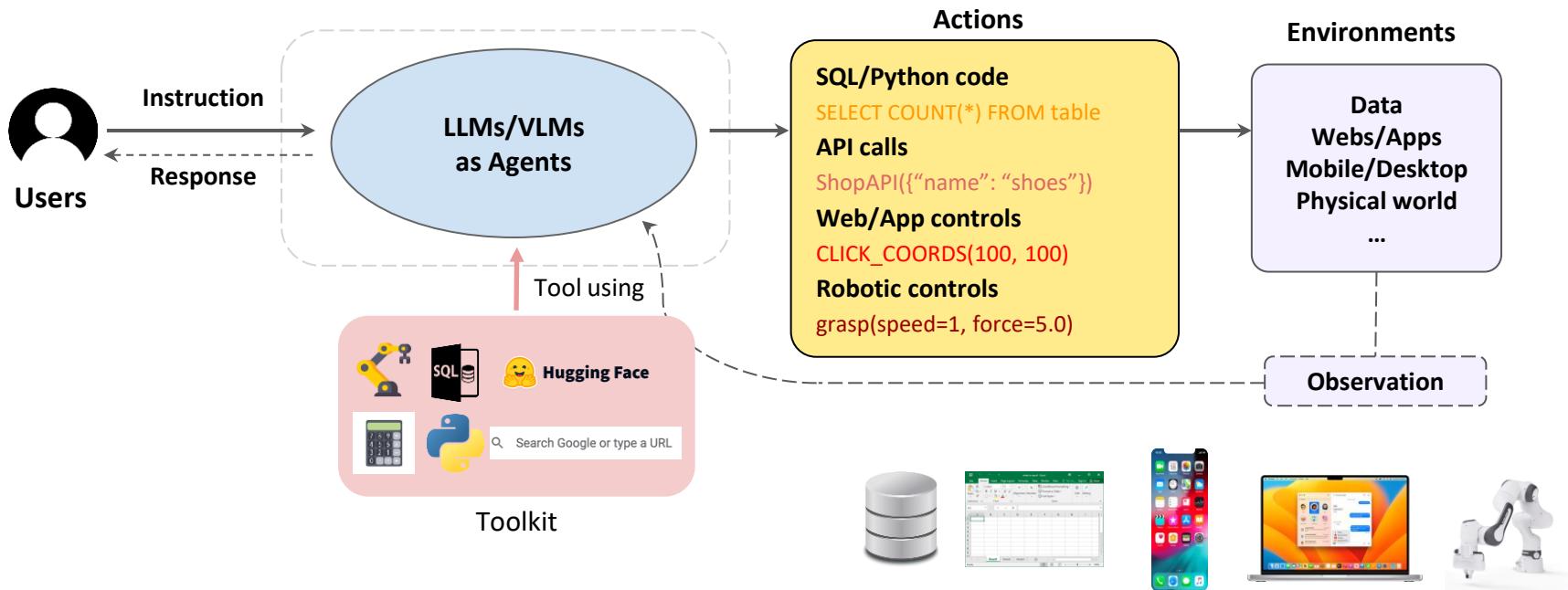
2.

Check the Instructions: Locate the assembly instructions within the package. Take a moment to review these instructions thoroughly before starting, to familiarize yourself with the process and the parts involved.

3.

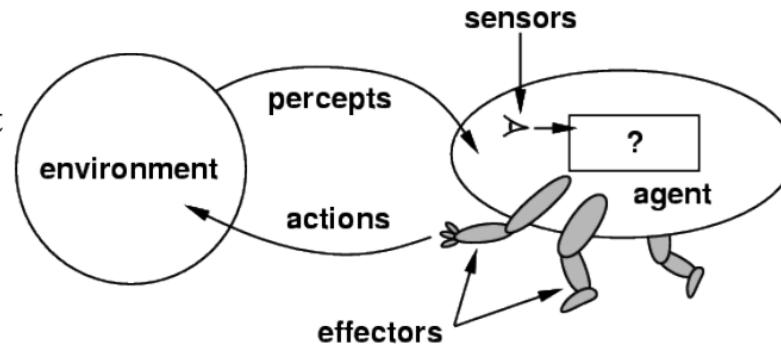


LLMs/VLMs as agents



Wait...what is an intelligent agent?

- Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.
- A discrete agent receives **percepts** one at a time, and maps this percept sequence to a sequence of discrete **actions**.
- Properties
 - Autonomous
 - Reactive to the environment
 - Pro-active (goal-directed)
 - Interacts with other agents via the environment



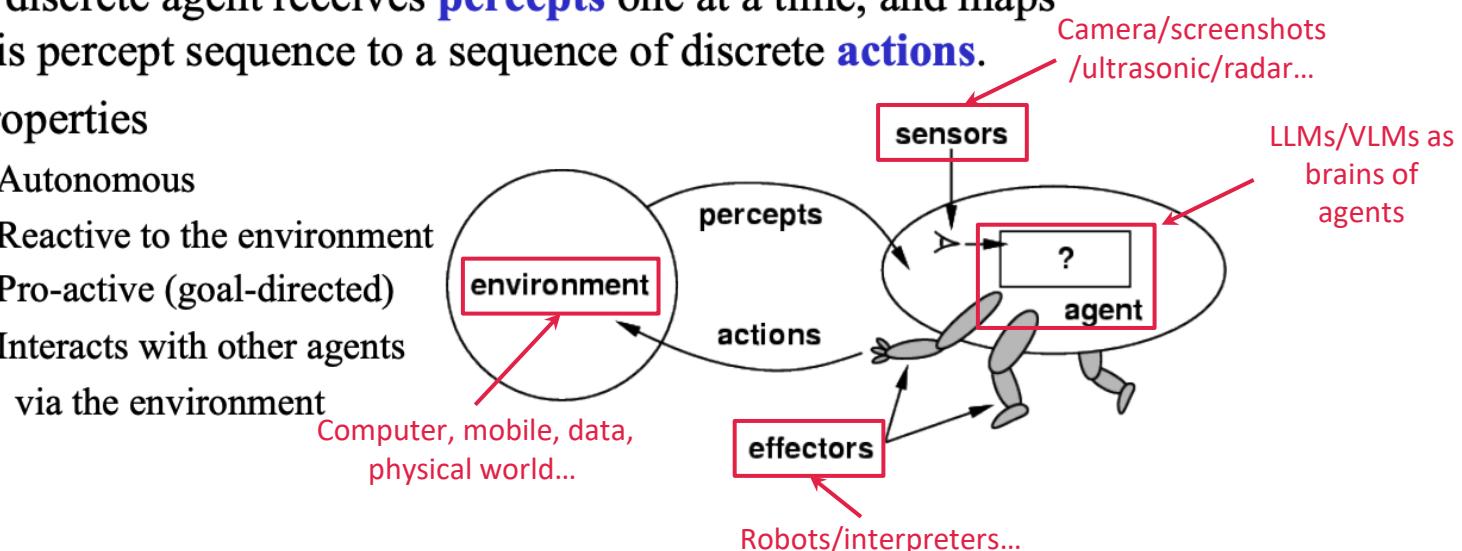
LLMs/VLMs as agents

- Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.
- A discrete agent receives **percepts** one at a time, and maps this percept sequence to a sequence of discrete **actions**.
- Properties

- Autonomous
- Reactive to the environment
- Pro-active (goal-directed)
- Interacts with other agents

via the environment

Computer, mobile, data,
physical world...



LLMs/VLMs as agents

- Definition: An **intelligent agent** acts via **sensors** and acts rationally using its **effectors**.
- A discrete agent receives **percepts** from the environment and maps this percept sequence to a sequence of **actions**.
- Properties
 - Autonomous
 - Reactive to the environment
 - Pro-active (goal-directed)
 - Interacts with other agents via the environment

Actions

SQL/Python code

`SELECT COUNT(*) FROM table`

API calls

`ShopAPI({"name": "shoes"})`

Web/App controls

`CLICK_COORDS(100, 100)`

Robotic controls

`grasp(speed=1, force=5.0)`

Environments

Data
Webs/Apps
Mobile/Desktop
Physical world
...

percepts

actions

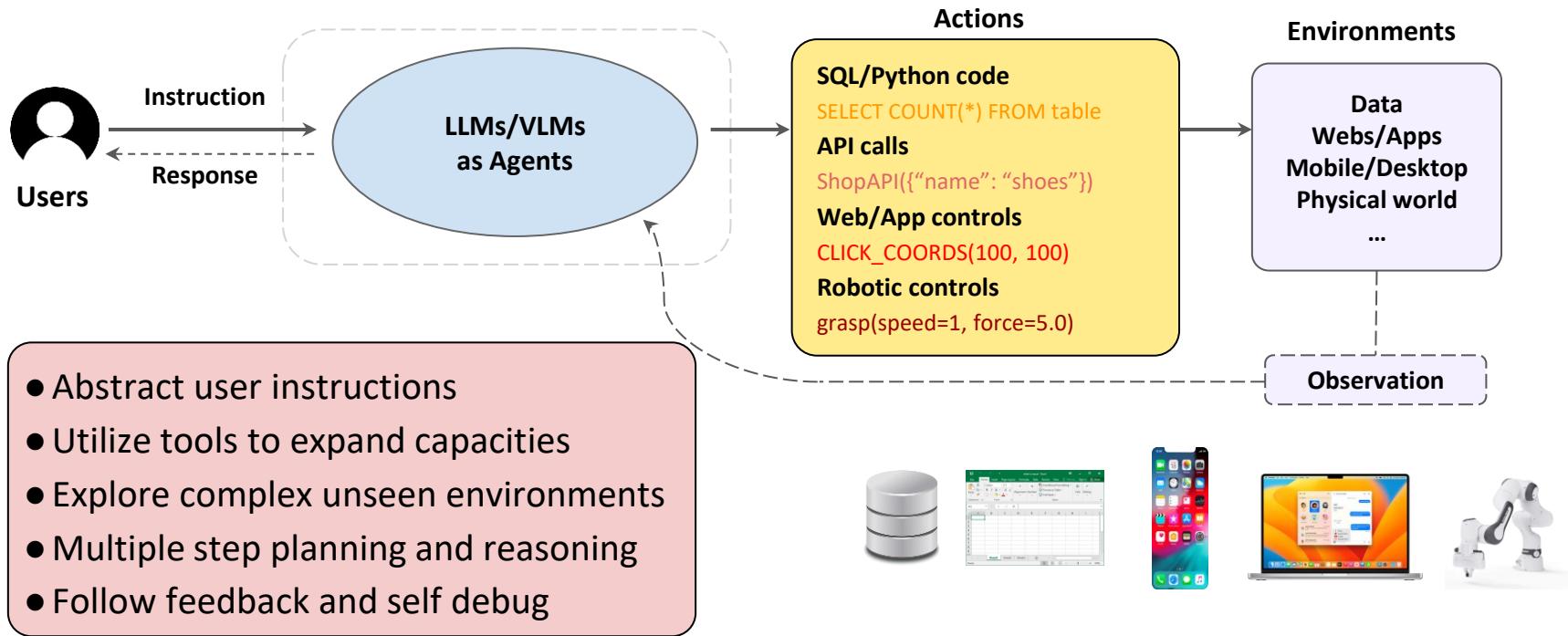
sensors

effectors

?

agent

LLM/VLM agent task

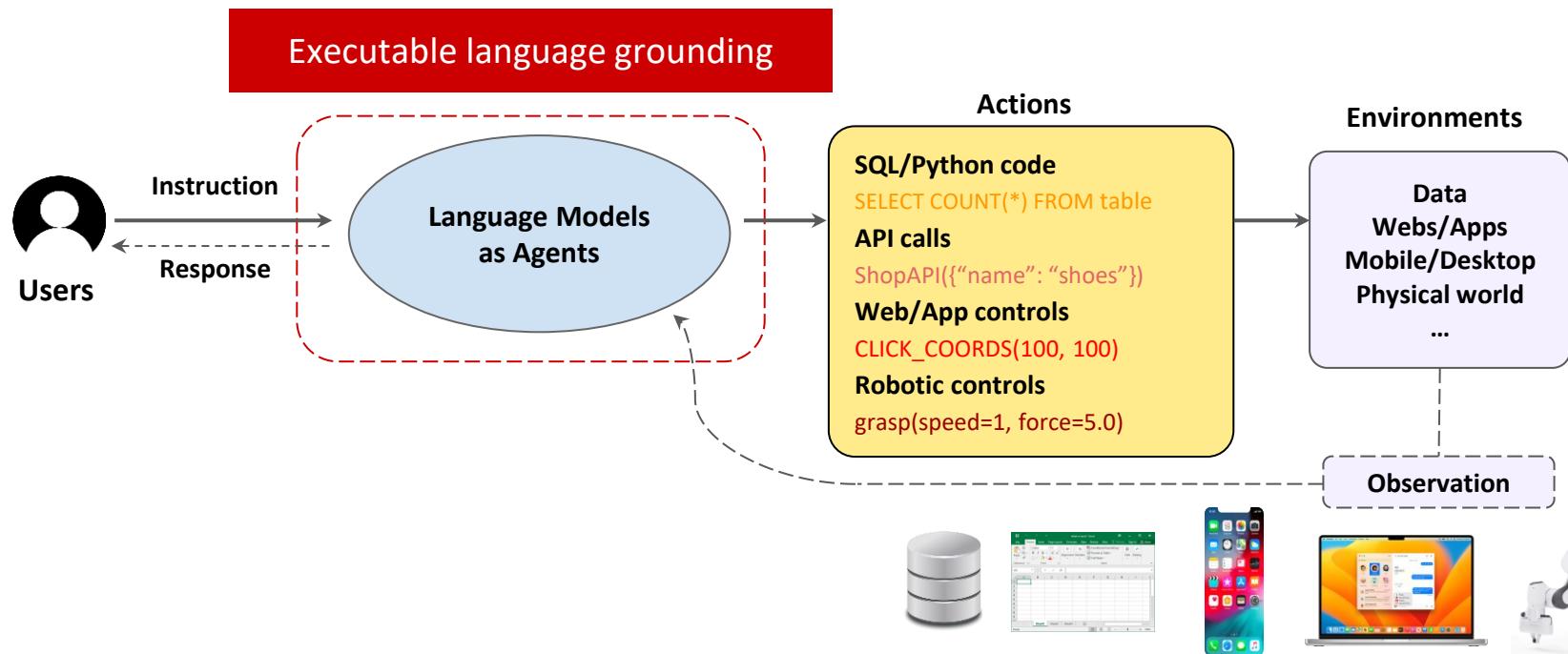




XLANG Lab executable language grounding

<https://xlang.ai>

Key research direction: transforming (“grounding”) language instructions into code or actions executable in real-world environments



Our recent work

Open source projects for advancing NLIs with language models as agents

- **Instructor:** adapt to various agent environments by simply providing instructions
- **Binder:** one of the first LLM + tool use studies
- **Lemur:** open state-of-art LLMs for language agents
- **OpenAgents:** an open platform for language agents in the wild
- **Text2Reward:** connect LLM agents to the physical world
- **OSWorld:** Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

Tianbao Xie¹, Danyang Zhang¹, Jixuan Chen¹, Xiaochuan Li¹,
Siheng Zhao¹, Ruisheng Cao¹, Toh Jing Hua¹, Zhoujun Cheng¹, Dongchan Shin¹, Fangyu Lei¹, Yitao Liu¹,
Yiheng Xu¹, Shuyan Zhou³, Silvio Savarese², Caiming Xiong², Victor Zhong⁴, Tao Yu¹

¹The University of Hong Kong, ²Salesforce Research, ³Carnegie Mellon University, ⁴University of Waterloo

 Paper

 Code

 Data

 Discord

 Data Viewer

Project website: <https://os-world.github.io>

Autonomous agents for computer tasks

Computer tasks often involve multiple apps and interfaces.

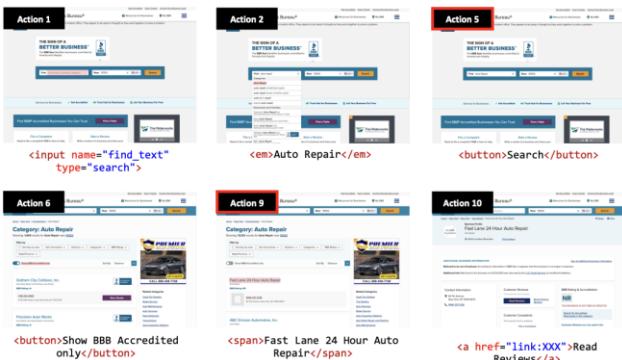
Task instruction 1: Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.



Major challenge: no real, scalable interactive environments

The absence of a real-world benchmark with a scalable interactive environment for multimodal agents hinders their task scope and agent scalability.

Mind2Web



Only demos *without* executable environment

- No execution based evaluation
- Cannot support interactive learning & real-world exploration

WebArena

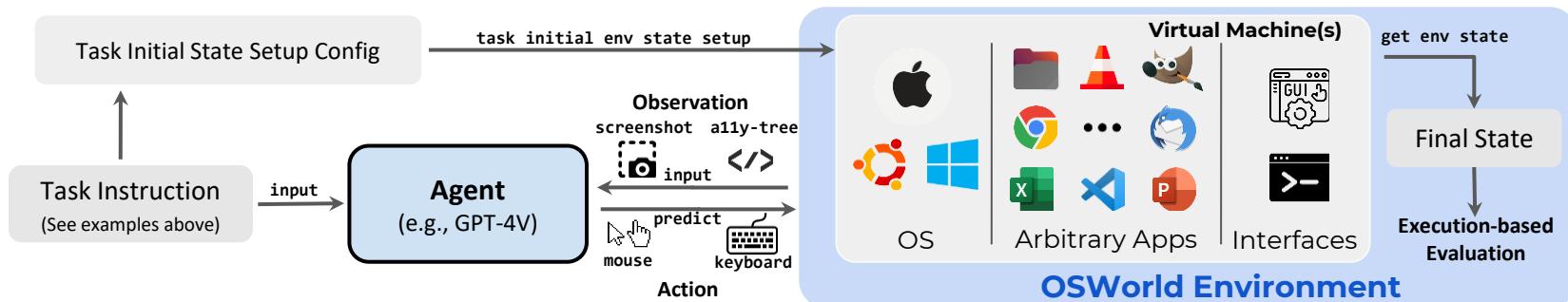


Environments limited to specific apps or domains

- Simplify agent's observation and action spaces
- Limit task scope, cannot support the evaluation of complex, real-world computer tasks

OSWorld: the first scalable, real computer environment

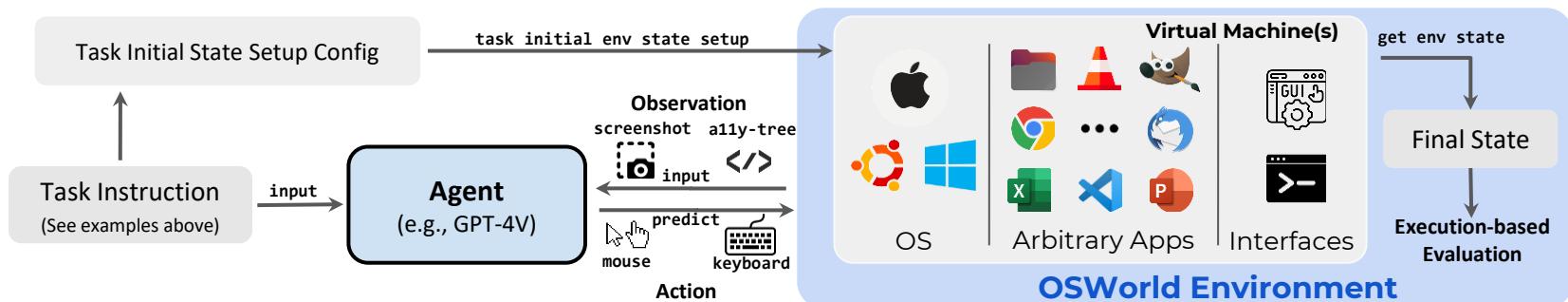
OSWorld can serve as a unified multimodal agent environment for evaluating open-ended computer tasks that involve arbitrary apps and interfaces across operating systems.



OSWorld agent task definition

An autonomous agent task can be formalized as a partially observable Markov decision process $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R})$

- State space \mathcal{S} (e.g., current Desktop environment)
- Observation space \mathcal{O} (e.g., task instruction, screenshot, a11y tree)
- Action space \mathcal{A} (e.g., clicking on the certain pixel of the screen `.click(300, 540, button='right')`)
- Transition function: $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$
- Reward function: $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$



OSWorld agent task definition

An autonomous agent task can be formalized as a partially observable Markov decision process $(\mathcal{S}, \mathcal{O}, \mathcal{A}, \mathcal{T}, \mathcal{R})$

- State space \mathcal{S} (e.g., current Desktop environment)
- Observation space \mathcal{O} (e.g., task instruction, screenshot, a11y tree)
- Action space \mathcal{A} (e.g., clicking on the certain pixel of the screen `.click(300, 540, button='right')`)
- Transition function: $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$
- Reward function: $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$

Given a computer task instruction:

- “Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.”

Task Instruction
(See examples above)

OSWorld agent task setup config

Each computer task in OSWorld has a task initial state setup and evaluation config file.

Task Initial State Setup Config



Task Instruction
(See examples above)

Task Config

```
{ "instruction": "Please update my bookkeeping sheet with  
the recent transactions from the provided folder, detailing  
my expenses over the past few days.",  
  "config": [{"type": "download",  
    "parameters": {"files": [  
      {"path": "/home/user/Desktop/my_bookkeeping.xlsx",  
       "url": "https://drive.google.com/uc?id=xxxx"},  
      {"path": "/home/user/Desktop/receipt_0.jpeg",  
       "url": "https://drive.google.com/uc?id=xxxx"},...]}],  
    {"type": "open",  
      "parameters": { "path":  
        "/home/user/Desktop/my_bookkeeping.xlsx"}]},  
    "evaluator": {"postconfig": [{"type": "activate_window",  
      "parameters": {"window_name": "my_bookkeeping.xlsx -  
LibreOffice Calc",...}},  
      "result": {"type": "vm_file",  
        "path": "/home/user/Desktop/my_bookkeeping.xlsx",  
        "dest": "my_bookkeeping.xlsx"},  
        "expected": {"type": "cloud_file",  
          "path": "https://drive.google.com/uc?id=xxx",  
          "dest": "my_bookkeeping_gold.xlsx"},  
          "func": "compare_table",  
          "options": {  
            "rules": [{  
              "type": "sheet_fuzzy",  
              "sheet_idx0": "RNSheet1",  
              "sheet_idx1": "ENSheet1",  
              "rules": [ {"range": ["A1:A8",... ]}]}]  
        }]
```

OSWorld agent task setup

The task initial state setup config is used to create a virtual machine instance, and initializes intermediate state for each computer task.

Task Config

```
{ "instruction": "Please update my bookkeeping sheet with the recent transactions from the provided folder, detailing my expenses over the past few days.",  
  "config": [{"type": "downloader"},  
    {"parameters": {"files": [  
      {"path": "/home/user/Desktop/my_bookkeeping.xlsx",  
       "url": "https://drive.google.com/uc?id=xxxxxx"},  
      {"path": "/home/user/Desktop/receipt_0.jpeg",  
       "url": "https://drive.google.com/uc?id=xxxxx"},...]}],  
    {"type": "open",  
     "parameters": [{"path":  
      "/home/user/Desktop/my_bookkeeping.xlsx"}]},  
    "evaluator": "postconfig": [{"type": "activate_window",  
      "parameters": {"window_name": "my_bookkeeping.xlsx - LibreOffice Calc",...}},  
      {"result": {"type": "vm_file",  
        "path": "/home/user/Desktop/my_bookkeeping.xlsx",  
        "dest": "my_bookkeeping.xlsx"},  
      "expected": {"type": "cloud_file",  
        "path": "https://drive.google.com/uc?id=xxx",  
        "dest": "my_bookkeeping_gold.xlsx"},  
      "func": "compare_table"},  
      "options": {  
        "rules": [{  
          "type": "sheet_fuzzy",  
          "sheet_idx": "RNSheet1",  
          "sheet_idxl": "ENSheet1",  
          "rules": [ {"range": ["A1:A8",... ]}]}]  
      }]}  
}
```

Task Initial State Setup Config

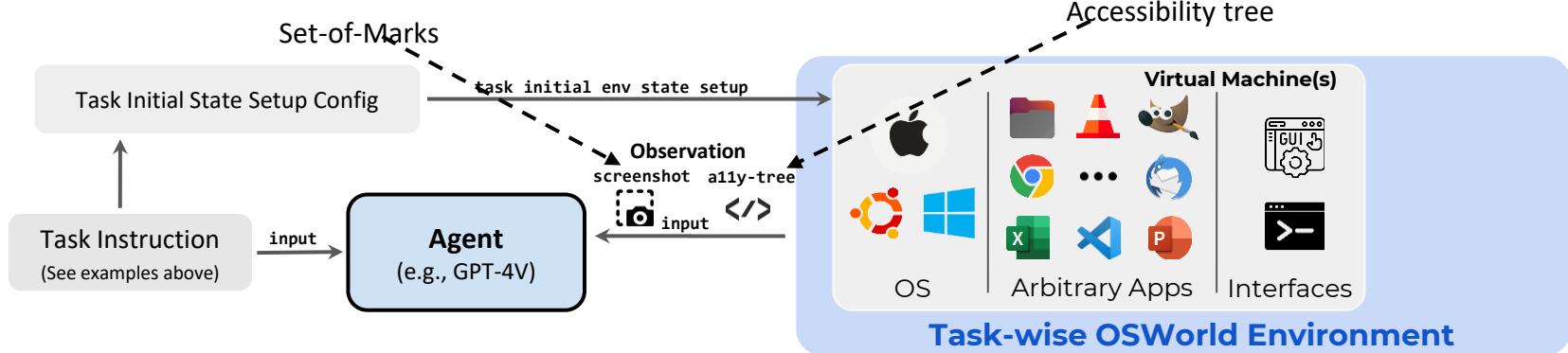
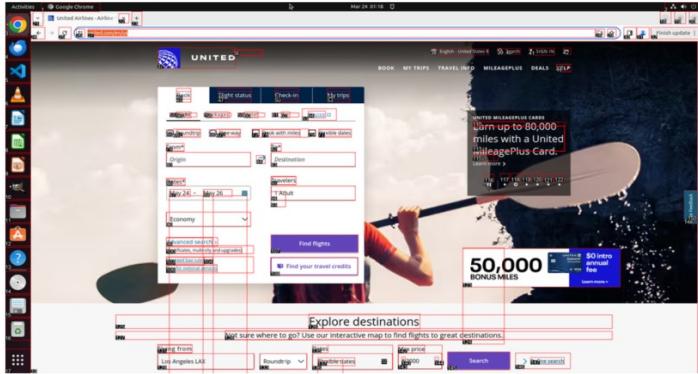
Task Instruction
(See examples above)

task initial env state setup



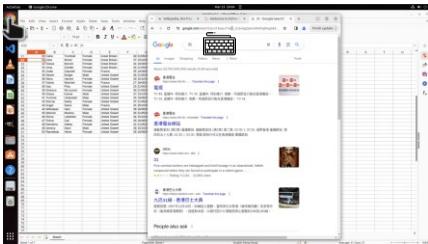
OSWorld agent task observation space

Given current observation $o_t \in \mathcal{O}$, task instruction, screenshot, a11y tree, or their combination...

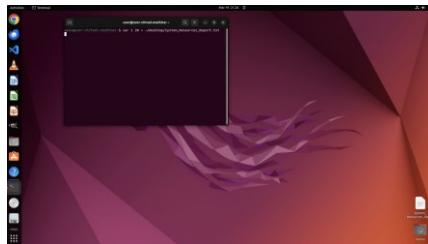


OSWorld agent task action space

An agent generates action $a_t \in \mathcal{A}$ which results in a new state $s_{t+1} \in \mathcal{S}$ and a new partial observation $o_{t+1} \in \mathcal{O}$



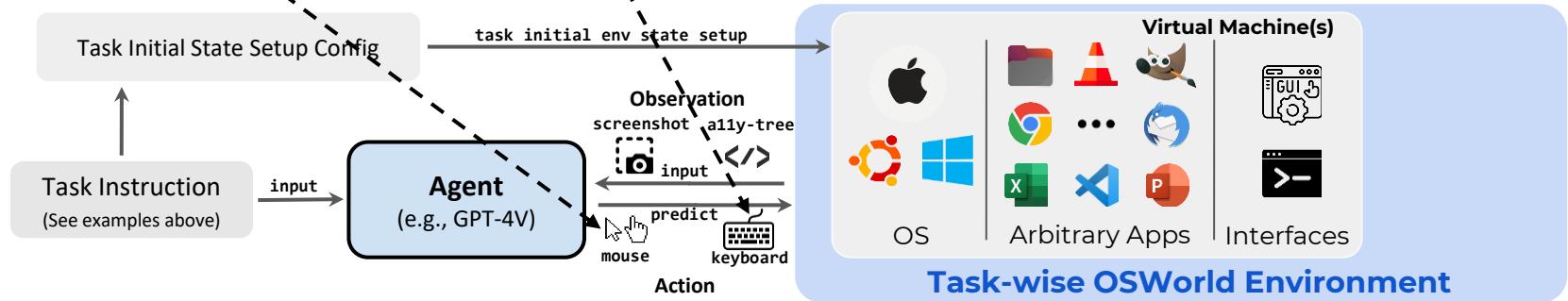
`pyautogui.click(chrome_x, chrome_y) ...`



`pyautogui.typewrite('sar 1 30 > ...', interval=0.5)`

Function	Description
<code>moveTo(x, y)</code>	Moves the mouse to the specified coordinates.
<code>click(x, y)</code>	Clicks at the specified coordinates.
<code>write('text')</code>	Types the specified text at the current cursor location.
<code>press('enter')</code>	Presses the Enter key.
<code>hotkey('ctrl', 'c')</code>	Performs the Ctrl+C hotkey combination (copy).
<code>scroll(200)</code>	Scrolls up by 200 units.
<code>scroll(-200)</code>	Scrolls down by 200 units.
<code>dragTo(x, y)</code>	Drags the mouse to the specified coordinates.
<code>keyDown('shift')</code>	Holds down the Shift key.
<code>keyUp('shift')</code>	Releases the Shift key.
<code>WAIT</code>	Agent decides it should wait.
<code>FAIL</code>	Agent decides the task is infeasible.
<code>DONE</code>	Agent decides the task is finished.

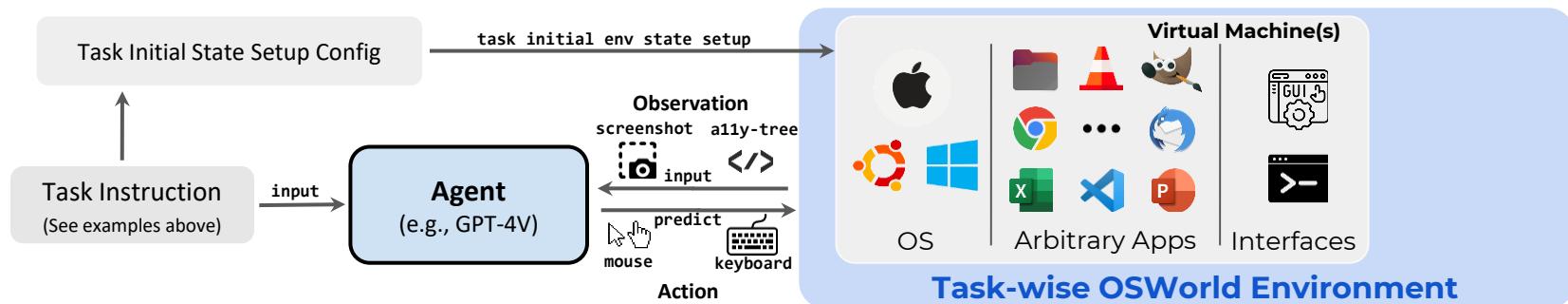
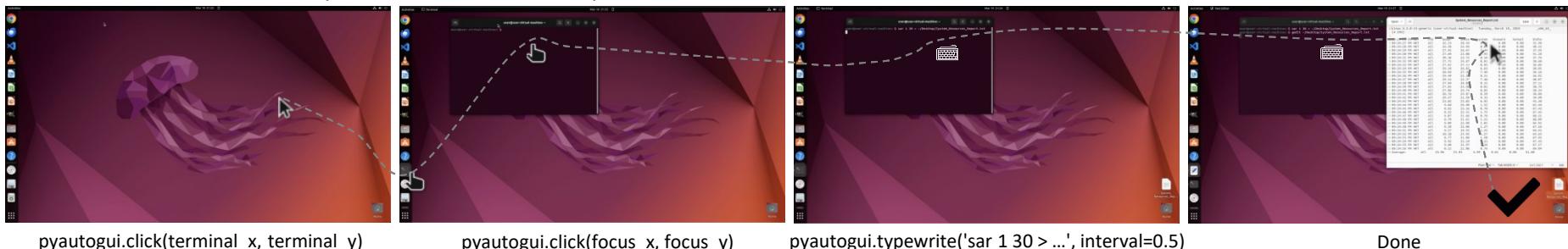
Some examples of the mouse and keyboard actions



OSWorld agent task interactive learning

The interaction loop between the agent and the environment repeats until an action that marks termination.

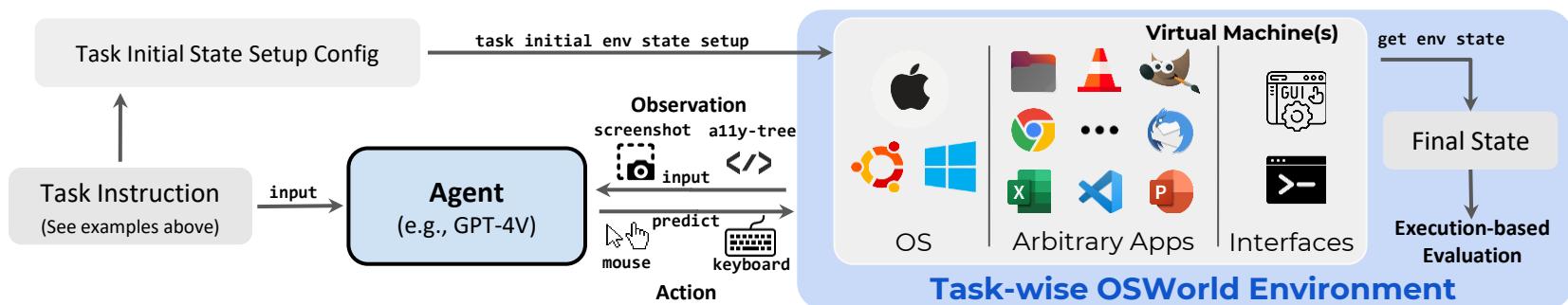
Task Instruction: monitor the system CPU for 30s and output the results



OSWorld agent task evaluation

In OSWorld, we implement an execution-based reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$

Initial State	Task Instruction	Evaluation Script (Simplified)
	<i>Can you help me clean up my computer by getting rid of all the tracking things that Amazon might have saved?</i>	<pre>cookie_data = get_cookie_data(env) rule = {"type": "domains", "domains": [".amazon.com"]} is_cookie_deleted(cookie_data, rule)</pre>
	<i>Rename "Sheet 1" to "LARS Resources". Then make a copy of it. Place the copy before "Sheet 2" and rename it by appending a suffix "(Backup)". ...</i>	<pre>result = get_file(env) expected = get_file(cloud) rules = [{"type": "sheet_name"}, {"type": "sheet_data", "sheet_idx0": 0, "sheet_idx1": 1}...] compare_table(result, expected, rules)</pre>



OSWorld benchmark dataset

369 real-world computer tasks that involve real web and desktop apps in open domains, OS file I/O, and multi-app workflows through both GUI and CLI. Each task example is carefully annotated with

- A real-world task instruction from real users
- An initial state setup config to simulate human work in progress
- A custom execution-based evaluation script

Table 3: Key statistics in OSWORLD. The “Supp. tasks” refers to the Windows-based tasks, that could only be used after activation due to copyright restrictions.

Statistic	Number
Total tasks (Ubuntu)	369 (100%)
- Multi-App Workflow	101 (27.4%)
- Single-App	268 (72.6%)
- Integrated	84 (22.8%)
- Infeasible	30 (8.1%)
Supp. tasks (Windows)	43
Initial States	302
Eval. Scripts	134

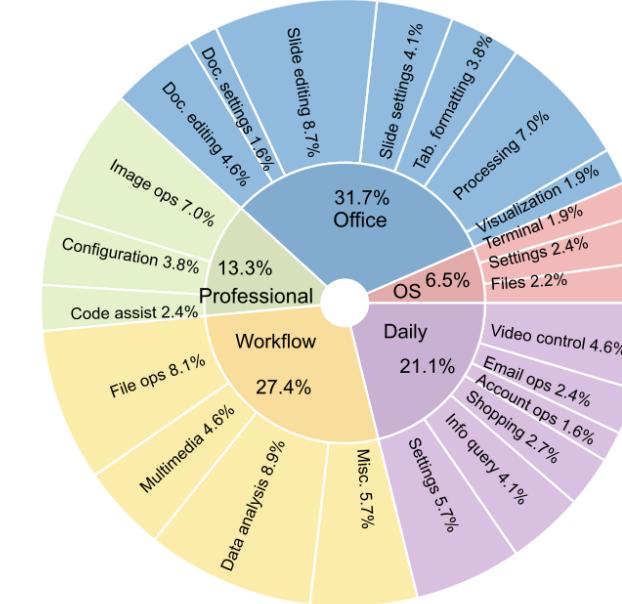
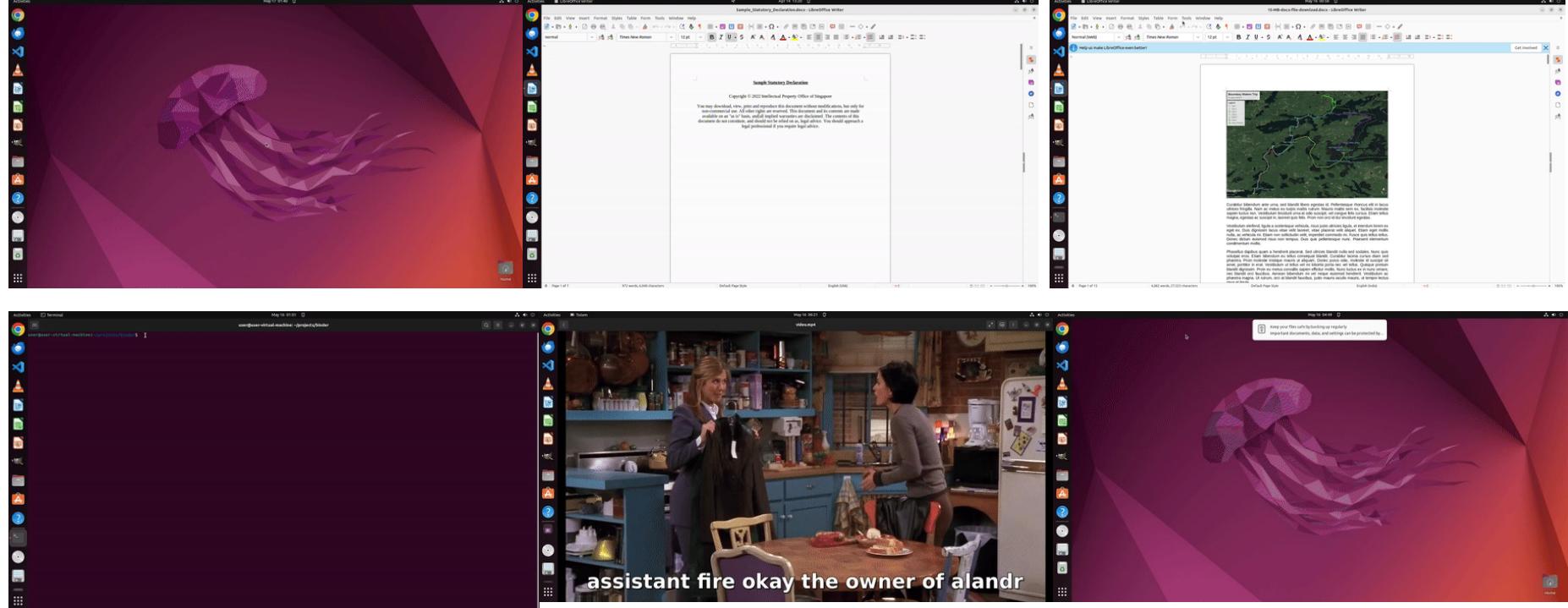


Figure 3: Distribution of task instructions in OSWORLD based on the app domains and operation types to showcase the content intuitively.



OSWorld benchmark dataset

	# Instances (# Templates)	Control. Exec. Env.?	Environment Scalability?	Multimodal Support?	Cross- App?	Intermediate Init. State?	# Exec.-based Eval. Func.
GAIA [35]	466	✗	-	✗	✗	✗	0
MIND2WEB [9]	2350	✗	-	✓	✗	✓	0
WEBLINX [33]	2337	✗	-	✓	✗	✓	0
PIXELHELP [27]	187	✗	-	✓	✗	✗	0
METAGUI [45]	1125	✗	-	✓	✗	✗	0
AITW [39]	30k	✗	-	✓	✗	✓	0
OMNIACT [21]	9802	✗	-	✓	✗	✓	0
AGENTBENCH [32]	1091	Multi-isolated	✗	✗	✗	✗	7
INTERCODE [54]	1350 (3)	Code	✗	✗	✗	✗	3
MINIWOB++ [30]	125	Web	✗	✓	✗	✗	125
WEBSHOP [55]	12k (1)	Web	✗	✓	✗	✗	1
WEBARENA [63]	812 (241)	Web	✗	✓	✗	✗	5
VWEBARENA [22]	910 (314)	Web	✗	✓	✗	✗	6
WORKARENA [10]	23k (29)	Web	✗	✓	✗	✓	7
WIKIHOW [58]	150 (16)	Mobile	✗	✓	✗	✗	16
ASSISTGUI [13]	100	✗	✗	✓	✗	✓	2
OSWORLD	369	Computer	✓	✓	✓	✓	134

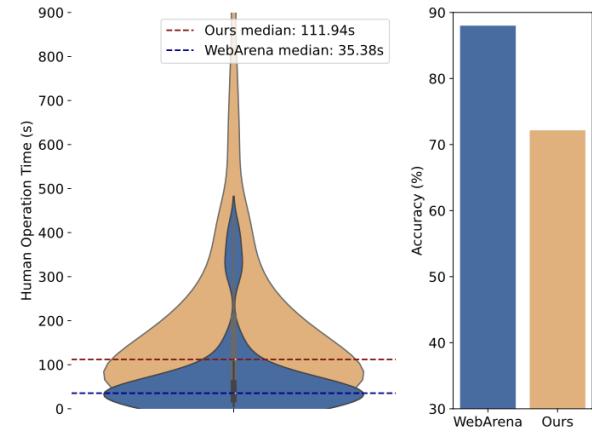


Figure 4: Human operation time and accuracy on OSWorld and WebArena.

Benchmarking LLM/VLM agent baselines

- LLMs and VLMs from Mixtral and CogAgent (open-source), and GPT4, Gemini-pro, and Claude-3 (closed-source) as agents.
- Prompt details (see right - much more complex prompting pipelines)
- Temperature of 1.0 and top-p of 0.9
- Providing the most recent 3 observations and actions as history context for each step.

You are an agent which follow my instruction and perform desktop computer
→ tasks as instructed.

You have good knowledge of computer and good internet connection and assume
→ your code will run on a computer for controlling the mouse and keyboard.
For each step, you will get an observation of an image, which is the
→ screenshot of the computer screen and you will predict the action of the
→ computer based on the image.

You are required to use `pyautogui` to perform the action grounded to the
→ observation, but DONOT use the `pyautogui.locateCenterOnScreen` function
→ to locate the element you want to operate with since we have no image of
→ the element you want to operate with. DONOT USE `pyautogui.screenshot()`
→ to make screenshot.

Return one line or multiple lines of python code to perform the action each
→ time, be time efficient. When predicting multiple lines of code, make
→ some small sleep like `time.sleep(0.5);` interval so that the machine
→ could take; Each time you need to predict a complete code, no variables
→ or function can be shared from history

You need to to specify the coordinates of by yourself based on your
→ observation of current observation, but you should be careful to ensure
→ that the coordinates are correct.

You ONLY need to return the code inside a code block, like this:

```
```python  
your code here
```
```

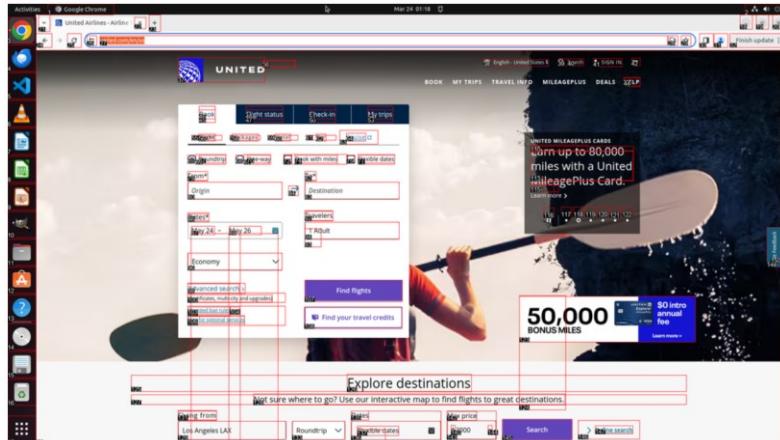
Specially, it is also allowed to return the following special code:
When you think you have to wait for some time, return ```WAIT```;
When you think the task can not be done, return ```FAIL```, don't easily say
→ ```FAIL```, try your best to do the task;
When you think the task is done, return ```DONE```.

My computer's password is 'password', feel free to use it when you need sudo
→ rights.

First give the current screenshot and previous things we did a short
→ reflection, then RETURN ME THE CODE OR SPECIAL CODE I ASKED FOR. NEVER
→ EVER RETURN ME ANYTHING ELSE.

Input settings for LLM/VLM agent baselines

- Accessibility tree
- Screenshot
- Screenshot + accessibility tree
- Set-of-Marks



Set-of-Marks

```
tag_name  text  position (top-left x&y) size (w&h)
menu Google Chrome --> (99, 0) (162, 27)
menu System --> (1814, 0) (106, 27)
label Please download waiting software updates.. --> (810, 73) (291, 17)
push-button Google Chrome --> (0, 33) (70, 64)
push-button Home --> (0, 169) (70, 64)
push-button Visual Studio Code --> (0, 169) (70, 64)
push-button VLC media player --> (0, 237) (70, 64)
push-button LibreOffice Writer --> (0, 305) (70, 64)
push-button LibreOffice Calc --> (0, 373) (70, 64)
push-button LibreOffice Impress --> (0, 441) (70, 64)
push-button GNU Image Manipulation Program --> (0, 509) (70, 64)
push-button Files --> (0, 577) (70, 64)
push-button LibreOffice Software --> (0, 645) (70, 64)
push-button Ubuntu 22.04.3 LTS amd64 --> (0, 784) (70, 64)
push-button Floppy Disk --> (0, 852) (70, 64)
push-button Trash --> (0, 920) (70, 64)
toggle-button Show Applications --> (0, 1010) (70, 70)
label Home Home (1833, 1037) (40, 17)
push-button Minimise Minimise (1389, 51) (30, 30)
push-button Maximise Maximise (1438, 51) (30, 30)
push-button Close Close (1478, 51) (30, 30)
push-button Search tabs Search tabs (656, 46) (28, 41)
push-button Close Close (892, 52) (28, 28)
push-button New Tab New Tab (928, 46) (28, 41)
push-button Back Back (655, 92) (34, 34)
push-button Reload Reload (727, 92) (34, 34)
push-button View site information View site information (97, 97) (24, 24)
entry Address and search bar Recreation.gov (807, 97) (353, 24)
push-button Install Recreation.gov --> (1162, 97) (24, 24)
push-button Bookmark this tab --> (1194, 97) (24, 24)
push-button Side panel Side panel (1239, 92) (34, 34)
push-button You You (1275, 92) (34, 34)
push-button New Chrome available New Chrome available (1314, 92) (196, 34)
document-web Recreation.gov - Camping, Cabins, RVS, Permits, Passes & More --> (650, 133) (866, 922)
link Recreation.gov - Camping, Cabins, RVS, Permits, Passes & More --> (650, 133) (866, 922)
```

Simplified Accessibility Tree

Results of LLM/VLM agent baselines

- LLMs and VLMs are still far from being digital agents on real computers.
- Agent performance fluctuations vs. consistent human performance across different types of computer tasks.
- A11y tree and SoM's effectiveness varies by models.
- VLM agents with screenshot-only setting show lower performance, but it should be the ultimate configuration in the long run.

| Inputs | Model | Success Rate (\uparrow) | | | | | |
|---------------------------|----------------|-----------------------------|--------|--------|----------|----------|---------|
| | | OS | Office | Daily | Profess. | Workflow | Overall |
| A11y tree | Mixtral-8x7B | 12.50% | 1.01% | 4.79% | 6.12% | 0.09% | 2.98% |
| | Llama-3-70B | 4.17% | 1.87% | 2.71% | 0.00% | 0.93% | 1.61% |
| | GPT-3.5 | 4.17% | 4.43% | 2.71% | 0.00% | 1.62% | 2.69% |
| | GPT-4 | 20.83% | 3.58% | 25.64% | 26.53% | 2.97% | 12.24% |
| | Gemini-Pro | 4.17% | 1.71% | 3.99% | 4.08% | 0.63% | 2.37% |
| | Gemini-Pro-1.5 | 12.50% | 2.56% | 7.83% | 4.08% | 3.60% | 4.81% |
| | Qwen-Plus | 29.17% | 3.58% | 8.36% | 10.20% | 2.61% | 6.87% |
| | GPT-4o | 20.83% | 6.99% | 16.81% | 16.33% | 7.56% | 11.36% |
| Screenshot | CogAgent | 4.17% | 0.85% | 2.71% | 0.00% | 0.00% | 1.11% |
| | GPT-4V | 12.50% | 1.86% | 7.58% | 4.08% | 6.04% | 5.26% |
| | Gemini-ProV | 8.33% | 3.58% | 6.55% | 16.33% | 2.08% | 5.80% |
| | Gemini-Pro-1.5 | 12.50% | 6.99% | 2.71% | 6.12% | 3.60% | 5.40% |
| | Claude-3-Opus | 4.17% | 1.87% | 2.71% | 2.04% | 2.61% | 2.42% |
| | GPT-4o | 8.33% | 3.58% | 6.07% | 4.08% | 5.58% | 5.03% |
| Screenshot
+ A11y tree | CogAgent | 4.17% | 0.85% | 2.71% | 0.62% | 0.09% | 1.32% |
| | GPT-4V | 16.66% | 6.99% | 24.50% | 18.37% | 4.64% | 12.17% |
| | Gemini-ProV | 4.17% | 4.43% | 6.55% | 0.00% | 1.52% | 3.48% |
| | Gemini-Pro-1.5 | 12.50% | 3.58% | 7.83% | 8.16% | 1.52% | 5.10% |
| | Claude-3-Opus | 12.50% | 3.57% | 5.27% | 8.16% | 1.00% | 4.41% |
| | GPT-4o | 41.67% | 6.16% | 12.33% | 14.29% | 7.46% | 11.21% |
| Set-of-Mark | CogAgent | 4.17% | 0.00% | 2.71% | 0.00% | 0.53% | 0.99% |
| | GPT-4V | 8.33% | 8.55% | 22.84% | 14.28% | 6.57% | 11.77% |
| | Gemini-ProV | 4.17% | 1.01% | 1.42% | 0.00% | 0.63% | 1.06% |
| | Gemini-Pro-1.5 | 16.67% | 5.13% | 12.96% | 10.20% | 3.60% | 7.79% |
| | Claude-3-Opus | 12.50% | 2.72% | 14.24% | 6.12% | 4.49% | 6.72% |
| | GPT-4o | 20.83% | 3.58% | 3.99% | 2.04% | 3.60% | 4.59% |
| Human Performance | | 75.00% | 71.79% | 70.51% | 73.47% | 73.27% | 72.36% |

Result analysis of LLM/VLM agent baselines

- Higher screenshot resolution typically leads to improved performance

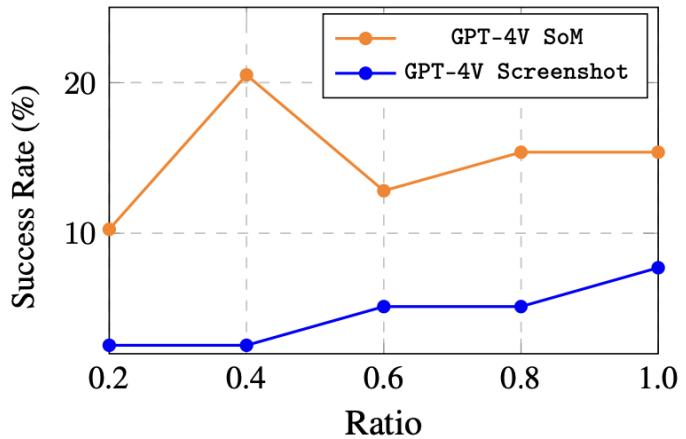


Figure 5: The effect of downsampling on the screenshot on performance with down-sampling ratios of 0.2, 0.4, 0.6 and 0.8 and run on a subset (10%) of examples.

Result analysis of LLM/VLM agent baselines

- Longer text-based trajectory history context improves performance, unlike screenshot- only history, but poses efficiency challenges

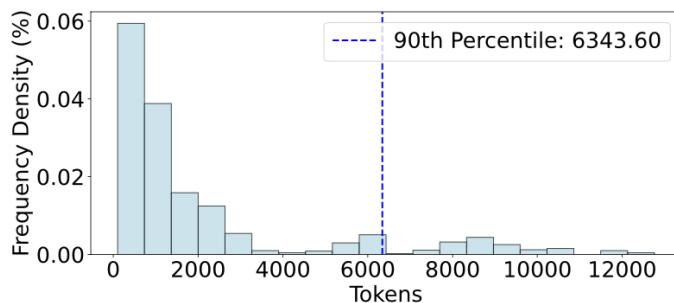


Figure 6: The length distribution of a11y tree as observation from sampled trajectories.

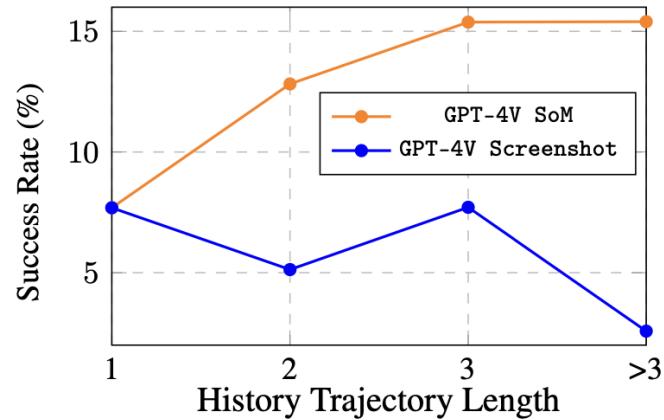


Figure 7: The effect of length of history on performance with the history encoding length of 1, 2, 3, and > 3 and run on a subset (10%) of examples.

Result analysis of LLM/VLM agent baselines

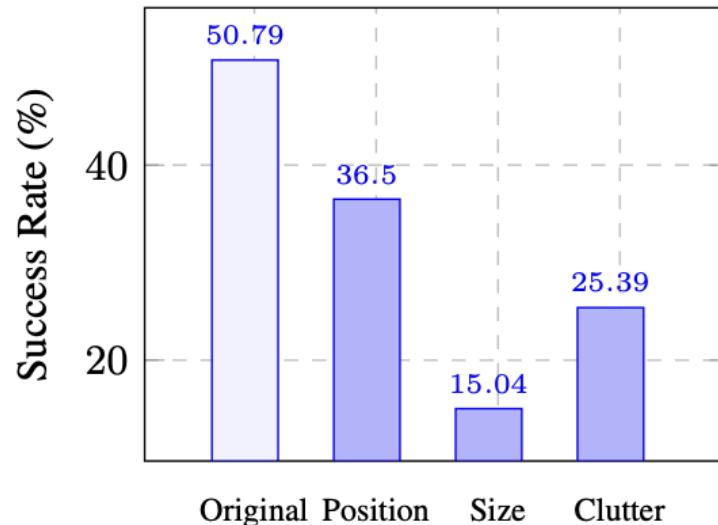
- The performance of VLM agents across different OS is in strong correlation.
- This implies that insights and methodologies developed within the OSWORLD framework can be effectively transferred to Windows environments with a high degree of reliability.

Table 7: Comparison of model performance and correlation across operating systems.

| OS | SR (%) | Correlation Coefficient |
|---------|--------|-------------------------|
| Ubuntu | 4.88 | |
| Windows | 2.55 | 0.7 |

Result analysis of LLM/VLM agent baselines

- Current VLM agents are not robust to UI layout and noise.
- See paper for more interesting analysis.

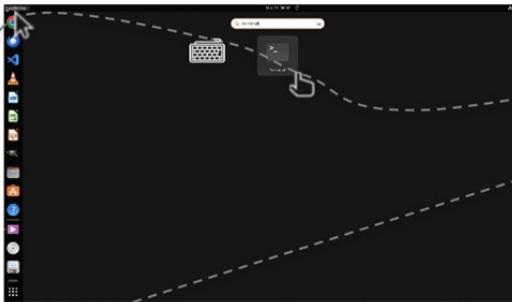


Success case of LLM/VLM agent baselines

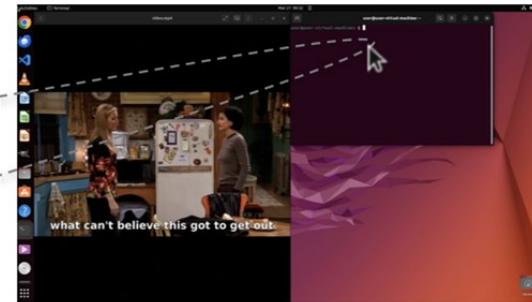
Task Instruction: I downloaded an episode of Friends to practice listening, but I don't know how to remove the subtitles. Please help me remove the subtitles from the video and export it as "subtitles.srt" and store it in the same directory as the video.



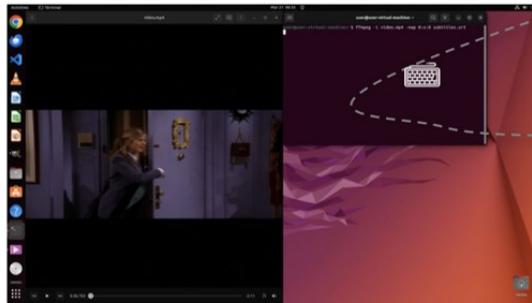
Step 1: pyautogui.click(activities_x, activities_y)



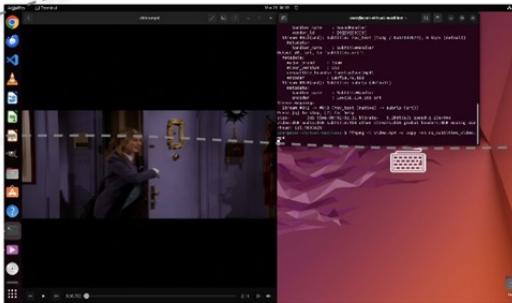
Step 2: pyautogui.typewrite('terminal', interval=0.5)



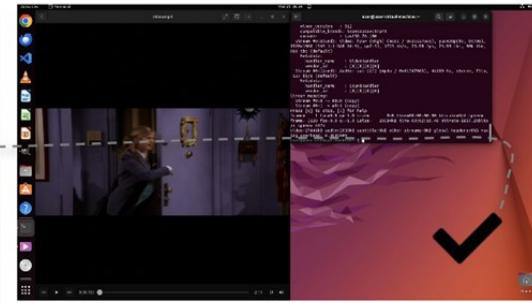
Step 3: pyautogui.click(focus_x, focus_y)



Step 4: pyautogui.typewrite('ffmpeg -i video.mp4 -map 0:s:0 subtitles.srt', interval=0.5)



Step 5: pyautogui.typewrite('ffmpeg -i video.mp4 -c copy -sn no_subtitles_video.mp4', interval=0.5)



Step 6: Done

Future Research

- **Capability:** new agent interactive learning paradigms, innovative foundation models, adjustment to complex environments with more tools
- **Reliability and security:** know when to abstain and verify, identify potential risks
- **Personalization:** customized agents
- **Efficiency:** compact models, real-time interactions

Acknowledgements

We thank the following institutions for their gift funds supporting our
open-source initiatives!



Research at Google



Thank you for listening!



<https://xlang.ai>



香港大學自然語言處理實驗室

Natural Language Processing Group, The University of Hong Kong