

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

# Computer Use Agents

JY Koh  
11-768: AI Agents  
15 Sep 2026



计算机使用智能体

JY Koh

11-768: AI 智能体

2026 年 9 月 15 日

计算机使用智能体

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

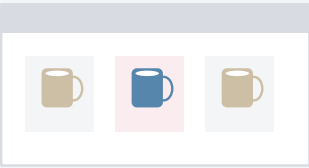
ZH / 中文译文 · 第 1/1 部分

01 / COMPUTER USE AGENTS

# Computer Use Agents (CUAs)

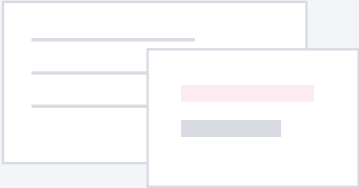
CUAs observe rendered GUIs, execute actions, and checks what changed.

BROWSER



Shop across pages

DESKTOP



Edit files across apps

MOBILE



Navigate screens and gestures

9

01 / 计算机使用智能体

计算机使用智能体（CUA）

CUA 观察渲染出来的图形用户界面（GUI）、执行动作，并检查发生了什么变化。

浏览器：跨页面购物。

桌面：跨应用编辑文件。

移动设备：浏览屏幕并执行手势。

计算机使用智能体（CUA）

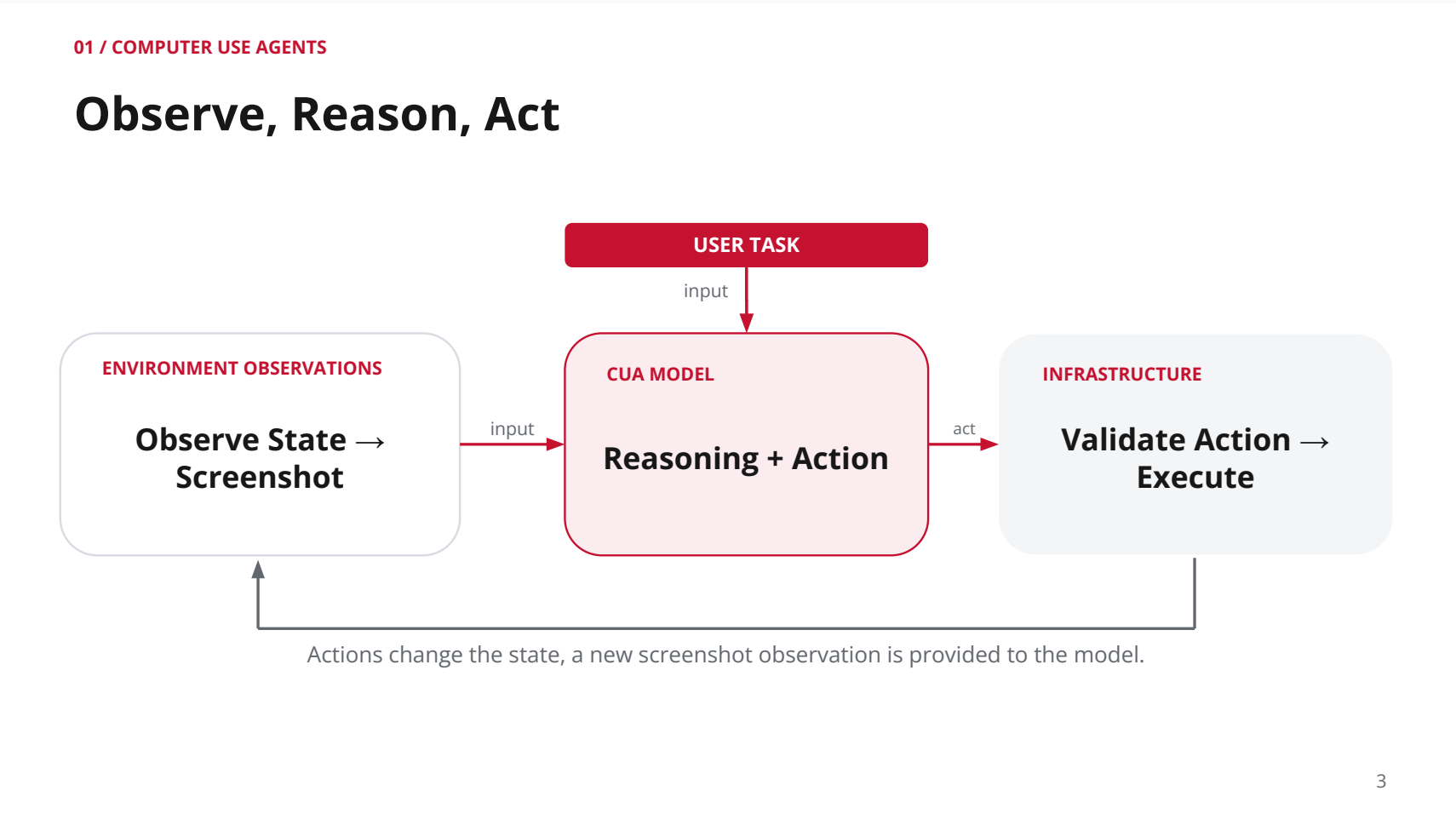
个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 2 | 原页 2

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



01 / 计算机使用智能体

观察、推理、行动

用户任务 → 输入 → CUA 模型（推理 + 动作） → 行动；基础设施负责验证动作并执行。

环境观测：观察状态 → 截图。

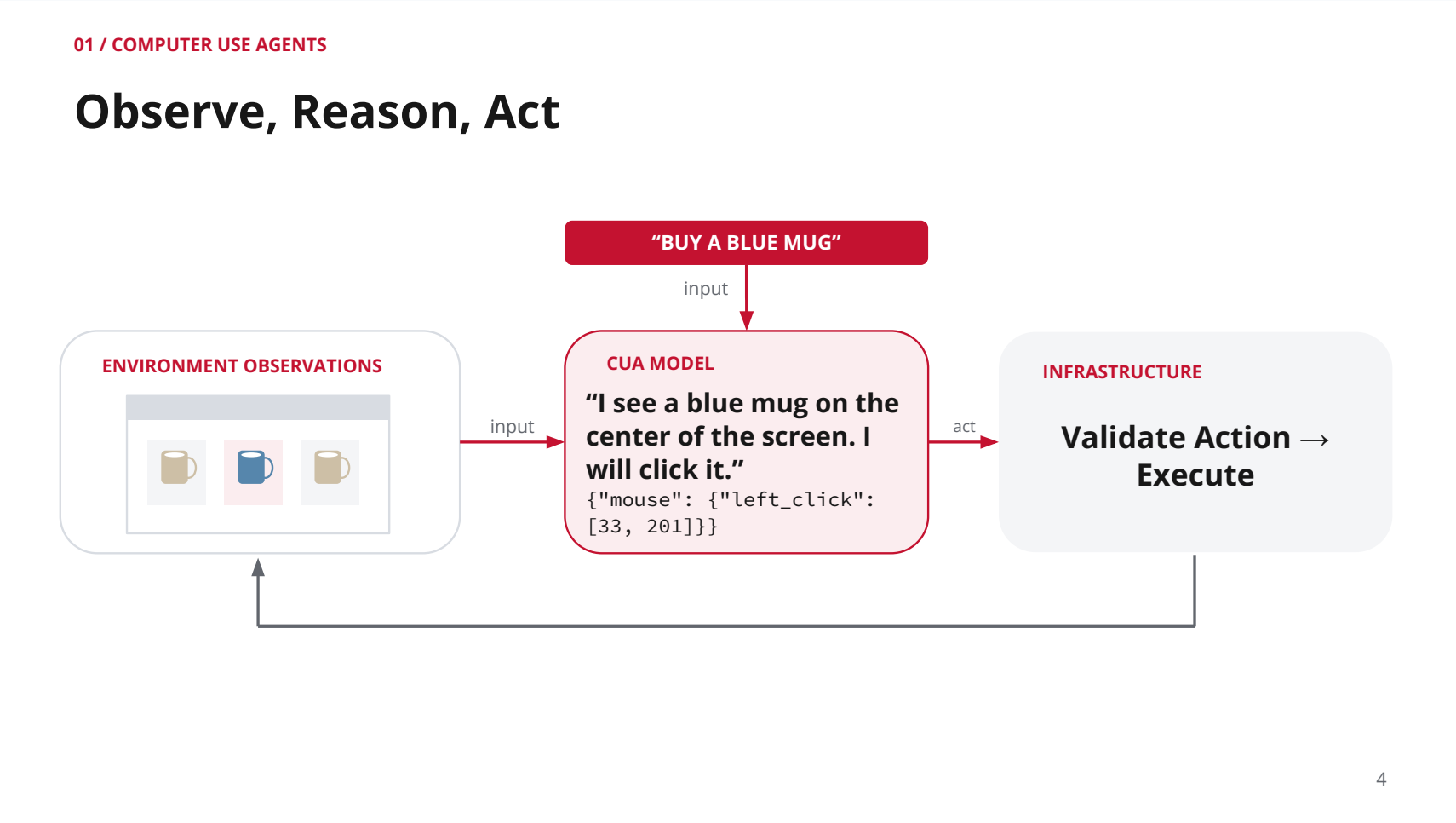
动作会改变状态，随后会向模型提供一张新的截图观测。

观察、推理、行动

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



01 / 计算机使用智能体

观察、推理、行动

用户任务：“买一个蓝色杯子”。

CUA 模型：“我看到屏幕中央有一个蓝色杯子。我会点击它。”

动作：{"mouse": {"left\_click": [33, 201]}}

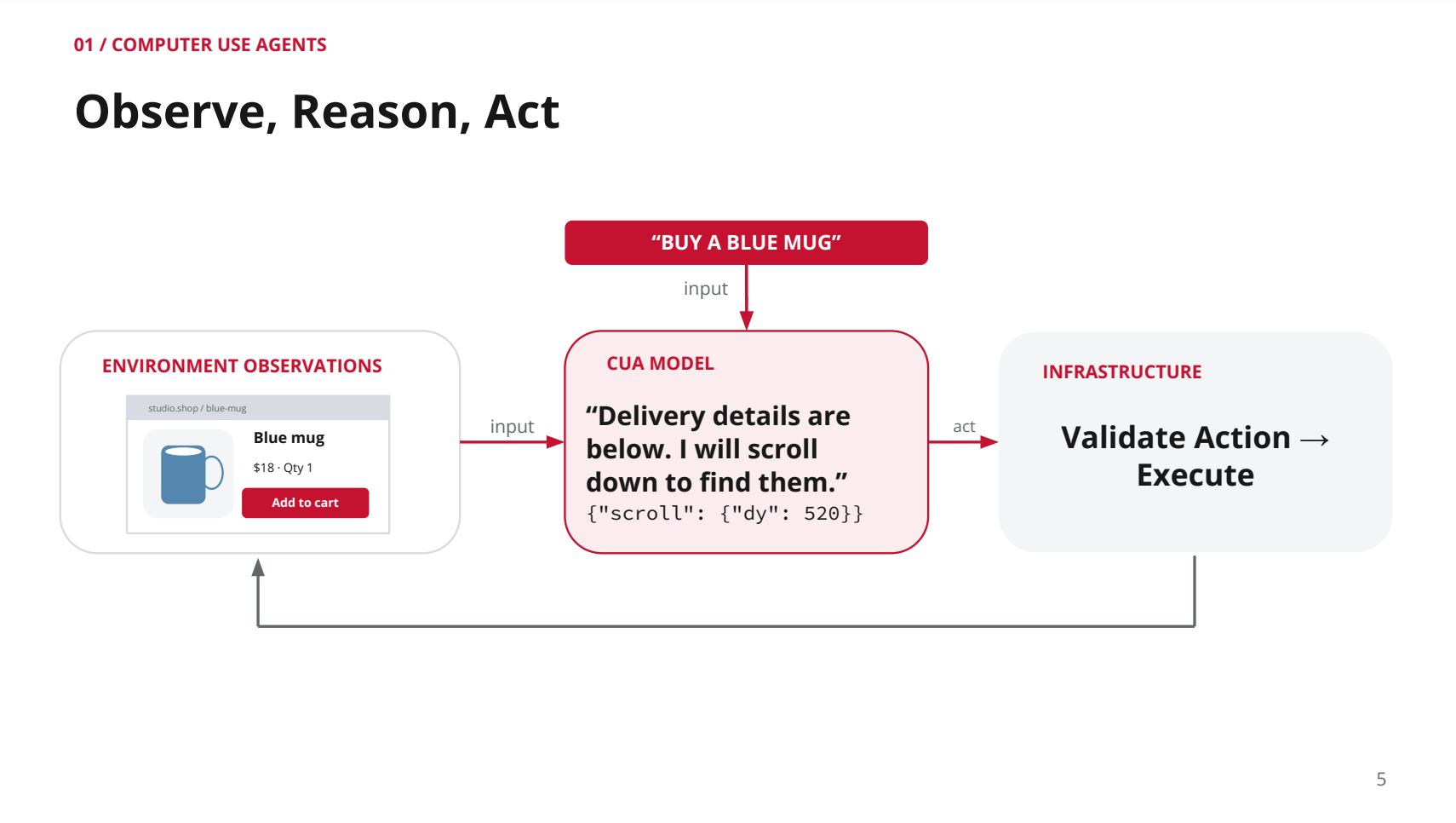
基础设施：验证动作 → 执行。

观察、推理、行动：点击蓝色杯子

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



01 / 计算机使用智能体

观察、推理、行动

用户任务：“买一个蓝色杯子”。

环境观测：studio.shop / blue-mug；蓝色杯子；\$18 · 数量 1；添加到购物车。

CUA 模型：“配送信息在下面。我会向下滚动查找。”

动作：{"scroll": {"dy": 520}}

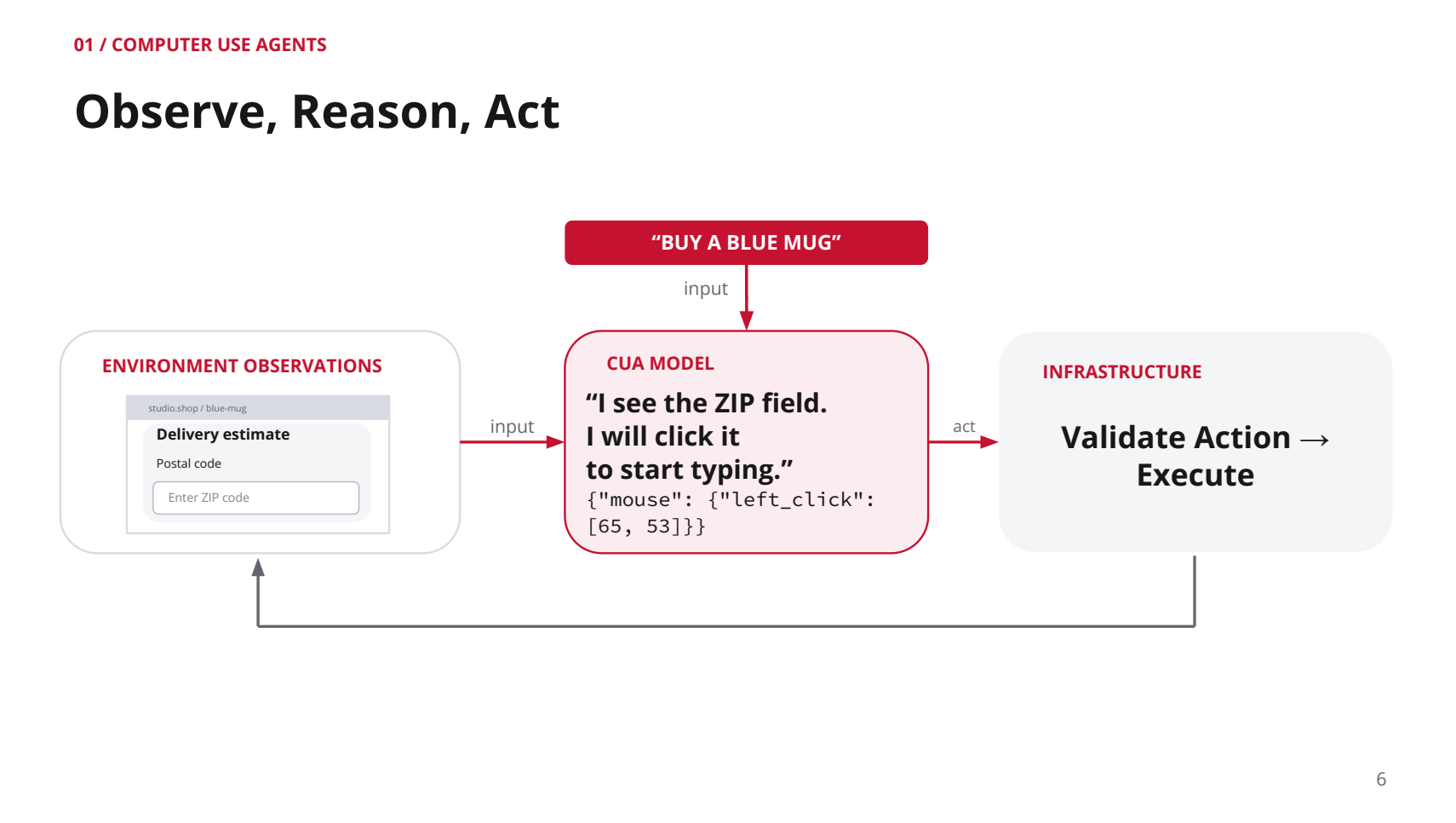
基础设施：验证动作 → 执行。

观察、推理、行动：查看配送信息

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



01 / 计算机使用智能体

观察、推理、行动

用户任务：“买一个蓝色杯子”。

环境观测：studio.shop / blue-mug；预计送达时间；邮政编码；输入 ZIP code。

CUA 模型：“我看到 ZIP 栏。我会点击它并开始输入。”

动作：{"mouse": {"left\_click": [65, 53]}}

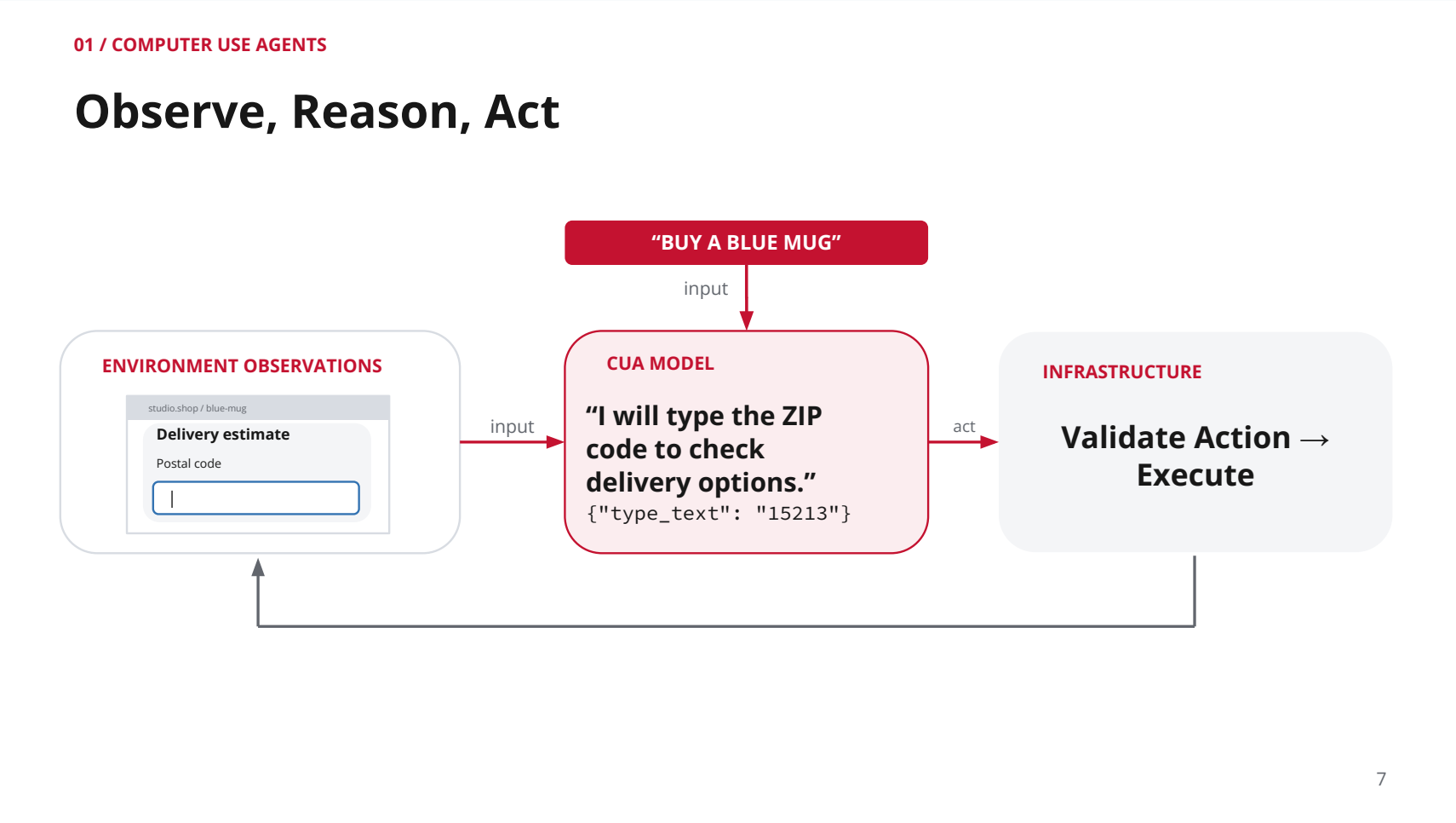
基础设施：验证动作 → 执行。

观察、推理、行动：点击邮政编码栏

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



01 / 计算机使用智能体

观察、推理、行动

用户任务：“买一个蓝色杯子”。

环境观测：studio.shop / blue-mug；预计送达时间；邮政编码。

CUA 模型：“我会输入邮政编码以查询配送选项。”

动作：{"type\_text": "15213"}

基础设施：验证动作 → 执行。

观察、推理、行动：输入邮政编码

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

01 / COMPUTER USE AGENTS

# Observe, Reason, Act

STEP 1

studio.shop / mugs

STUDIO / Everyday objects

Cream mug  
\$16

Blue mug  
\$18

Blue set  
\$28

click the blue mug

STEP 2

studio.shop / blue-mug

STUDIO / Everyday objects

Blue mug  
\$18 · Quantity 1

Add to cart

scroll down

STEP 3

studio.shop / blue-mug

Delivery estimate

Postal code

15213

type "15213"

REWARD

1.0

Task success

The loop continues until task completion, then a reward is computed.

8

01 / 计算机使用智能体

第 1 步：studio.shop / mugs → 点击蓝色杯子。商品：STUDIO / 日常用品；奶油色杯子 \$16；蓝色杯子 \$18；蓝色套装 \$28。

第 2 步：studio.shop / blue-mug → 向下滚动。蓝色杯子 \$18 · 数量 1；添加到购物车。

第 3 步：studio.shop / blue-mug → 输入 “15213” 。配送估算；邮政编码 15213 → 奖励 1.0，任务成功。

循环持续进行直到任务完成，随后计算奖励。

浏览器交互循环示例

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 8 | 原页 8

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

01 / COMPUTER USE AGENTS

A Brief History of CUAs: Toy Environments

Learning to use greatly simplified toy web interfaces

Tell Anita I'll "Orci, sit magnis." in response to her email.

Primary

Alyson  
Cursus.  
Turpis molestie..

Katherine  
Sollicitudin.  
Faucibus quam i..

Marlee  
Rhoncus.  
Nisl. Sed. Quis..

Geralda

2017-2022

2022-2024

2025-2026

Sep 2026

MiniWoB: Shi et al., ICML 2017 · MiniWoB++: Liu et al., ICLR 2018 · WebShop: Yao et al., NeurIPS 2022

CUA 简史：玩具环境

01 / 计算机使用智能体

CUA 简史：玩具环境

学习使用极度简化的玩具网页界面。

时间线：2017–2022 → 2022–2024 → 2025–2026 → 2026 年 9 月。

MiniWoB: Shi 等, ICML 2017 · MiniWoB++: Liu 等, ICLR 2018 · WebShop : Yao 等, NeurIPS 2022

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 9 | 原页 9

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

01 / COMPUTER USE AGENTS

Shopping with  
real products and  
language goals

WebShop

Instruction:  
I am looking for x-large, red color women faux fur  
lined winter warm jacket coat, and price lower than  
70.00 dollars

Search... I

Search

Project Site

Task Instructions

Current Action: search[women faux jacket coat]

2017-2022

2022-2024

2025-2026

Sep 2026

WebShop: Yao et al., NeurIPS 2022 · Official environment demo

CUA 简史：在真实商品中购物

01 / 计算机使用智能体

CUA 简史：玩具环境

使用真实商品并根据自然语言目标购物。

时间线：2017–2022 → 2022–2024 → 2025–2026 → 2026 年 9 月。

WebShop: Yao 等, NeurIPS 2022 · 官方环境演示

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 10 | 原页 10

第 7 讲 · 计算机使用智能体

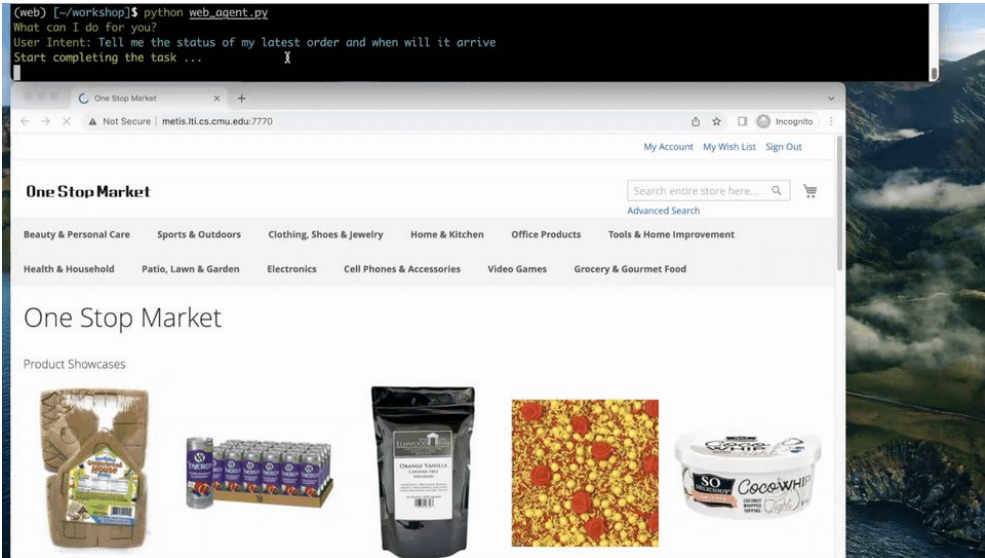
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

01 / COMPUTER USE AGENTS

A Brief History of CUAs: Realistic Environments

From toy tasks  
to realistic  
web interfaces



2017–2022

2022–2024

2025–2026

Sep 2026

WebArena: Zhou et al., ICLR 2024 · VisualWebArena: Koh et al., ACL 2024

01 / 计算机使用智能体

CUA 简史：逼真的环境

从玩具任务走向逼真的网页界面。

时间线：2017–2022 → 2022–2024 → 2025–2026 → 2026 年 9 月。

WebArena: Zhou 等, ICLR 2024 · VisualWebArena: Koh 等, ACL 2024

CUA 简史：逼真的环境

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

01 / 计算机使用智能体

CUA 简史：逼真的环境

完整的（真实）桌面工作流。

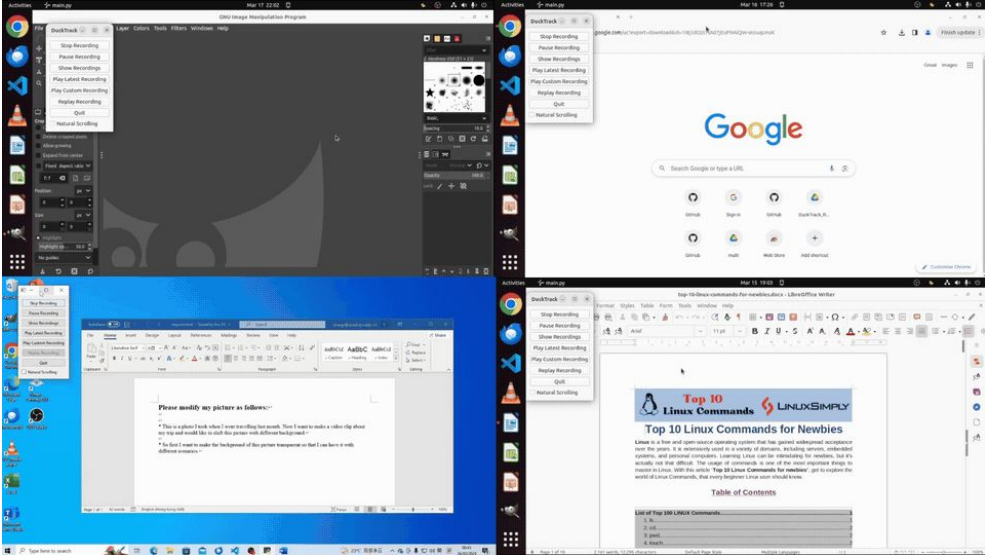
时间线：2017–2022 → 2022–2024 → 2025–2026 → 2026 年 9 月。

OSWorld: Xie 等, NeurIPS 2024 · 官方概览视频

01 / COMPUTER USE AGENTS

A Brief History of CUAs: Realistic Environments

Full (realistic) desktop workflows



2017–2022

2022–2024

2025–2026

Sep 2026

OSWorld: Xie et al., NeurIPS 2024 · Official overview video

CUA 简史：完整桌面工作流

## 第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

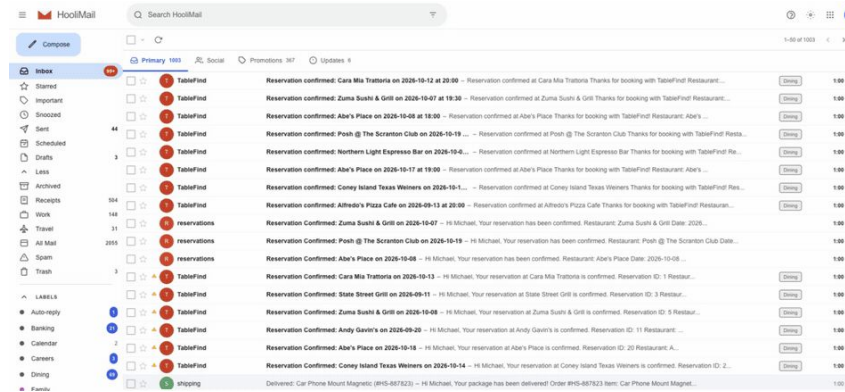
## 01 / COMPUTER USE AGENTS

## A Brief History of CUAs: Realistic Envs and Data

## MYPCBENCH

## Realistic apps with personal history

Explore mail, calendar events,  
and project issues in one  
shared world.



2017-2022

2022-2024

2025-2026

Sep 2026

Jang et al., 2026 · MyPCBench · Recorded environment walkthrough (33 s)

## CUA 简史：逼真的环境与数据

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 13 | 原页 13

## 01 / 计算机使用智能体

## CUA 简史：逼真的环境与数据

MYPCBENCH

带有个人历史信息的真实应用：在同一个共享世界中探索邮件、日历事件和项目问题。

时间线: 2017-2022 → 2022-2024 → 2025-2026 → 2026 年 9 月。

Jang 等, 2026 · MyPCBench · 录制的环境演练 (33 秒)

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

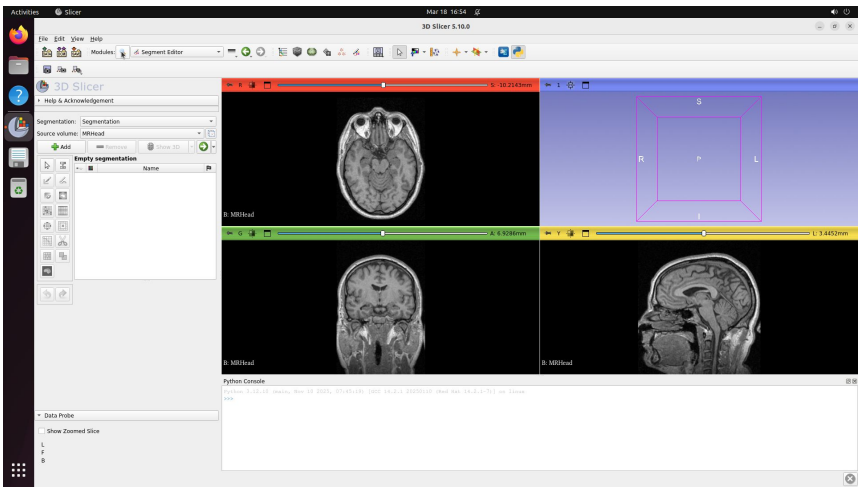
01 / COMPUTER USE AGENTS

A Brief History of CUAs: Realistic Envs and Data

CUA-WORLD

Real software, realistic tasks

Specialized desktop software with loaded data and realistic tasks.



01 / 计算机使用智能体

CUA 简史：逼真的环境与数据

CUA-WORLD

真实软件、真实任务：带有已载入数据和真实任务的专业桌面软件。

时间线：2017-2022 → 2022-2024 → 2025-2026 → 2026 年 9 月。

Aggarwal 等, 2026 · Gym-Anything / CUA-World

CUA 简史：真实软件与真实任务

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

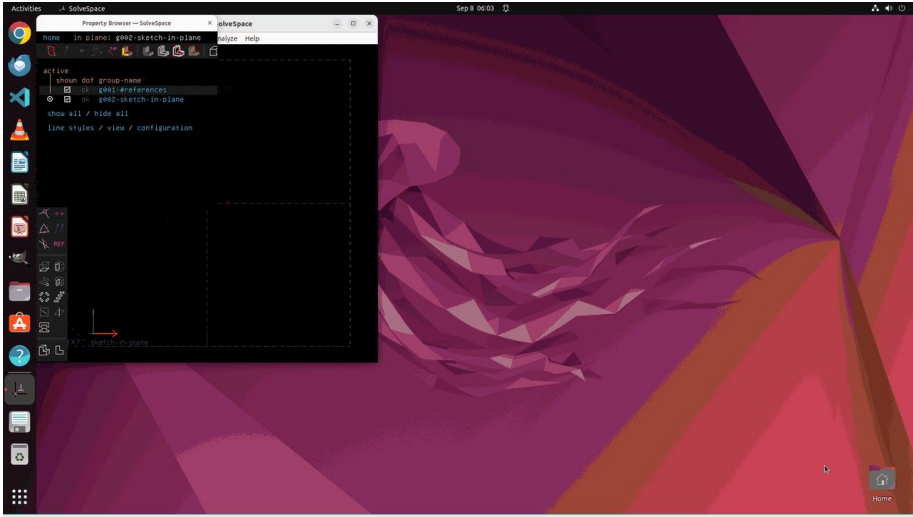
ZH / 中文译文 · 第 1/1 部分

01 / COMPUTER USE AGENTS

A Brief History of CUAs: Real World Use

GPT-6 ASTRA

Very capable and fast,  
still (very) expensive.



2017–2022

2022–2024

2025–2026

Sep 2026

Koh, Sep 2026 · What's the Point of Computer Use Agents? · OSWorld 2.0 SolveSpace demo (30×)

01 / 计算机使用智能体

CUA 简史：现实世界中的使用

GPT-6 ASTRA

能力强、速度快，但仍然（非常）昂贵。

时间线：2017–2022 → 2022–2024 → 2025–2026 → 2026 年 9 月。

Koh, 2026 年 9 月 · What’ s the Point of Computer Use Agents? · OSWorld 2.0 SolveSpace 演示 (30 倍速)

CUA 简史：现实世界中的使用

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



基准与评估

基准与评估

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02 / BENCHMARKS

What should count as success?

The same task can be evaluated in multiple different ways.

ACTION

Did it click the target?

Fast grounding / imitation tests

OUTCOME

Is the right item in the cart?

State checks or evidence judging

PROCESS

Did it respect the constraints?

No checkout; no unwanted changes

Evaluators should match the capability you want to measure.

17

02 / 基准

什么才算成功？

同一个任务可以通过多种不同方式评估。

动作：它是否点击了目标？——快速定位 / 模仿测试。

结果：购物车里是否有正确商品？——状态检查或证据判断。

过程：它是否遵守了约束？——不结账；不做不需要的更改。

评估器应与想要测量的能力相匹配。

什么才算成功？

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02 / BENCHMARKS

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

18

02 / 基准

基准

从预测单个动作走向完成延伸的工作流。

02A 静态评估：在固定观测上预测一个动作。

02B 端到端评估（网页）：通过浏览器交互完成任务。

02C 端到端评估（桌面、移动）：跨应用和系统状态完成任务。

02D 长时程计算机使用：在延伸的工作流中持续推进。

长时程评估将端到端评估扩展到网页和桌面。

基准：从预测动作到完成长工作流

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

### Static evals: Was this action correct?

TASK:

Add a blue mug under \$20 to the cart. Do not purchase.

SCREENSHOT



REFERENCE RULE

$(x, y) \in \text{target box}$

For recorded actions:  
Match target, operation,  
and input value.

STEP SCORE

1.0

correct point / step

No environment rollout

A correct step is not a completed task.

BENCHMARK EXAMPLES

[ScreenSpot-Pro · Li et al., 2025](#)

[AITW · Rawles et al., 2023](#)

[Mind2Web · Deng et al., 2023](#)

[AndroidControl · Li et al., 2024](#)

静态评估：这个动作正确吗？

02A / 静态评估

静态评估：这个动作正确吗？

任务：将一个低于 \$20 的蓝色杯子加入购物车。不要购买。

截图：studio.shop / mugs。

参考规则：(x, y) ∈ 目标框。

对于记录下来的动作：匹配目标、操作和输入值。

步骤得分：1.0；正确点击点数 / 步骤；不运行环境轨迹。

一个正确步骤不等于完成任务。

基准示例：ScreenSpot-Pro · Li 等，2025；Mind2Web · Deng 等，2023；AITW · Rawles 等，2023；AndroidControl · Li 等，2024。

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 19 | 原页 19

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

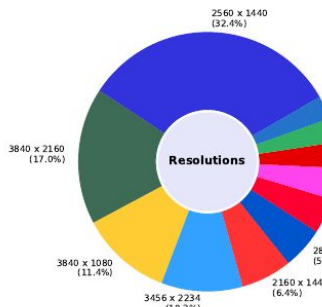
ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

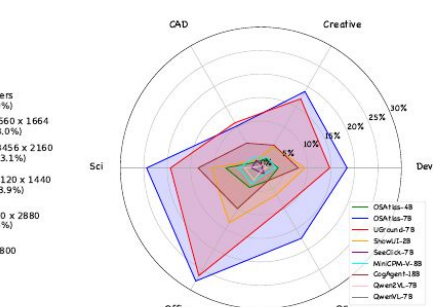
ScreenSpot-Pro: Grounding in professional UIs



(a) Software categories and the number of tasks.



(b) Resolution distribution.



(c) Results of representative GUI Grounding models across 6 categories of applications.

BENCHMARK STATS

1,581  
screenshots

23  
applications

3  
operating systems

EVALUATION CRITERION

Point / region matching

High-resolution professional UIs make grounding a benchmark of its own.

Source: Li et al., 2025, Fig. 1

ScreenSpot-Pro: 专业界面中的定位

02A / 静态评估

ScreenSpot-Pro: 专业界面中的定位

基准统计: 1,581 张截图; 23 个应用; 3 种操作系统。

评估标准: 点 / 区域匹配。

高分辨率的专业界面使定位本身成为一个独立基准。

来源: Li 等, 2025, 图 1

第 7 讲 · 计算机使用智能体

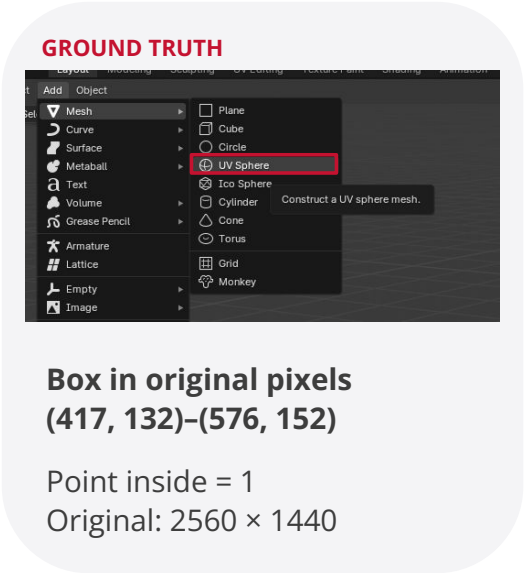
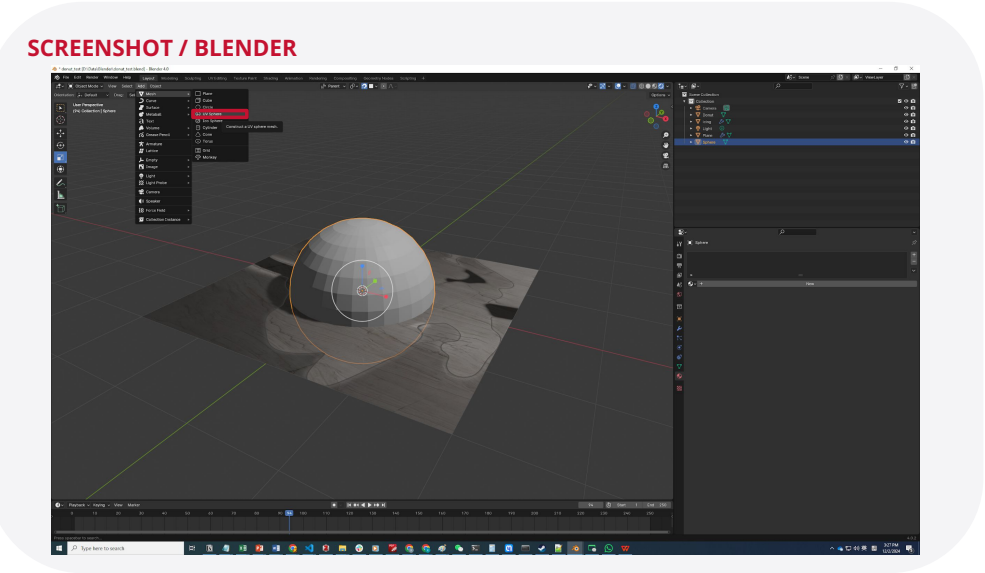
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

ScreenSpot-Pro: Grounding in professional UIs

**TASK:** Construct a UV sphere mesh.



Li et al., 2025 · ScreenSpot-Pro, blender\_windows\_1

02A / 静态评估

ScreenSpot-Pro：专业界面中的定位

任务：构建一个 UV 球体网格。

截图 / Blender。

真实标注：原始像素坐标中的框 (417, 132)–(576, 152)；点落在框内 = 1；原始图像大小：2560 × 1440。

Li 等，2025 · ScreenSpot-Pro, blender\_windows\_1

ScreenSpot-Pro：专业界面中的定位示例

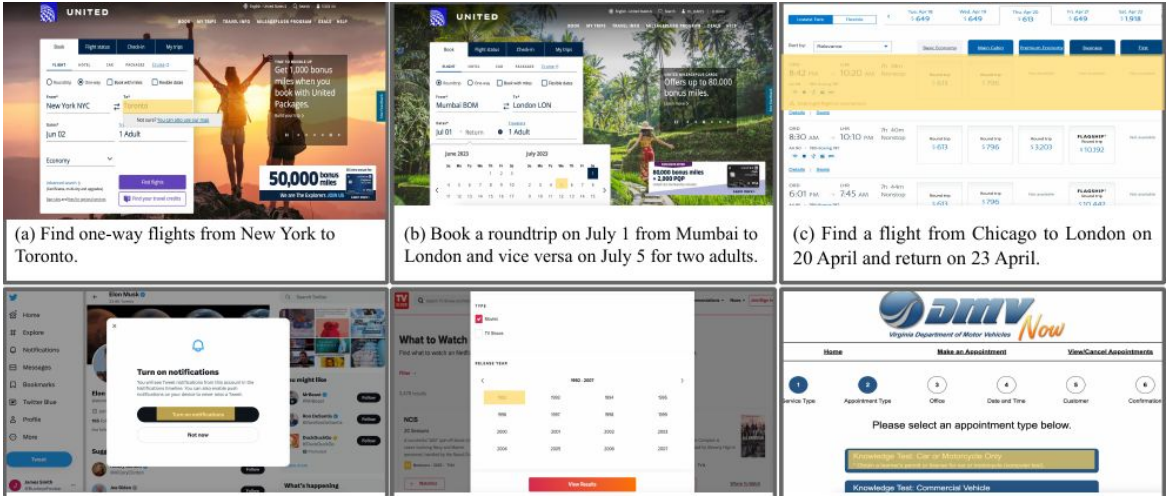
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

Mind2Web: Offline generalization across the web



BENCHMARK STATS

2,350 tasks

137 websites

31 domains

EVALUATION CRITERION

Offline action matching

Next-action imitation is not the same thing as task completion.

Source: Deng et al., NeurIPS 2023, Fig. 1

02A / 静态评估

Mind2Web：跨网页的离线泛化

基准统计：2,350 个任务；137 个网站；31 个领域。

评估标准：离线动作匹配。

模仿下一步动作与完成任务不是一回事。

来源：Deng 等，NeurIPS 2023，图 1

Mind2Web：跨网页的离线泛化

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

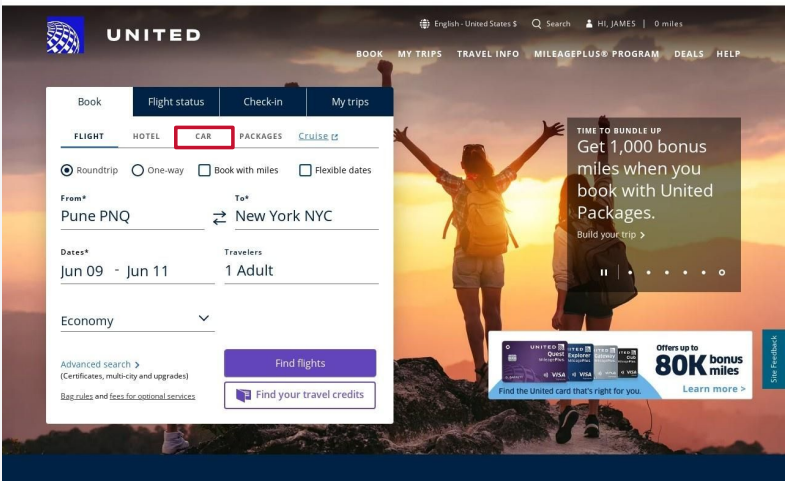
ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

Mind2Web: Offline generalization across the web

**TASK:** Rent a car in Brooklyn–Central, NY, for April 9–15.

RECORDED SCREENSHOT / UNITED



Ground-truth target outlined in red

Deng et al., NeurIPS 2023 · Multimodal-Mind2Web, training example

GROUND-TRUTH ACTION

**CLICK**  
**Target: CAR tab**  
**Value: none**

Element: bookCarTab

Match the target element and operation. No task rollout is performed.

02A / 静态评估

Mind2Web：跨网页的离线泛化

任务：4 月 9 日至 15 日在纽约州布鲁克林—中央区租车。

记录截图 / United。

真实动作：点击 CAR 标签页；值：无；元素：bookCarTab。

匹配目标元素和操作，不运行任务轨迹。

红框标出真实目标。

Deng 等，NeurIPS 2023 · Multimodal-Mind2Web，训练样例

Mind2Web：离线网页任务示例

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02 / BENCHMARKS

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

24

02 / 基准

基准

从预测单个动作走向完成延伸的工作流。

02A 静态评估：在固定观测上预测一个动作。

02B 端到端评估（网页）：通过浏览器交互完成任务。

02C 端到端评估（桌面、移动）：跨应用和系统状态完成任务。

02D 长时程计算机使用：在延伸的工作流中持续推进。

长时程评估将端到端评估扩展到网页和桌面。

基准：从预测动作到完成长工作流

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

### Programmatic state: Test the resulting world

TASK:

Add a blue mug under \$20 to the cart. Do not purchase.

FINAL APP STATE

studio.shop / cart



Blue mug

\$18 · Qty 1

In cart

EXECUTABLE CHECKS

```
cart.item == blue_mug
cart.price < 20
orders.count == 0
```

assert all(checks)

TASK REWARD

PASS

reward = 1.0

Database / file / app state

Repeatable and precise—but only for conditions the tests cover.

BENCHMARK EXAMPLES

[WebArena · Zhou et al., 2024](#)

[OSWorld · Xie et al., 2024](#)

[AndroidWorld · Rawles et al., 2024](#)

[WorkArena · Drouin et al., 2024](#)

36

程序化状态：检查最终世界状态

02B / 端到端评估（网页）

程序化状态：检查最终世界状态

任务：将一个低于 \$20 的蓝色杯子加入购物车。不要购买。

最终应用状态：studio.shop / cart；蓝色杯子；\$18 · 数量 1；已在购物车。

可执行检查：cart.item == blue\_mug；cart.price < 20；orders.count == 0；assert all(checks)。

任务奖励：通过；reward = 1.0；数据库 / 文件 / 应用状态。

结果可重复且精确，但只覆盖测试所检查的条件。

基准示例：WebArena · Zhou 等, 2024；OSWorld · Xie 等, 2024；AndroidWorld · Rawles 等, 2024；WorkArena · Drouin 等, 2024。

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 25 | 原页 25

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

### WebArena: Reproducible sites, executable goals

BENCHMARK STATS

812

long-horizon tasks

4

self-hosted domains

functional

success checks

EVALUATION CRITERION

Programmatic end state

Evaluators check environment state, not whether the trace “looks right”.

Source: Zhou et al., ICLR 2024, Fig. 1

02B / 端到端评估（网页）

WebArena：可复现网站与可执行目标

基准统计：812 个长时程任务；4 个自托管领域；功能性成功检查。

评估标准：程序化最终状态。

评估器检查环境状态，而不是轨迹“看起来是否正确”。

来源：Zhou 等, ICLR 2024, 图 1

WebArena：可复现网站与可执行目标

第 7 讲 · 计算机使用智能体

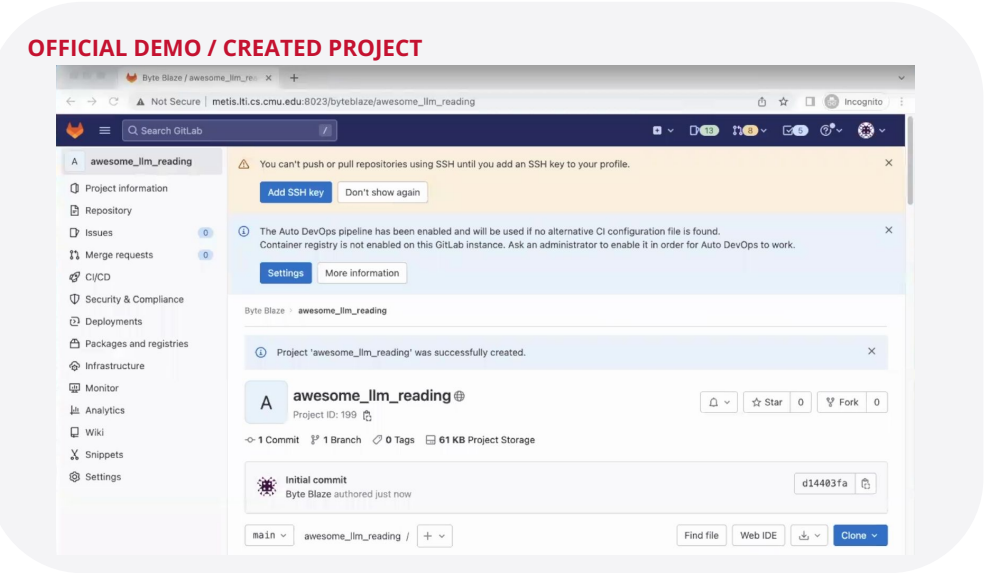
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

WebArena: Reproducible sites, executable goals

**TASK:** Create an empty repository named awesome\_llm\_reading.



PROGRAMMATIC VERIFIER

```
name = task.project_name
p = "/byteblaze/"
page.goto(GITLAB+p+name)
html = page.content()
reward = int(
    name in html.lower()
)
```

Task 476 (simplified).

Checks name at target URL.  
Emptiness is not checked.

Zhou et al., ICLR 2024 · Official demo + task 476 / program\_html evaluator

02B / 端到端评估（网页）

WebArena：可复现网站与可执行目标

任务：创建一个名为 awesome\_llm\_reading 的空仓库。

官方演示 / 已创建项目。

程序化验证器：name = task.project\_name; p = "/byteblaze/"; page.goto(GITLAB+p+name); html = page.content(); reward = int(name in html.lower())。

简化后的任务 476：在目标 URL 检查名称，不检查仓库是否为空。

Zhou 等, ICLR 2024 · 官方演示 + 任务 476 / program\_html 评估器

WebArena：程序化验证器示例

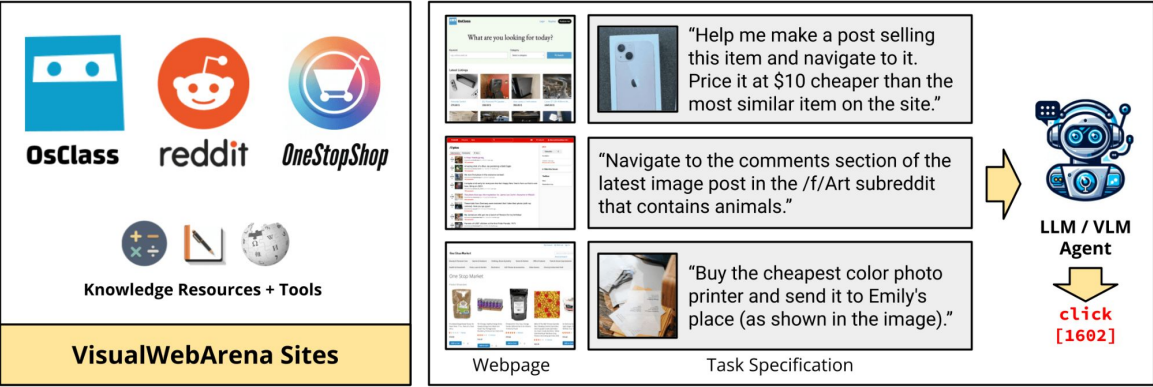
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

VisualWebArena: Vision is part of the task



Visual understanding changes which tasks are even expressible.

Source: Koh et al., ACL 2024, Fig. 1

02B / 端到端评估（网页）

VisualWebArena：视觉是任务的一部分

基准统计：910 个任务；3 个自托管网站；25.2% 的任务需要图像输入。

评估标准：多模态 + 最终状态。

视觉理解会改变哪些任务能够被表达出来。

来源：Koh 等，ACL 2024，图 1

VisualWebArena：视觉是任务的一部分

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

# LLM / VLM judging: Evaluate the evidence

TASK:

Add a blue mug under \$20 to the cart. Do not purchase.

OBSERVED EVIDENCE

studio.shop / cart



Blue mug

\$18 · Qty 1

In cart

Screenshots + action history

LLM / VLM + RUBRIC

✓ Blue mug present

✓ Price below \$20

✓ No purchase shown

Rubric → evidence → verdict

JUDGMENT

PASS

"The visible cart meets the requested criteria."

Flexible evidence judging needs calibration against human labels.

BENCHMARK EXAMPLES

[WebVoyager · He et al., 2024](#)

[CUA-World-Long · Aggarwal et al., 2026](#)

[Online-Mind2Web / WebJudge · Xue et al., 2025](#)

40

02B / 端到端评估（网页）

LLM / VLM 评判：评估证据

任务：将一个低于 \$20 的蓝色杯子加入购物车。不要购买。

观测证据：studio.shop / cart；蓝色杯子；\$18 · 数量 1；已在购物车。证据包括截图 + 动作历史。

LLM / VLM + 评分规则：✓ 蓝色杯子存在；✓ 价格低于 \$20；✓ 未显示购买。评分规则 → 证据 → 判定。

判定：通过。“可见的购物车符合所请求的标准。”

灵活的证据评判需要人工标签校准。

基准示例：WebVoyager · He 等, 2024；Online-Mind2Web / WebJudge · Xue 等, 2025；CUA-World-Long · Aggarwal 等, 2026。

LLM / VLM 评判：评估证据

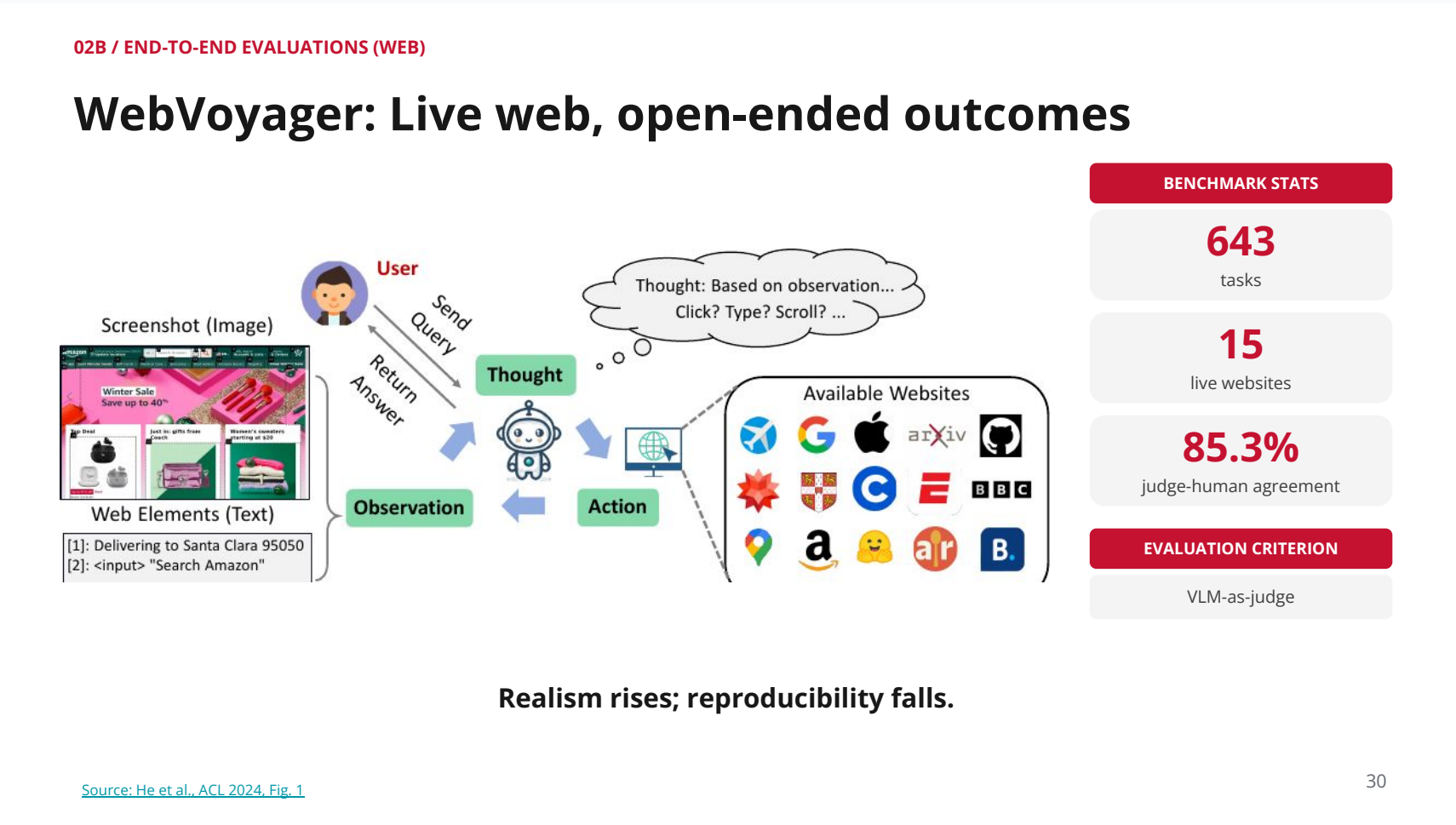
个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 29 | 原页 29

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



02B / 端到端评估（网页）

WebVoyager：开放式结果的实时网页任务

基准统计：643 个任务；15 个实时网站；评判者与人工的一致率为 85.3%。

评估标准：VLM 评判。

真实感上升，可复现性下降。

来源：He 等，ACL 2024，图 1

WebVoyager：开放式结果的实时网页任务

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

WebVoyager: Live web, open-ended outcomes

**TASK:** Find the most-starred GitHub project for climate-change data visualization.

**PUBLISHED TRACE / GITHUB SEARCH**

Historical trace, not current GitHub rankings

**VLM JUDGE INPUTS**

**Task + final answer  
+ last k screenshots**

Recorded answer  
resource-watch/  
resource-watch: 63 stars

Judge checks the evidence  
and returns **SUCCESS** or  
**NOT SUCCESS.**

He et al., ACL 2024 · Released GitHub--0 trace and automatic evaluator

43

02B / 端到端评估（网页）

WebVoyager：实时网站与开放式结果

任务：找出 GitHub 上最受欢迎的气候变化数据可视化项目。

已发布轨迹 / GitHub 搜索。

VLM 评判输入：任务 + 最终回答 + 最近 k 张截图；记录的回答：  
resource-watch/ resource-watch：63 颗星。

评判器检查证据并返回“成功”或“不成功”。

这是历史轨迹，不代表当前 GitHub 排名。

He 等，ACL 2024 · 已发布的 GitHub--0 轨迹与自动评估器

WebVoyager：实时网站与开放式结果示例

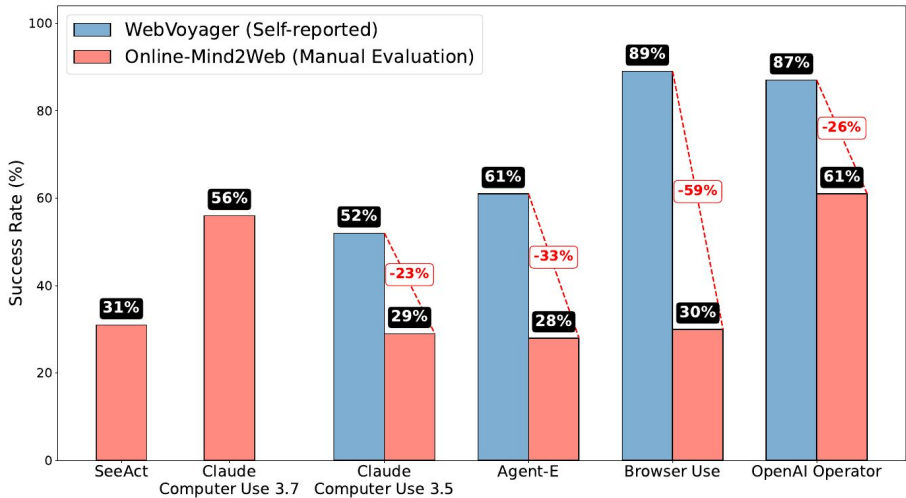
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

Online-Mind2Web: Live sites expose benchmark drift



BENCHMARK STATS

300 tasks

136 live websites

85.7% WebJudge-human agreement

EVALUATION CRITERION

Human + LLM judge

Live benchmarks drift—and judge quality becomes part of the score.

Source: Xue et al., COLM 2025, Fig. 1 (cropped)

02B / 端到端评估（网页）

Online-Mind2Web：实时网站带来基准漂移

基准统计：300 个任务；136 个实时网站；WebJudge 与人工的一致率为 85.7%。

评估标准：人工 + LLM 评判。

实时基准会发生漂移，而评判质量也会成为分数的一部分。

来源：Xue 等，COLM 2025，图 1（裁剪）

Online-Mind2Web：实时网站带来基准漂移

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02B / END-TO-END EVALUATIONS (WEB)

### Human review: Resolve ambiguous outcomes

TASK:

Add a blue mug under \$20 to the cart. Do not purchase.

REVIEW EVIDENCE

studio.shop / orders



Blue mug

\$18 · Qty 1

ORDER PLACED

INDEPENDENT REVIEW

Reviewer A

PASS

Reviewer B

FAIL

ADJUDICATION

FAIL

Order confirmation shows a forbidden purchase.

Resolve disagreement using evidence

Human review can score tasks or audit an automatic evaluator.

BENCHMARK EXAMPLES

[WebVoyager · human scoring & judge audit · He et al., 2024](#)

[Online-Mind2Web · human labels & review · Xue et al., 2025](#)

02B / 端到端评估（网页）

人工复核：解决结果模糊的情况

任务：将一个低于 \$20 的蓝色杯子加入购物车。不要购买。

复核证据：studio.shop / orders；蓝色杯子；\$18 · 数量 1；订单已提交。

独立复核：复核者 A 判通过；复核者 B 判失败。

裁决：失败。订单确认信息显示发生了禁止的购买。根据证据解决分歧。

人工复核可以给任务打分，也可以审计自动评估器。

基准示例：WebVoyager · 人工评分与评判审计 · He 等，2024；Online-Mind2Web · 人工标签与复核 · Xue 等，2025。

人工复核：解决结果模糊的情况

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 33 | 原页 33

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02 / BENCHMARKS

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

34

02 / 基准

基准

从预测单个动作走向完成延伸的工作流。

02A 静态评估：在固定观测上预测一个动作。

02B 端到端评估（网页）：通过浏览器交互完成任务。

02C 端到端评估（桌面、移动）：跨应用和系统状态完成任务。

02D 长时程计算机使用：在延伸的工作流中持续推进。

长时程评估将端到端评估扩展到网页和桌面。

基准：从预测动作到完成长工作流

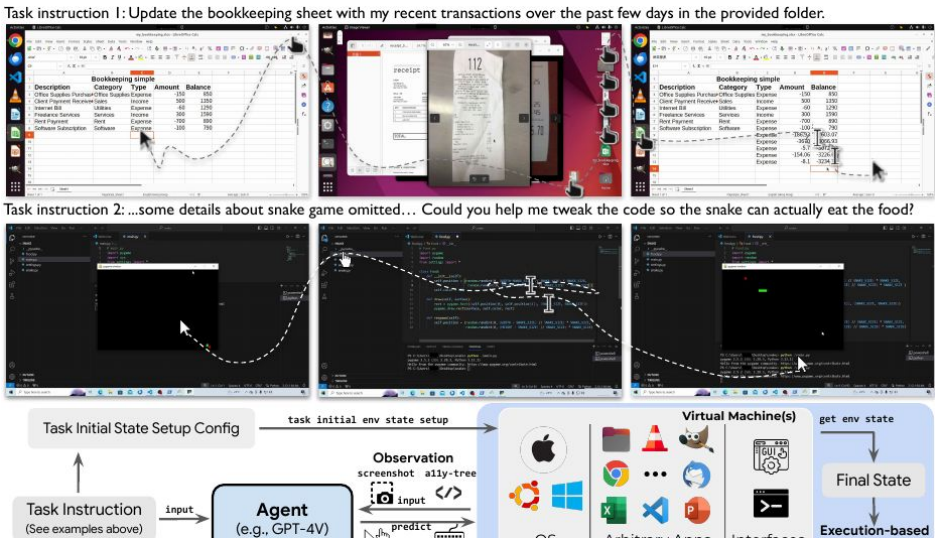
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

OSWorld: Real desktops, custom end-state checks



BENCHMARK STATS

369

tasks

9

applications

custom

task evaluators

EVALUATION CRITERION

Custom end-state checks

Every task ships with setup logic and application-specific state checks.

Source: Xie et al., NeurIPS 2024, Fig. 1

35

02C / 端到端评估（桌面、移动）

OSWorld：真实桌面与定制的最终状态检查

基准统计：369 个任务；9 个应用；每项任务都有定制评估器。

评估标准：定制的最终状态检查。

每项任务都附有设置逻辑和针对特定应用的状态检查。

来源：Xie 等，NeurIPS 2024，图 1

OSWorld：真实桌面与定制的最终状态检查

第 7 讲 · 计算机使用智能体

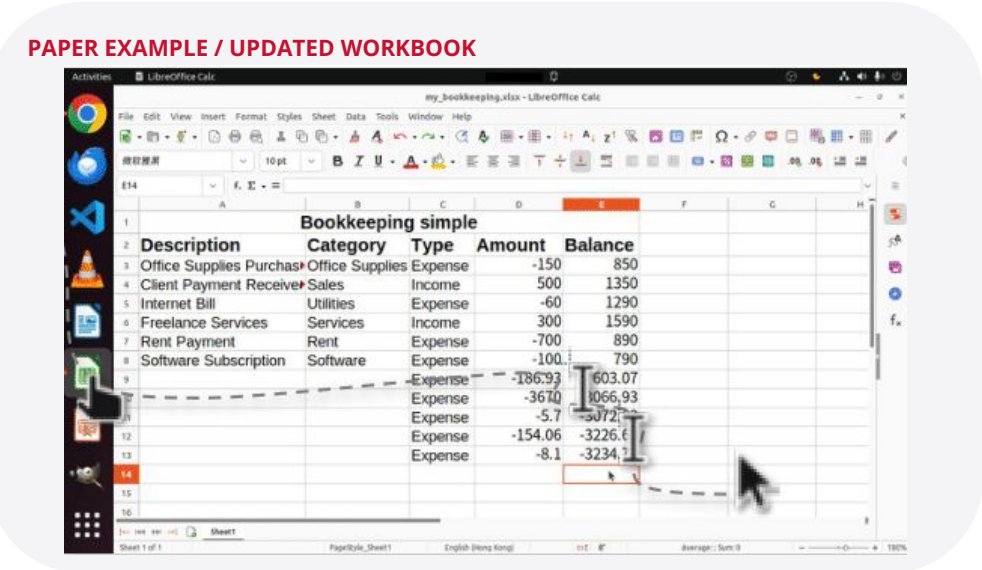
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

OSWorld: Real desktops, custom end-state checks

**TASK:** Update the bookkeeping sheet using the receipts in the provided folder.



REFERENCE FILE + RULES

```
compare_table(
  saved_workbook,
  gold_workbook,
  rules
)
```

Reference checks  
A1:E8 unchanged  
C9:C13 expense labels  
Amounts + running balances

D9 ≈ -186.93  
E9 ≈ 603.07

Xie et al., NeurIPS 2024 · Fig. 1 + bookkeeping task 8e116af7

02C / 端到端评估（桌面、移动）

OSWorld：真实桌面与定制的最终状态检查

任务：使用提供文件夹中的收据更新记账表格。

纸面示例 / 更新后的工作簿。

参考文件 + 规则：compare\_table(saved\_workbook, gold\_workbook, rules)。

参考检查：A1:E8 保持不变；C9:C13 为费用标签；金额 + 累计余额。

D9 ≈ -186.93；E9 ≈ 603.07。

Xie 等, NeurIPS 2024 · 图 1 + 记账任务 8e116af7

OSWorld：更新账务工作簿

第 7 讲 · 计算机使用智能体

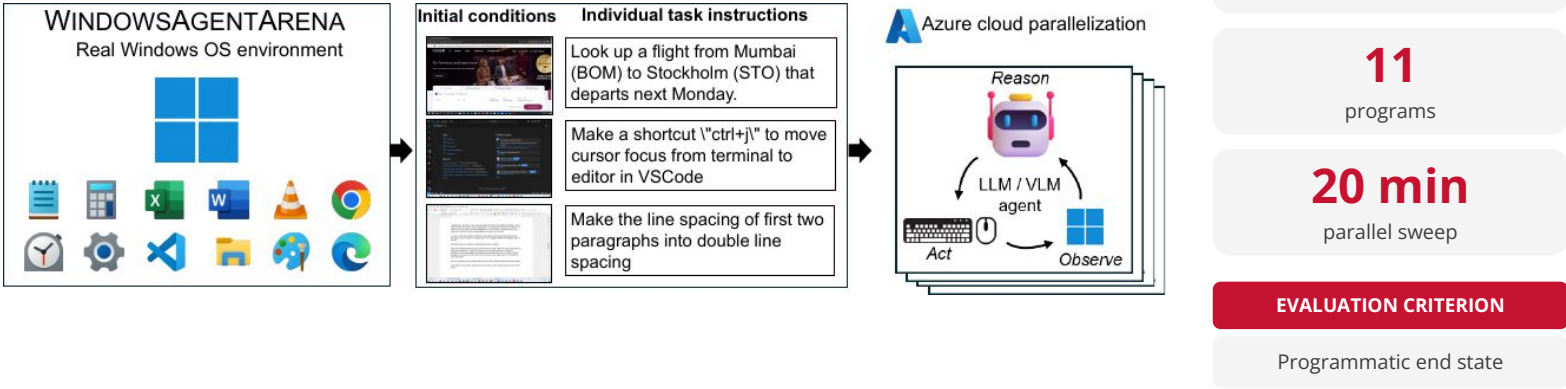
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

WindowsAgentArena: Windows at cloud scale

WINDOWSAGENTARENA  
Real Windows OS environment



Parallel VMs turn multi-day desktop evaluations into a cloud-scale sweep.

Source: Bonatti et al., 2024, Fig. 1

37

02C / 端到端评估（桌面、移动）

WindowsAgentArena：云规模的 Windows 评估

基准统计：154 个任务；11 个程序；20 分钟并行扫描。

评估标准：程序化最终状态。

并行虚拟机将原本耗时多日的桌面评估转化为云规模扫描。

来源：Bonatti 等，2024，图 1

WindowsAgentArena：云规模的 Windows 评估

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 37 | 原页 37

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02 / BENCHMARKS

Benchmarks

From predicting one action to completing extended workflows.

02A

Static evaluations

Predict an action on a fixed observation.

02B

End-to-end evaluations (web)

Complete a task through browser interaction.

02C

End-to-end evaluations (desktop, mobile)

Complete tasks across apps and system state.

02D

Long horizon computer use

Sustain progress through extended workflows.

Long horizon extends end-to-end evaluation across both web and desktop.

38

02 / 基准

基准

从预测单个动作走向完成延伸的工作流。

02A 静态评估：在固定观测上预测一个动作。

02B 端到端评估（网页）：通过浏览器交互完成任务。

02C 端到端评估（桌面、移动）：跨应用和系统状态完成任务。

02D 长时程计算机使用：在延伸的工作流中持续推进。

长时程评估将端到端评估扩展到网页和桌面。

基准：从预测动作到完成长工作流

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

### Hybrid / trajectory: Check outcome and process

TASK:

Add a blue mug under \$20 to the cart. Do not purchase.

ACTION TRACE

1   click(mug)

2   click(add\_to\_cart)

3   click(place\_order)

TWO KINDS OF CHECK

STATE CHECK

✓ Correct item

PROCESS CHECK

✗ Purchase was made

FINAL VERDICT

FAIL

Outcome-only grading could miss the violation.

A plausible deliverable is not enough if the process violated the task.

BENCHMARK EXAMPLES

[WeaveBench · trajectory-aware judge · Li et al., 2026](#)

[OSWorld 2.0 · completion + separate safety audit, 2026](#)

55

02D / 长时程计算机使用

混合 / 轨迹评估：同时检查结果与过程

任务：将一个低于 \$20 的蓝色杯子加入购物车。不要购买。

动作轨迹：1 click(mug); 2 click(add\_to\_cart); 3 click(place\_order)。

两类检查：状态检查 ✓ 商品正确；过程检查 ✗ 发生了购买。

最终判定：失败。只按结果评分可能会漏掉这一违规行为。

如果过程违反了任务要求，仅有看似合理的交付物是不够的。

基准示例：WeaveBench · 轨迹感知评判 · Li 等, 2026; OSWorld 2.0 · 任务完成 + 单独的安全审计, 2026。

混合 / 轨迹评估：同时检查结果与过程

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 39 | 原页 39

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

Odysseys: Long-horizon tasks on the live web

Example Odysseys Task (8 rubrics)

Plan a trip to a Palm Springs wedding — search round-trip flights PIT→LAX & PIT→PSP, check drive time (9am–4pm), book hotel if needed, rent car, assess Soban & Holbox detours, build CryptPad itinerary.

Step 1

Search flights PIT → LAX & PSP

Step 2

Check drive time LAX → PS

Step 3

Find hotel near LAX

Step 4

Rent a car at airport

Step 5

Write itinerary on CryptPad

Score: 88%, 7/8 rubrics satisfied

|                                                             |                                                             |
|-------------------------------------------------------------|-------------------------------------------------------------|
| ✓ <b>R1:</b> PIT→LAX non-stop flights searched, best opened | ✓ <b>R5:</b> Car rental listing with vehicle type & rate    |
| ✓ <b>R2:</b> PIT→PSP flights searched, best opened          | ✗ <b>R6:</b> Soban & Holbox detour feasibility assessed     |
| ✓ <b>R3:</b> LAX→PSP drive time & 9–4pm window assessed     | ✓ <b>R7:</b> CryptPad day-by-day itinerary document created |
| ✓ <b>R4:</b> LAX vs PSP compared, airport recommended       | ✓ <b>R8:</b> Final integrated trip summary provided         |

BENCHMARK STATS

200

live-web tasks

6.1

rubric items per task (avg.)

100

steps · main evaluation cap

EVALUATION CRITERION

Rubric-based LLM judge

Grade progress across the workflow—not just a single final screenshot.

[Odysseys: Benchmarking Web Agents on Realistic Long Horizon Tasks.](#) Jang\*, Koh\*, Fried & Salakhutdinov, 2026.

40

02D / 长时程计算机使用

Odysseys：实时网页上的长时程任务

基准统计：200 个实时网页任务；每项任务平均 6.1 个评分规则条目；主要评估上限为 100 步。

评估标准：基于评分规则的 LLM 评判。

评估整个 workflow 中的进展，而不是只看最后一张截图。

Odysseys：Benchmarking Web Agents on Realistic Long Horizon Tasks。Jang\*、Koh\*、Fried 与 Salakhutdinov，2026。

Odysseys：实时网页上的长时程任务

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

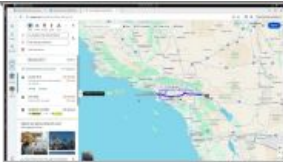
ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

## Odysseys: Score progress across the whole trip

TASK:

Plan a Palm Springs wedding trip, compare airports, and build an itinerary.



Flights

Drive window

Hotel if needed

Car rental

CryptPad itinerary

RUBRIC CHECKPOINTS / PUBLISHED ILLUSTRATION

|    |                                |   |    |                                    |   |
|----|--------------------------------|---|----|------------------------------------|---|
| R1 | Compare PIT-LAX flights        | ✓ | R5 | Open a priced car-rental option    | ✓ |
| R2 | Compare PIT-PSP flights        | ✓ | R6 | Evaluate both restaurant detours   | ✗ |
| R3 | Check the 9am-4pm drive window | ✓ | R7 | Create an editable daily itinerary | ✓ |
| R4 | Recommend LAX or PSP           | ✓ | R8 | Summarize the complete trip plan   | ✓ |

7 / 8 satisfied = 0.875 average

Perfect completion = 0.0

Jang et al., 2026 · Published trip-planning illustration; rubric text condensed

57

02D / 长时程计算机使用

Odysseys：沿整个行程评估进展

任务：规划一趟前往棕榈泉参加婚礼的行程，比较机场，并制定行程表。

行程阶段：航班；驾车时间窗口；必要时预订酒店；租车；CryptPad 行程表。

评分检查点 / 已发布示意：R1 比较 PIT-LAX 航班 ✓；R2 比较 PIT-PSP 航班 ✓；R3 检查上午 9 点至下午 4 点的驾车时间窗口 ✓；R4 推荐 LAX 或 PSP ✓；R5 打开一个标有价格的租车选项 ✓；R6 评估两条餐厅绕行路线 ✗；R7 创建可编辑的每日行程表 ✓；R8 总结完整旅行计划 ✓。

8 项中满足 7 项 = 平均分 0.875；完美完成 = 0.0。

Jang 等, 2026 · 已发布的行程规划示意图；评分规则文本经压缩

Odysseys：沿整个行程评估进展

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 41 | 原页 41

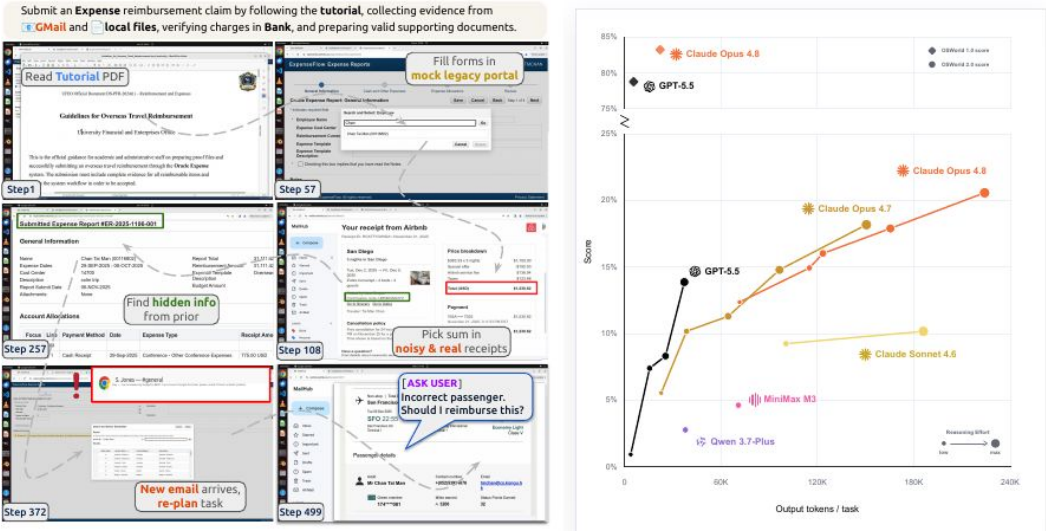
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

OSWorld 2.0: Hours-long workflows



The frontier shifts from clicks to hidden state, changing requirements, and verification.

Source: Yuan, Zhou, Xiong et al., 2026, Fig. 1

02D / 长时程计算机使用

OSWorld 2.0：持续数小时的工作流

基准统计：108 个工作流；人工完成时间中位数为 1.6 小时；智能体调用平均数为 318 次。

评估标准：部分检查 + 二元检查。

前沿从点击转向隐藏状态、不断变化的要求和验证。

来源：Yuan、Zhou、Xiong 等，2026，图 1

OSWorld 2.0：持续数小时的工作流

第 7 讲 · 计算机使用智能体

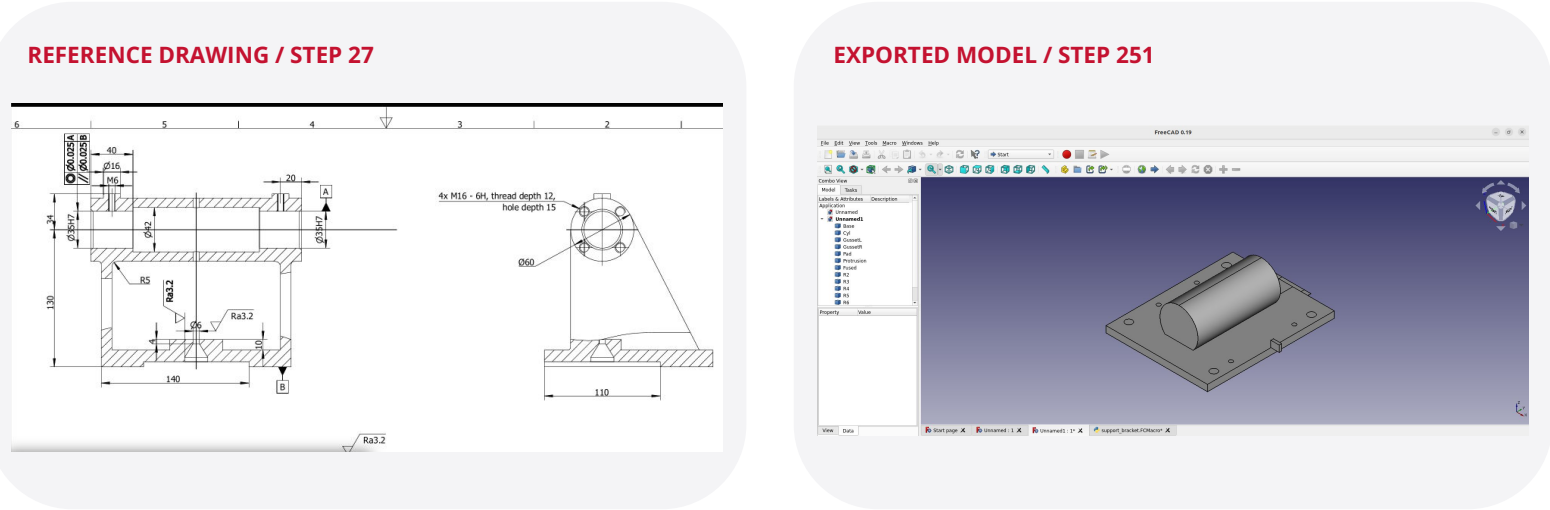
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

OSWorld 2.0: A file can exist and still be wrong

**TASK:** Build a FreeCAD support bracket and export a STEP model matching the drawing.



Yuan, Zhou, Xiong et al., 2026 · Official Task103 failure case (cropped)

02D / 长时程计算机使用

OSWorld 2.0：文件存在仍可能不正确

任务：构建一个 FreeCAD 支撑架，并导出与工程图匹配的 STEP 模型。

参考工程图 / STEP 27；导出模型 / STEP 251。

Yuan、Zhou、Xiong 等，2026 · 官方 Task103 失败案例（裁剪）

OSWorld 2.0：文件存在仍可能不正确

第 7 讲 · 计算机使用智能体

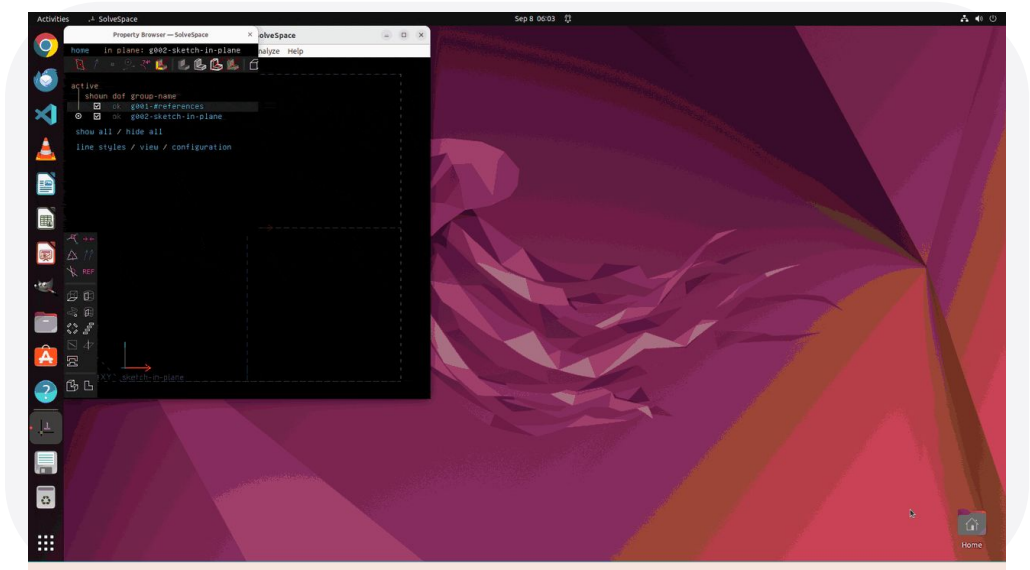
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

OSWorld 2.0: Modeling a part in SolveSpace

**TASK:** Recreate a mechanical part from the reference video's dimensions and steps, then save it as a SolveSpace file.



[Koh, Sep 2026 · What's the Point of Computer Use Agents? · GPT-6 Astra demo \(30x\)](#)

OSWorld 2.0：在 SolveSpace 中建模零件

02D / 长时程计算机使用

OSWorld 2.0：在 SolveSpace 中建模零件

任务：根据参考视频中的尺寸和步骤复现机械零件，然后将其保存为 SolveSpace 文件。

Koh, 2026 年 9 月 · What’ s the Point of Computer Use Agents? · GPT-6 Astra 演示 (30 倍速)

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

# CUA-World-Long: One hard task per software

Healthcare

Engineering

Design

Enterprise

Science

Finance

GeoScience

Education

BENCHMARK STATS

200  
long tasks

>500  
steps often needed

8  
quality criteria

EVALUATION CRITERION

Checklist VLM verifier

One long task per application exposes failures across real professional software.

Source: Aggarwal et al., 2026, Fig. 1 (cropped)

45

02D / 长时程计算机使用

CUA-World-Long：每种软件设置一项困难任务

基准统计：200 项长任务；通常需要超过 500 步；8 项质量标准。

评估标准：清单式 VLM 验证器。

每个应用中的一项长任务可以揭示真实专业软件中的各类失败。

来源：Aggarwal 等，2026，图 1（裁剪）

CUA-World-Long：每种软件设置一项困难任务

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 45 | 原页 45

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



建模

建模

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

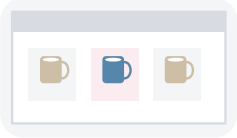
03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP 0

"Buy a blue mug < \$20"

GOAL



SCREENSHOT 0

64

03 / 建模

VLM 根据视觉历史预测动作

第 0 步的上下文：目标 “购买一个低于 \$20 的蓝色杯子” + 截图 0。

VLM 根据视觉历史预测动作

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP 0

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

$$\pi_{\theta}(a_0 \mid \text{goal}, o_0)$$

ACTION 0

click(x<sub>0</sub>, y<sub>0</sub>)

65

VLM 根据视觉历史预测动作

03 / 建模

VLM 根据视觉历史预测动作

第 0 步的上下文：目标 “购买一个低于 \$20 的蓝色杯子” + 截图 0。

动作 0: click(x<sub>0</sub>, y<sub>0</sub>)。

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

APPEND ACTION 0 TO THE TRAJECTORY

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

click(x<sub>0</sub>, y<sub>0</sub>)

ACTION 0

$\pi_{\theta}(a_0 \mid \text{goal}, o_0)$

ACTION 0

click(x<sub>0</sub>, y<sub>0</sub>)

66

03 / 建模

VLM 根据视觉历史预测动作

将动作 0 添加到轨迹。

目标 + 截图 0 + 动作 0: click(x<sub>0</sub>, y<sub>0</sub>)。

模型预测的动作 0: click(x<sub>0</sub>, y<sub>0</sub>)。

将动作 0 添加到轨迹

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP 1

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

click(x<sub>0</sub>, y<sub>0</sub>)

ACTION 0

studio.shop / blue-mug

Blue mug

\$18 · Qty 1

Add to cart

SCREENSHOT 1

$\pi_{\theta}(a_1 \mid \text{goal}, o_0, a_0, o_1)$

67

VLM 根据视觉历史预测动作：第 1 步

03 / 建模

VLM 根据视觉历史预测动作

第 1 步的上下文：目标 “购买一个低于 \$20 的蓝色杯子” + 截图 0 + 动作 0 + 截图 1。

截图 1 中显示 studio.shop / blue-mug；蓝色杯子；\$18 · 数量 1；添加到购物车。

动作 0: click(x<sub>0</sub>, y<sub>0</sub>)。

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP 1

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

click(x<sub>0</sub>, y<sub>0</sub>)

ACTION 0

studio.shop / blue-mug

Blue mug

\$18 · Qty 1

Add to cart

SCREENSHOT 1

$$\pi_{\theta}(a_1 \mid \text{goal}, o_0, a_0, o_1)$$

ACTION 1

click(x<sub>1</sub>, y<sub>1</sub>)

68

VLM 根据视觉历史预测动作：第 1 步

03 / 建模

VLM 根据视觉历史预测动作

第 1 步的上下文：目标 “购买一个低于 \$20 的蓝色杯子” + 截图 0 + 动作 0 + 截图 1。

截图 1：studio.shop / blue-mug；蓝色杯子；\$18 · 数量 1；添加到购物车。

动作 0：click(x<sub>0</sub>, y<sub>0</sub>)。

动作 1：click(x<sub>1</sub>, y<sub>1</sub>)。

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

APPEND ACTION 1 TO THE TRAJECTORY

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

ACTION 0

click(x<sub>0</sub>, y<sub>0</sub>)

SCREENSHOT 1

ACTION 1

click(x<sub>1</sub>, y<sub>1</sub>)

$$\pi_{\theta}(a_1 \mid \text{goal}, o_0, a_0, o_1)$$

ACTION 1

click(x<sub>1</sub>, y<sub>1</sub>)

69

03 / 建模

VLM 根据视觉历史预测动作

将动作 1 添加到轨迹。

目标 + 截图 0 + 动作 0 + 截图 1 + 动作 1。

动作 1: click(x<sub>1</sub>, y<sub>1</sub>)。

截图 1 显示蓝色杯子 \$18 · 数量 1, 操作为添加到购物车。

将动作 1 添加到轨迹

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 52 | 原页 52

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP t

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

ACTION 0

click(x<sub>0</sub>, y<sub>0</sub>)

SCREENSHOT 1

ACTION 1

click(x<sub>1</sub>, y<sub>1</sub>)

...

SCREENSHOT t

$$\pi_{\theta}(a_t \mid \text{goal}, o_{\leq t}, a_{< t})$$

70

VLM 根据视觉历史预测动作：第 t 步

03 / 建模

VLM 根据视觉历史预测动作

第 t 步的上下文：目标 “购买一个低于 \$20 的蓝色杯子” + 截图 0 + 动作 0 + 截图 1 + 动作 1 + 截图 t。

截图 1：studio.shop / blue-mug；蓝色杯子；\$18 · 数量 1；添加到购物车。

截图 t：studio.shop / cart；蓝色杯子；\$18 · 数量 1；结账。

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 53 | 原页 53

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP t

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

ACTION 0

click(x0, y0)

SCREENSHOT 1

ACTION 1

click(x1, y1)

...

SCREENSHOT t

$$\pi_{\theta}(a_t \mid \text{goal}, o_{\leq t}, a_{< t})$$

ACTION t

click / type / tool / stop

71

03 / 建模

VLM 根据视觉历史预测动作

第 t 步的上下文包括目标、截图 0、动作 0、截图 1、动作 1 和截图 t。

动作 t: 点击 / 输入 / 工具 / 停止。

截图 t 显示 studio.shop / cart; 蓝色杯子; \$18 · 数量 1; 结账。

VLM 根据视觉历史预测动作：后续动作

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 54 | 原页 54

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

# A VLM predicts actions from visual history

CONTEXT AT STEP  $t$

"Buy a blue mug < \$20"

GOAL

SCREENSHOT 0

ACTION 0

click(x<sub>0</sub>, y<sub>0</sub>)

SCREENSHOT 1

ACTION 1

click(x<sub>1</sub>, y<sub>1</sub>)

...

SCREENSHOT t

$$\pi_{\theta}(a_t \mid \text{goal}, o_{\leq t}, a_{< t})$$

ACTION t

click / type / tool / stop

Interleaved visual and action history lets the policy condition on past progress.

Generic VLM policy · o denotes observations, a actions · tokenization and history compression vary by system

71

03 / 建模

VLM 根据视觉历史预测动作

第  $t$  步的上下文：目标、截图 0、动作 0、截图 1、动作 1、截图  $t$ ；之后模型预测动作  $t$ ：点击 / 输入 / 工具 / 停止。

交错的视觉与动作历史让策略能够依据过去的进展作出判断。

通用 VLM 策略 ·  $o$  表示观测， $a$  表示动作 · 不同系统的词元化与历史压缩方式各不相同。

交错的视觉与动作历史

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 55 | 原页 55

第 7 讲 · 计算机使用智能体

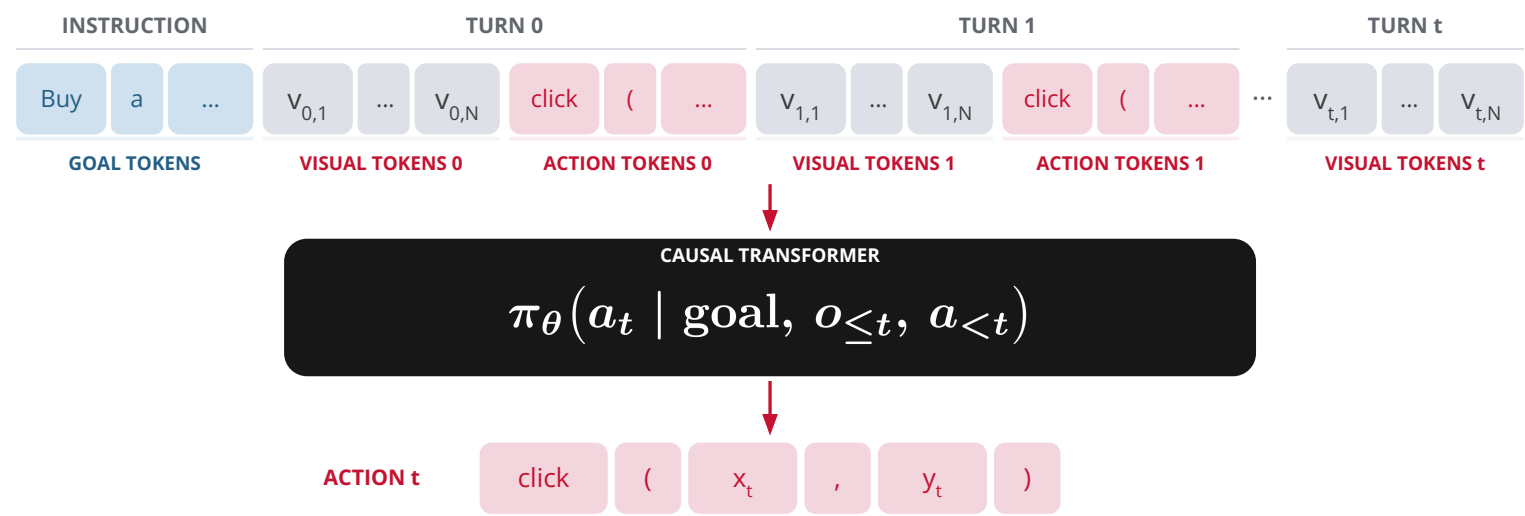
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

A VLM predicts actions from interleaved tokens

TOKENIZED CONTEXT AT STEP t



Text → tokenizer + embeddings; screenshots → vision encoder + projector.

Illustrative token splits; visual tokens are continuous embeddings, not vocabulary IDs. Token counts / encoders vary.

72

03 / 建模

VLM 根据交错词元预测动作

第  $t$  步的词元化上下文：指令 → 轮次 0 → 轮次 1 → 轮次  $t$ 。

指令词元：Buy a ...；视觉词元： $v_{0,1} \dots v_{0,N}$ ；动作词元：click (；视觉词元： $v_{1,1} \dots v_{1,N}$ ；动作词元：click (；视觉词元： $v_{t,1} \dots v_{t,N}$ 。

目标词元 · 视觉词元 0 · 动作词元 0 · 视觉词元 1 · 动作词元 1 · 视觉词元  $t$  → 因果 Transformer → 动作  $t$ ：click (  $x_t$  ,  $y_t$  )。

文本 → 词元化器 + 嵌入；截图 → 视觉编码器 + 投影器。

词元切分仅作示意；视觉词元是连续嵌入，不是词表 ID。词元数量 / 编码器会随系统而异。

VLM 根据交错词元预测动作

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

03 / MODELING

### Frontier CUAs: Similar observations, different actions

| MODEL            | ACTION INTERFACE                                      | OUTPUT FORMAT / SCHEMATIC EXAMPLE                                                                  |
|------------------|-------------------------------------------------------|----------------------------------------------------------------------------------------------------|
| GPT-6 Astra      | Python / PyAutoGUI or native computer tool            | <code>pyautogui.click(x, y)</code><br><code>computer_call: actions=[...]</code>                    |
| Fable / Opus 5   | Native computer tool calls (one or more per response) | <code>tool_use: left_click</code><br><code>input: {coordinate: [x, y]}</code>                      |
| Gemini 3.8 Flash | Computer-use function calls (normalized coordinates)  | <code>functionCall: click</code><br><code>args: {x, y} ∈ 0–999</code>                              |
| Qwen 3.8         | Function calls (app-defined GUI schema)               | <code>&lt;tool_call&gt;&lt;function=...&gt;...</code><br>XML-style tool serialization <sup>1</sup> |
| Muse Spark       | Scripts + direct GUI actions including action batches | Tool / script output <sup>2</sup><br>Exact GUI schema is app-specific                              |
| Kimi K3          | Function calls (app-defined GUI schema)               | <code>tool_calls[ ].function</code><br>name + JSON arguments                                       |

Output format and schema differs between models (reproducibility is a nightmare).

Sources: [OpenAI](#) · [Anthropic](#) · [Google](#) · [Qwen](#)<sup>1</sup> · [Meta](#)<sup>2</sup> · [Moonshot](#) | <sup>1</sup>27B template; <sup>2</sup>1.1 report | Sep 2026

73

前沿 CUA：相似观测，不同动作

03 / 建模

前沿 CUA：相似观测，不同动作

模型 | 动作接口 | 输出格式 / 示意样例

GPT-6 Astra | Python / PyAutoGUI 或原生计算机工具 | `pyautogui.click(x, y)`;  
`computer_call: actions=[...]`

Fable / Opus 5 | 原生计算机工具调用（每次响应一个或多个） | `tool_use: left_click`;  
`input: {coordinate: [x, y]}`

Gemini 3.8 Flash | 计算机使用函数调用（归一化坐标） | `functionCall: click`;  
`args: {x, y} ∈ 0–999`

Qwen 3.8 | 函数调用（应用自定义 GUI 模式） | `<tool_call> <function=...> ...`;  
XML 风格工具序列化<sup>1</sup>

Muse Spark | 脚本 + 直接 GUI 动作（包括动作批次） | 工具 / 脚本输出<sup>2</sup>；精确  
GUI 模式由应用决定

Kimi K3 | 函数调用（应用自定义 GUI 模式） | `tool_calls[ ].function` name +  
JSON 参数

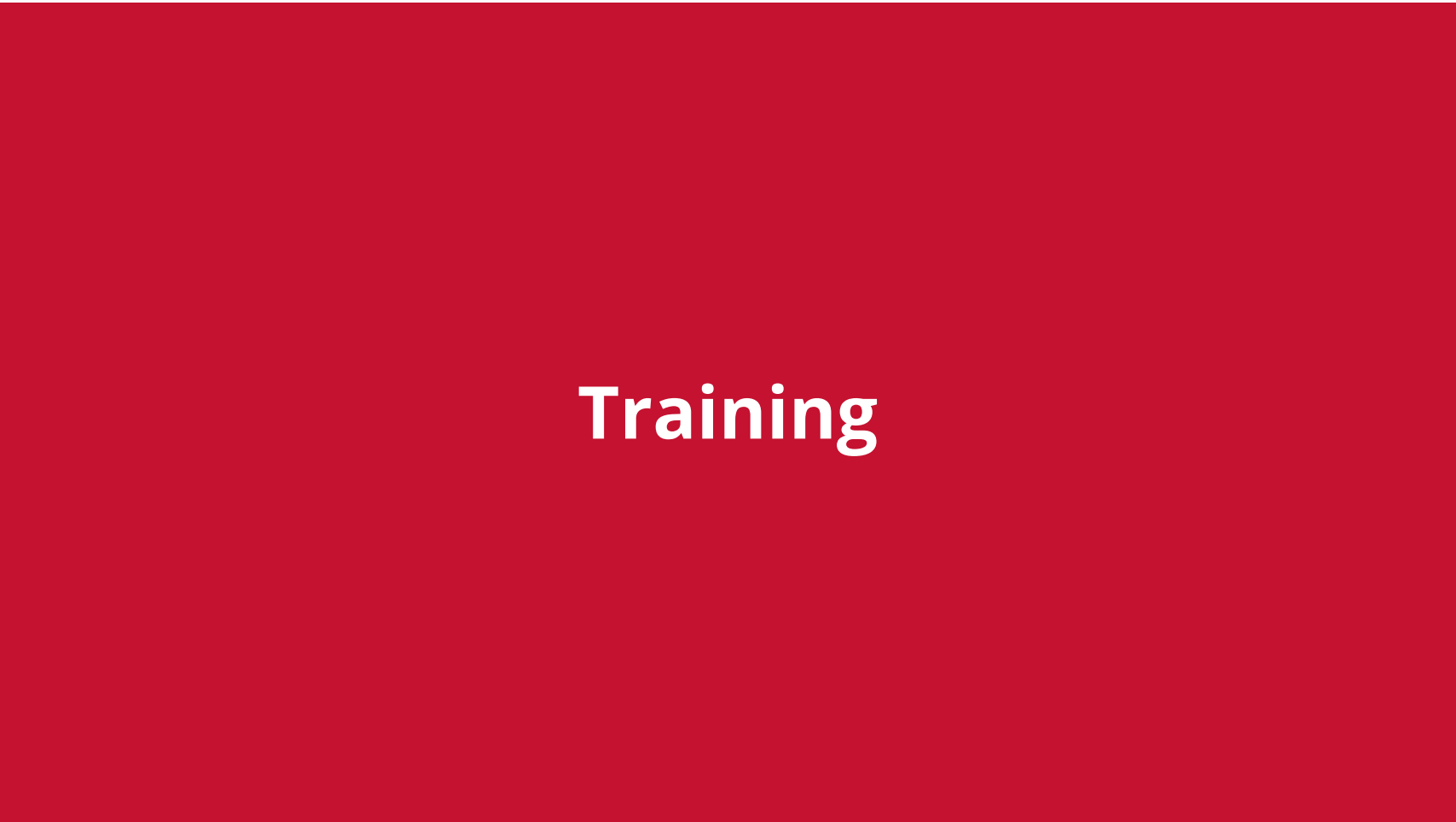
不同模型的输出格式和模式各不相同（可复现性会变得极其困难）。

来源：OpenAI · Anthropic · Google · Qwen<sup>1</sup> · Meta<sup>2</sup> · Moonshot | <sup>1</sup>27B 模板  
； <sup>2</sup>1.1 报告 | 2026 年 9 月

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



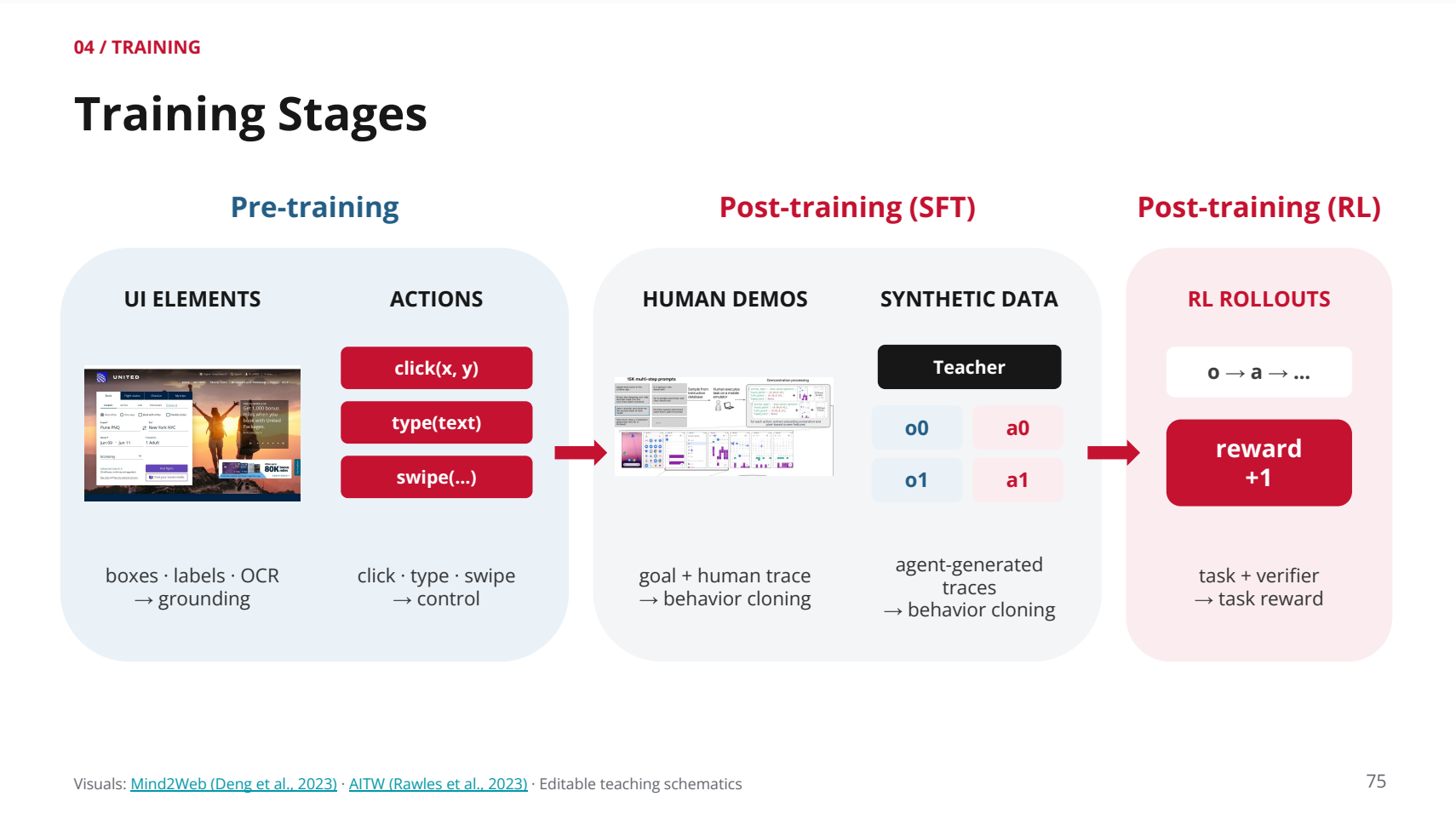
训练

训练

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



第 7 讲 · 计算机使用智能体

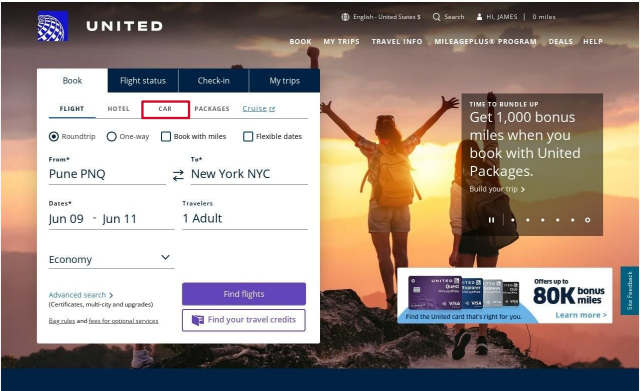
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

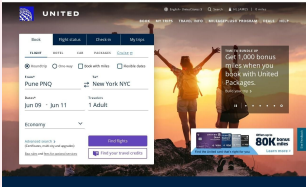
Pre-training: UI elements and actions

UI ELEMENTS / GROUNDING



“CAR” ↔ element box / screen coordinates

ACTION PREDICTION



CLICK

element: bookCarTab  
value: none

Other labels: type(text), swipe(start, end)

Learn what is on the screen, and how to address it.

[Deng et al., 2023 · Mind2Web training example](#); [Wu et al., 2025 · OS-Atlas](#)

04 / 训练

预训练：UI 元素与动作

UI 元素 / 定位： “CAR” ↔ 元素框 / 屏幕坐标。

动作预测：点击；element: bookCarTab；value: none。

其他标注：type(text)、swipe(start, end)。

学习屏幕上有什么，以及如何指向它。

Deng 等, 2023 · Mind2Web 训练样例；Wu 等, 2025 · OS-Atlas

预训练：UI 元素与动作

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

Post-training (SFT): Imitate useful trajectories

HUMAN DATA

“find a birria taco recipe I could make in under 3 hours”

go to: foodnetwork.com

thought: “I need to go to the food network website”

action: goto(“foodnetwork.com”)

thought: “I need to search ‘tacos birria’. I will click on the search bar”

action: click(80.5, 8.2)

search: tacos birria

thought: “The search bar is selected. I will type ‘tacos birria’.”

action: type(“tacos birria”)

MolmoWebMix (example above)  
36K human + 105K synthetic task trajectories

AgentNet  
22,625 released task trajectories · 3 OS

SYNTHETIC DATA

AxTree teacher → Roll out → Filter

“find a birria taco recipe I could make in under 3 hours”

go to: foodnetwork.com

thought: “I need to go to the food network website”

action: goto(“foodnetwork.com”)

thought: “I need to search ‘tacos birria’. I will click on the search bar”

action: click(80.5, 8.2)

search: tacos birria

thought: “The search bar is selected. I will type ‘tacos birria’.”

action: type(“tacos birria”)

Record format (Fig. 2) · 105K synthetic tasks

Retain a useful trajectory → SFT

Same next-action loss; demonstrations can come from humans or agents.

[Gupta et al., 2026 · MolmoWeb Fig. 2 excerpt: Table 2; Wang et al., 2025 · OpenCUA Table 2; §2.3](#)

04 / 训练

后训练（SFT）：模仿有用轨迹

人类数据：MolmoWebMix（上方示例），36K 条人类任务轨迹 + 105K 条合成任务轨迹；AgentNet，22,625 条已发布任务轨迹 · 3 种操作系统。

合成数据：可访问性树（AxTree）教师 → 轨迹采样 → 筛选；记录格式（图 2） · 105K 个合成任务。

保留有用轨迹 → SFT。

下一动作损失相同；示范可以由人类或智能体生成。

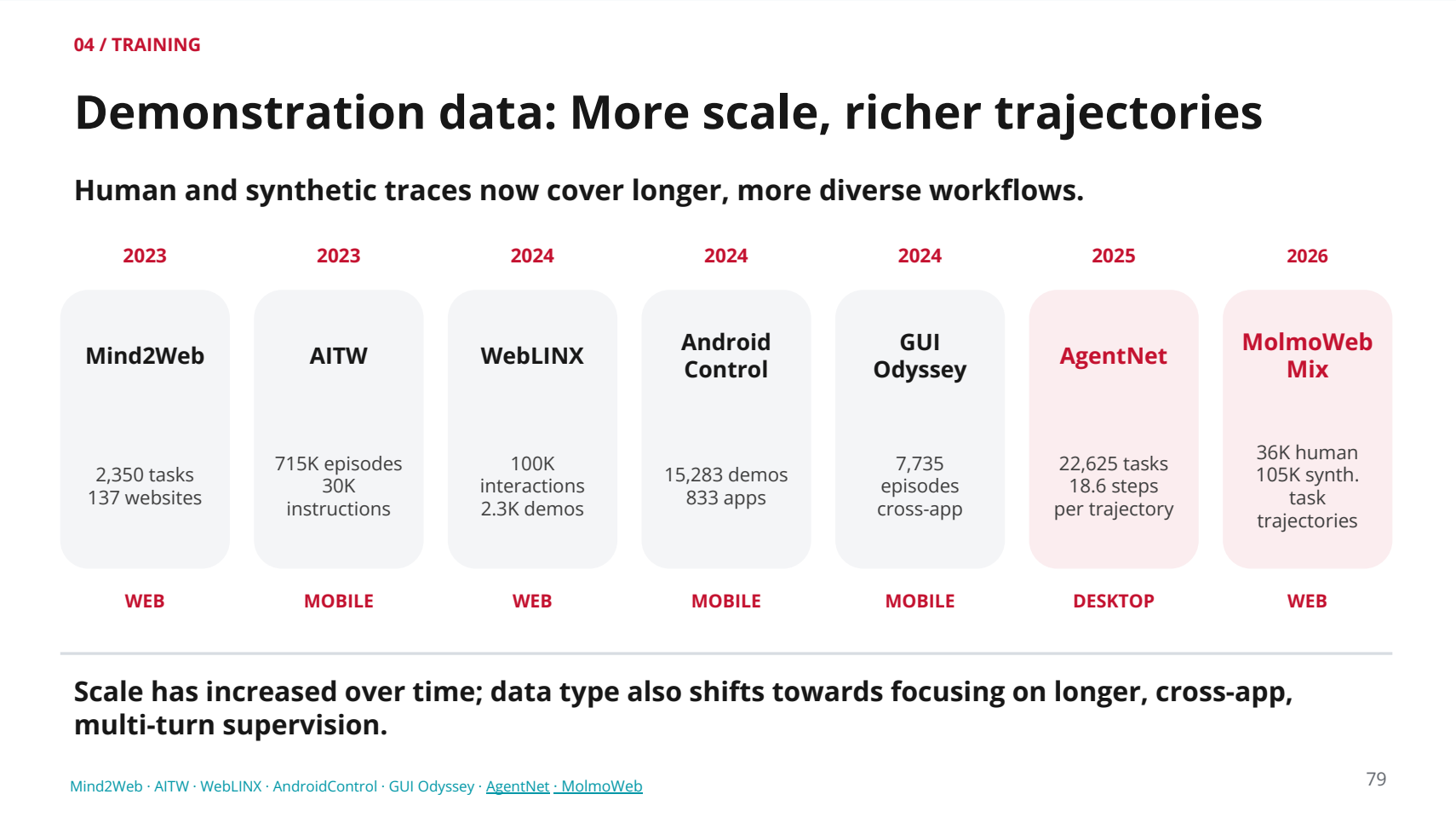
Gupta 等, 2026 · MolmoWeb 图 2 摘录、表 2；Wang 等, 2025 · OpenCUA 表 2、第 2.3 节

后训练（SFT）：模仿有用轨迹

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



示范数据：规模更大，轨迹更丰富

04 / 训练

示范数据：规模更大，轨迹更丰富

人类与合成轨迹如今覆盖更长、更多样的工作流。

2023 Mind2Web：2,350 个任务；137 个网站。

2023 AITW：715K 个回合；30K 条指令。

2024 WebLINX：100K 次交互；2.3K 个示范。

2024 AndroidControl：15,283 个示范；833 个应用。

2024 GUI Odyssey：7,735 个回合；跨应用。

2025 AgentNet：22,625 个任务；每条轨迹 18.6 步。

2026 MolmoWebMix：36K 条人类轨迹；105K 条合成任务轨迹。

平台：网页 · 移动端 · 网页 · 移动端 · 移动端 · 桌面端 · 网页。

随着时间推移，数据规模扩大；数据类型也逐渐转向更长、跨应用、多轮的监督。

Mind2Web · AITW · WebLINX · AndroidControl · GUI Odyssey · AgentNet · MolmoWeb

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

# Post-training (RL): Learn from scored rollouts

Goal: Find a blue mug, check delivery, and add it to the cart.

POLICY ACTION 0

studio.shop / mugs

STUDIO / Everyday objects

Cream mug  
\$16

Blue mug  
\$18

Blue set  
\$28

POLICY ACTION 1

studio.shop / blue-mug

STUDIO / Everyday objects

Blue mug  
\$18 · Quantity 1

Add to cart

POLICY ACTION 2

studio.shop / blue-mug

Delivery estimate

Postal code

15213

...

REWARD

1.0

Task success

click the blue mug

scroll down

type "15213"

Update the policy using trajectory rewards

Generate experience → score the outcome → improve the policy.

[Illustrative rollout · CUA-Gym / Gym-Anything: Interactive tasks and verifiers](#)

78

04 / 训练

后训练（RL）：从评分轨迹中学习

目标：找到蓝色杯子、查询配送情况，并将它加入购物车。

策略动作 0：点击蓝色杯子。策略动作 1：向下滚动。策略动作 2：输入“15213”。

页面与状态：STUDIO / 日常用品；奶油色杯子 \$16、蓝色杯子 \$18、蓝色套装 \$28；蓝色杯子 \$18 · 数量 1；预计送达时间；邮政编码 15213。

奖励 1.0：任务成功。

使用轨迹奖励更新策略。

生成经验 → 评估结果 → 改进策略。

示意轨迹 · CUA-Gym / Gym-Anything：交互任务与验证器

后训练（RL）：从评分轨迹中学习

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

# RL Training

TASK:

Add one blue mug under \$20 to the cart; do not check out.

studio.shop / mugs

STUDIO / Everyday objects

Cream mug  
\$16

Blue mug  
\$18

Blue set  
\$28

history + screen

ACTION 0

click product

studio.shop / blue-mug

STUDIO / Everyday objects

Blue mug  
\$18 · Quantity 1

Add to cart

history + screen

ACTION 1

add to cart

studio.shop / cart

Your cart

Blue mug  
Qty 1  
\$18

Subtotal \$18

Checkout

history + screen

ACTION 2

verify / stop

SFT learns next action prediction; RL evaluates trajectories based on outcomes.

Teaching example · screenshots, actions, and optional rationale labels are interleaved

80

04 / 训练

RL 训练

任务：将一个低于 \$20 的蓝色杯子加入购物车；不要结账。

状态序列：studio.shop / mugs → studio.shop / blue-mug → studio.shop / cart。购物车显示蓝色杯子 \$18、数量 1、Subtotal \$18。

历史 + 屏幕 → 动作 0：点击商品 → 历史 + 屏幕 → 动作 1：添加到购物车 → 历史 + 屏幕 → 动作 2：验证 / 停止。

SFT 学习下一动作预测；RL 根据结果评估轨迹。

教学示例 · 截图、动作和可选理由标签交错排列

RL 训练

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 64 | 原页 64

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

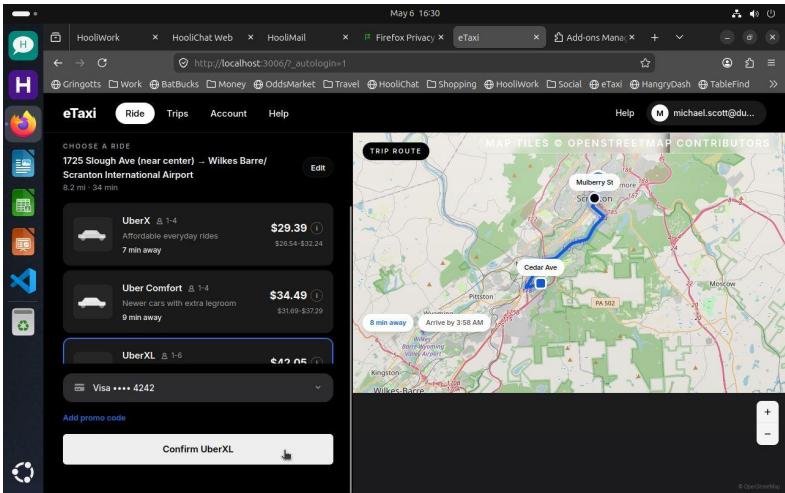
ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

We need simulated environments for RL

TASK:

Book an airport ride for my upcoming Jamaica flight.



REAL MONEY

Repeated attempts can create duplicate paid bookings.

REAL PEOPLE

A booking can dispatch a driver; exploration affects other people.

NO CLEAN RESET

Cancellation may incur a fee; spent time cannot be restored.

Explore in resettable replicas, not repeated real world interactions.

[MyPCBench · Jang et al., 2026 · Task bounded\\_action-f041 · Published trajectory, step 53](#)

04 / 训练

我们需要用于 RL 的模拟环境

任务：为我即将搭乘的牙买加航班预订机场接送。

真实金钱：反复尝试可能会产生重复的付费预订。

真实的人：预订可能派遣司机；探索会影响其他人。

无法干净地重置：取消可能产生费用；已经花掉的时间无法恢复。

应在可重置的副本中探索，而不是反复与真实世界交互。

MyPCBench · Jang 等，2026 · 任务 bounded\_action-f041 · 已发布轨迹，第 53 步

我们需要用于 RL 的模拟环境

第 7 讲 · 计算机使用智能体

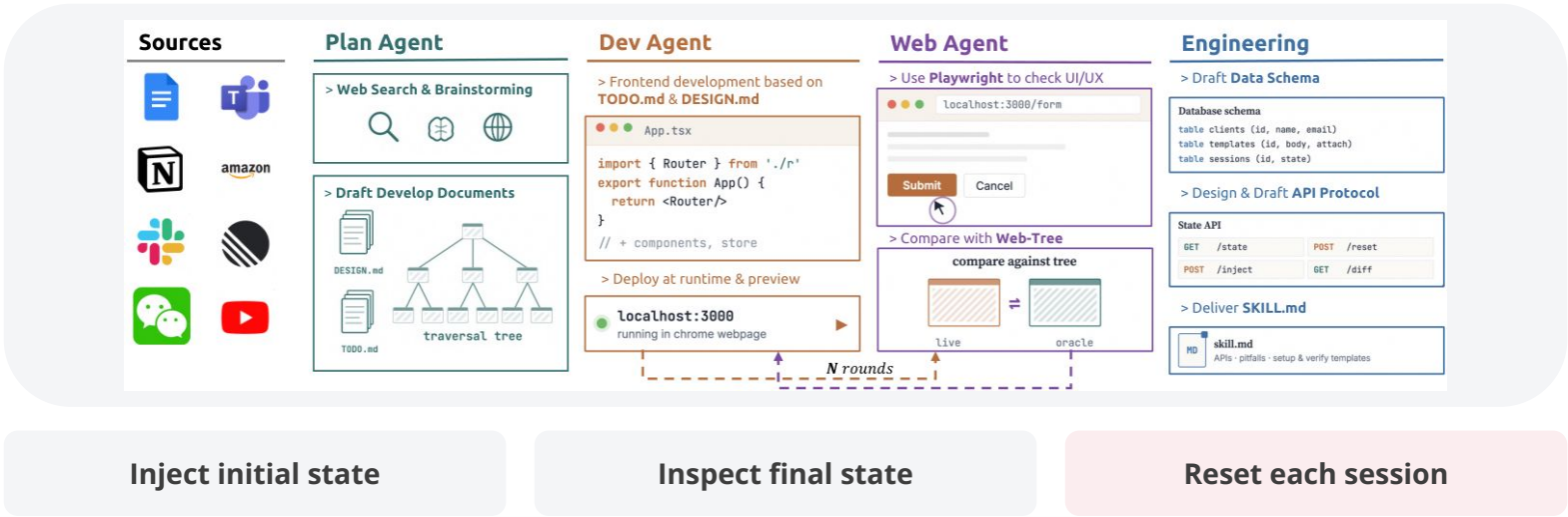
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

CUA-Gym: Build reusable, resettable mock applications

94 MOCK WEB APPS / PLAN → IMPLEMENT → TEST



One mock app supports many tasks through controlled, isolated state.

[Wang et al., 2026 · CUA-Gym Fig. 2: §2.2](#)

83

04 / 训练

CUA-Gym：构建可复用且可重置的模拟应用

94 个模拟网页应用 / 规划 → 实现 → 测试。

注入初始状态；检查最终状态；重置每次会话。

同一个模拟应用可通过受控、隔离的状态支持多个任务。

Wang 等, 2026 · CUA-Gym 图 2、第 2.2 节

CUA-Gym：构建可复用且可重置的模拟应用

第 7 讲 · 计算机使用智能体

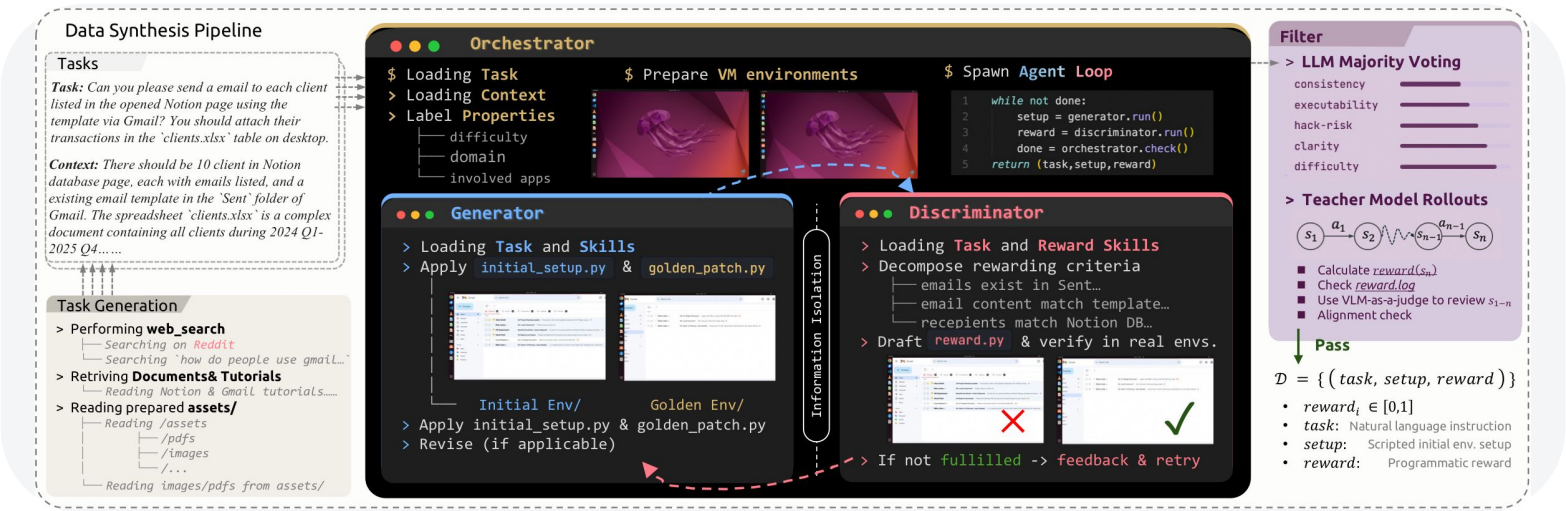
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

CUA-Gym: Co-generate tasks, states, and rewards

~32K VERIFIED TUPLES / 110 ENVIRONMENTS



Generate the world and its reward independently—then test that they agree.

[Wang et al., 2026 · CUA-Gym Fig. 1: §2.1](#)

04 / 训练

CUA-Gym：共同生成任务、状态与奖励

约 32K 个已验证元组 / 110 个环境。

独立生成世界与奖励，再测试二者是否一致。

Wang 等，2026 · CUA-Gym 图 1、第 2.1 节

CUA-Gym：共同生成任务、状态与奖励

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

# CUA-Gym: How much training data?

94 MOCK WEB APPS + 16 DESKTOP APPS

Computer & Mathematical · DevOps

aws

AWS

Azure

Azure

GCP

GCP

Vercel

Vercel

Cloud...

Cloud...

CircleCI

CircleCI

Analytics & Monitoring

Sentry

Sentry

Datadog

Datadog

Mixpa...

Mixpa...

Analyti...

Analyti...

Tableau

Tableau

W&B

W&B

Office & Admin · Communication

Slack

Slack

Teams

Teams

Discord

Discord

WeChat

WeChat

DingTalk

DingTalk

Feishu

Feishu

Office & Admin · Project Mgmt

Asana

Asana

Jira

Jira

Trello

Trello

Notion

Notion

Monday

Monday

Linear

Linear

Business · CRM & Finance

Salesf...

Salesf...

HubSp...

HubSp...

Quick...

Quick...

Stripe

Stripe

SAP

SAP

Business · HR & Recruiting

Bambo...

Bambo...

Green...

Green...

Gusto

Gusto

Lattice

Lattice

Workday

Workday

Sales · E-Commerce

Amazon

Amazon

Shopify

Shopify

eBay

eBay

Etsy

Etsy

Arts & Media · Content

YouTube

YouTube

Pintere...

Pintere...

Drive

Drive

Docs

Docs

Sheets

Sheets

Healthcare · Legal · Realty

Epic

Epic

Clio

Clio

Zillow

Zillow

DocuS...

DocuS...

Chinese / International

12306

12306

Weibo

Weibo

RED

RED

知乎

知乎

Desktop Applications

Calc

Calc

Writer

Writer

Impress

Impress

GIMP

GIMP

VSCode

VSCode

Chrome

Chrome

VLC

VLC

PDF

PDF

Grafana

Grafana

drawio

drawio

Penpot

Penpot

Excalid...

Excalid...

Overleaf

Overleaf

OpenS...

OpenS...

Blender

Blender

110

environments

32,112

verified RLVR tuples

38%

cross-application tasks

A broad environment pool supports many independently verified tasks.

Wang et al., 2026 · CUA-Gym Fig. 8(a); §2.3

84

CUA-Gym：需要多少训练数据？

04 / 训练

CUA-Gym：需要多少训练数据？

94 个模拟网页应用 + 16 个桌面应用。

110 个环境；32,112 个已验证 RLVR 元组；38% 为跨应用任务。

广泛的环境池支持大量经过独立验证的任务。

Wang 等, 2026 · CUA-Gym 图 8(a)、第 2.3 节

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 68 | 原页 68

第 7 讲 · 计算机使用智能体

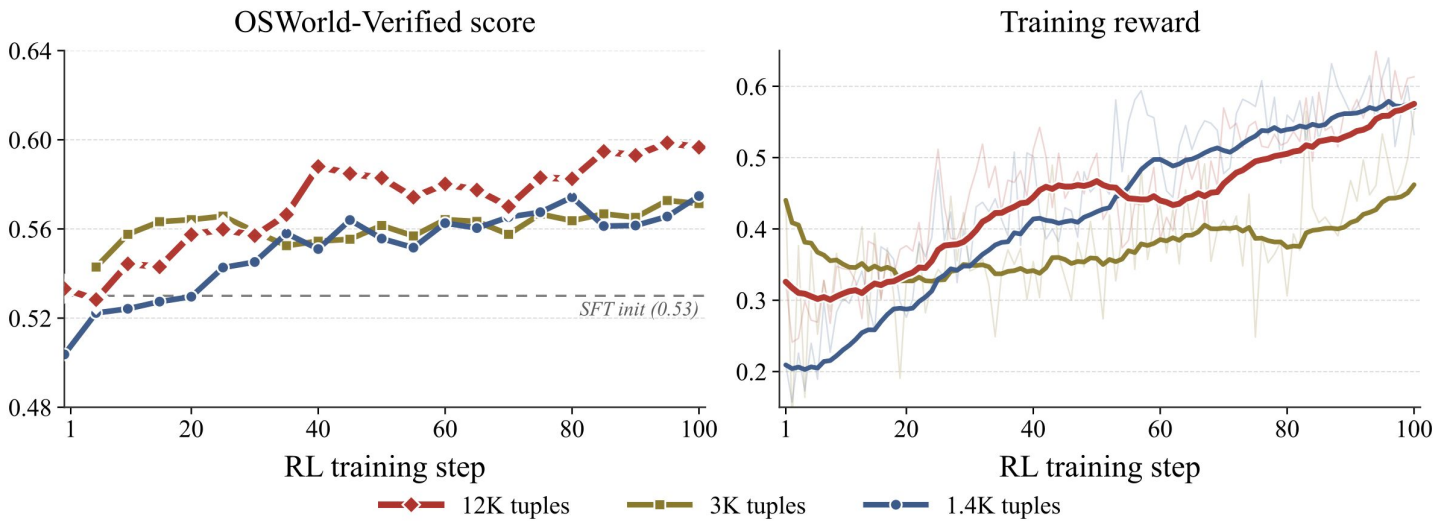
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

CUA-Gym: Verified rollouts improve the policy

3,578 SFT DEMOS → GSPO ON 10,858 VERIFIED TUPLES



OSWorld Performance: 62.2 → 72.6%

[Wang et al., 2026 · Fig. 9 data-scale ablation: §3 main results](#)  
[OSWorld-Verified · GUI calls · 100 model steps · programmatic checks](#)

04 / 训练

CUA-Gym: 已验证轨迹采样提升策略

3,578 条 SFT 示范 → 在 10,858 个已验证元组上运行 GSPO。

OSWorld 表现: 62.2 → 72.6%。

Wang 等, 2026 · 图 9 数据规模消融; 第 3 节主要结果。

OSWorld-Verified · GUI 调用 · 100 个模型步骤 · 程序化检查

CUA-Gym: 已验证轨迹采样提升策略

第 7 讲 · 计算机使用智能体

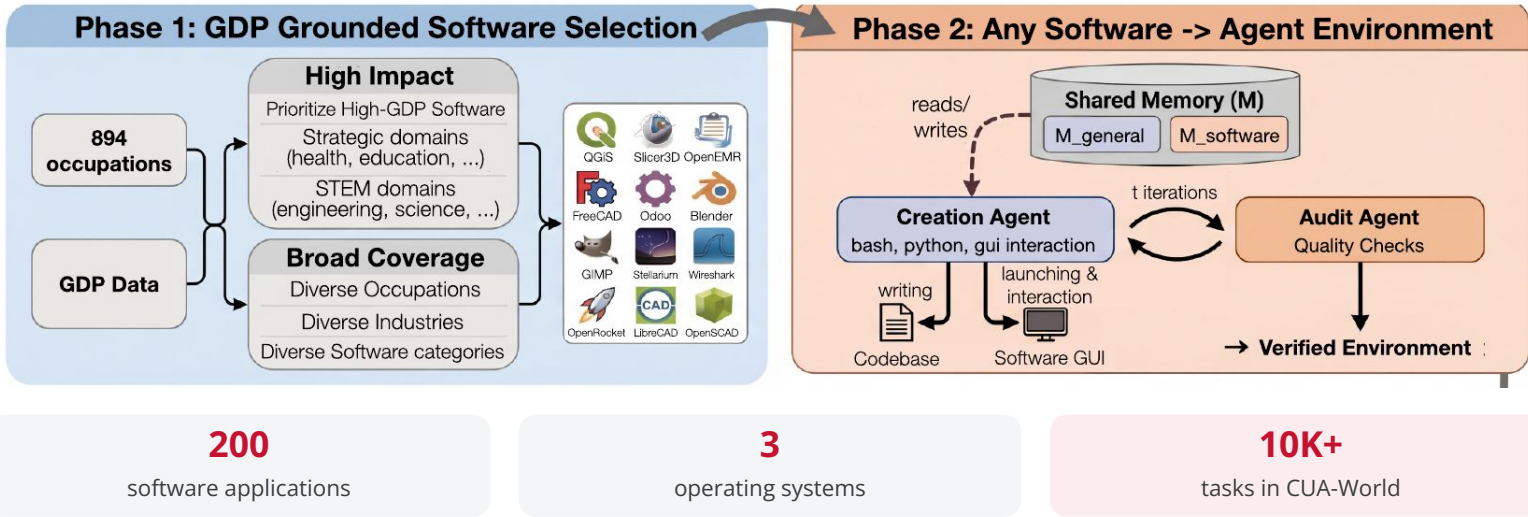
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

Gym-Anything: Create and audit real software setups

SELECT SOFTWARE / CREATE ↔ AUDIT



The creation agent builds, an independent audit verifies the evidence.

Aggarwal et al., 2026 · Gym-Anything Fig. 2, phases 1–2; §3

86

04 / 训练

Gym-Anything：创建并审计真实软件环境

选择软件 / 创建 ↔ 审计。

200 个软件应用；3 种操作系统；CUA-World 中 10K+ 个任务。

创建智能体负责构建，独立审计负责验证证据。

Aggarwal 等，2026 · Gym-Anything 图 2，阶段 1–2；第 3 节

Gym-Anything：创建并审计真实软件环境

第 7 讲 · 计算机使用智能体

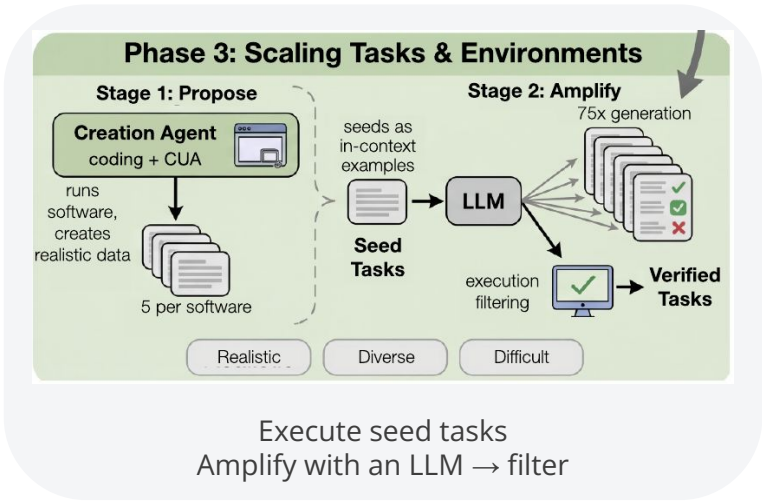
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

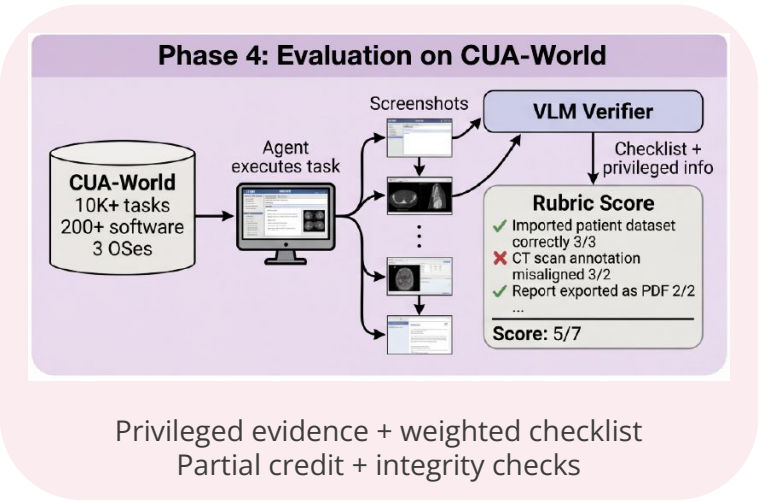
04 / TRAINING

Gym-Anything: Scale tasks, then verify trajectories

PROPOSE + AMPLIFY



CHECKLIST-BASED VERIFICATION



[Aggarwal et al., 2026 · Gym-Anything Fig. 2, phases 3-4; §4.1](#)

87

04 / 训练

Gym-Anything: 扩展任务，再验证轨迹

提出 + 扩增：执行种子任务；使用 LLM 扩增 → 筛选。

基于清单的验证：特权证据 + 加权清单；部分得分 + 完整性检查。

Aggarwal 等, 2026 · Gym-Anything 图 2, 阶段 3-4; 第 4.1 节

Gym-Anything: 扩展任务，再验证轨迹

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

04 / TRAINING

# Gym-Anything: Environments across real software

REAL INSTALLED SOFTWARE / LINUX · WINDOWS · ANDROID

Healthcare

Engineering

Design

Enterprise

Science

Finance

GeoScience

Education

200

software applications

12,103

tasks and environments

22 / 22

SOC occupation groups covered

The generator creates task-ready setups across professional software.

[Aggarwal et al., 2026 · Gym-Anything Fig. 1: \\$5](#)

88

04 / 训练

Gym-Anything：覆盖真实软件环境

真实已安装软件 / Linux · Windows · Android。

200 个软件应用；12,103 个任务与环境；22 / 22 个 SOC 职业类别均有覆盖。

生成器会创建适用于专业软件的任务环境。

Aggarwal 等, 2026 · Gym-Anything 图 1；第 5 节

Gym-Anything：覆盖真实软件环境

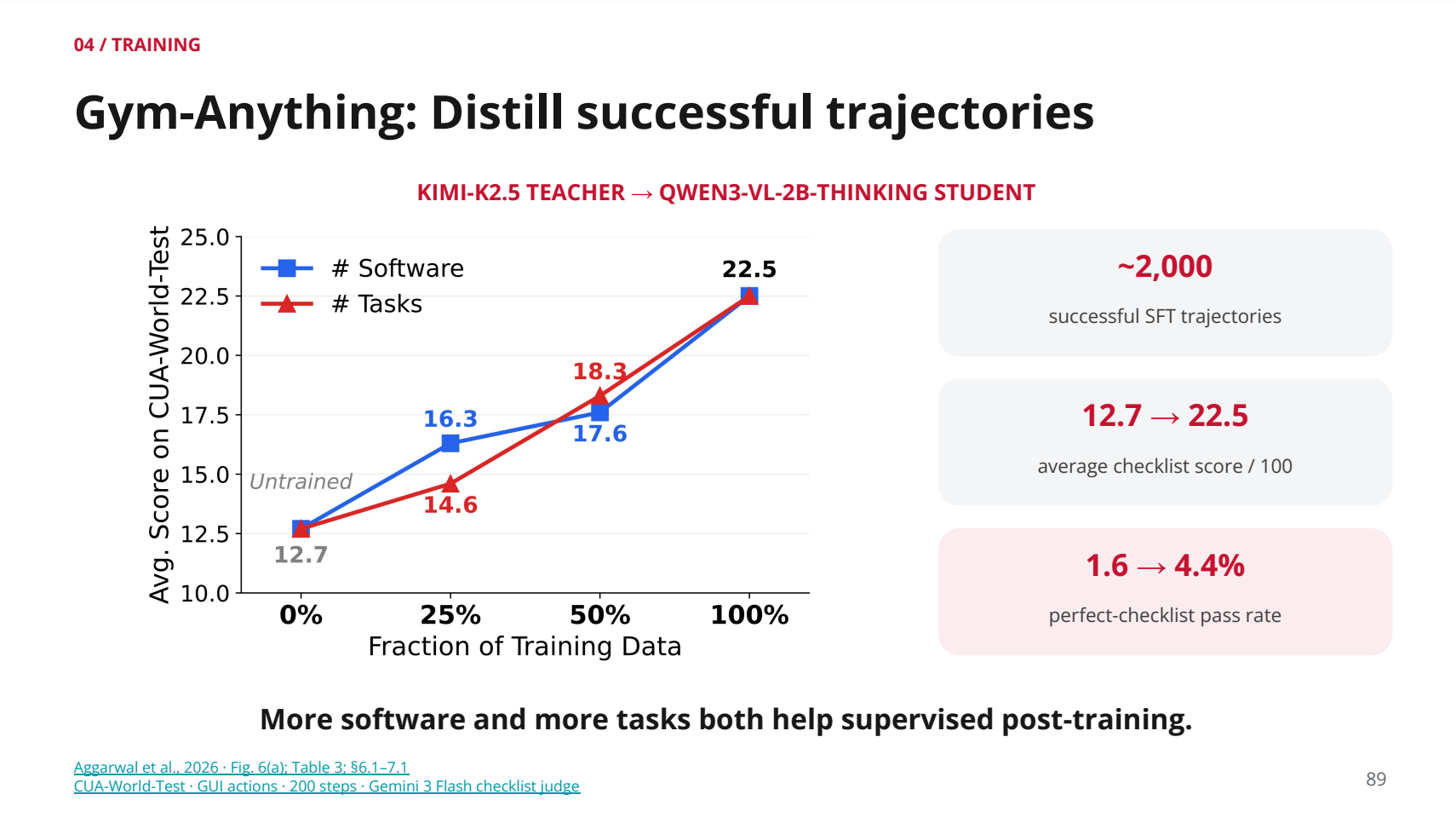
个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 72 | 原页 72

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



04 / 训练

Gym-Anything：蒸馏成功轨迹

KIMI-K2.5 教师 → QWEN3-VL-2B-THINKING 学生。

约 2,000 条成功的 SFT 轨迹。

平均清单得分 / 100：12.7 → 22.5。

清单全对率：1.6 → 4.4%。

更多软件与更多任务都有助于监督式后训练。

Aggarwal 等, 2026 · 图 6(a)、表 3、第 6.1–7.1 节。

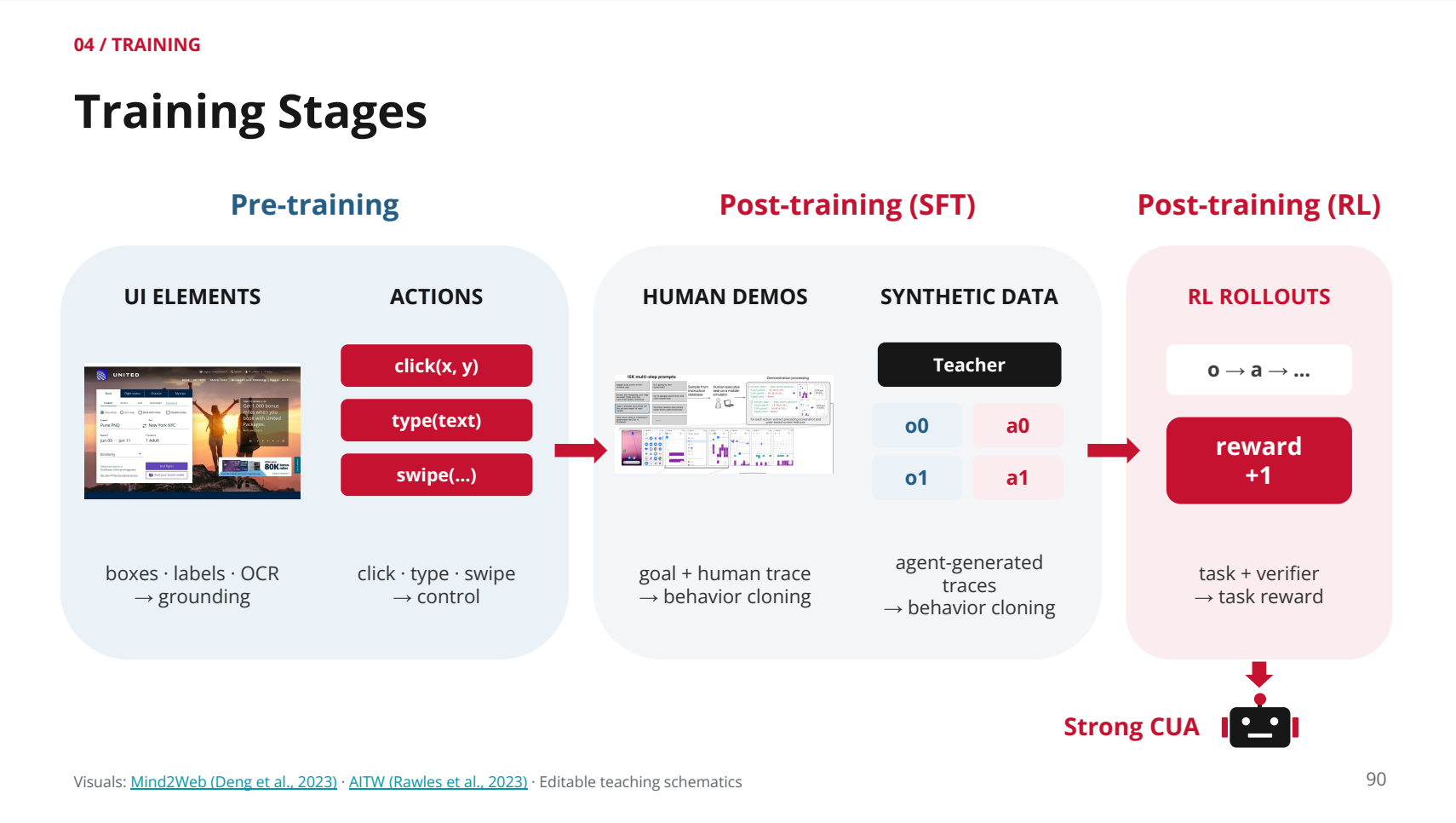
CUA-World-Test · GUI 动作 · 200 步 · Gemini 3 Flash 清单评判

Gym-Anything：蒸馏成功轨迹

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



04 / 训练

训练阶段

预训练：UI 元素（框 · 标签 · OCR → 定位）；动作（click · type · swipe → 控制）。

后训练（SFT）：人类示范——目标 + 人类轨迹 → 行为克隆；合成数据——教师生成轨迹、智能体生成轨迹 → 行为克隆。

后训练（RL）：RL 轨迹采样——观测  $o \rightarrow$  动作  $a \rightarrow \dots$ ；任务 + 验证器 → 任务奖励，例如 +1。

强大的 CUA。

视觉资料：Mind2Web（Deng 等, 2023）· AITW（Rawles 等, 2023）· 可编辑教学示意图

训练阶段

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

05 / SUMMARY

Summary

- 1

**CUAs operate directly on the GUI**  
They observe and act within the same interface as humans.
- 2

**Evaluations should target the right metrics and level of behavior**  
Static actions, end-to-end tasks, and long horizon workflows are measured differently.
- 3

**Models predict actions from interleaved visual history**  
CUAs are vision-language models that emit GUI actions, some also use code and tools.
- 4

**Pretraining → SFT → RL**  
Models are trained through large scale pretraining, SFT on human annotated and synthetic trajectories, and RL within simulated environments.

总结

05 / 总结

总结

1. CUA 直接操作 GUI：它们在与人类相同的界面中观察并采取行动。
2. 评估应针对正确的指标和行为层级：静态动作、端到端任务与长时程工作流的测量方式不同。
3. 模型根据交错的视觉历史预测动作：CUA 是输出 GUI 动作的视觉语言模型，有些也会使用代码和工具。
4. 预训练 → SFT → RL：模型先经过大规模预训练，再基于人类标注和合成轨迹进行 SFT，最后在模拟环境中进行 RL。

11-768 · GUI 智能体

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



还有哪些工作？

还有哪些工作？

第 7 讲 · 计算机使用智能体

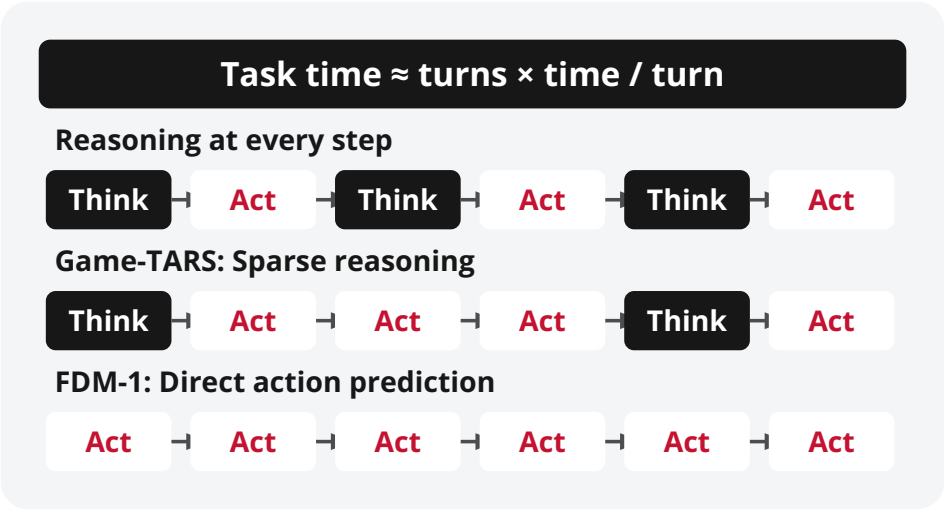
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

05 / WHAT'S REMAINING

Speed + Cost

FASTER CUAS REQUIRE BOTH SYSTEMS AND MODEL IMPROVEMENTS



Long contexts are expensive

Screenshots and reasoning add prefill and generation cost.

Game-TARS: Sparse Thinking

Reason at key decisions; act directly on routine steps.

Fewer / cheaper turns

Improve planning and batching; explore direct action models (FDM-1).

[Game-TARS · ByteDance Seed, 2025, §3.3 / §4.3.2](#) · [FDM-1 · Standard Intelligence, 2026](#)

93

05 / 尚待解决

速度与成本

更快的 CUA 需要系统和模型同时改进。

任务耗时  $\approx$  轮次数  $\times$  每轮耗时。

每一步都进行推理：Game-TARS 稀疏推理——在关键决策处推理，在常规步骤直接行动。

直接动作预测：FDM-1。

长上下文成本高：截图和推理会增加预填充与生成成本。

更少 / 更便宜的轮次：改进规划与批处理；探索直接动作模型（FDM-1）。

Game-TARS · ByteDance Seed, 2025, 第 3.3 / 4.3.2 节 · FDM-1 · Standard Intelligence, 2026

速度与成本

第 7 讲 · 计算机使用智能体

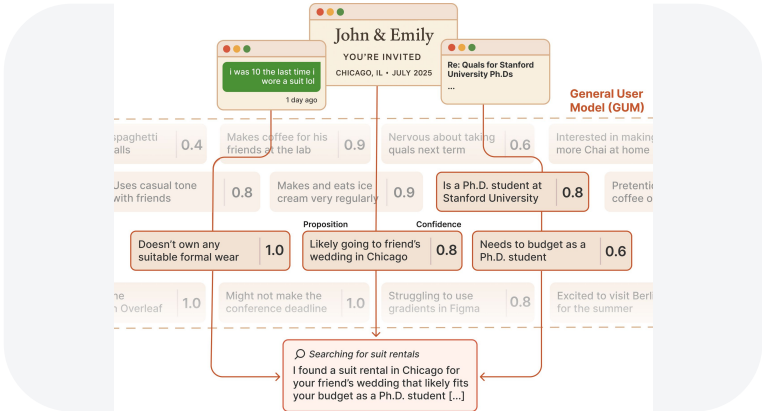
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

05 / WHAT'S REMAINING

Personalization: Adapt to the user

GUM: INFER THE USER'S CONTEXT



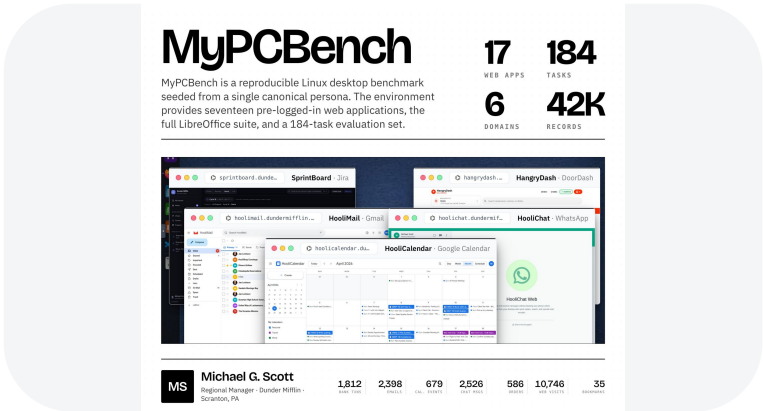
Personal context changes

Use the user's files, history and preferences—not generic defaults.

Memory must be revisable

GUM retrieves and updates confidence-weighted user facts.

MYPCBENCH: A PERSONAL DIGITAL LIFE



User control is essential

MyPCBench tests personal context. Consent and forgetting are essential.

[GUM · Shaikh et al., 2025, Fig. 1](#) · [MyPCBench · Jang et al., 2026, project overview](#)

05 / 尚待解决

个性化：适应用户

GUM：推断用户上下文。个人上下文会变化：使用用户的文件、历史记录和偏好，而不是通用默认值。

记忆必须可修订：GUM 检索并更新带置信度权重的用户事实。

MyPCBench：个人数字生活。用户控制至关重要：MyPCBench 测试个人上下文；同意与遗忘至关重要。

GUM · Shaikh 等, 2025, 图 1 · MyPCBench · Jang 等, 2026, 项目概览

个性化：适应用户

第 7 讲 · 计算机使用智能体

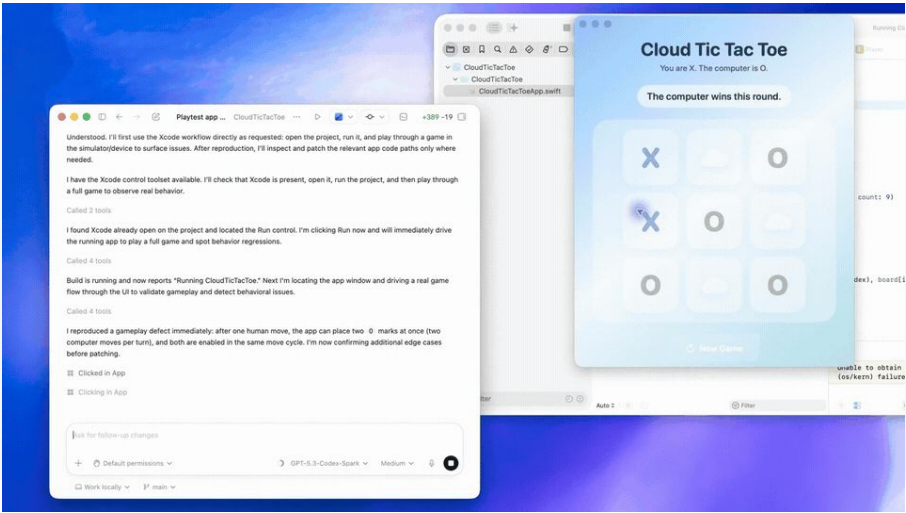
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

05 / WHAT'S REMAINING

Infrastructure + UX

CODEx: BACKGROUND COMPUTER USE



Don't block the user

Run in isolated background sessions while their work continues.

Collaborate in real time

Show progress; accept corrections, approval and interruption.

Novel UI / UX for CUAs

Much to be explored still!

Proactivity needs consent, visibility, and an easy off switch.

[OpenAI, 2026 · Codex for \(almost\) everything · demo excerpt](#)

05 / 尚待解决

基础设施与用户体验

CODEx：后台计算机使用。

不要阻塞用户：在隔离的后台会话中运行，让用户继续工作。

实时协作：展示进度；接受更正、批准与中断。

CUA 的新型 UI / UX：仍有大量探索空间。

主动性需要取得同意、让过程可见，并提供简单易用的关闭开关。

OpenAI, 2026 · Codex for (almost) everything · 演示摘录

基础设施与用户体验

第 7 讲 · 计算机使用智能体

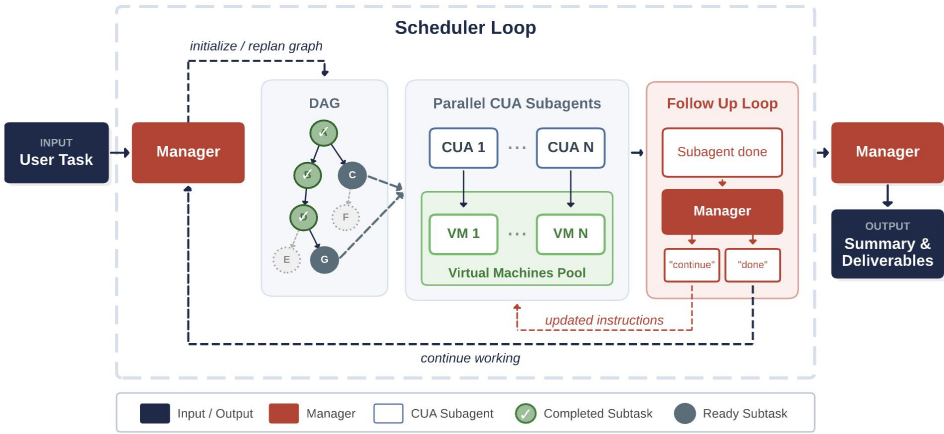
EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

05 / WHAT'S REMAINING

Multi-Agent Systems

MACU: PLAN → PARALLELIZE → REPLAN



Divide work by dependencies

MACU dispatches ready subtasks to CUAs in isolated environments.

Share evidence, then replan

Hand off files and results; verify work and recover from failures.

Beyond homogeneous CUAs

Combine CUA, code and tool agents; control conflicts and overhead.

Parallelism helps only when coordination preserves correctness.

[MACU · Koh, Salakhutdinov & Fried, 2026, Fig. 1 / §3 \(adapted\)](#)

05 / 尚待解决

多智能体系统

MACU：规划 → 并行化 → 重新规划。

按依赖关系拆分工作：MACU 将就绪的子任务分派给隔离环境中的 CUA。

共享证据，然后重新规划：交接文件和结果；验证工作并从失败中恢复。

超越同质 CUA：组合 CUA、代码智能体与工具智能体；控制冲突与开销。

只有当协调能够保证正确性时，并行才有帮助。

MACU · Koh、Salakhutdinov 与 Fried，2026，图 1 / 第 3 节（改编）

多智能体系统

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分



谢谢！

谢谢！

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

APPENDIX / REFERENCE

Appendix: Benchmarks

WEB + STATIC EVALUATION

[MiniWoB](#) / [MiniWoB++](#)  
[WebShop](#)  
[ScreenSpot-Pro](#)  
[Mind2Web](#)  
[Android in the Wild](#)  
[AndroidControl](#)  
[WebArena](#)  
[VisualWebArena](#)  
[WebVoyager](#)  
[Online-Mind2Web](#)  
[WorkArena](#) / [WorkArena++](#)

DESKTOP, MOBILE + LONG HORIZON

[OSWorld](#) / [Verified](#)  
[WindowsAgentArena](#)  
[AndroidWorld](#)  
[MobileWorld](#)  
[Odysseys](#)  
[OSWorld 2.0](#)  
[MyPCBench](#)  
[WeaveBench](#)

AgentNetBench and CUA-World-Long share their papers with training datasets/environments (p. 99).

98

附录 / 参考资料

附录：基准

网页 + 静态评估：MiniWoB / MiniWoB++、WebShop、ScreenSpot-Pro、Mind2Web、Android in the Wild、AndroidControl、WebArena、VisualWebArena、WebVoyager、Online-Mind2Web、WorkArena / WorkArena++。

桌面、移动端 + 长时程：OSWorld / Verified、WindowsAgentArena、AndroidWorld、MobileWorld、Odysseys、OSWorld 2.0、MyPCBench、WeaveBench。

AgentNetBench 与 CUA-World-Long 的论文还包含训练数据集 / 环境（第 99 页）。

附录：基准

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

APPENDIX / REFERENCE

Appendix: Modeling / Training

MODELS + GROUNDING

[OS-Atlas](#)  
[UI-TARS](#)  
[Qwen-UI-Agent](#)  
[Qwen3.6-27B \(model card\)](#)  
[Game-TARS](#)

DATA + TRAINING ENVIRONMENTS

[OpenCUA / AgentNet / AgentNetBench](#)  
[MolmoWeb / MolmoWebMix](#)  
[CUA-Gym](#)  
[Gym-Anything / CUA-World](#)

Grounding and demonstration datasets that also serve as static benchmarks are listed on p. 98.

99

附录 / 参考资料

附录：建模 / 训练

模型 + 定位：OS-Atlas、UI-TARS、Qwen-UI-Agent、Qwen3.6-27B（模型卡）、Game-TARS。

数据 + 训练环境：OpenCUA / AgentNet / AgentNetBench、MolmoWeb / MolmoWebMix、CUA-Gym、Gym-Anything / CUA-World。

同时作为静态基准的定位和示范数据集见第 98 页。

附录：建模与训练

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

APPENDIX / REFERENCE

Appendix: Systems

AGENT LOOPS + ORCHESTRATION

[ReAct](#)  
Interleaved reasoning and actions

[MACU: Multi-Agent Computer Use](#)  
Orchestrating multiple computer-use agents

[GUM: General User Models](#)  
User modeling from computer activity

FRAMEWORKS + EFFICIENCY

[BrowserGym ecosystem](#)  
Shared environments and experiment tooling

[Cua \(open-source framework\)](#)  
Computer-use environments and infrastructure

[FDM-1 \(technical blog\)](#)  
High-frequency computer action modeling

Papers and technical resources for agent orchestration, personalization, infrastructure, and efficiency.

100

附录 / 参考资料

附录：系统

智能体循环 + 编排：ReAct——交错推理与动作；MACU：多智能体计算机使用——协调多个计算机使用智能体；GUM：通用用户模型——根据计算机活动建模用户。

框架 + 效率：BrowserGym 生态系统——共享环境与实验工具；Cua（开源框架）——计算机使用环境与基础设施；FDM-1（技术博客）——高频计算机动作建模。

智能体编排、个性化、基础设施与效率相关论文及技术资源。

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

APPENDIX / REFERENCE

# Appendix: Products

OPENAI + ANTHROPIC

[Codex Computer Use \(OpenAI blog\)](#)  
[Operator \(OpenAI launch\)](#)  
[OpenAI computer-use API guide](#)  
[Claude computer use \(launch blog\)](#)  
[Developing computer use \(Claude blog\)](#)  
[Claude computer-use tool docs](#)

OTHER PRODUCTS + RELEASES

[Grok Bot \(xAI launch\)](#)  
[Gemini computer-use API guide](#)  
[Muse Spark 1.1 \(Meta release\)](#)  
[Navigator n2 \(Yutori release\)](#)

Official product announcements, technical blogs, and API documentation. Operator is a historical launch.

101

附录 / 参考资料

附录：产品

OpenAI + Anthropic: Codex Computer Use (OpenAI 博客)、Operator (OpenAI 发布公告)、OpenAI 计算机使用 API 指南、Claude 计算机使用 (发布博客)、Developing computer use (Claude 博客)、Claude 计算机使用工具文档。

其他产品 + 发布: Grok Bot (xAI 发布)、Gemini 计算机使用 API 指南、Muse Spark 1.1 (Meta 发布)、Navigator n2 (Yutori 发布)。

官方产品公告、技术博客与 API 文档。Operator 是历史发布内容。

附录：产品

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 85 | 原页 85

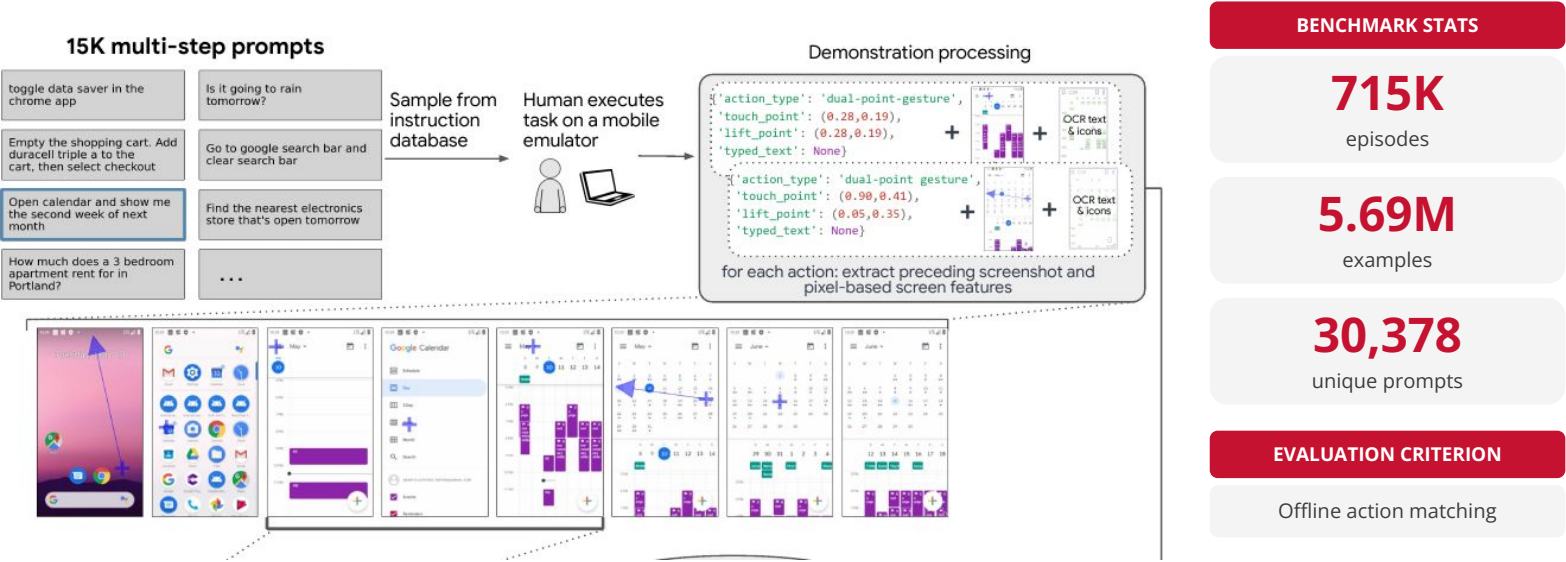
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

Android in the Wild: Scale without end-state checks



Scale buys coverage; the evaluator still matches recorded actions.

Source: Rawles et al., 2023, Fig. 1

86

02A / 静态评估

Android in the Wild：规模大，但没有最终状态检查

基准统计：715K 个回合；5.69M 个样例；30,378 个独特提示词。

评估标准：离线动作匹配。

大规模数据带来了覆盖范围，但评估器仍然只是匹配记录下来的动作。

来源：Rawles 等，2023，图 1

Android in the Wild：规模大，但没有最终状态检查

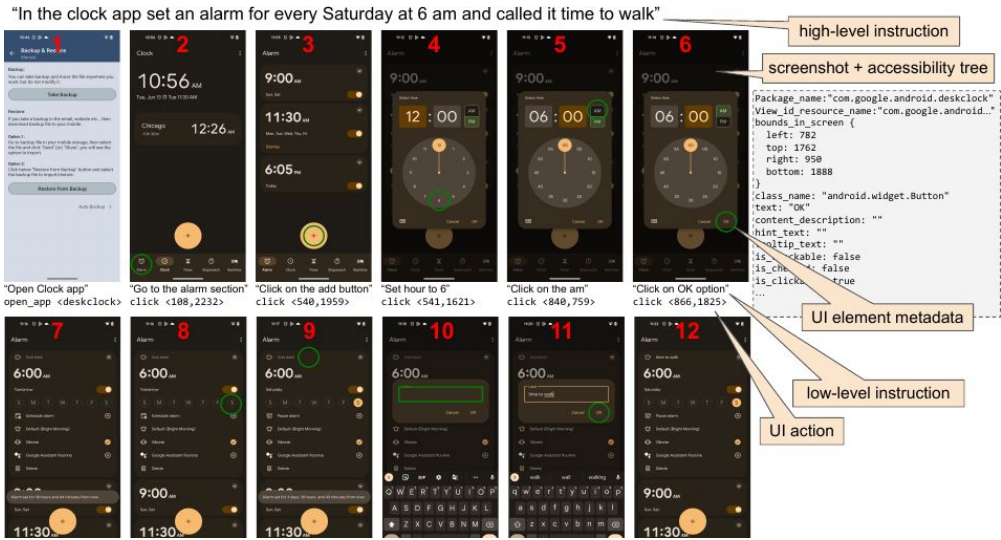
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

AndroidControl: Diverse apps, paired instructions



**BENCHMARK STATS**

**15,283**  
demonstrations

**14,548**  
unique tasks

**833**  
Android apps

**EVALUATION CRITERION**

Offline action matching

Paired high- and low-level instructions separate planning from motor control.

Source: Li et al., NeurIPS 2024, Fig. 1

02A / 静态评估

AndroidControl：多样化应用与配对指令

基准统计：15,283 个示范；14,548 个独特任务；833 个 Android 应用。

评估标准：离线动作匹配。

高层与低层指令的配对有助于将规划与运动控制区分开。

来源：Li 等，NeurIPS 2024，图 1

AndroidControl：多样化应用与配对指令

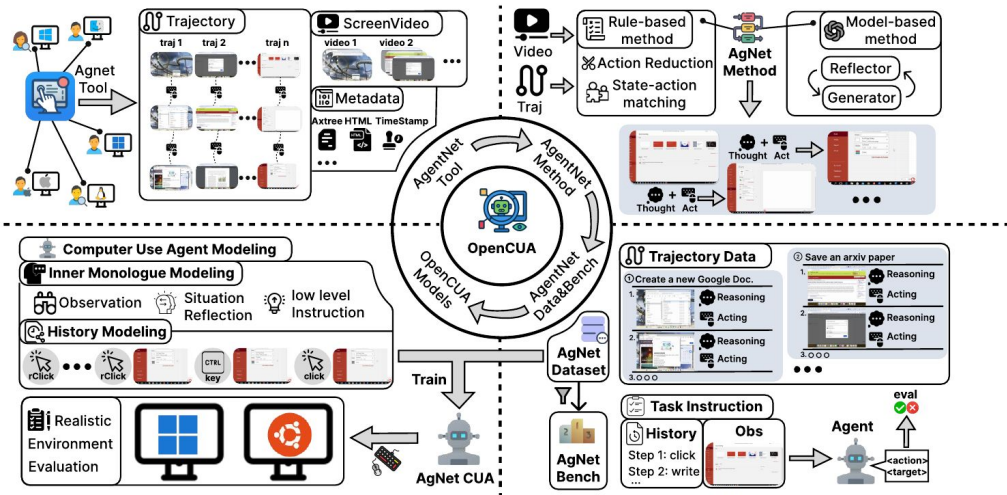
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02A / STATIC EVALUATIONS

AgentNetBench: An offline desktop proxy



BENCHMARK STATS

100  
tasks

17.63  
average steps

2,143  
total actions

EVALUATION CRITERION

Offline step matching

AgentNetBench turns human desktop trajectories into a fast offline proxy.

Source: Wang et al., NeurIPS 2025, Fig. 2 (cropped)

AgentNetBench：离线桌面代理基准

02A / 静态评估

AgentNetBench：离线桌面代理基准

基准统计：100 个任务；平均 17.63 步；总计 2,143 个动作。

评估标准：离线步骤匹配。

AgentNetBench 将人工桌面轨迹转化为快速离线代理基准。

来源：Wang 等，NeurIPS 2025，图 2（裁剪）

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

WorkArena: Enterprise tasks with validators

BENCHMARK STATS

33

task templates

19,912

instances

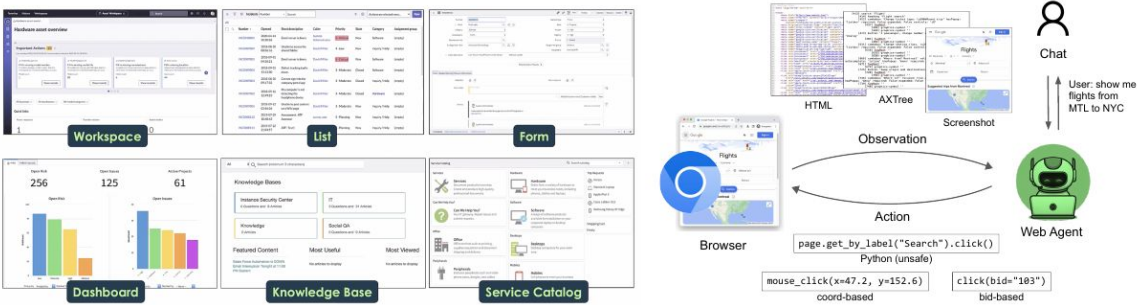
ServiceNow

environment

EVALUATION CRITERION

Programmatic checks

WorkArena: Web Agents for Common Knowledge Work Tasks



Parameterized tasks make enterprise state validation reproducible.

Source: Drouin et al., ICML 2024, Fig. 1

WorkArena：带验证器的企业任务

02C / 端到端评估（桌面、移动）

WorkArena：带验证器的企业任务

基准统计：33 个任务模板；19,912 个实例；ServiceNow 环境。

评估标准：程序化检查。

参数化任务使企业系统状态验证可复现。

来源：Drouin 等, ICML 2024, 图 1

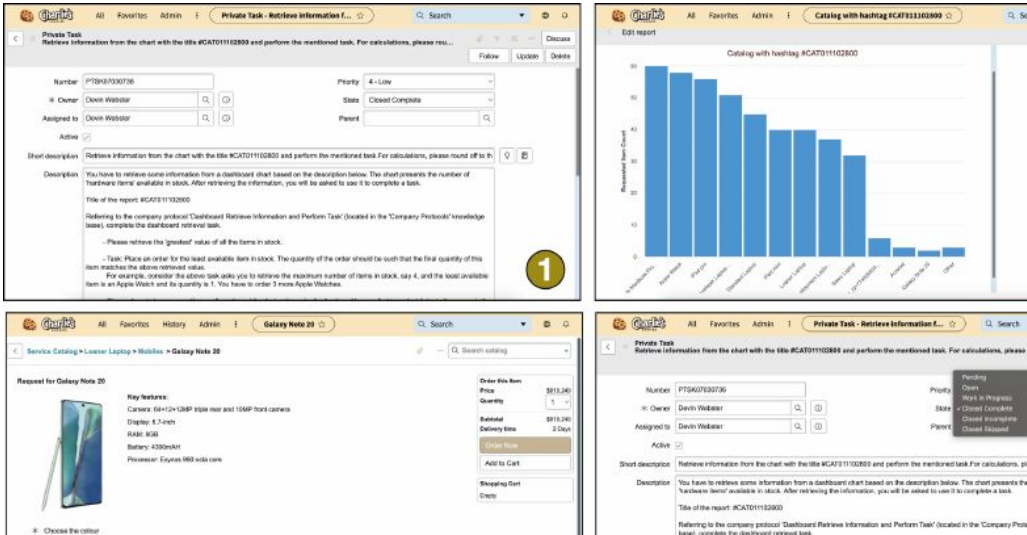
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

WorkArena++: Compositional enterprise workflows



BENCHMARK STATS

682  
composite tasks

5  
skill families

10  
company themes

EVALUATION CRITERION

Compositional end state

Composition converts atomic UI skills into knowledge-work workflows.

Source: Boisvert et al., NeurIPS 2024, Fig. 1

WorkArena++：组合式企业工作流

02C / 端到端评估（桌面、移动）

WorkArena++：组合式企业工作流

基准统计：682 个组合任务；5 类技能；10 个公司主题。

评估标准：组合式最终状态。

组合将原子级 UI 技能转化为知识工作流。

来源：Boisvert 等，NeurIPS 2024，图 1

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

# AndroidWorld: Parameterized mobile state

The diagram illustrates the AndroidWorld environment. It features an Android Emulator on the left, which interacts with an Agent (represented by a robot icon) in the center. The Agent sends actions to the emulator and receives rewards. A TaskEval module is shown, which defines goals, generates random parameters, initializes tasks, checks for success, and tears down the environment. The Agent also receives task goals from a set of 116 Android and 92 MiniWeb++ tasks. The environment includes various mobile apps and OS state. Screenshot examples show tasks like creating a calendar event and adding a location marker. A UI tree is also shown.

BENCHMARK STATS

116

task families

20

applications

∞

parameterized instances

EVALUATION CRITERION

System-state reward

Task generators and system-state rewards create repeatable mobile episodes.

Source: Rawles et al., ICLR 2025, Fig. 1

02C / 端到端评估（桌面、移动）

AndroidWorld：参数化移动端状态

基准统计：116 类任务；20 个应用； $\infty$  个参数化实例。

评估标准：系统状态奖励。

任务生成器与系统状态奖励可创建可重复的移动端回合。

来源：Rawles 等，ICLR 2025，图 1

AndroidWorld：参数化移动端状态

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 91 | 原页 91

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02C / END-TO-END EVALUATIONS (DESKTOP, MOBILE)

MobileWorld: Cross-app, user, and MCP tasks

Agent User Interaction Task Example

Send an email to Kevin with the message "Hello"

.....

Current Screen

I did not find Kevin in contact app, I need to ask user what is the Kevin's email address.

Action: ask\_user

Text: what is the Kevin's email address?

System Prompt:

You are acting as a mobile user.

...

The relevant information for the task is: Kevin's email address: kevin\_zhang@example.com

...

Simulated User Response

Kevin's email address is kevin\_zhang@example.com

MCP Tool Use Task Example

Check README from android\_world GitHub repo and send a Mattermost message to Sam about running it on MacOS...

.....

GitHub MCP Tools:

github\_create\_or\_update\_file

github\_create\_repository

github\_get\_file

...

MCP Service

MCPs from Tongyi Bailian Marketplace

Among the provided tools, I should use github\_get\_file to fetch the README.md content so I can ...

Action: MCP(github\_get\_file)

args: {"owner": "google-research", "repo": "android\_world", "path": "README.md"}

MCP Server Response

{"name": "README.md", "path": "README.md", "type": "file", "size": 10849, "sha": "a19a6770e4b813..."}

BENCHMARK STATS

201

tasks

27.8

average steps

62.2%

cross-app tasks

EVALUATION CRITERION

Programmatic + hybrid

Modern mobile use mixes GUI control, clarification, and structured tools.

Source: Kong et al., ACL 2026, Fig. 2 (cropped)

92

02C / 端到端评估（桌面、移动）

MobileWorld：跨应用、用户与 MCP 任务

基准统计：201 个任务；平均 27.8 步；62.2% 为跨应用任务。

评估标准：程序化 + 混合评估。

现代移动设备使用结合了 GUI 控制、澄清提问与结构化工具。

来源：Kong 等, ACL 2026, 图 2（裁剪）

MobileWorld：跨应用、用户与 MCP 任务

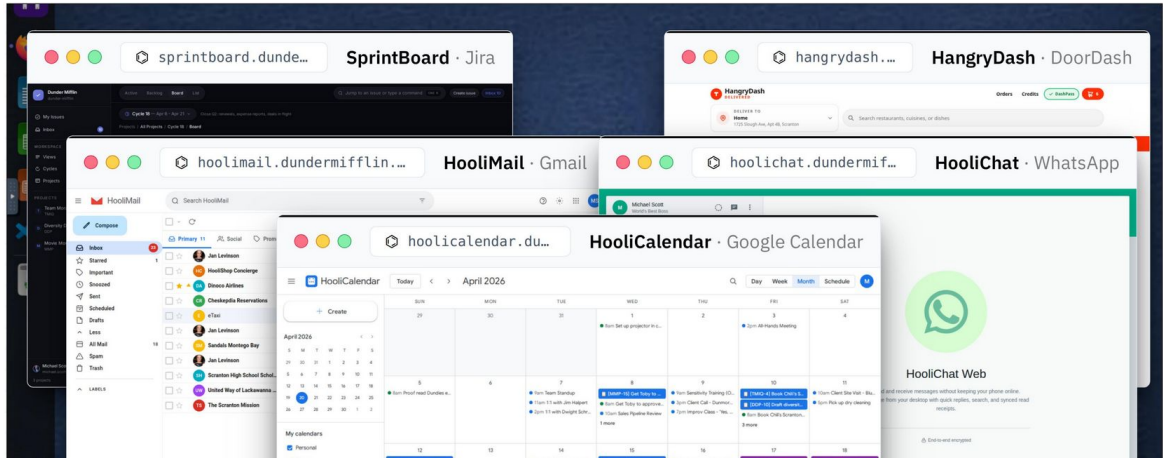
第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

MyPCBench: One person's cross-app digital life



BENCHMARK STATS

184 tasks

17 logged-in apps

68% multi-app tasks

EVALUATION CRITERION

Per-rubric LLM judge

One seeded identity turns separate apps into one reproducible digital life.

Source: lang et al., arXiv 2026, Fig. 1 (cropped)

02D / 长时程计算机使用

MyPCBench: 一个人的跨应用数字生活

基准统计：184 个任务；17 个已登录应用；68% 为多应用任务。

评估标准：逐条评分规则的 LLM 评判。

一个预置身份将彼此独立的应用连接成可复现的数字生活。

来源：Jang 等，arXiv 2026，图 1（裁剪）

MyPCBench: 一个人的跨应用数字生活

第 7 讲 · 计算机使用智能体

EN / 英文官方原页 · 保持原样

ZH / 中文译文 · 第 1/1 部分

02D / LONG HORIZON COMPUTER USE

WeaveBench: Hybrid GUI + CLI trajectories

The visualization includes a donut chart showing 114 tasks categorized into 14 domains: WEB, DSK, SPA, GAM, DES, DOC, OPS, DAV, Pipeline, Observability, Debug Toolchain, Cluster/Monitor, Print-ready Doc, Office Suite, and Markup/LaTeX. Two bar charts show channel switches and tool calls per task, with medians of 16 and 76 respectively.

BENCHMARK STATS

114 tasks

16 median switches

76 median tool calls

EVALUATION CRITERION

Trajectory-aware judge

Outcome-only grading can overcredit agents that take invalid shortcuts.

Source: Li et al., EMNLP 2026, Fig. 3

94

WeaveBench: GUI + CLI 混合轨迹

02D / 长时程计算机使用

WeaveBench: GUI + CLI 混合轨迹

基准统计：114 个任务；切换次数中位数 16；工具调用次数中位数 76。

评估标准：轨迹感知评判。

只按结果评分，可能会过度奖励采用无效捷径的智能体。

来源：Li 等，EMNLP 2026，图 3

个人学习译文 · 非 CMU 官方译本 · 译注独立标记 · 原始课件来自 cmu-agents.com

双语 PDF 94 | 原页 94