

FlashAdventure: A Benchmark for GUI Agents

Solving Full Story Arcs in Diverse Adventure Games



(* Equal Contribution)

Jaewoo Ahn*, Junseo Kim*, Heeseung Yun, Jaehyeon Son,
Dongmin Park, Jaewoong Cho, Gunhee Kim



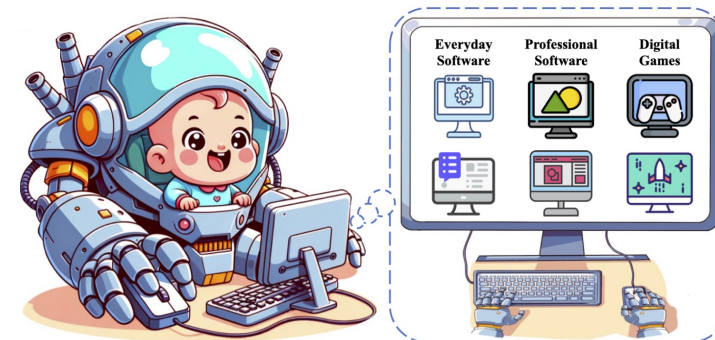
SEOUL NATIONAL UNIV.
VISION & LEARNING

KRAFTON

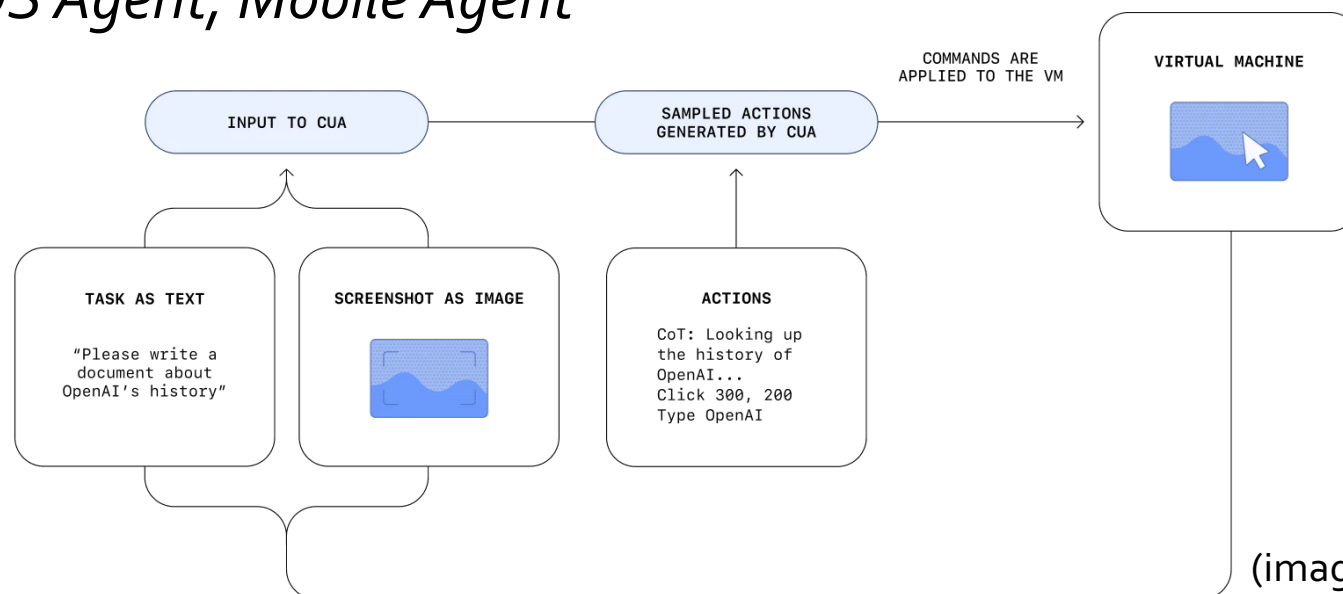


Intro: LLM-Powered GUI Agents

- GUI Agents, or Computer-Using Agents (CUA)
 - Input: Task description, Screenshot + α
 - Output: Actions (e.g., *mouse click*, *keystroke*)
- Work on diverse digital environments
 - *Web Agent, OS Agent, Mobile Agent*



(image credit: Cradle^[1])



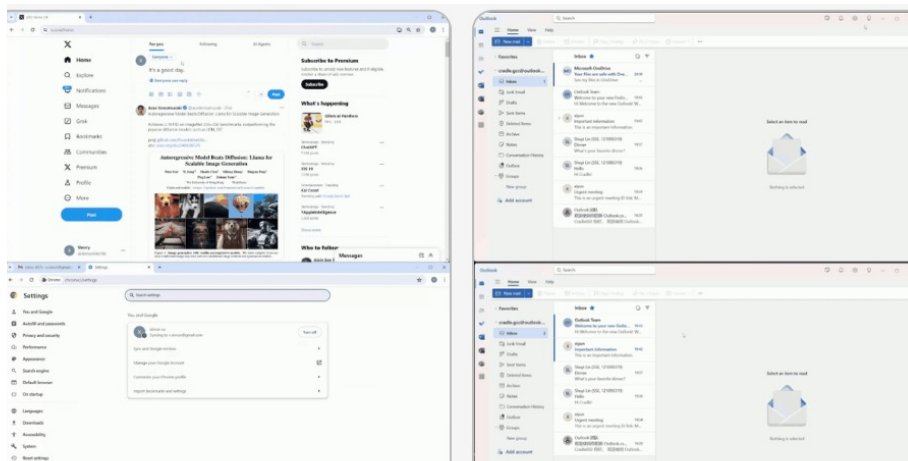
(image credit: OpenAI^[2])

[1] Tan et al., *Cradle: Empowering Foundation Agents towards General Computer Control*, ICML 2025

[2] OpenAI, *Computer-Using Agent*, <https://openai.com/index/computer-using-agent/>

Intro: GUI Agents for Video GamePlaying

- Game environment?
 - Offer **balanced env.** btw {device UI} vs. {real-world perception}
 - {Standard GUIs} vs. {Non-standard layouts & complex interaction}
 - Excellent testbed for evaluating **generalizability** of GUI agents
 - Especially **Adventure games** → introduce **additional complexity**
 - Diverse layouts & interactions: *inventory management, dialog tree w/ NPC, ...*
 - Story arcs: *Players need to appropriately use items to understand/solve overall story*



VS



RDR2: Main Storyline



RDR2: Open-ended World

(image credit: Cradle^[1])

[1] Tan et al., *Cradle: Empowering Foundation Agents towards General Computer Control*, ICML 2025

Motivation: Task Diversity & Full Story Arcs

- Existing game benchmarks for GUI agents
 - OOTB^[1] – Hearthstone, Honkai: Star Rail
 - Perform low-level tasks (*e.g., rename card deck*)
 - VARP^[2] – Black Myth: Wukong (AAA game)
 - Execute sequence of actions (*e.g., combat boss monster*)
 - Cradle^[3] – RDR2 (AAA game), Stardew Valley, Cities: Skylines, Dealer's Life 2
 - Completes a part of missions (*e.g., kill wolves & ride horse*)
- Despite advances, there exists **critical gaps**
 - **Low task/game diversity**
 - Hard to evaluate agents in diverse gaming scenarios beyond solving specific tasks
 - **Absence of completing full story arcs**
 - None of them evaluate agents on completing entire story arcs

[1] Hu et al., *The Dawn of GUI Agent: A Preliminary Case Study with Claude 3.5 Computer Use*, arXiv 2024

[2] Chen et al., *Can VLMs Play Action Role-Playing Games? Take Black Myth Wukong as a Study Case*, arXiv 2024

[3] Tan et al., *Cradle: Empowering Foundation Agents towards General Computer Control*, ICML 2025

The FlashAdventure Benchmark

- A benchmark to evaluate GUI agents solving full story arcs

- Consist of **34 Flash adventure** games

- **Largest** video game benchmark

- 15 Room Escape
- 11 Point-and-Click Adventure (Mystery/Detective)
- 4 Visual Novel
- 2 Hidden Object
- 2 Life/Management Simulation

- **Compact play-time** (0.5 - 1 hour)

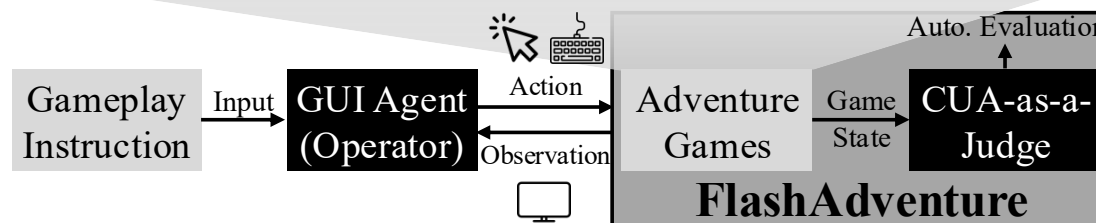
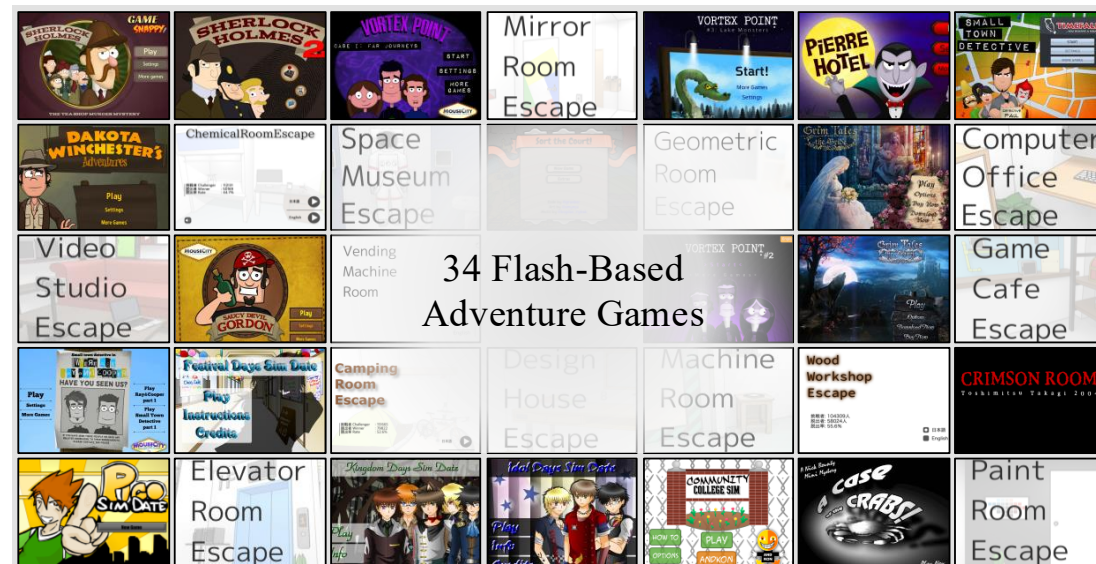
- **Free-to-play**

- Input:

- Gameplay instruction & Screenshot images

- Output:

- Direct Mouse & Keyboard actions



The FlashAdventure Benchmark: Key Challenge

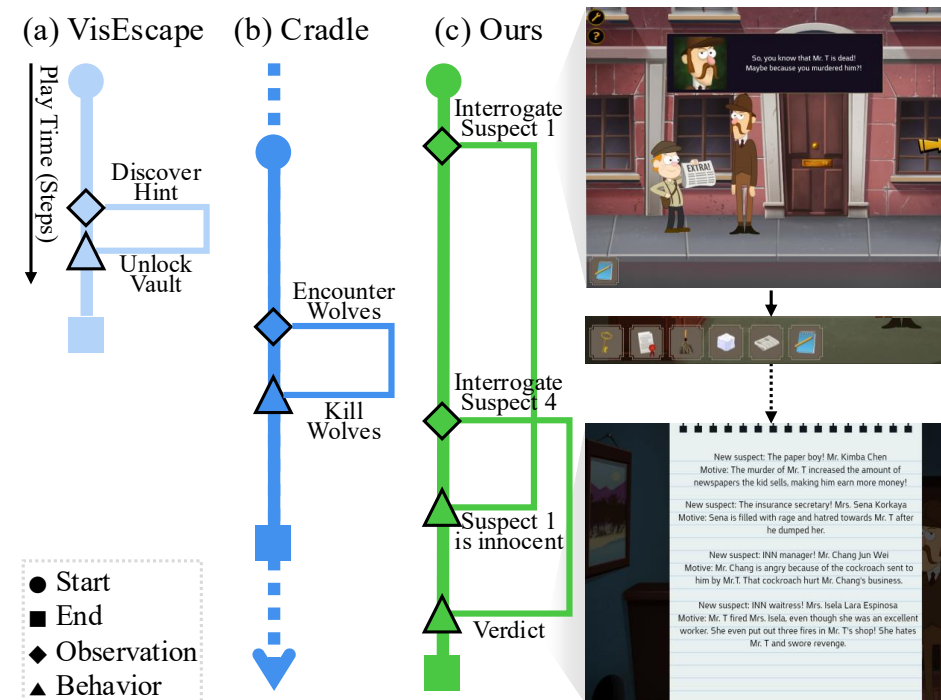
- Long-term “**Observation-Behavior Gap**”
 - Time lag btw {Whe agent *observe information*} vs. {when it *acts upon it*}
 - Ex) Detective {interrogates suspect} vs. {later discover their guilt / innocence}
 - It's crucial when solving full story arcs
 - + Tolman's theory on “**latent learning**”^[1]
 - Humans retrieve & apply clues after long delay
 - Question: “*Do agents have this ability, too?*”

Prior benchmarks:

(a) Include **short story arcs**, or (b) focus on **short-term objectives**

FlashAdventure:

(c) Emphasize completion of full story arcs involving long-term objectives



[1] Tolman, *Cognitive maps in rats and men*, Psychological review 2025

The FlashAdventure Benchmark: Gameplay

- Game Selection

- Utilize *FlashPoint Archive*^[1] for secure playback of Flash-based browser games
- Selected **34 adventure games** according to:
 - (1) Free-to-play, (2) games prioritizing reasoning over speed, (3) validated human walkthrough

- Human Gameplay

- 13 participants to play through all 34 games
- Stats.
 - Completed full story acrs w/ **1,142 steps** & **26 minutes**
 - Much longer than prior room escape benchmark^[2] (52.8 steps)
 - Success rate of avg **97.1%**
 - Long-term observation-behavior gap of **251.1 steps**

(1-a) Aquired the newspaper [118 steps]



(1-b) Gave it to a man and received his voucher [217 steps]



(2-a) Aquired the duster [43 steps]



(2-b) Gave it to a woman and unlocked new suspects [464 steps]



[1] <https://flashpointarchive.org/>

[2] Lim et al., *VisEscape: A Benchmark for Evaluating Exploration-driven Decision-making in Virtual Escape Rooms*, arXiv 2025

The FlashAdventure Benchmark: Evaluation

- Evaluation Metrics

- Success Rate: 0% or 100% (=Completed)
- Milestone Completion Rate: 0 ~ 100%
 - Authors manually defined milestones (either discrete/continuous) to measure progress

- (Optional) Steps

- Automatic Evaluation

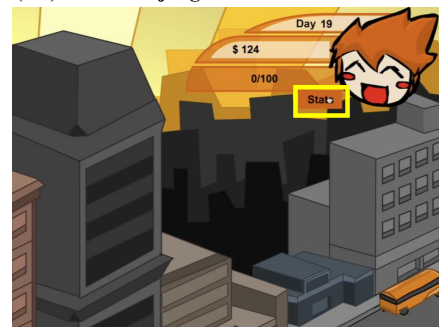
- Problem: Prior studies **lacked automatic eval.** and **relied on human eval.**

- Solution: **CUA-as-a-Judge**

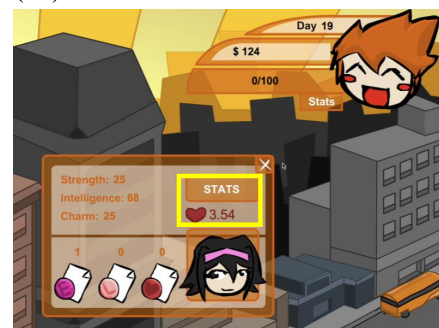
- CUA-as-a-Judge **interacts with game environment & executes actions** to verify whether **milestones have been achieved**, simulating human judging process

- CUA vs. Human Judge: Acc. **94%**, Spearman **0.9912**, Pearson **0.9999**

(1-a) CUA-as-a-judge clicks Stats.



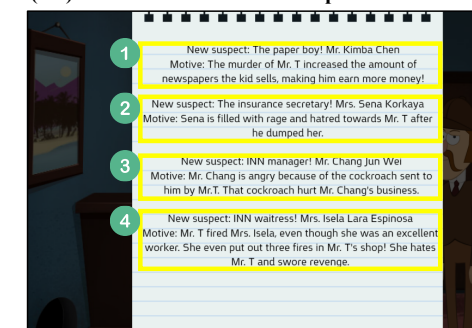
(1-b) It counts the affection score.



(2-a) CUA-as-a-judge clicks Notebook.



(2-b) It counts the number of suspect.



The FlashAdventure Benchmark: Comparison

- Largest benchmark to evaluate GUI agents solving full story arcs

Benchmark / Framework	# Games	Environment	Free?	Automatic Evaluation	Complete Story Arc	Featured Games
Code/API-based						
TextStarCraft II	1	API	✓	✓	✗	StarCraft II
BALROG	6	API & Screen	✓	✓	✗	MiniHack, NLE, Baba Is AI, Crafter, BabyAI, TextWorld
LVLM-Playground	6	API & Screen	✓	✓	✗	TicTacToe, Reversi, Sudoku, Minesweeper, Gomoku, Chess
VisEscape	-*	API & Screen	✓	✓	✓	*Room escape game created for research, instead of adapting existing ones
Orak	12	API & Screen	✗	✓	✗	2 Games × 6 Genres (Action, Adventure, RPG, Simulation, Strategy, Puzzle)
Pixel/Screenshot-based						
OOTB	2	Screen Only	✓	✗	✗	Hearthstone, Honkai: Star Rail
VARP	1	Screen Only	✗	✗	✗	Black Myth: Wukong (AAA game)
Cradle	4	Screen Only	✗	✗	✗	RDR2 (AAA game), Stardew Valley, Cities: Skylines, Dealer's Life 2
FlashAdventure	34	Screen Only	✓	✓	✓	Classic Adventure Games (Mystery/Detective, Hidden Object, Room Escape, Visual Novel, Life/Management Simulation)

Approach

- **COAST (Clue-Oriented Agent for Sequential Tasks)**
 - (1) Clue Memory & (2) Seek-Map-Solve cycle

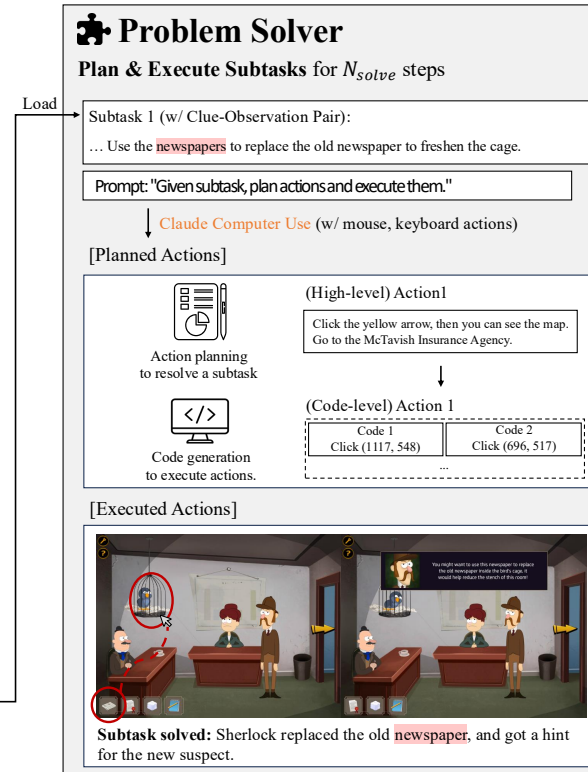
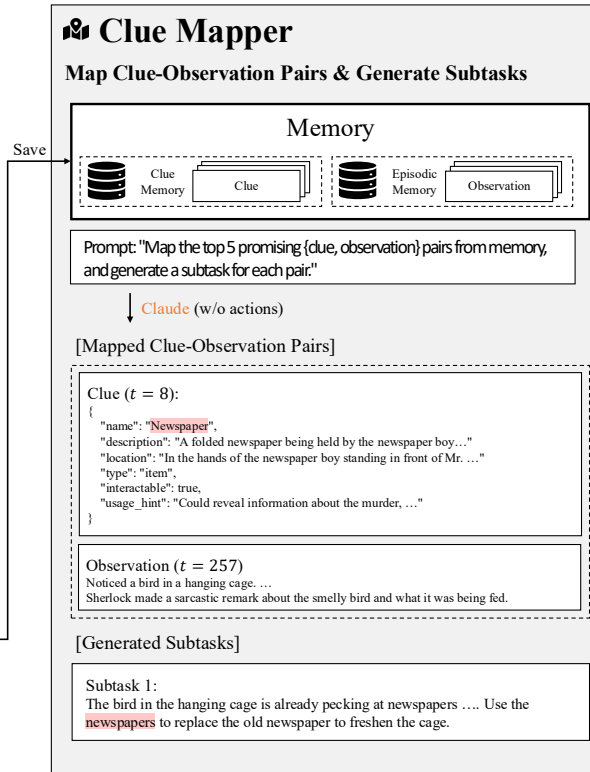
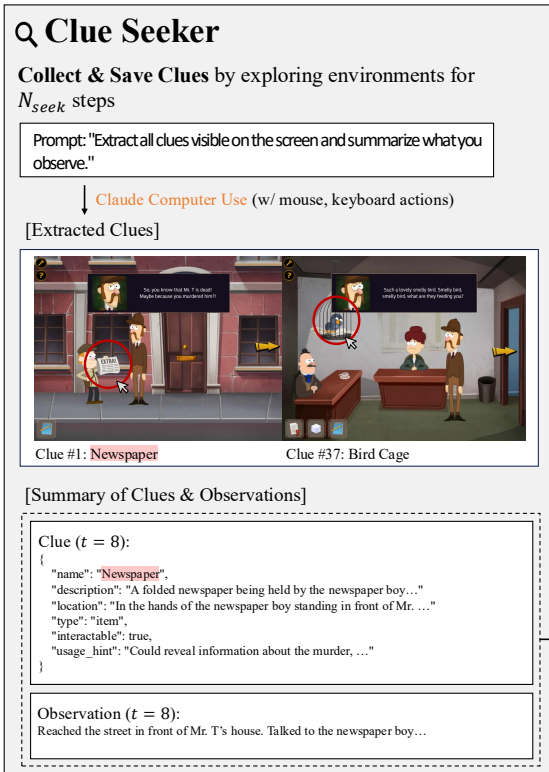
Algorithm 1: COAST Framework with Seek-Map-Solve Cycle.

Input: task query q , maximum step T
Data: $\mathcal{M}$ (clue memory), τ (trajectory), $\mathcal{G}_R$ (resolved-goal set)

```

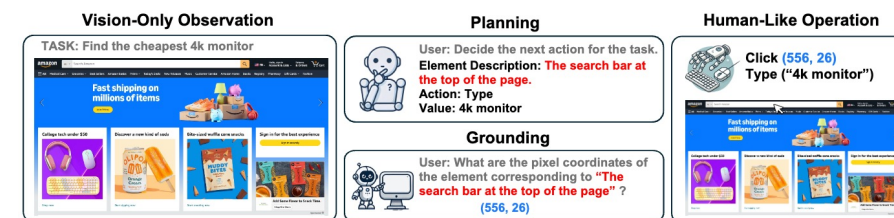
1  $\mathcal{M} \leftarrow \emptyset, \tau \leftarrow \emptyset, \mathcal{G}_R \leftarrow \emptyset, t \leftarrow 0$ 
2 while  $t < T$  do
    // 1. Clue Seeker
3 for  $i = 1$  to  $N_{seek}$  do
4     if  $t \geq T$  then
5         break
6     Observe frame  $o_t$ 
7      $a_t \leftarrow \pi_{\theta}^{seek}(o_t, \mathcal{M}, q)$ 
8      $\Delta \mathcal{M} \leftarrow \text{Execute action } a_t$ 
9      $\mathcal{M} \leftarrow \mathcal{M} \cup \Delta \mathcal{M}$ 
10     $\tau \leftarrow \tau \oplus (o_t, a_t)$  // append  $(o_t, a_t)$  to  $\tau$ 
11
12     $t \leftarrow t + 1$ 
13 if  $t \geq T$  then
14     break
15 // 2. Clue Mapper
16  $\mathcal{G} \leftarrow \pi_{\phi}^{map}(\mathcal{M}, \tau, q)$ 
17  $\mathcal{G} \leftarrow \mathcal{G} \setminus \mathcal{G}_R$  // filter out resolved goals
18 if  $\mathcal{G} = \emptyset$  then
19     continue // restart Seek block
20 else
21     // 3. Problem Solver
22     foreach  $g \in \mathcal{G}$  do
23         for  $j = 1$  to  $N_{solve}$  do
24             if  $t \geq T$  then
25                 break
26              $a_t \leftarrow \pi_{\psi}^{solve}(o_t, g, q)$ 
27             success  $\leftarrow \text{Execute action } a_t$ 
28              $\tau \leftarrow \tau \oplus (o_t, a_t)$ 
29             if success then
30                  $\mathcal{G}_R \leftarrow \mathcal{G}_R \cup \{g\}$ 
31              $t \leftarrow t + 1$ 

```

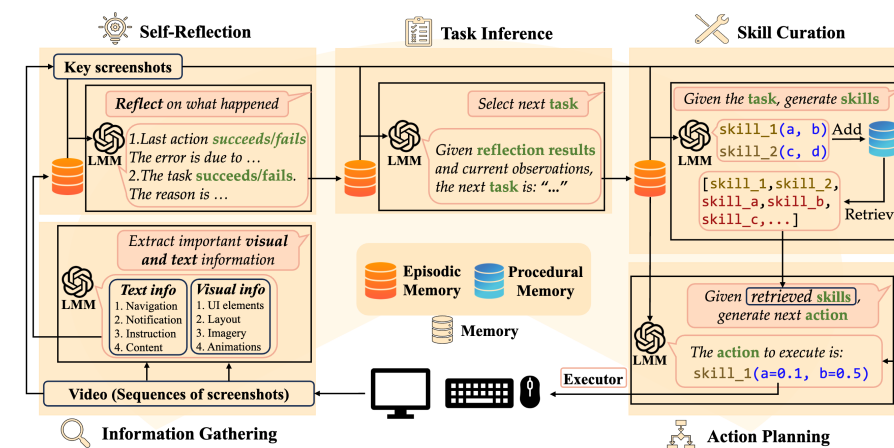


Experiments: Evaluation Protocol (1)

- Baseline agents
 - Modular (Model + GUI grounding module + Agentic Framework)
 - GPT-4o^[1] + UGround-V1-7B^[2] + Cradle^[3]
 - Claude-3.7-Sonnet^[4] + UGround-V1-7B + Cradle
 - Claude-3.7-Sonnet + Claude-3.7-Sonnet + Cradle
 - Claude-3.7-Sonnet + Claude-3.7-Sonnet + Agent S2^[5]
 - End-to-end
 - UI-TARS-1.5-7B^[6]
 - OpenAI CUA^[7]
 - Claude-3.7-Sonnet Computer-Use^[8]
 - Claude-3.7-Sonnet Computer-Use + COAST (Ours)



(image credit: Uground^[2])



(image credit: Cradle^[3])

[1] OpenAI, GPT-4o system card, <https://openai.com/index/gpt-4o-system-card/>

[2] Gou et al., Navigating the Digital World as Humans Do: Universal Visual Grounding for GUI Agents, ICLR 2025

[3] Tan et al., Cradle: Empowering Foundation Agents towards General Computer Control, ICML 2025

[4] Anthropic, Claude 3.7 Sonnet and Claude Code, <https://www.anthropic.com/news/claude-3-7-sonnet>

[5] Agashe et al., Agent s2: A compositional generalist-specialist framework for computer use agent, COLM 2025

[6] Qin et al., Ui-tars: Pioneering automated gui interaction with native agents, arXiv 2025

[7] OpenAI, Computer-Using Agent, <https://openai.com/index/computer-using-agent/>

[8] Anthropic, Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku, <https://www.anthropic.com/news/3-5-models-and-computer-use>

Experiments: Evaluation Protocol (2)

- Experimental settings
 - Max # steps per game = 1,000
 - Evaluated on all 34 games

Subgenres	Games
Point-and-Click Adventure (Mystery/Detective)	Sherlock Holmes: The Tea Shop Murder Mystery
	Sherlock Holmes 2
	Vortex Point 1
	Vortex Point 2
	Vortex Point 3
	Pierre Hotel
	Small Town Detective
	Dakota Winchester's Adventures
	Saucy Devil Gordon
	Ray and Cooper 2
	Nick Bounty: A Case of the Crabs
Hidden Object	Grim Tales: The Bride
	Grim Tales: The Legacy Collector's Edition

Subgenres	Games
Room Escape	Computer Office Escape
	Crimson Room
	Camping Room Escape
	Chemical Room Escape
	Space Museum Escape
	Vending Machine Room
	Wood Workshop Escape
	Geometric Room Escape
	Game Cafe Escape
	Machine Room Escape
	VideoStudio Escape
	Design House Escape
	Paint Room Escape
Visual Novel (Dating Sim)	Mirror Room Escape
	Elevator Room Escape
	Pico Sim Date
	Festival Days Sim Date
Simulation: Life	Kingdom Days
	Idol Days Sim Date
Simulation: Management	Community College Sim
	Sort the Court

Experiments: Results (1)

- Current GUI agents **struggle with full story arc completion**
 - Best: **5.88% SR**

Model	GUI Grounding / Action Execution	Agentic Framework	Success Rate↑ (%)	Milestone Completion Rate↑ (%)	# Steps
GPT-4o	Uground-V1-7B / pyautogui	Cradle	0.00	4.56	1000.0
	Uground-V1-7B / pyautogui	Cradle	0.00	6.59	1000.0
Claude-3.7-Sonnet		Cradle	0.00	10.60	1000.0
	Claude-3.7-Sonnet / pyautogui	Agent S2	0.00	1.20	1000.0
	UI-TARS-1.5-7B		0.00	6.93	1000.0
	OpenAI CUA		5.88	15.39	954.1
	Claude-3.7-Sonnet Computer-Use		0.00	<u>17.11</u>	992.4
	Claude-3.7-Sonnet Computer-Use + COAST (Ours)		5.88	19.89	966.8
	Human Performance (max 1,000 steps)		50.98	78.98	815.5
	Human Performance (unlimited)		97.06	100.00	1142.0

Experiments: Results (2)

- COAST achieves the **highest SR & MCR**
 - Improves SR / MCR by 5.88 / 2.78 % point

Model	GUI Grounding / Action Execution	Agentic Framework	Success Rate↑ (%)	Milestone Completion Rate↑ (%)	# Steps
GPT-4o	Uground-V1-7B / pyautogui	Cradle	0.00	4.56	1000.0
	Uground-V1-7B / pyautogui	Cradle	0.00	6.59	1000.0
Claude-3.7-Sonnet		Cradle	0.00	10.60	1000.0
	Claude-3.7-Sonnet / pyautogui	Agent S2	0.00	1.20	1000.0
	UI-TARS-1.5-7B		0.00	6.93	1000.0
	OpenAI CUA		5.88	15.39	954.1
	Claude-3.7-Sonnet Computer-Use		0.00	<u>17.11</u>	992.4
	Claude-3.7-Sonnet Computer-Use + COAST (Ours)		5.88	19.89	966.8
	Human Performance (max 1,000 steps)		50.98	78.98	815.5
	Human Performance (unlimited)		97.06	100.00	1142.0

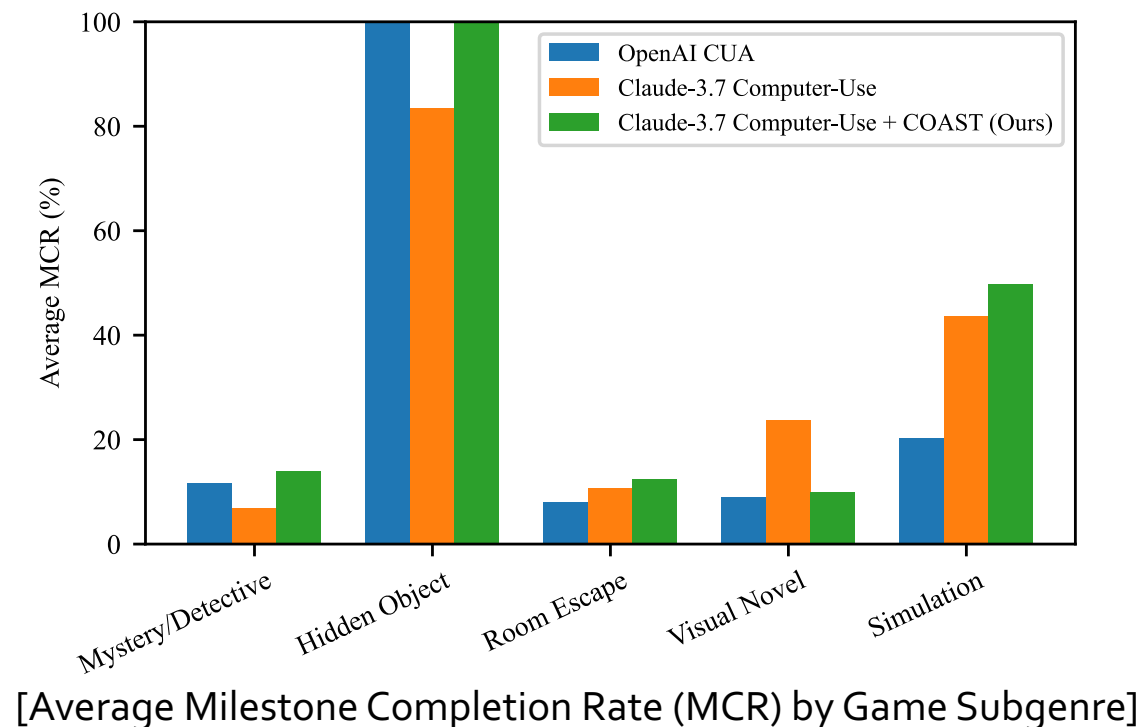
Experiments: Results (3)

- **Significant gap** btw agents (**5.88%**) vs. human (**97.06%**)
 - Agents exhibit **weak planning, poor visual perception, deficient lateral thinking**

Model	GUI Grounding / Action Execution	Agentic Framework	Success Rate↑ (%)	Milestone Completion Rate↑ (%)	# Steps
GPT-4o	Uground-V1-7B / pyautogui	Cradle	0.00	4.56	1000.0
	Uground-V1-7B / pyautogui	Cradle	0.00	6.59	1000.0
Claude-3.7-Sonnet		Cradle	0.00	10.60	1000.0
	Claude-3.7-Sonnet / pyautogui	Agent S2	0.00	1.20	1000.0
	UI-TARS-1.5-7B		0.00	6.93	1000.0
	OpenAI CUA		5.88	15.39	954.1
	Claude-3.7-Sonnet Computer-Use		0.00	<u>17.11</u>	992.4
	Claude-3.7-Sonnet Computer-Use + COAST (Ours)		5.88	19.89	966.8
	Human Performance (max 1,000 steps)		50.98	78.98	815.5
	Human Performance (unlimited)		97.06	100.00	1142.0

Experiments: Analysis (1)

- COAST improves over games requiring *long-term memory & planning*
 - **Mystery/Detective & Room Escape:** Benefit from *clue-based reasoning*
 - **Visual novels:** Inconsistent trends, as observation–behavior gap is less pronounced



Experiments: Analysis (2)

- Ablation study

- Seeker

- Merely identifying clues
w/o subsequent planning
is insufficient

- Seeker + Solver

- Suboptimal MCR in clue-reach
*mystery/detective &
room escape* games

- COAST

- Highest overall performance
 - Mapper helps effective subtask
plans by connecting
raw observations and clues.

Game (Subgenre)	Metric	Seeker	Seeker + Solver	COAST
Sherlock Holmes 2 (Mystery/Detective)	Success (%)	0.0	0.0	0.0
	Milestone (%)	37.5	37.5	62.5
	# steps	1000	1000	1000
Grim Tales: The Bride (Hidden Object)	Success (%)	0.0	100.0	100.0
	Milestone (%)	66.7	100.0	100.0
	# steps	1000	790	225
Camping Room Escape (Room Escape)	Success (%)	0.0	0.0	0.0
	Milestone (%)	22.2	33.3	44.4
	# steps	1000	1000	1000
Idol Days Sim Date (Visual Novel)	Success (%)	0.0	0.0	0.0
	Milestone (%)	2.3	25.6	25.6
	# steps	1000	1000	1000
Sort the Court (Simulation)	Success (%)	0.0	0.0	0.0
	Milestone (%)	83.2	93.0	95.5
	# steps	1000	1000	1000
Total	Success (%)	0.0	20.0	20.0
	Milestone (%)	42.4	57.9	65.6
	# steps	1000	958	845

Experiments: Analysis (3)

- Failure Analysis & Mitigation

- COAST improves planning in *hidden object*
- It alleviates lateral thinking in *mystery/detective* & *room escape*
- Mixed results for resource management
- Fails to mitigate poor visual perception

Game (Subgenre)	GPT-4o + UGround + Cradle	Claude-3.7 Computer-Use	Claude-3.7 Computer-Use + COAST
Sherlock Holmes 2 (Mystery/Detective)	1,2,3	1,2,3	1,2
Grim Tales: The Bride (Hidden Object)	1,2	1	-
Camping Room Escape (Room Escape)	1,2,3	1,2,3	1,2
Idol Days Sim Date (Visual Novel)	4	4	4
Sort the Court (Simulation)	4	4	-

[Failure patterns:

1 = planning, 2 = perception, 3 = lateral thinking, 4 = resource management]

Concluding Remarks

- **FlashAdventure** Benchmark

- Largest benchmark to evaluate GUI agents solving full story arcs
- Tackles challenge of long-term ***observation-behavior gap***
- Automated gameplay evaluation via **CUA-as-a-Judge**

- **COAST** Framework

- Long-term **clue memory** to address observation-behavior gap
- **Seek-Map-Solve** cycle for better planning & solving

- Empirical findings: **Current GUI agents struggle w/ full story arcs**

- COAST **bridges observation-behavior gap**, yet far below human performance

Thank you

Code <https://github.com/ahnjaewoo/FlashAdventure>
Paper <https://arxiv.org/abs/2509.01052>
Webpage <https://ahnjaewoo.github.io/flashadventure/>
Contact jaewoo.ahn@vision.snu.ac.kr,
junseo.kim@vision.snu.ac.kr

