



Project



Jaewoo Ahn



Junseo Kim

FlashAdventure: A Benchmark for GUI Agents Solving Full Story Arcs in Diverse Adventure Games

Jaewoo Ahn*, Junseo Kim*, Heeseung Yun, Jaehyeon Son, Doingmin Park, Jaewoong Cho, Gunhee Kim (* Equal Contribution)

Can GUI agents solve adventure games?

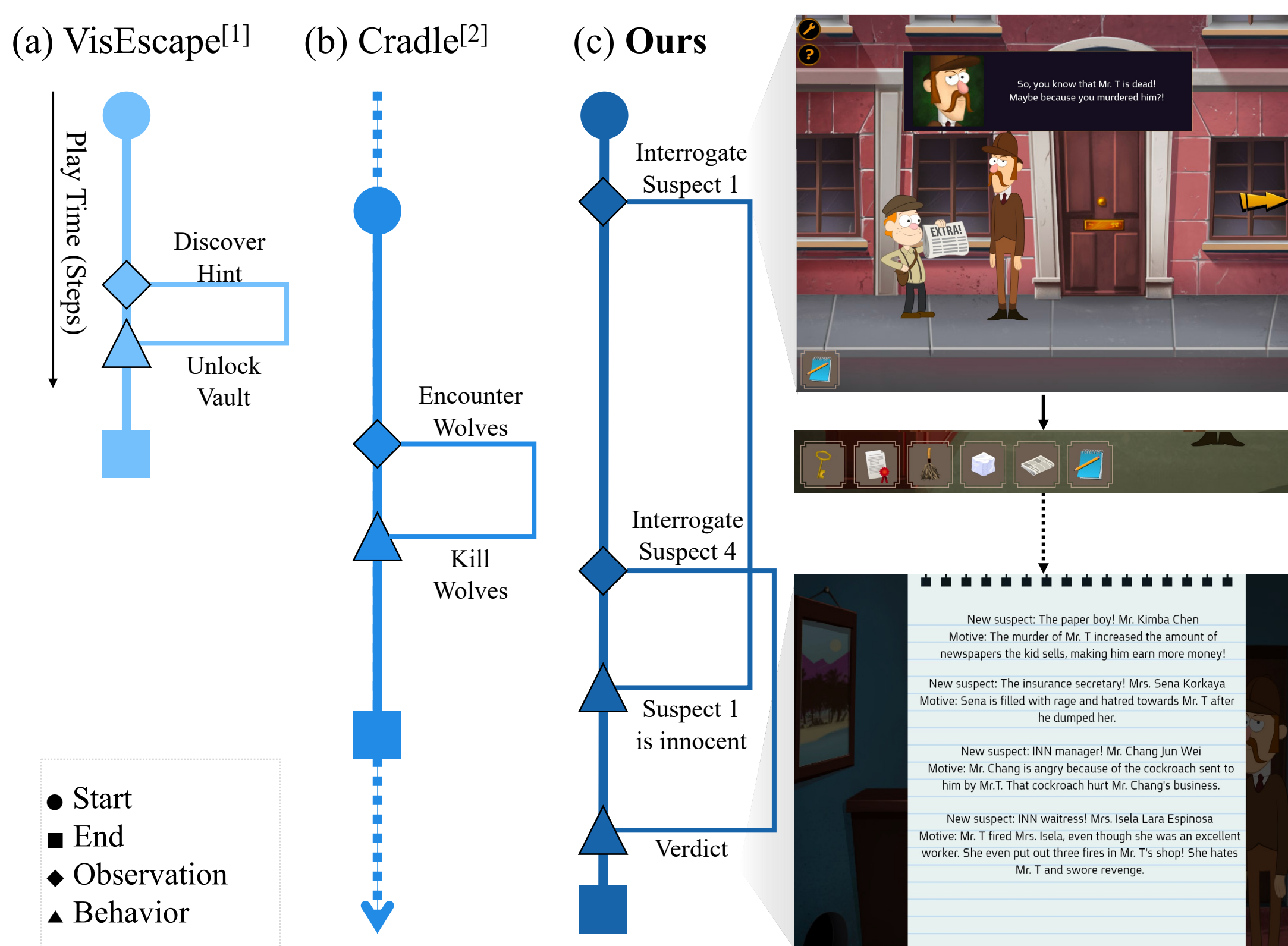
(🕵️ Mystery/Detective, 🔍 Hidden Object, 🏠 Room Escape, 📖 Visual Novel, 🏠 Simulation)

Motivation

Evaluation of game-playing GUI agent (or Computer-Using Agent, CUA)'s **Completion of Full Story Arcs** → *But why is it important?*

Challenge: Long-term Observation–Behavior Gap

Critical delay between {seeing a clue} and {acting on it} later

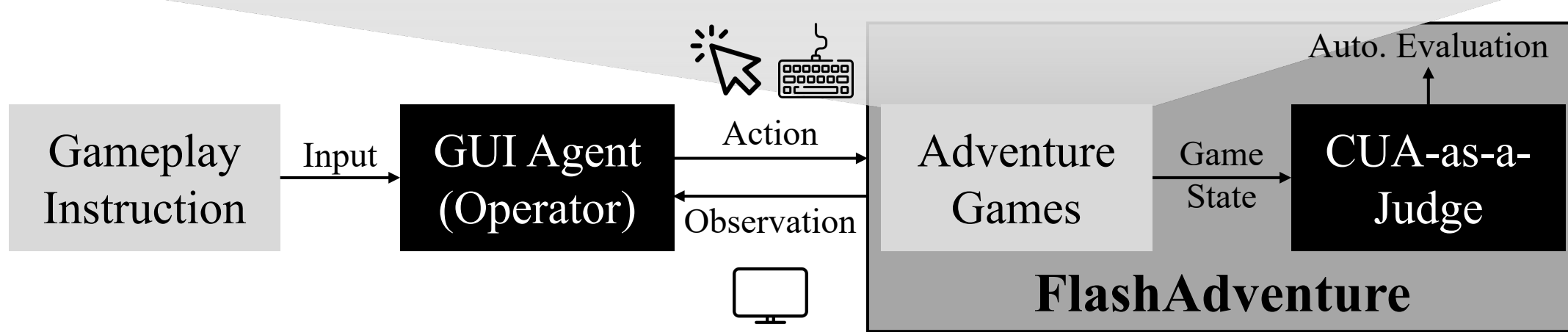


Prior benchmarks: (a) short story arcs (52.8 steps), (b) short-term objectives

Ours: (c) long story arcs (**1142 steps**, 26 min) & long obs.-beh. gap (**251.1 steps**)

FlashAdventure

Benchmark of **34** Flash-based adventure games designed to test agents' **full story arcs completion** & tackle **observation-behavior gap**



Benchmark / Framework	# Games	Environment	Free?	Automatic Evaluation	Complete Story Arc
Cradle ^[2]	4	Screen Only	✗	✗	✗
BALROG ^[3]	6	API & Screen	✓	✓	✗
VisEscape ^[1]	-*	API & Screen	✓	✓	✓
FlashAdventure	34	Screen Only	✓	✓	✓

*Room escape game created for research, instead of adapting existing ones

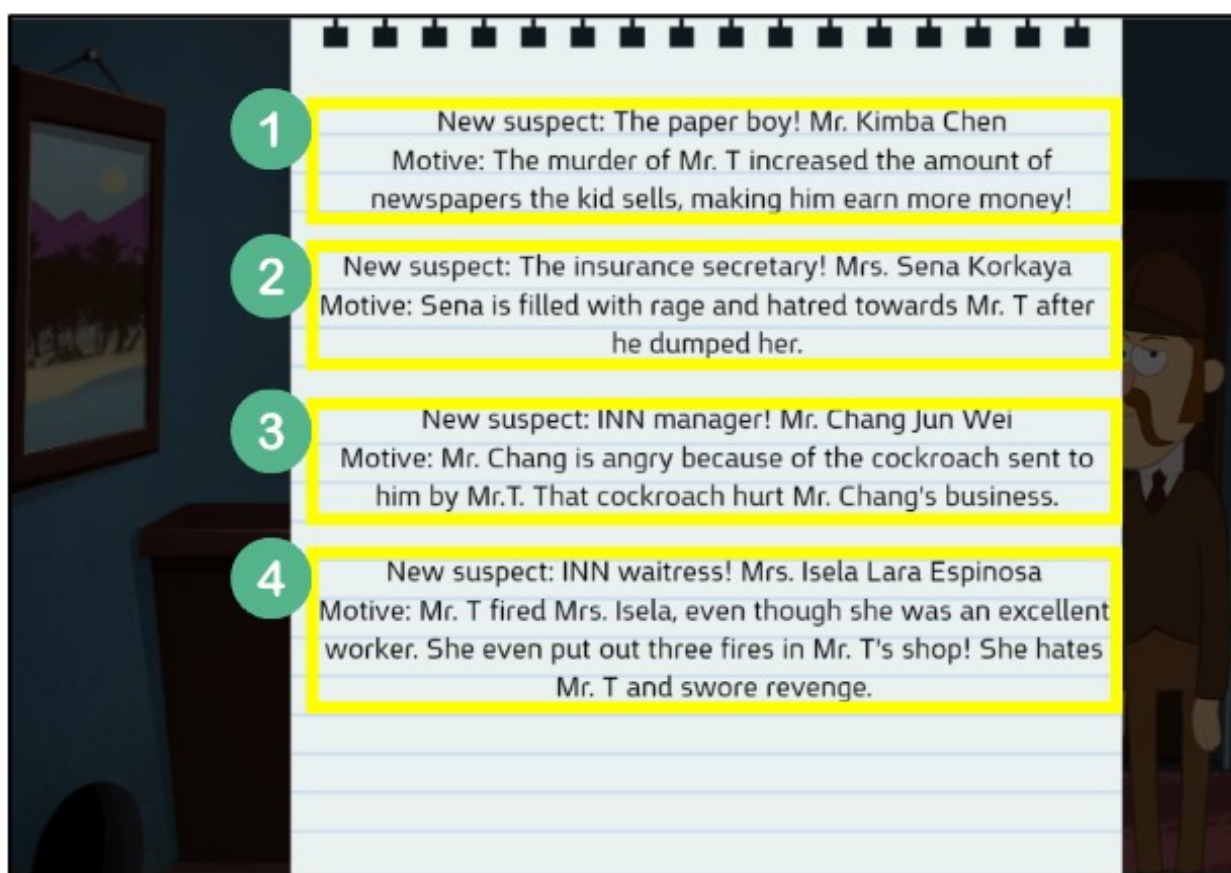
CUA-as-a-Judge

Agentic judge that interacts with environment to verify whether predefined milestones have been achieved (Accuracy: **94%**, Spearman's ρ : **0.9912**)

1. CUA-as-a-judge clicks Notebook.



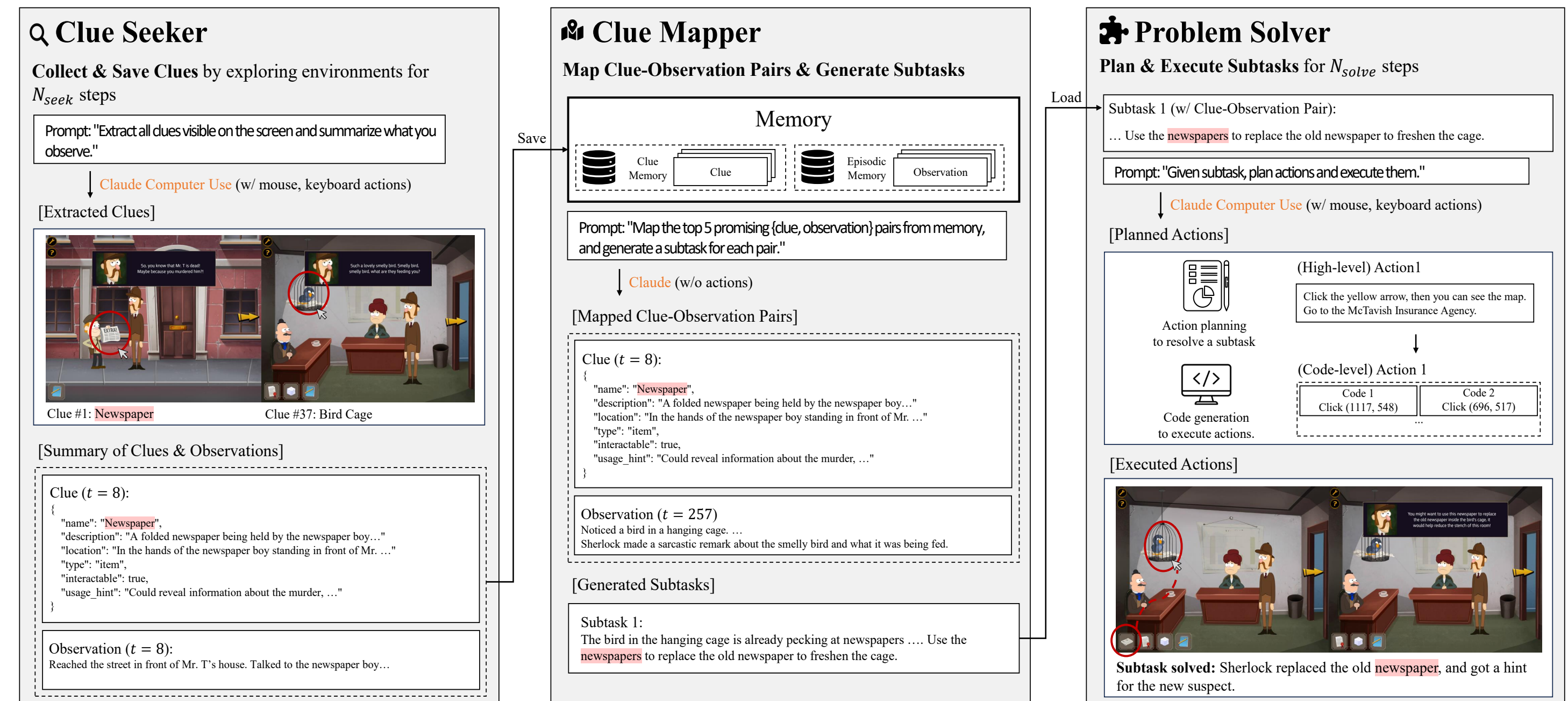
2. It counts the number of suspect.



COAST (Clue-Oriented Agent for Sequential Tasks)

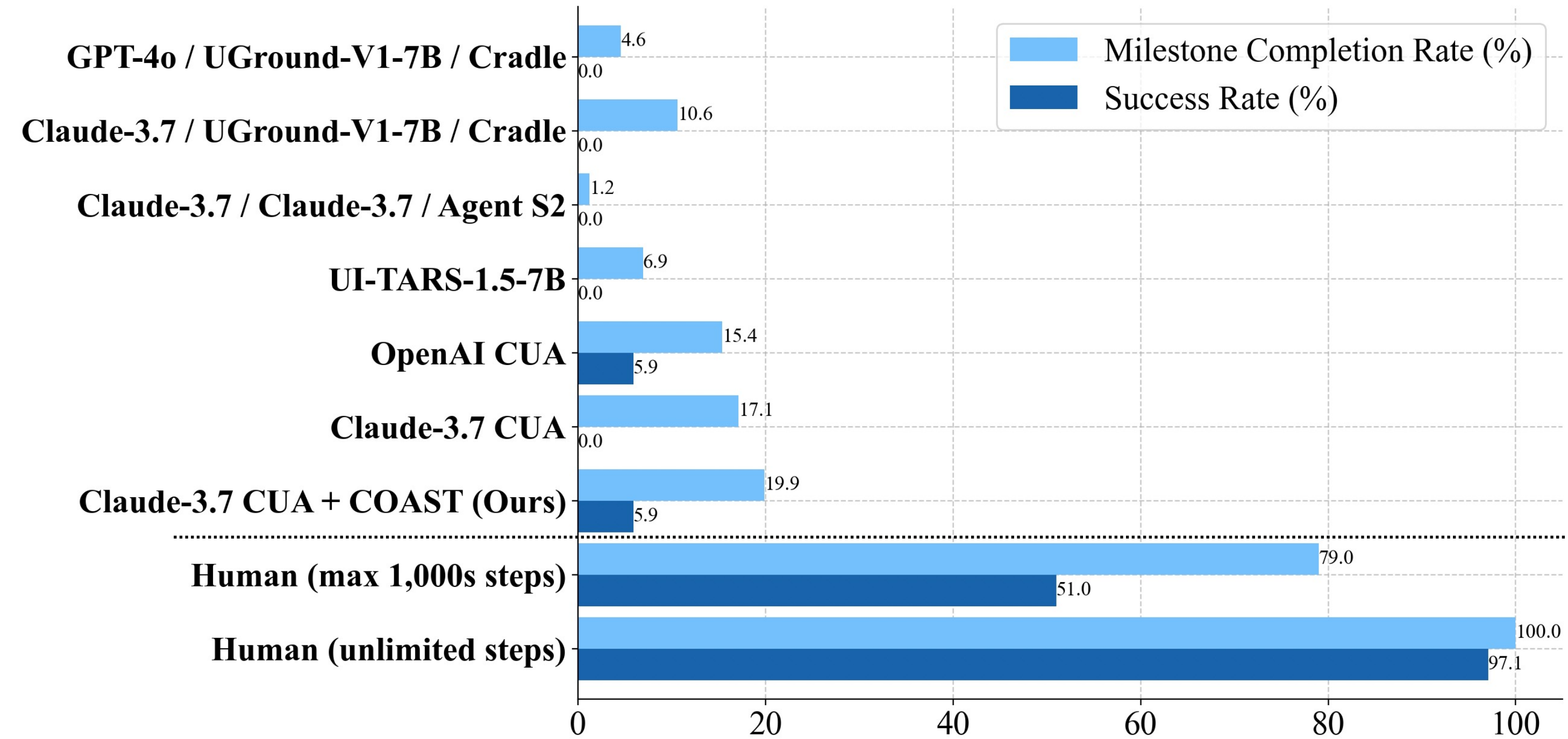
Agentic framework that addresses observation–behavior gap through

(1) **Clue Memory** & (2) **Seek–Map–Solve Cycle**



Experiments

Comparison of GUI agents across all 34 video tasks games (max 1,000 steps)



- GUI agents **struggle with full story arc completion** (< 6% success rate)
- COAST improves SR / MCR by **5.9 / 2.8% points**
- Still, **Significant gap** between GUI agents & human (97.1% vs 5.9%)

Failure Analysis & Mitigation

Comparison of failure patterns:

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

Game (Subgenre)	GPT-4o + UGround + Cradle	Claude-3.7 CUA	Claude-3.7 CUA + 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	-

Using **COAST**,

- Strengths:** Mitigation of planning & lateral thinking failures
- Limitations:** Unresolved issues in perception & resource management

Key Findings

While COAST improved SR/MCR, GUI agents **lag far behind humans in full story arcs** due to diverse limitations: (1) **weak planning**, (2) **poor visual perception**, (3) **deficient lateral thinking**, (4) **lack of resource management**

Reference

[1] Lim et al., *VisEscape: A Benchmark for Exploration-Driven Decision-Making in Virtual Escape Rooms*. EMNLP 2025

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

[3] Paglieri et al., *BALROG: Benchmarking Agentic LLM and VLM Reasoning on Games*. ICLR 2025