



ECCV 2026 | MALMÖ | SEPT 8-12



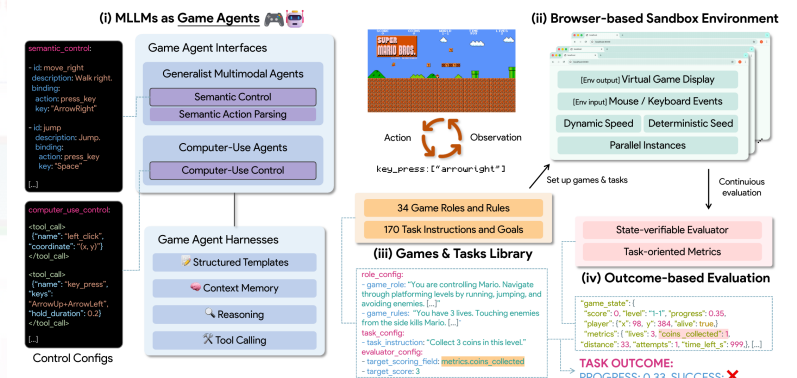
GameWorld: Towards Standardized and Verifiable Evaluation of Multimodal Game Agents

Mingyu Ouyang^{1*} Siyuan Hu^{1*} Kevin Qinghong Lin²
Hwee Tou Ng^{1†} Mike Zheng Shou^{1†}

* Equal contribution † Corresponding authors

¹ National University of Singapore ² University of Oxford

ECCV 2026 Workshop on Multimodal Digital Agents



Why Games?

From game-playing AI to general-purpose agents

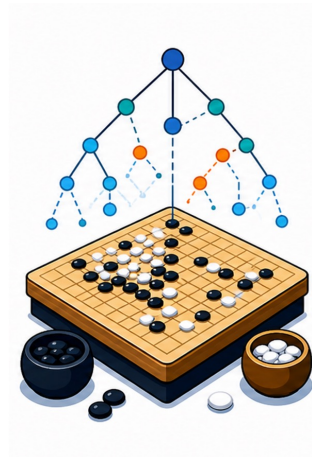
From Atari to Real World



Atari Games

DQN, 2015

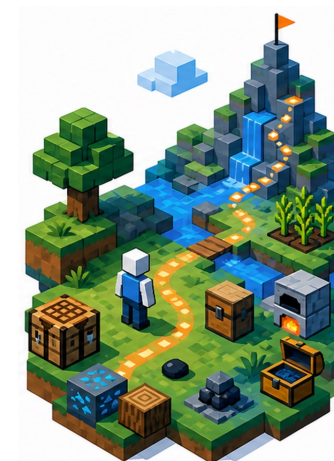
Learning from pixels



Go

AlphaGo, 2016

Search and self-play



Game Agents

2026

General multimodal
interaction in digital world

[1] Mnih, V., et al. Human-level control through deep reinforcement learning. Nature, 2015.

[2] Silver, D., et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016.

GameWorld

A shared protocol across diverse visual interaction games and tasks

34

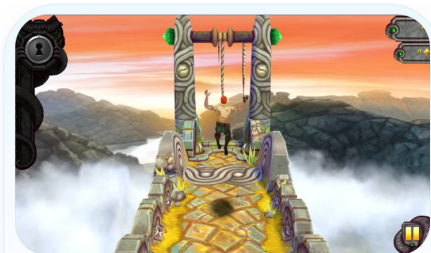
Browser games

170

Verifiable tasks

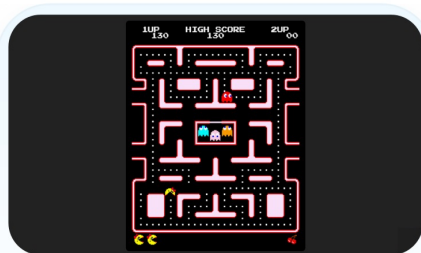
5

Game genres



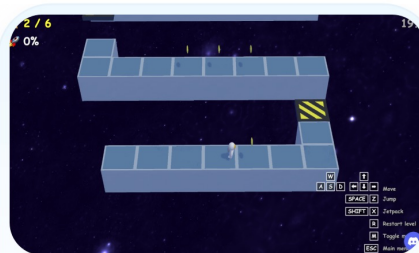
Runner

8 GAMES · 40 TASKS



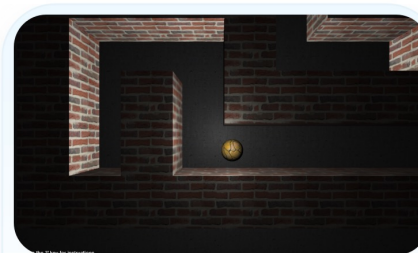
Arcade

7 GAMES · 35 TASKS



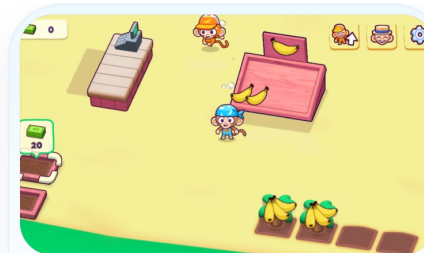
Platformer

8 GAMES · 40 TASKS



Puzzle

7 GAMES · 35 TASKS

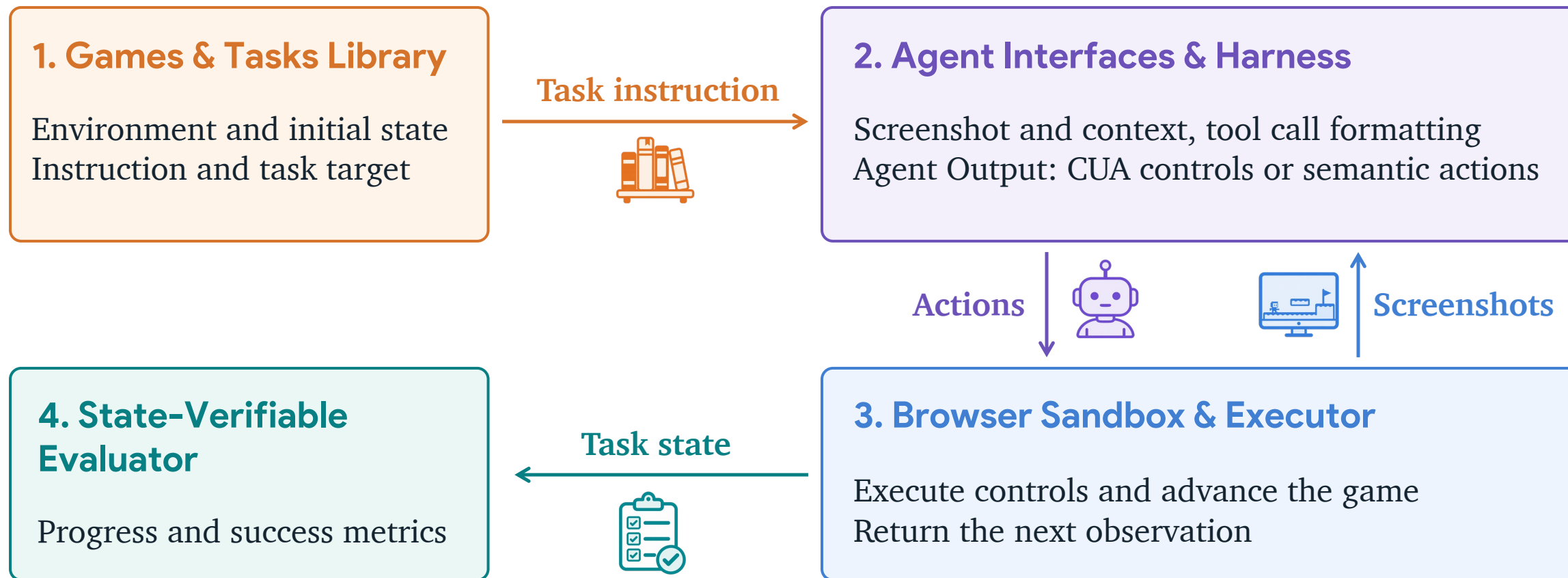


Simulation

4 GAMES · 20 TASKS

The GameWorld Evaluation Loop

Four modules connect task setup, agent interaction and verified outcomes

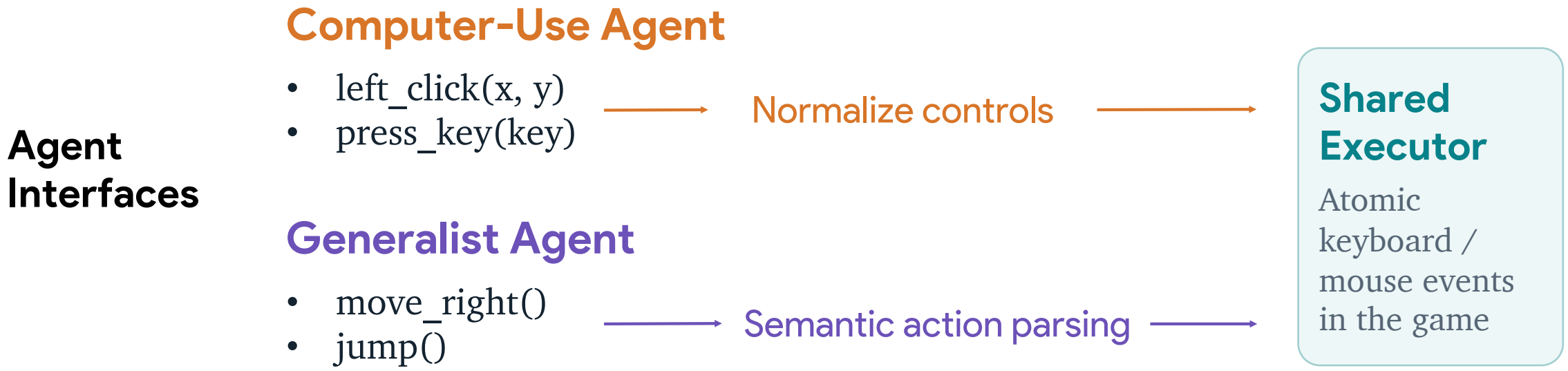


Other shared protocol:

- (i). 100 actions per task; (ii). The game resets when terminated (e.g., player gets killed by the enemy);
- (iii). The game pauses during model inference.

Build Game Agents: Two Action Interfaces, One Runtime

Model outputs become executable game controls



Evaluated Models

CUA-specialized models

- Gemini 2.5 Computer Use
- OpenAI Computer Use
- ...

Evaluated as **CUA only**

General models with CUA capability

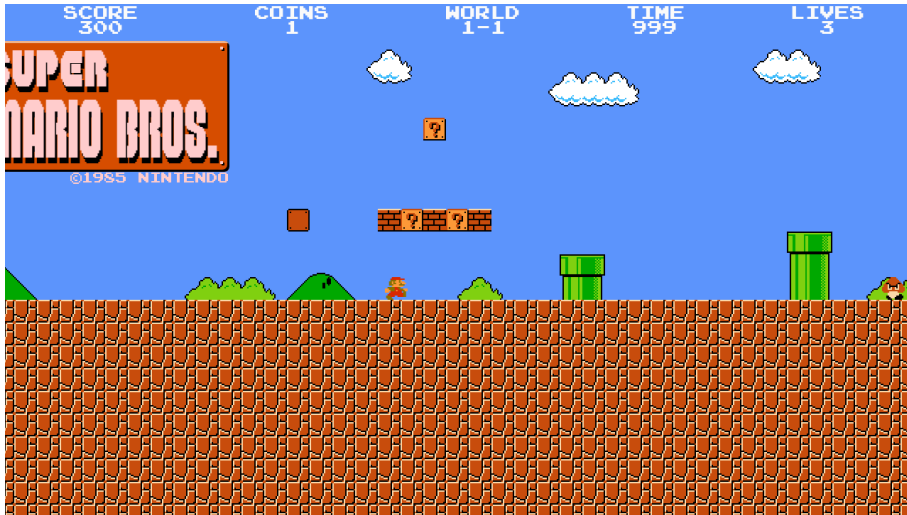
- Claude Sonnet 4.6, Seed-1.8,
- Qwen3-VL-Plus
- ...

Evaluated as **both CUA and Generalist**

Close-Loop Gameplay: Observation, Reasoning and Action

Browser Sandbox & Executor: a Mario interaction

Observation at step t

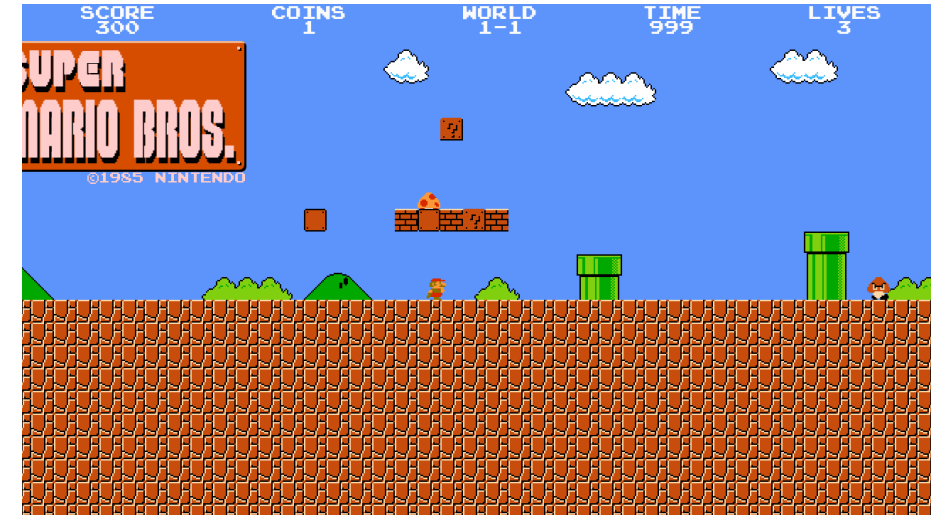


Browser Env
+ Executor:

→

`game_unpause()`
`keydown(ArrowUp)`
`wait(0.25 s)`
`keyup()`
`game_pause()`

Next observation at step $t+1$



Fresh visual feedback

Reasoning

“Mario is now in a good spot to jump up and hit the floating question block.”

Action

`jump()`

Reasoning

“A reward has appeared. I should move toward it and collect it.”

Next Action

`move_right()`

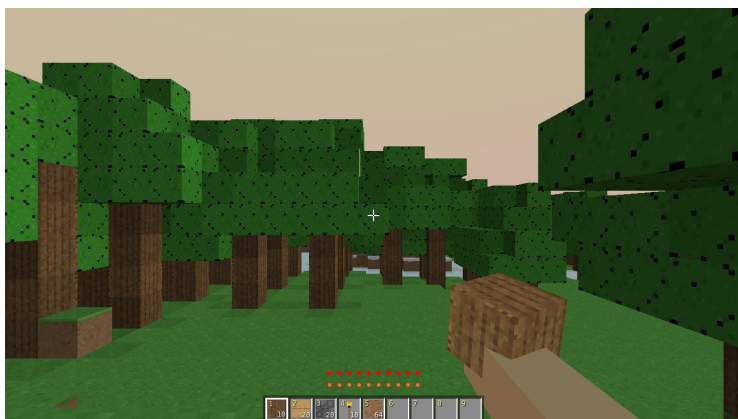
Evaluation: State-Verifiable Progress & Success

Success and progress from the task state

PG (progress): best normalized task progress.

SR (success rate): fraction of runs that reach the full task target.

Minecraft-clone task: "collect 20 dirt blocks"



...



...



Step 1

0 blocks -> 0% progress

Step 50

9 blocks -> 45% progress

Step 100

18 blocks -> 90% progress

Success = 0

100-action budget reached

Progress = 90%

18 of 20 blocks collected

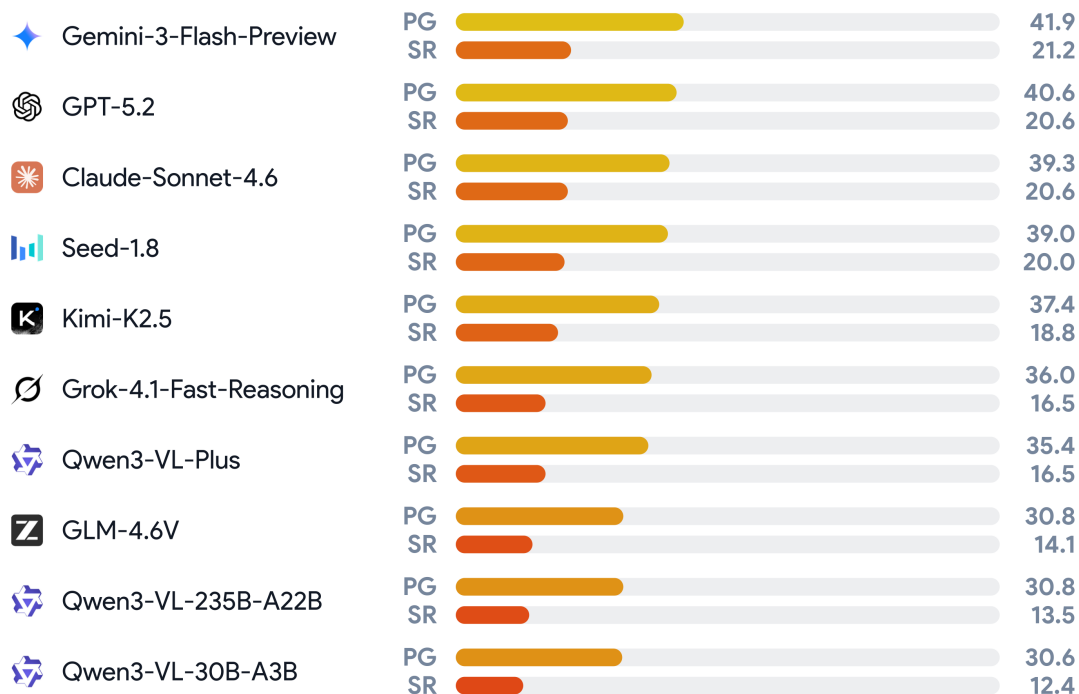
GameWorld Leaderboard

Model performance across two action interfaces

- 13 models, 18 model-interface pairs
- Ranked by PG

Generalist Multimodal Agents

Progress (PG) / Success Rate (SR) (%)



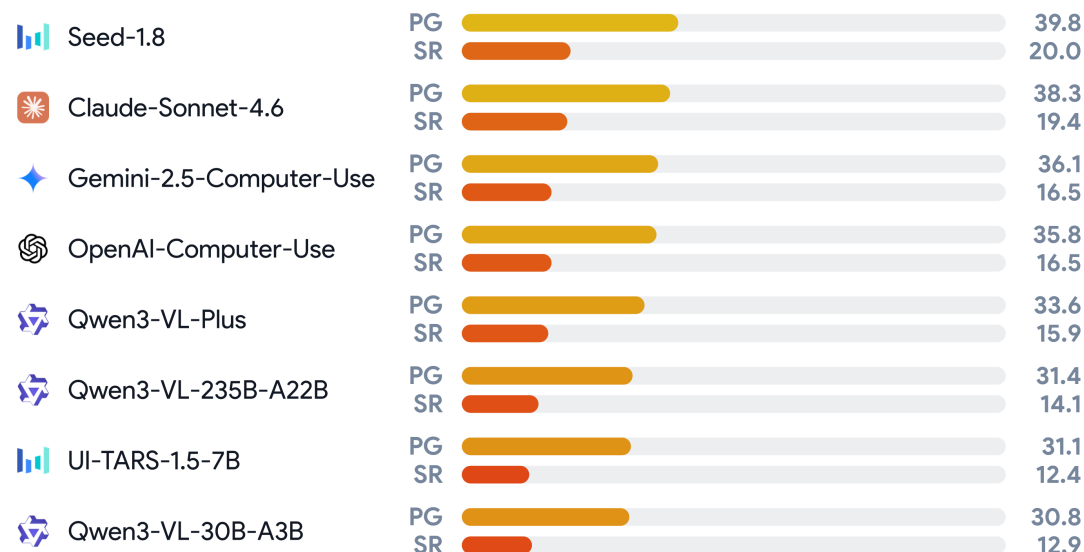
Human Baselines

Progress (PG) / Success Rate (SR) (%)



Computer-Use Agents

Progress (PG) / Success Rate (SR) (%)



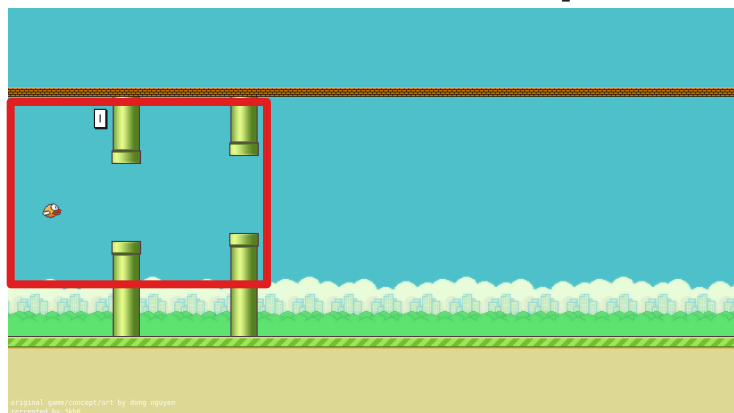
Models still far from novice humans

- Best model SR: 21.2%
- Novice human SR: 55.3%

Finding 1: Visual Similarity Obscures Timing

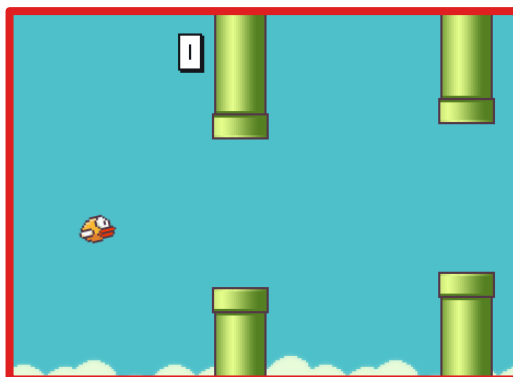
Small visual changes can conceal motion direction and the moment to act.

Full observation (step t)



Red box: region enlarged on the right

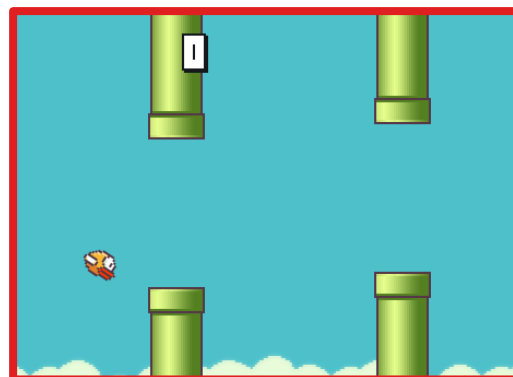
Step t



Wait..

0.2 s

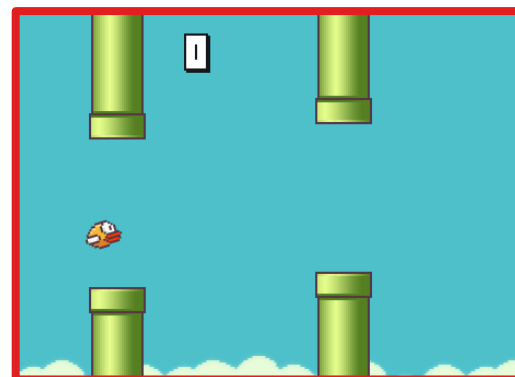
Step $t+1$



Flap!

0.2 s

Step $t+2$



Wait...

0.2 s

Similar position can hide opposite motion.

Models struggle to recover direction and timing from sparse screenshots.

Finding 2: Long-Horizon Goal Following is Fragile

A continuing task can be mistaken for a fresh start.

Minecraft-clone: collect 20 dirt blocks



Visual Hint: Dirt remains directly ahead

1 Continuing to mine

“Dirt is right ahead. I’ll keep mining here.”

2 Losing the active subgoal

*“I need to find dirt blocks before I can begin.
Let me navigate the terrain to find it.”*

3 Restarting the search

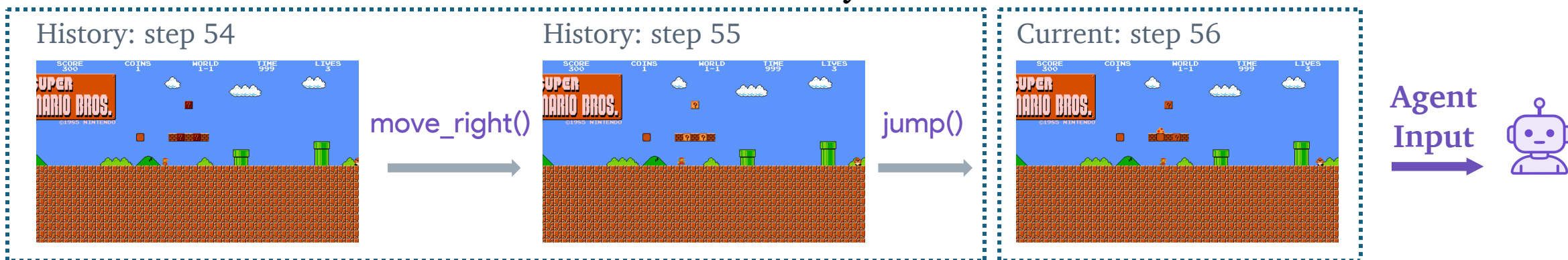
Turns away and searches again,
interrupting the mining strategy

Losing the active subgoal diverts actions away from progress.

Finding 3: More History Helps, but Selectively

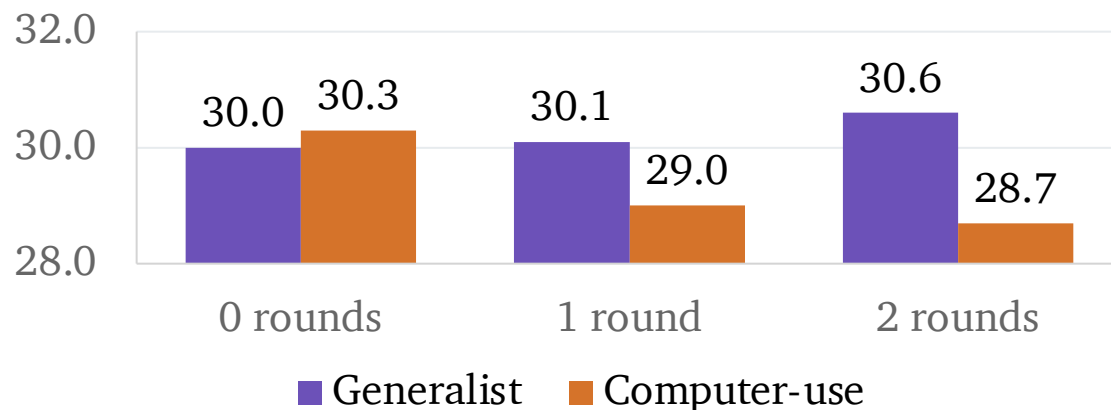
The benefit of additional context depends on the action interface

Agent input with 2 history rounds



Average task progress (%)

Qwen3-VL-235B-A22B



More context, more latency

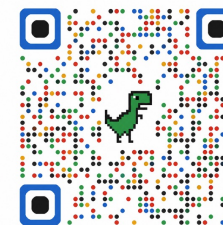
History	Generalist	CUA
0	5.5 s	7.2 s
1	6.8 s (+23.6%)	10.1 s (+40.3%)
2	8.6 s (+56.4%)	12.8 s (+77.8%)

Secs / step, increase vs. 0 history rounds, 1080P settings

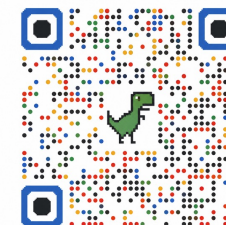
What's Next?

From digital agents to real-world embodiment

Future: Games as a bridge to real-world agents



Code



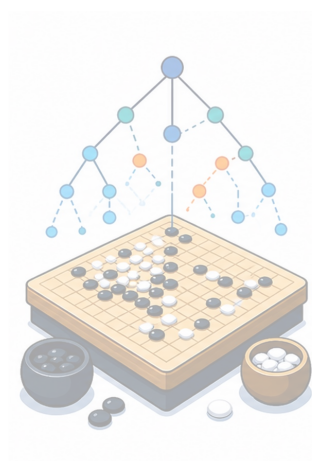
Project Page



Atari

DQN, 2015

Learning from pixels



Go

AlphaGo, 2016

Search and self-play

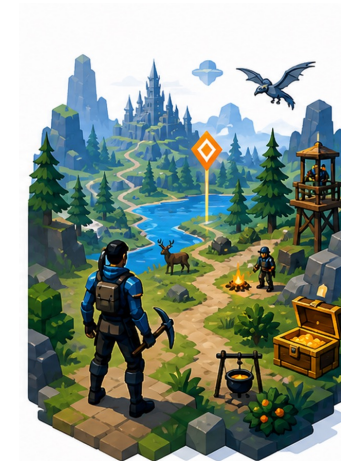


GameWorld

2026

Agent learning
in digital worlds

Generalize



Real-world agents

Future direction

Transfer and adaptation
to physical environments

More visit: gameworld-project.github.io

Thanks to the game creators who bring experience and imagination to life.