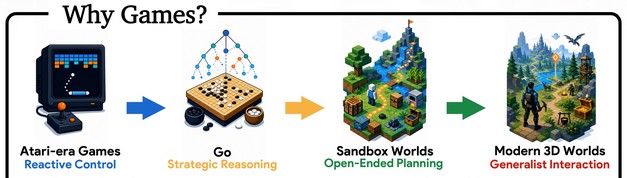
GameWorld

Towards Standardized and Verifiable Evaluation of Multimodal Game Agents

Mingyu Ouyang*[1], Siyuan Hu*[1], Kevin Qinghong Lin[2], Hwee Tou Ng†[1], Mike Zheng Shou†[1]
[1]National University of Singapore, [2]University of Oxford

ECCV 2026 | MALMÖ | SEPT 8–12

Why Games?



- From Atari to open-worlds, games have long driven AI advances.
- Unlike static visual QA, gameplay demands closed-loop perception, reasoning, action, long-horizon planning and recovery from mistakes.
- Yet existing benchmarks remain: (i). fragmented action interfaces, (ii). coupled to inference latency, and (iii). difficult to verify.

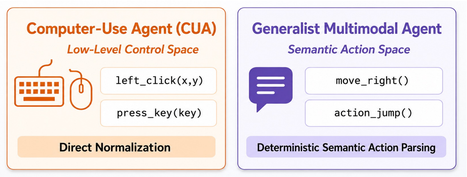
34 Games, 170 Verifiable Tasks

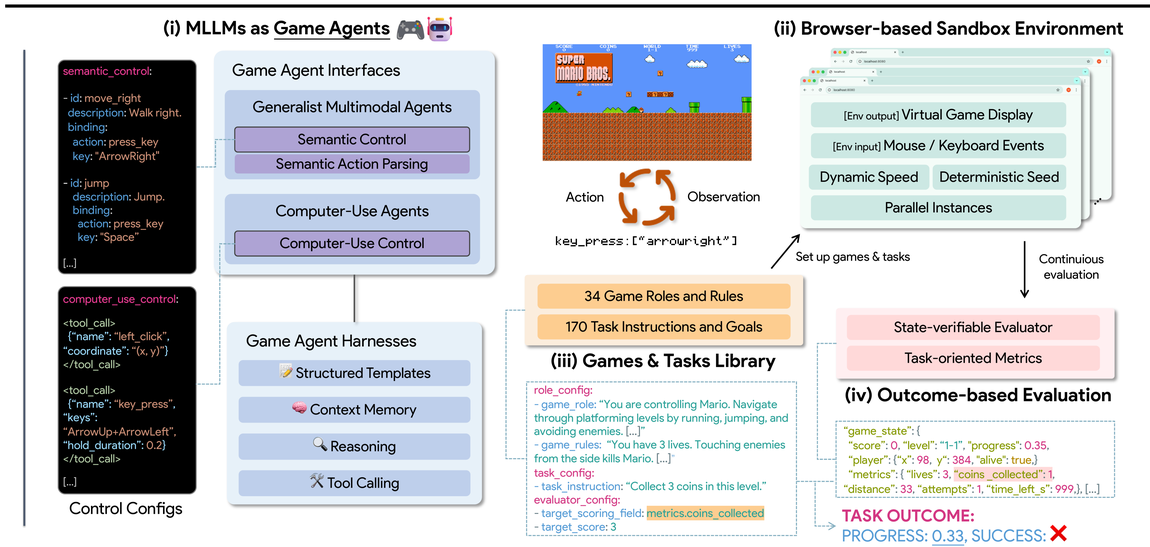
Runner	Arcade	Platformer	Puzzle	Simulation
8 GAMES · 40 TASKS	7 GAMES · 35 TASKS	8 GAMES · 40 TASKS	7 GAMES · 35 TASKS	4 GAMES · 20 TASKS
Continuous state progression with high-frequency reactive control and precise timing for obstacle avoidance.	Fast closed-loop interaction with dynamic multi-entity tracking, reactive evasion, and reward collection.	Spatiotemporal navigation with precise physics-based movement, localized planning, and hazard evasion.	Discrete state-space exploration focused on long-horizon strategy, rule tracking, and logical decision-making.	Open-ended environments that test coordination, resource management, strategic exploration, and error recovery.

- Comprises **34 games** and **170 verifiable tasks** across **five genres**, designed to evaluate complementary gameplay capabilities.

GameWorld: Standardized & Verifiable Evaluation

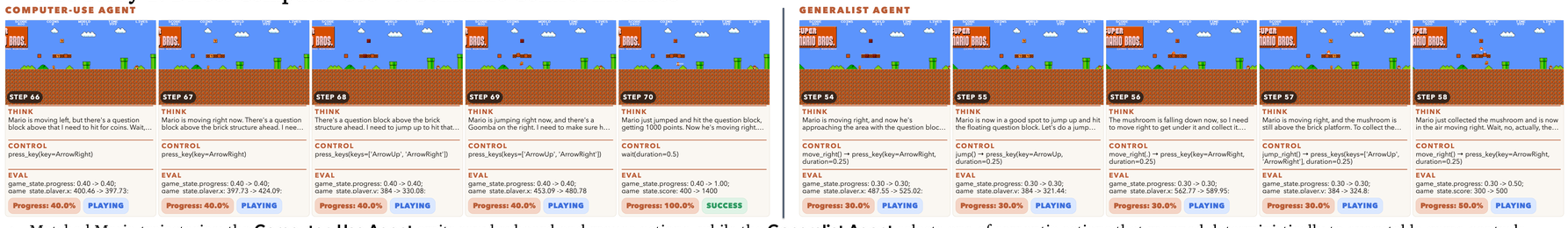
- **Comparable across interfaces.** GameWorld standardizes both Computer-Use Agents and Generalist multimodal agents under shared browser environment.
- **Environment diversity.** The suite spans 34 games across five genres to compare reactive control, spatial navigation, symbolic reasoning, and open-ended coordination under one protocol.
- **Isolate decision quality and response speed.** The browser sandbox pauses game execution during model inference, so agents face identical dynamics regardless of latency.
- **Verifiable outcome metrics.** Instead of visual heuristics or LLM-as-judge, GameWorld designs and reads serialized game state to compute progress and success directly from task-relevant variables.





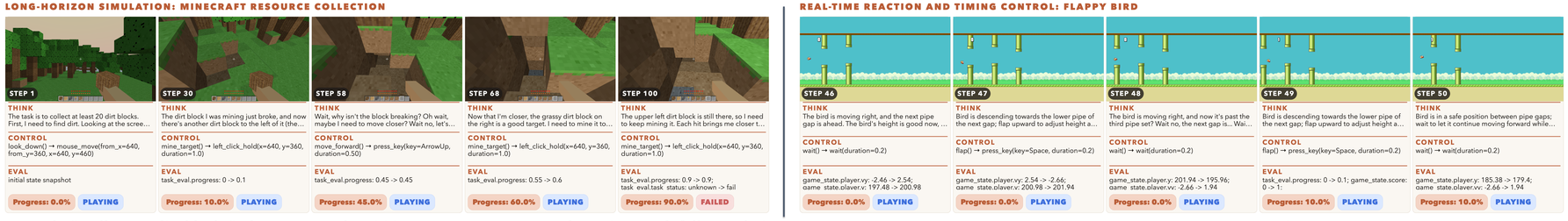
Framework Overview: **GameWorld** closes a controlled observation-action-verification loop for evaluating agents on games.

Case Study 1: Direct Computer-Use vs. Semantic Control Interface



- Matched Mario trajectories: the **Computer-Use Agent** emits raw keyboard and mouse actions, while the **Generalist Agent** selects one of semantic actions that mapped deterministically to executable game controls.

Case Study 2 & 3: Long-Horizon Goal & Real-Time Reactive Control



- Task: "*Collect 20 dirt blocks.*" The agent reaches **90% progress** (18/20 collected) but misses task full completion.
- Flappy Bird: visually consecutive frames require alternating wait/flap actions with precise timing.

Results & Analysis

Generalist Multimodal Agents

	Progress (PG)	Success Rate (SR) (%)
Gemini-3-Flash-Preview	41.9	21.2
GPT-5.2	40.6	20.6
Claude-Sonnet-4.6	39.3	20.6
Seed-1.8	39.0	20.0
Kimi-K2.5	37.4	18.8
Grok-4.1-Fast-Reasoning	36.0	16.5
Qwen3-VL-Plus	35.4	16.5
GLM-4.6V	30.8	14.1
Qwen3-VL-235B-A22B	30.8	13.5
Qwen3-VL-30B-A3B	30.6	12.4

Computer-Use Agents

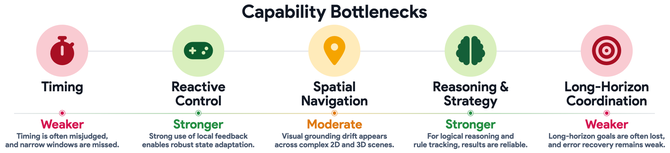
	Progress (PG)	Success Rate (SR) (%)
Seed-1.8	39.8	20.0
Claude-Sonnet-4.6	38.3	19.4
Gemini-2.5-Computer-Use	36.1	16.5
OpenAI-Computer-Use	35.8	16.5
Qwen3-VL-Plus	33.6	15.9
Qwen3-VL-235B-A22B	31.4	14.1
UI-TARS-1.5-7B	31.1	12.4
Qwen3-VL-30B-A3B	30.8	12.9

Human Baselines

	Progress (PG)	Success Rate (SR) (%)
Expert Player	82.6	77.1
Novice Player	64.1	55.3

- Best Generalist: **41.9% PG / 21.2% SR**; best CUA: **39.8% PG / 20.0% SR**.
- Both remain far below the novice human at **64.1% PG / 55.3% SR**.

Capability Bottlenecks



- Across games and agent interfaces, reactive control and symbolic reasoning are comparatively **stronger**; timing grounding, long-horizon completion, and coordination remain the **weaker** capabilities.

More detailed benchmark implementation and results:

Code

Project Page

ECCV 2026 | MALMÖ | SEPT 8–12
GameWorld