AtariBench: Evaluating Spatiotemporal Reasoning of Vision-Language Models in Atari Gameplay

Mingjia Huo, Yao Fu, Bo Chang, Yaqing Wang, Dawei Zhu, Dylan Zhang, Xinyang Yi, Lichan Hong, Ed H. Chi, Hao Zhang, Jiaxi Tang

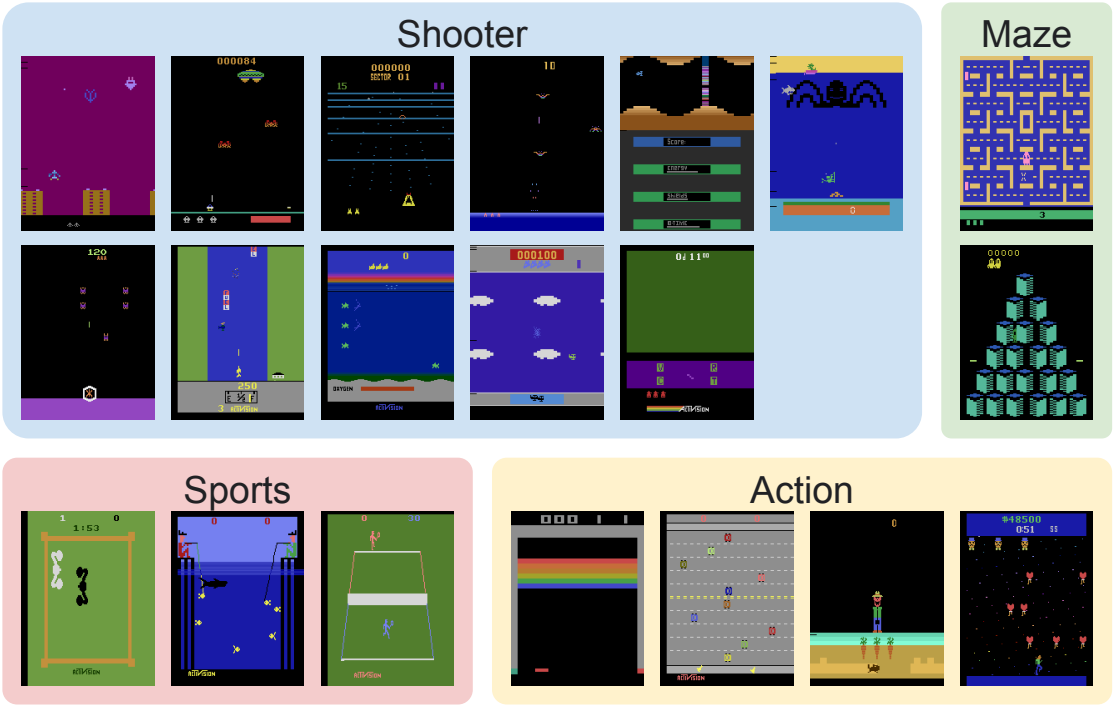
UC San Diego · Google DeepMind · Peking University · UIUC

NeurIPS 2026 · Evaluations & Datasets Track

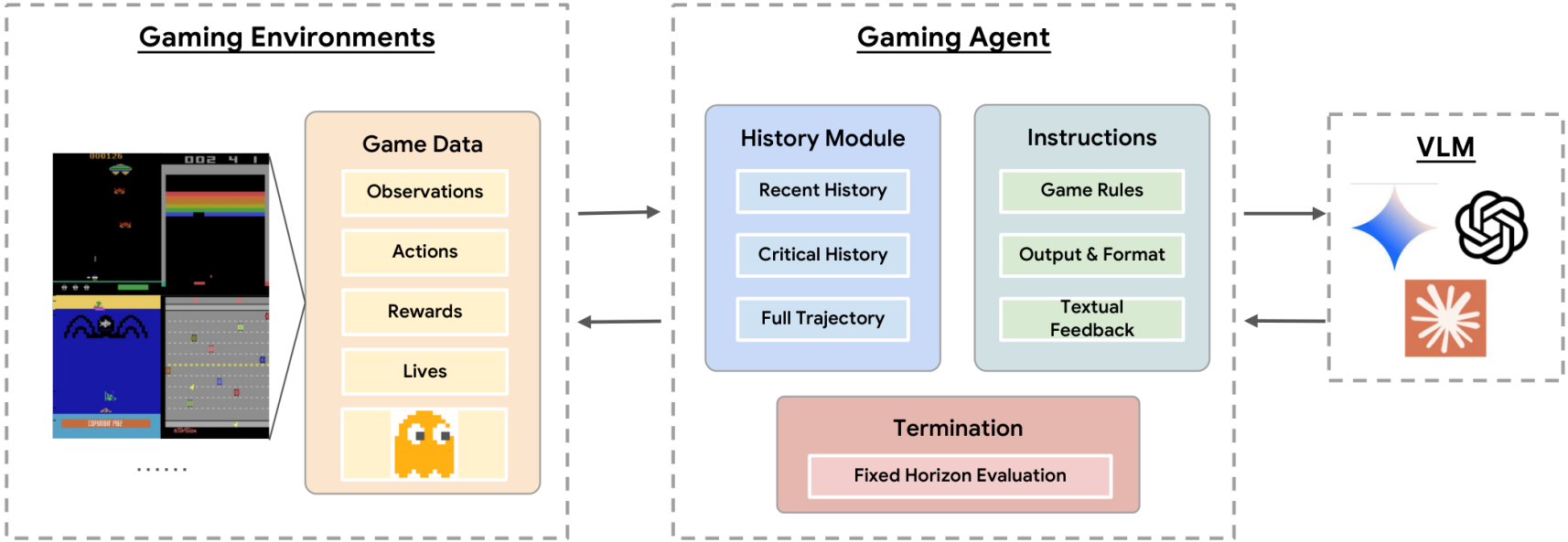
Google

Why Atari?

- VLMs excel at static images; dynamic scenes also need:
  - Spatiotemporal physics
  - Sequential decisions
  - Visual grounding
- Atari: fast dynamics, simple 2D physics, objective scores
- 20 games requiring motion perception, across 4 genres

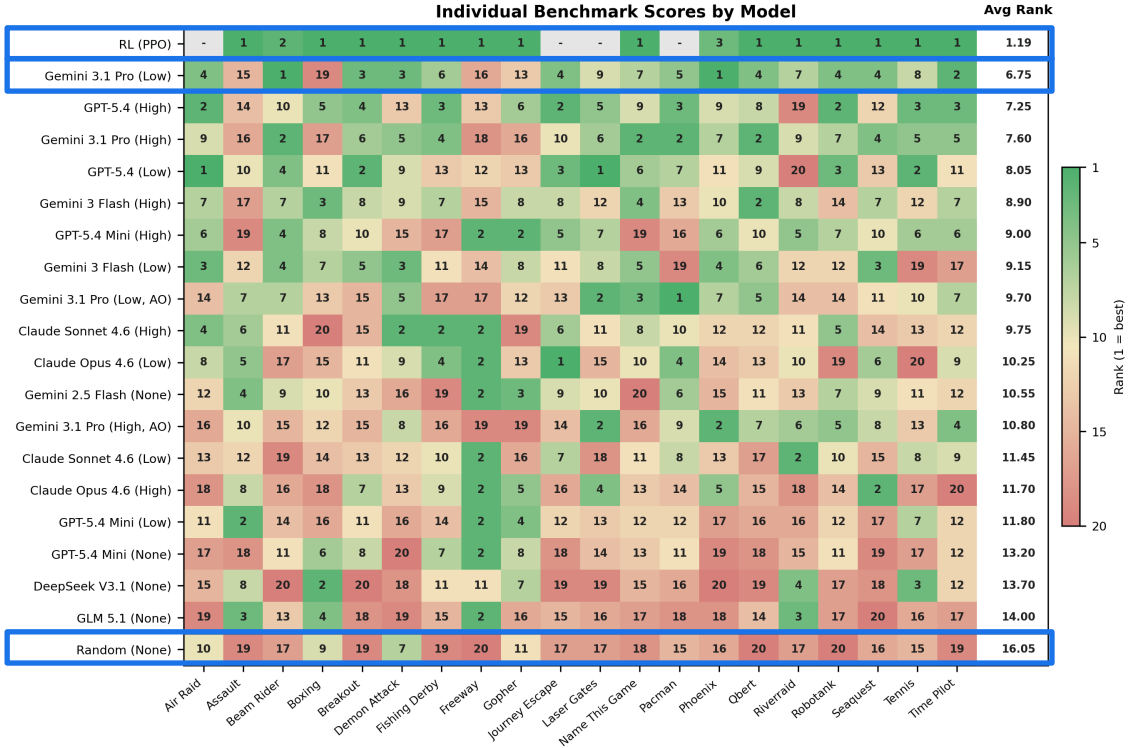


AtariBench: Closed-Loop Gameplay



- 10 Hz control, paused during inference
- Output: short thought + 1–10 actions
- Fixed 30 s horizon, 10 seeds per game
- Memory: recent + reward-relevant clips

Results: VLMs Beat Random, but Trail RL Agents



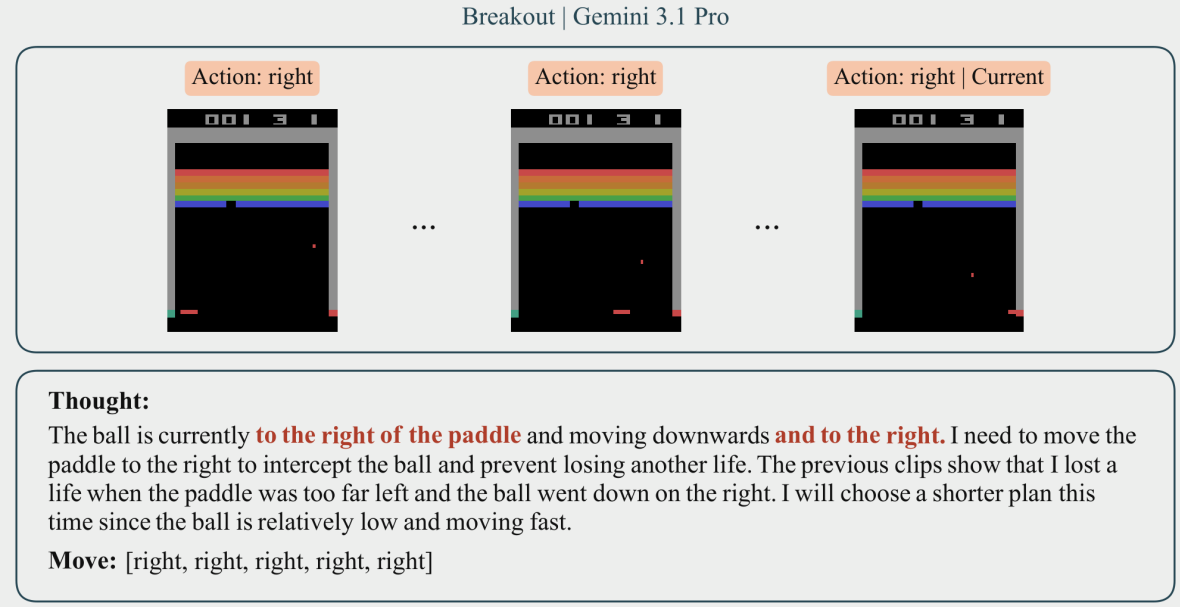
- RL (PPO): #1 on 14 of 16 games
- Best VLMs: Gemini 3.1 Pro, GPT-5.4
- Random: near the bottom, so scores reflect skill
- All 20 games separate models (18 strongly, 2 weakly)

Failure 1: Motion Is the Weakest Skill

Diagnostic QA

Skill	Accuracy
Perception	61–100%
Motion	47–74%
Decision	48–94%

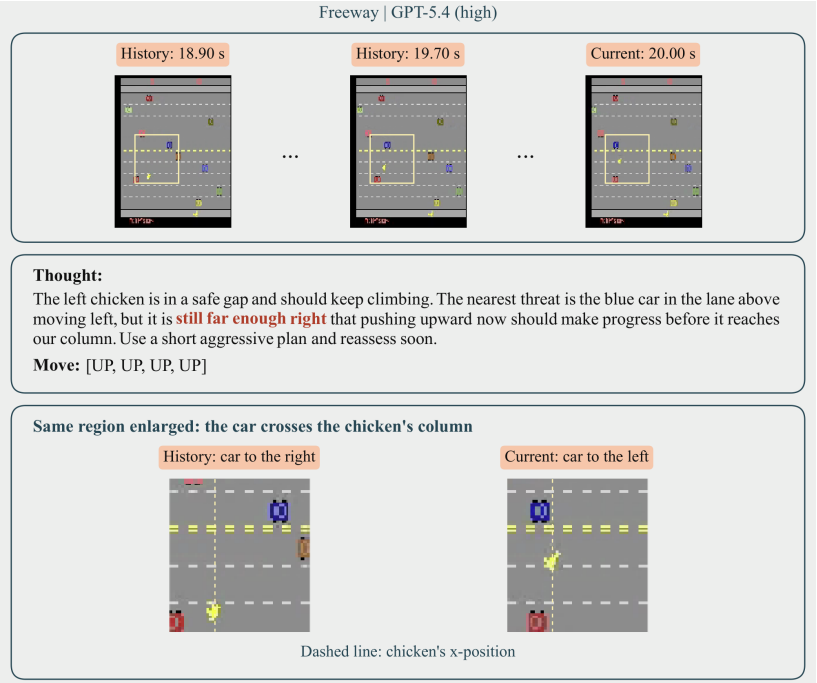
60 multiple-choice questions;
range over 11 model settings



Ground truth: the ball moves down-left; the model says down-right

Higher reasoning effort does not reliably fix motion errors.

Failure 2: More History ≠ Better Decisions



Spatial-relation errors

Full history	Current frame only
39.3%	10.7%

No-op-only turns (Gemini 2.5 Flash)

Full transcript (~100k)	Compact memory (~5k)
48.8%	0.7%

Context helps only if the model still grounds on the current frame.

Takeaways

- AtariBench: a lightweight, controlled testbed to evaluate VLMs in dynamic environments
- VLMs can see objects, but struggle to track motion and to ground decisions in the current state
- More context and more thinking have mixed effects: they can help or hurt



mignonjia.github.io/AtariBench