ROAR

Benchmarking LMM Agents under Real-Time Constraints

A multiplayer action-shooter benchmark



Dawei Wang

Newcastle University, UK

Newcastle University · UNIVERSITY OF AUCKLAND Waipapa Taumata Rau NEW ZEALAND · NeurIPS

Introduction

LMM agents in multi-agent environments

Multimodal models increasingly perceive, reason, and act in interactive environments, including robot teams, social simulations, and autonomous vehicle fleets.

A gap in existing benchmarks

Most multi-agent benchmarks pause the environment and other agents during inference. This hides the cost of delayed decisions when the world continues to evolve.

The real-time challenge

Useful actions require both reasoning quality and timely responses.
An agent must keep up with changing observations, teammates, and opponents.

Source: paper §1

Why ROAR?

Time-sensitive action-shooter gameplay

Movement, aiming, and shooting require 3D perception and situation-aware reasoning.
Brief opportunities to attack or evade make inference latency consequential.

A benchmark and asynchronous interface

ROAR covers cooperation, competition, mixed teams, and single-agent tasks.
Separate threads let the environment and each agent advance independently.

Evidence on latency and coordination

The experiments examine how real-time constraints change model performance,
and when centralized control or agentic frameworks help despite their added latency.

Source: paper §1

Four evaluation settings



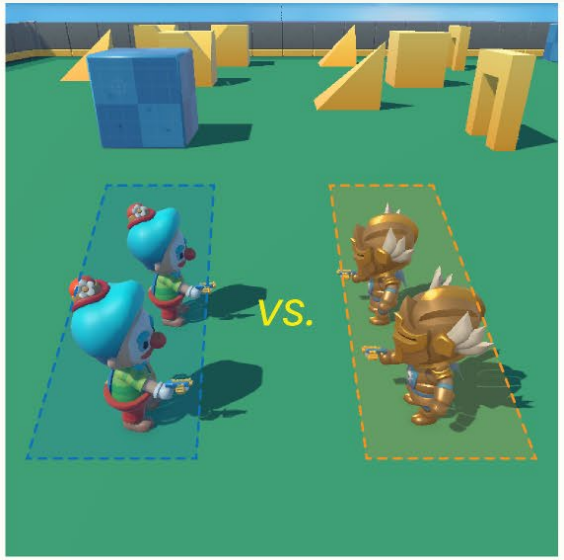
Cooperative

Four agents collect coins together



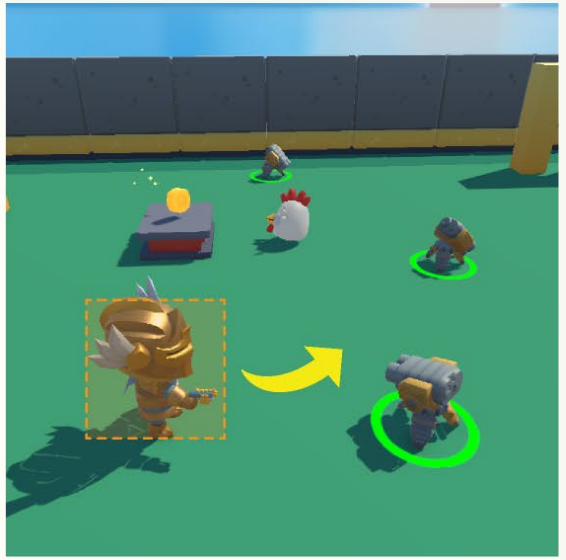
Competitive

Four agents compete to survive



Mixed

Two teams of two face off

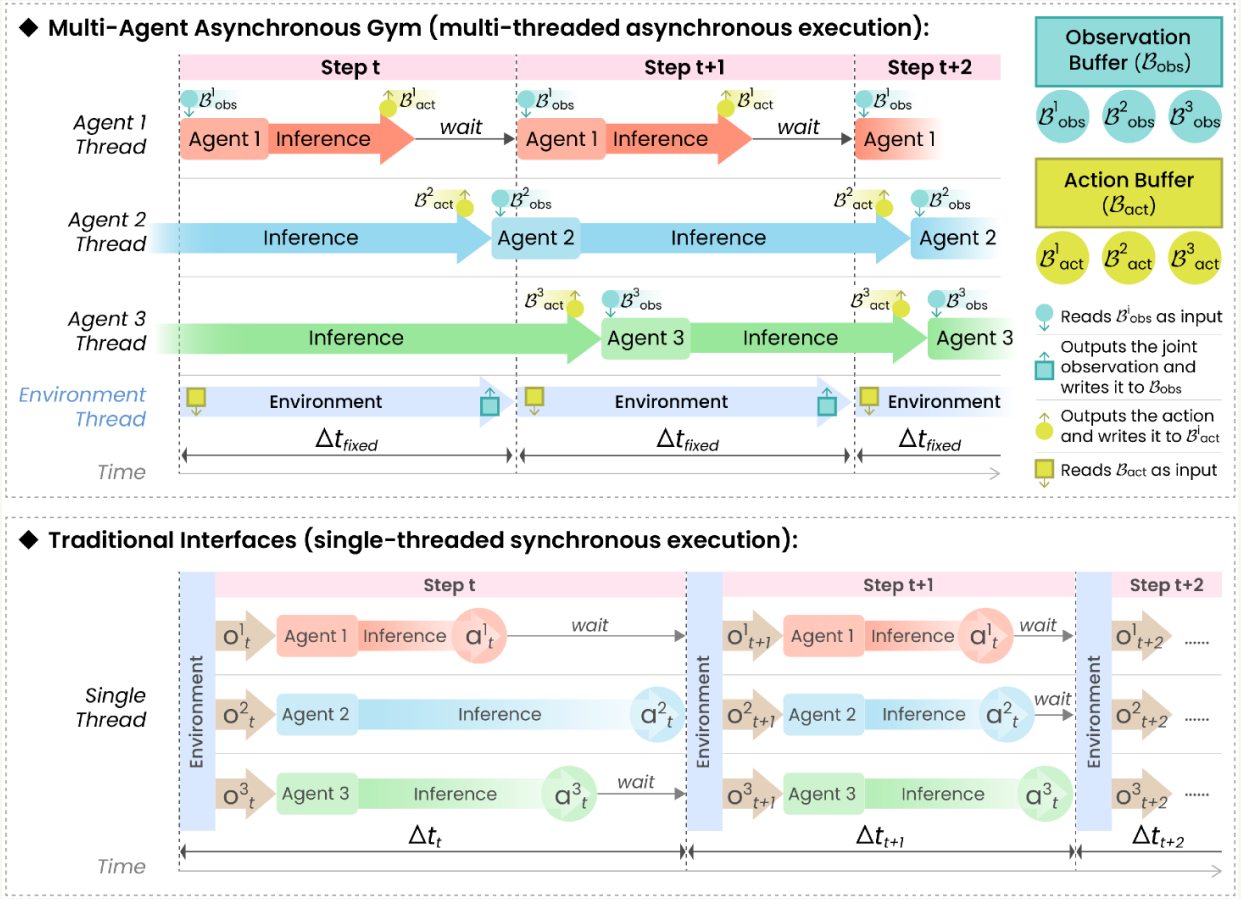


Single-agent

One agent collects coins alone

Source: paper §3.1 and poster, Four Evaluation Settings

Asynchronous Gym steps at fixed intervals



Separate threads

The environment and agents run independently.

Observation and action buffers

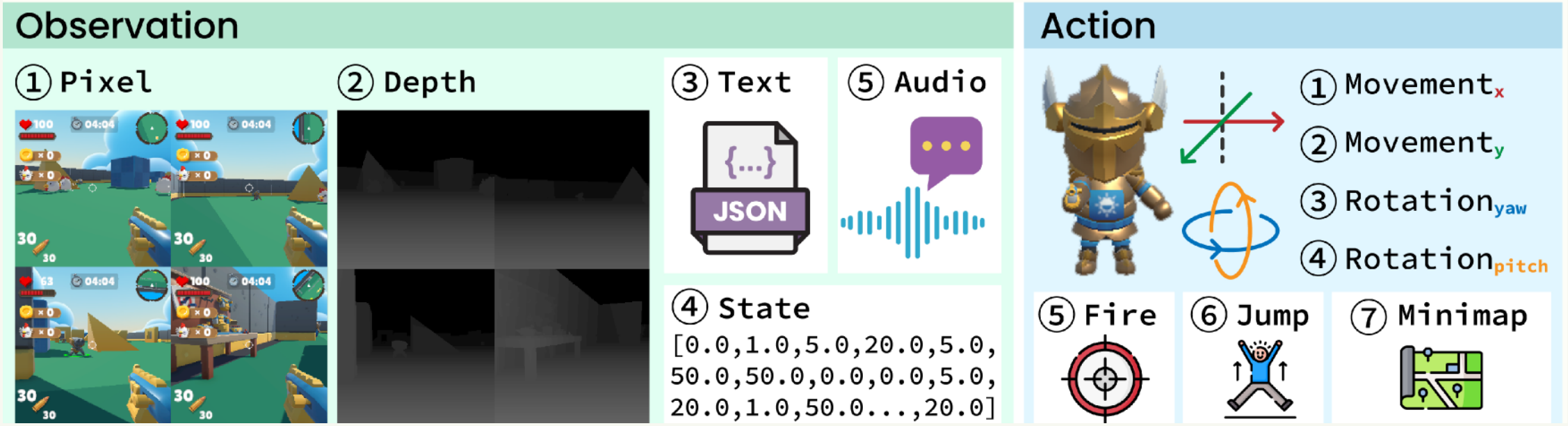
Agents read the latest observation and write their next action.

Default action

If no new action is ready, the environment applies a default and keeps stepping.

Source: paper §3.2 and poster, Multi-Agent Asynchronous Gym

What agents observe and how they act



5 observation modalities

RGB, depth, text, structured state, and audio

7 action dimensions

Two movement axes, yaw, pitch, fire, jump, minimap

Source: paper §3.3 and poster, Multimodal Observation & Action

Experimental results

Mean episode score across four levels

Model	Turn-based	250 ms	40 ms
Gemini-3.1 Pro	0.87	0.44	**0.34**
GPT-5.4	0.47	0.26	**0.19**
Claude Haiku 4.5	0.13	0.12	**0.11**
Qwen-3.5 9B	0.10	0.07	**0.07**
Gemma-4 E4B	0.08	0.07	**0.07**

All five models scored lower on average under real-time constraints.

Cooperation was especially sensitive to delay. Added coordination and reasoning helped only when their benefits outweighed the latency cost.

Source: paper §4.2–4.3, Tables 2–4

Summary

A benchmark for real-time multi-agent interaction

ROAR evaluates LMM agents in cooperative, competitive, mixed, and single-agent tasks.
Its asynchronous interface lets the environment and agents advance independently.

Latency changes performance and coordination

Real-time constraints lowered average scores across all five tested models.
Cooperation was especially sensitive, while some competitive score gaps narrowed.

Reasoning quality and timely action matter together

Centralized control and agentic frameworks add latency as well as capability.
Their benefits depend on the model and the environment’s timing constraints.

Source: paper §1, §4–5