Memory, Benchmark & Robots: A Benchmark for Solving Complex Tasks with Reinforcement Learning

Kachaev Nikita

Research Engineer, AIRI

NEURAL INFORMATION PROCESSING SYSTEMS

Spotlight paper from NeurIPS 2025

Memory, Benchmark & Robots: A Benchmark for Solving Complex Tasks with Reinforcement Learning

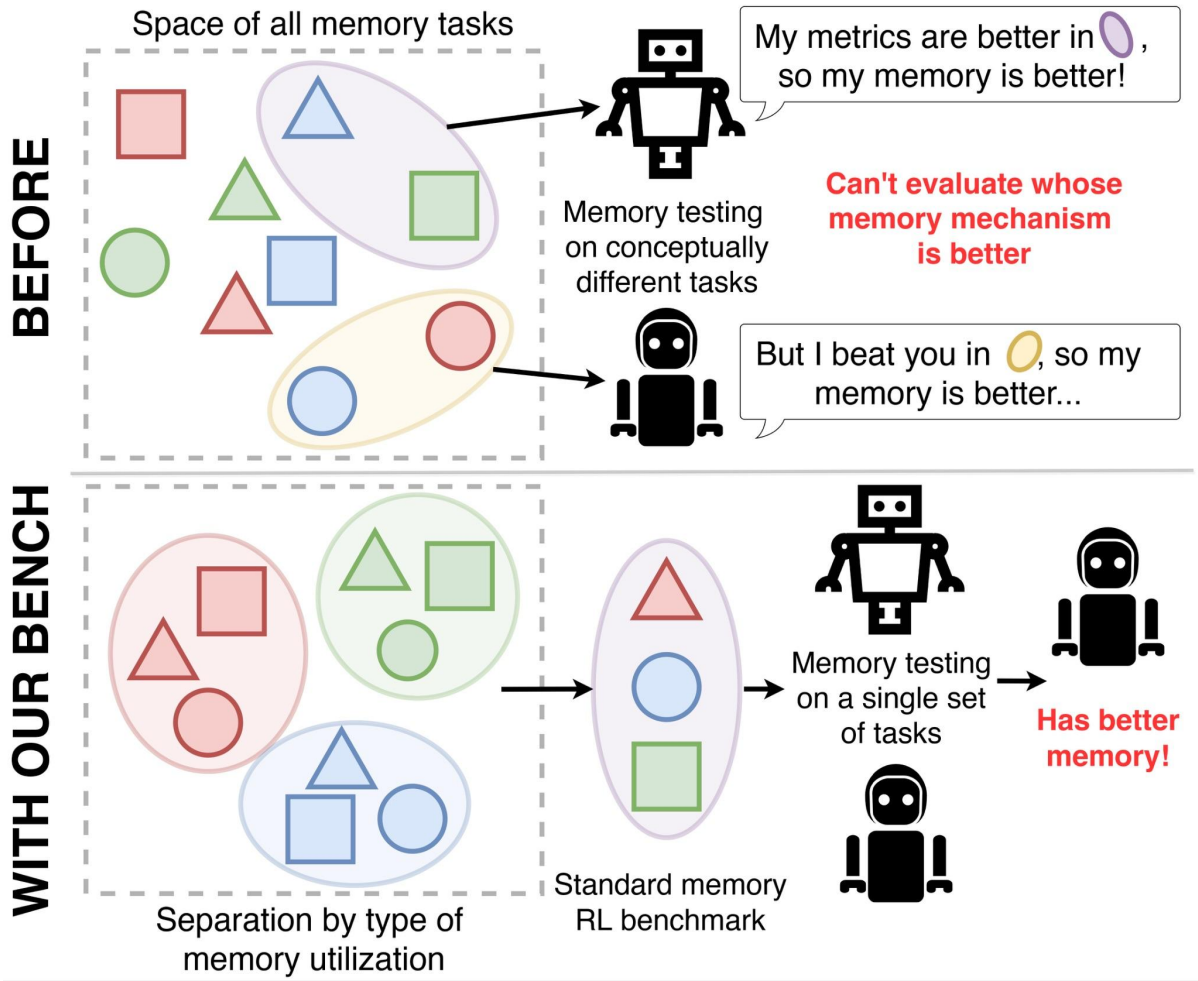
Egor Cherepanov[1,2] **Nikita Kachaev**[1,3] **Alexey K. Kovalev**[1,2] **Aleksandr I. Panov**[1,2]
[1]AIRI, Moscow, Russia [2]MIPT, Dolgoprudny, Russia [3]HSE University, Moscow, Russia
{cherepanov,kachaev,kovalev,panov}@airi.net

Abstract

Memory is crucial for enabling agents to tackle complex tasks with temporal and spatial dependencies. While many reinforcement learning (RL) algorithms incorporate memory, the field lacks a universal benchmark to assess an agent's memory capabilities across diverse scenarios. This gap is particularly evident in tabletop robotic manipulation, where memory is essential for solving tasks with partial observability and ensuring robust performance, yet no standardized benchmarks exist. To address this, we introduce **MIKASA** (**M**emory-**I**ntensive **Ska**lls **A**ssessment **S**uite for **A**gents), a comprehensive benchmark for memory RL, with three key contributions: (1) we propose a comprehensive classification framework for memory-intensive RL tasks, (2) we collect **MIKASA-Base** – a unified benchmark that enables systematic evaluation of memory-enhanced agents across diverse scenarios, and (3) we develop **MIKASA-Robo**[1] – a novel benchmark of 32 carefully designed memory-intensive tasks that assess memory capabilities in tabletop robotic manipulation. Our work introduces a unified framework to advance memory RL research, enabling more robust systems for real-world use. **MIKASA is available at** https://tinyurl.com/membenchrobots.

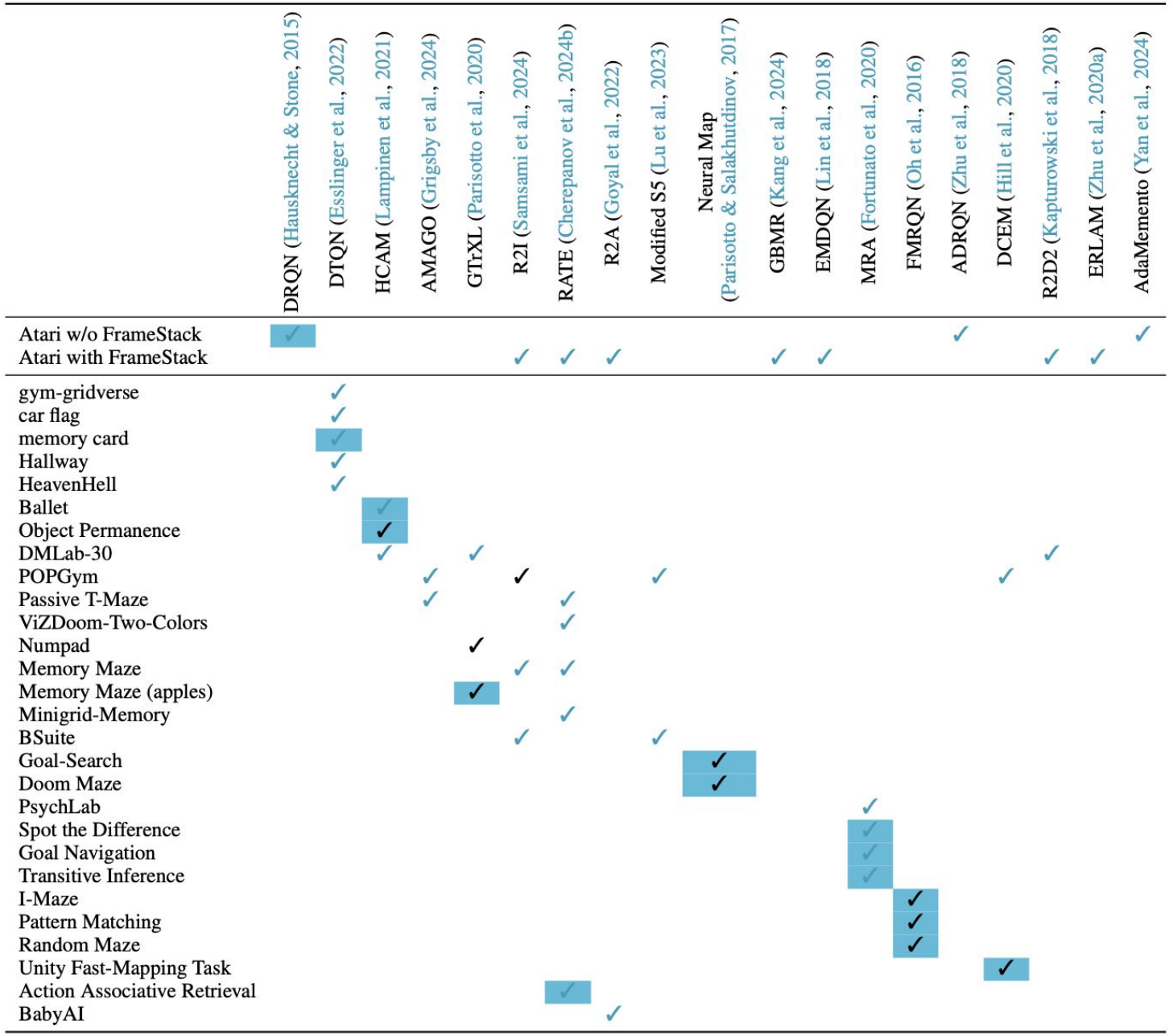
Problem 1: Most works evaluate memory only on a narrow subset of memory tasks

- Research studies use different sets of environments with minimal overlap, making it **difficult to compare memory-enhanced agents across studies**
- Even within individual studies, benchmarks may focus on testing similar memory aspects while neglecting others, **leading to incomplete evaluation of agents' memory**



Problem 1:

- Benchmark first introduced in the same work.
- Benchmark is open-sourced.
- The Atari (Bellemare et al., 2013) environment with frame stacking is included to illustrate that **many memory-enhanced agents are tested solely in MDP.**

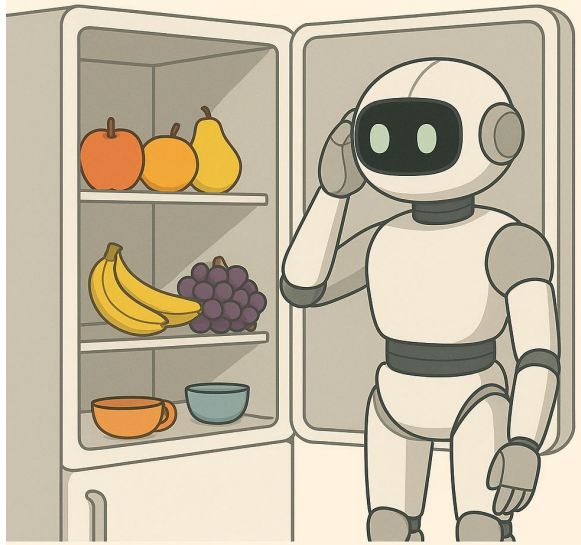


Memory Task Classification

- To address this gap, we propose a **unified task taxonomy for evaluating memory in RL**. It divides memory-demanding tasks into four core types:

OBJECT MEMORY

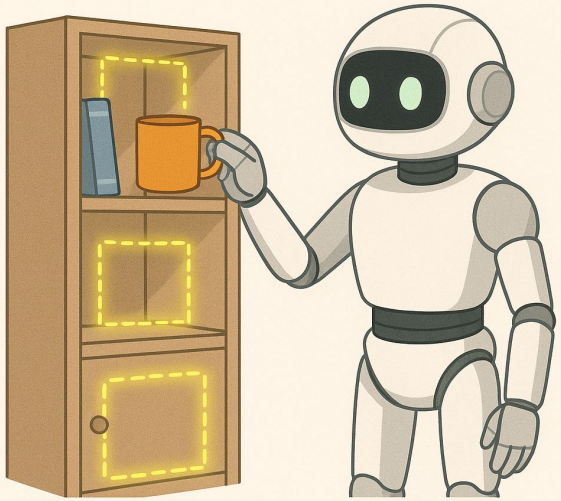
Tasks that evaluate an agent's ability to maintain object-related information over time, particularly when objects become temporarily unobservable



Example: a robot remembers which fruit it put in the fridge

SPATIAL MEMORY

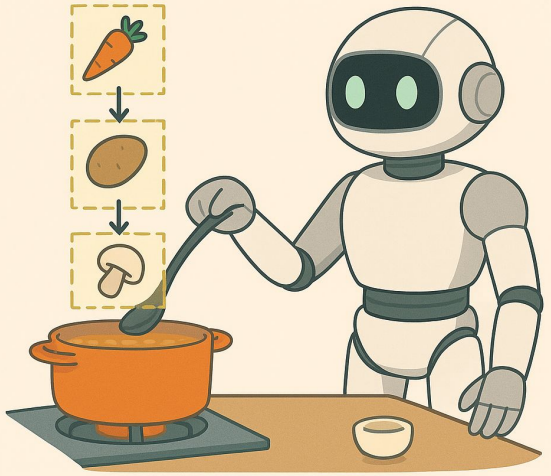
Tasks focused on environmental awareness and navigation, where agents must remember object locations, maintain mental maps of environment layouts, and navigate based on previously observed spatial information



Example: the robot remembers the position of a mug it moved while cleaning and returns it to its place

SEQUENTIAL MEMORY

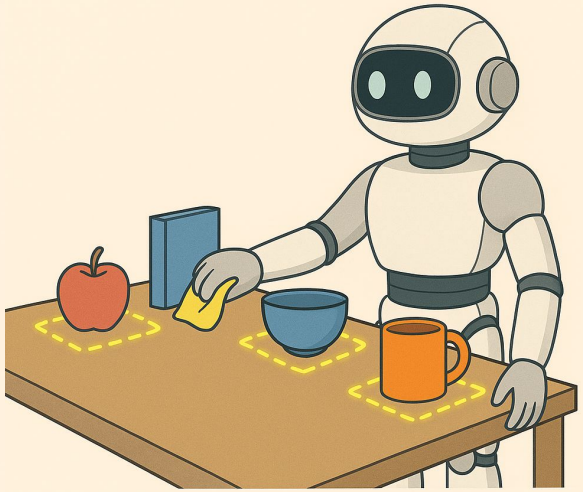
Tasks that test an agent's ability to process and utilize temporally ordered information, similar to human serial recall and working memory. These tasks require remembering action sequences, maintaining order-dependent information, and using past decisions to inform future actions



Example: a robot memorizes the order of the ingredients it has added to a soup

MEMORY CAPACITY

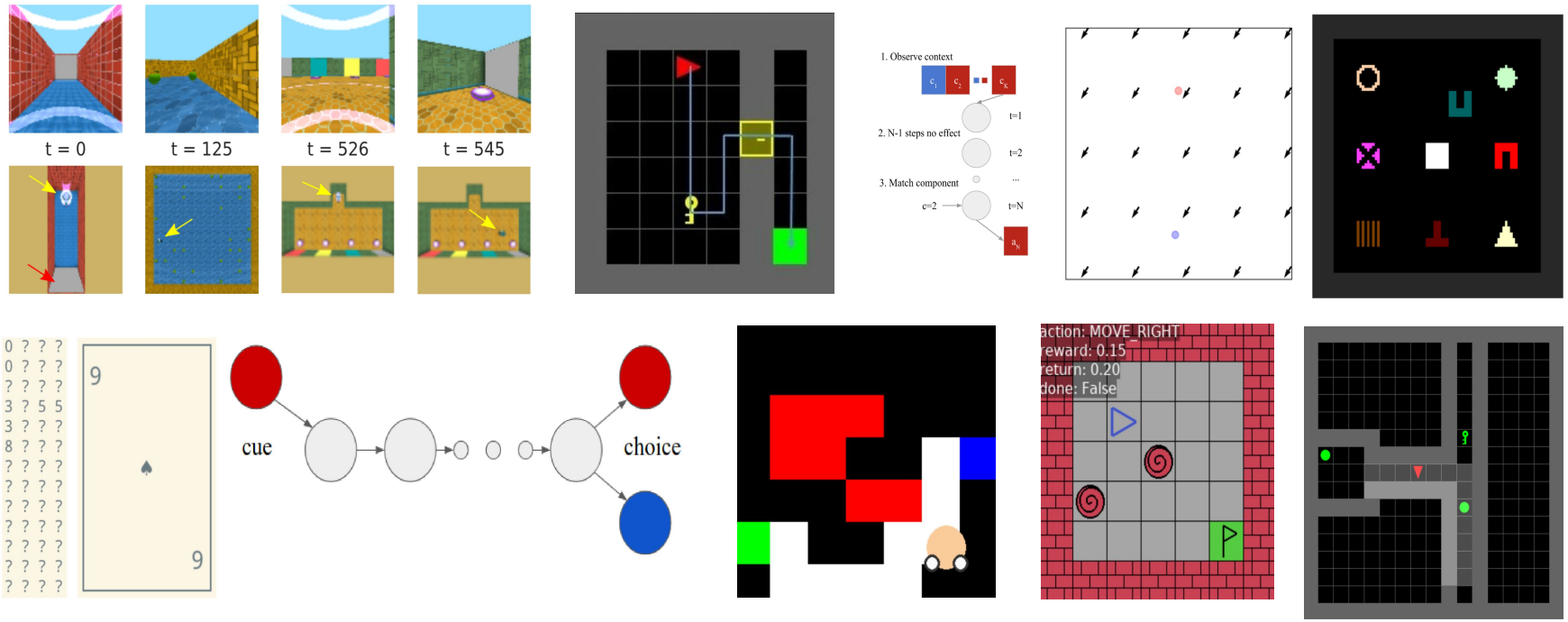
Tasks that challenge an agent's ability to manage multiple pieces of information simultaneously, analogous to human memory span. These tasks evaluate information retention limits and multi-task information processing



Example: a robot is able to memorize the positions of several different objects while cleaning a table

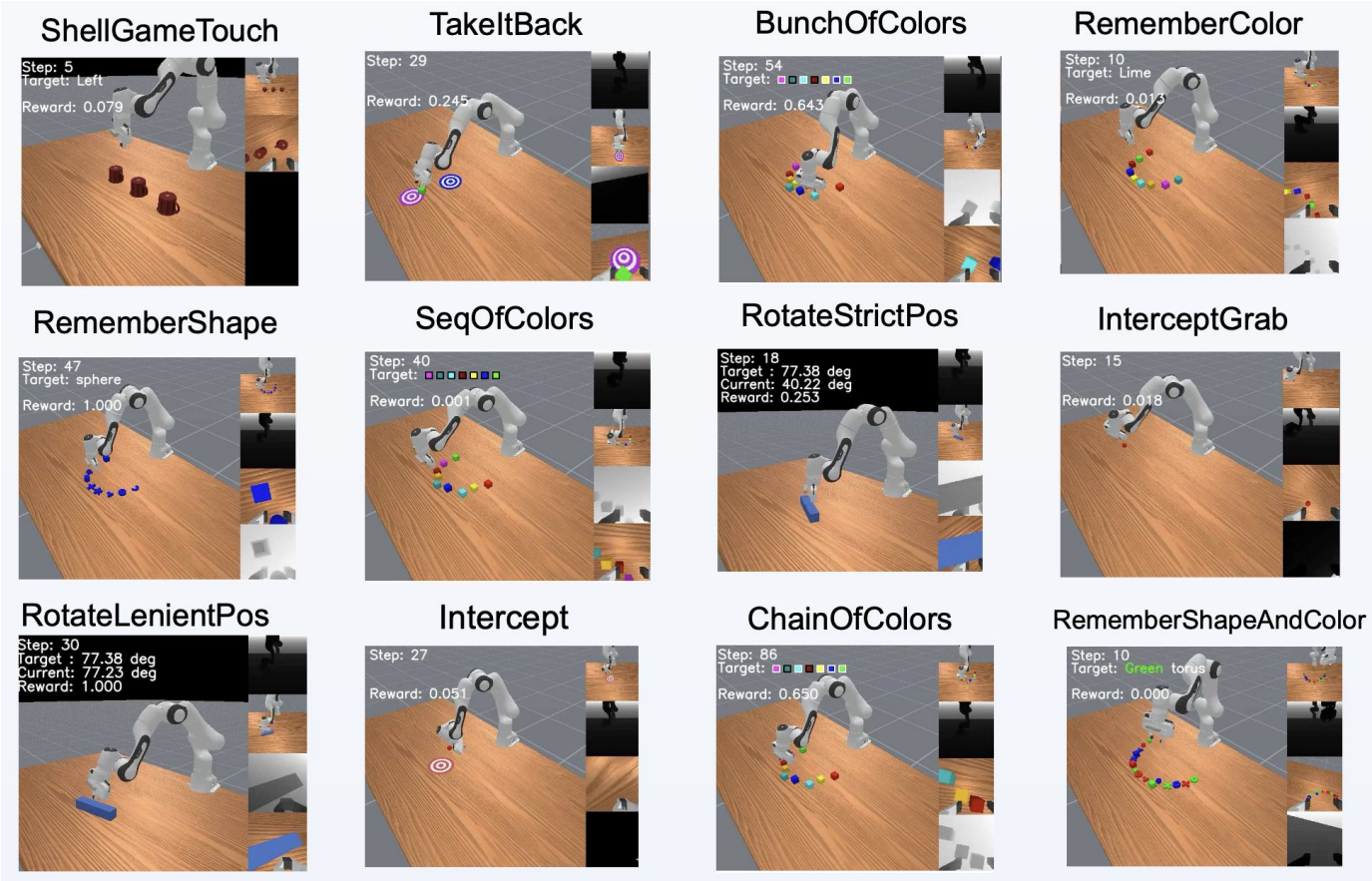
AIRI

Problem 2: Existing benchmarks are simplified



- Existing memory benchmarks remain **simplified and fail to capture the complexity of real-world scenarios**. They often rely on grid-world environments with discrete actions and low-fidelity visuals, which do not reflect the challenges faced by embodied agents operating in realistic and visually rich settings.

MIKASA-Robo Memory Tasks



- **32 robotic tasks** that target specific memory-dependent skills in realistic settings
- **Many baselines:** Online/Offline RL, VLA models
- Based on ManiSkill and support **parallel GPU Rendering**
- All dataset and data collection scripts are **open-sourced**

AIRI

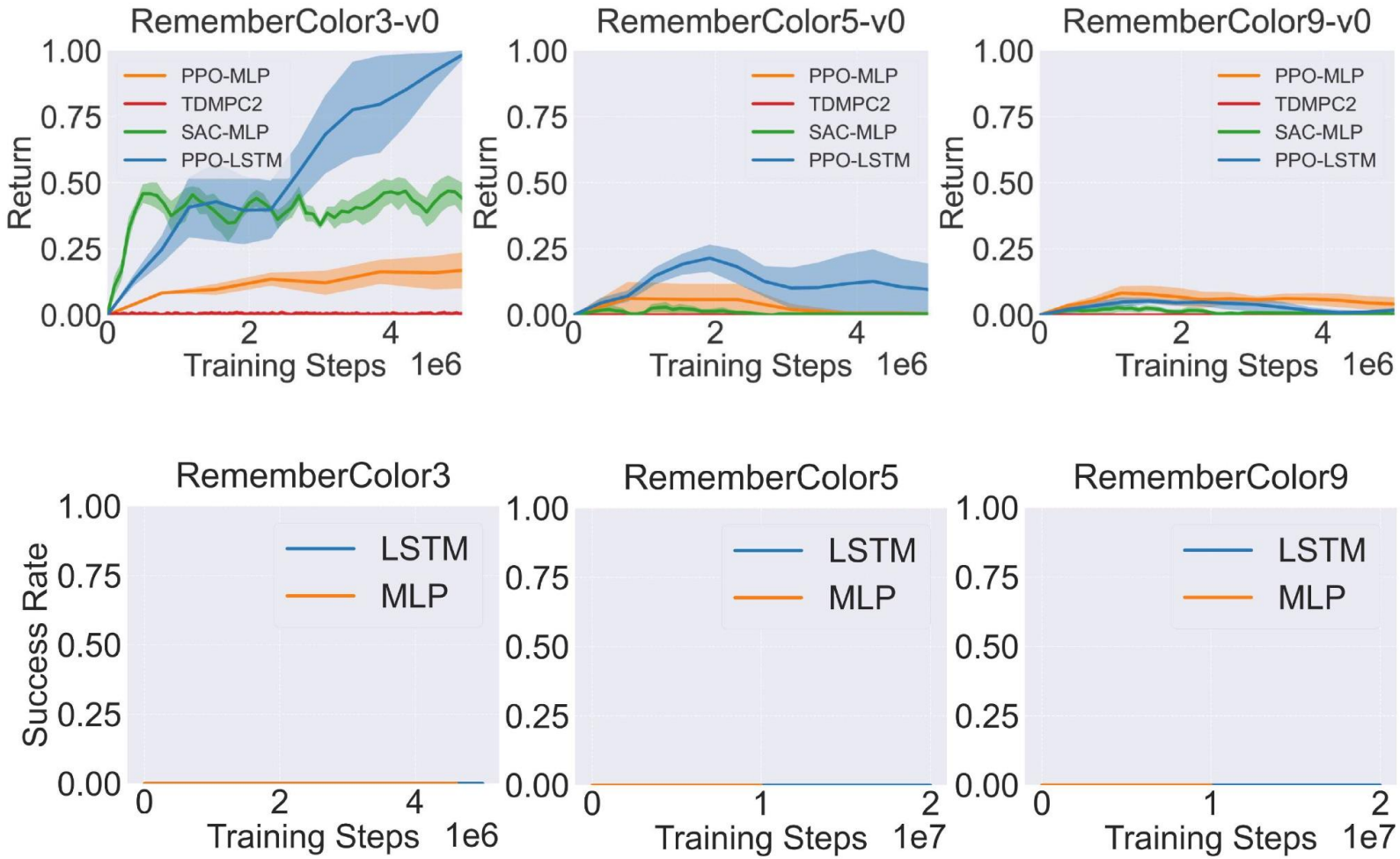
MIKASA-Robo Memory Tasks

```python
from mani_skill.utils.registration import register_env
from mikasa_robo_suite.remember_color import RememberColorBaseEnv
import gymnasium as gym

@register_env("RememberColor4Debug-v0", max_episode_steps=1000)
class RememberColorDebugEnv(RememberColorBaseEnv):
    COLORS         = 4     # 1-9 unique cubes
    TIME_OFFSET    = 200   # how long target cube is shown
    GOAL_THRESH    = 0.03  # more difficult goal threshold
    CUBE_HALFSIZE = 0.02   # cubes size
    DELTA_TIME     = 100   # empty table duration (seconds)

    env = gym.make("RememberColor4Debug-v0", num_envs=256, obs_mode="
        ↪ rgb", render_mode="all", delta_time=DELTA_TIME)}
```

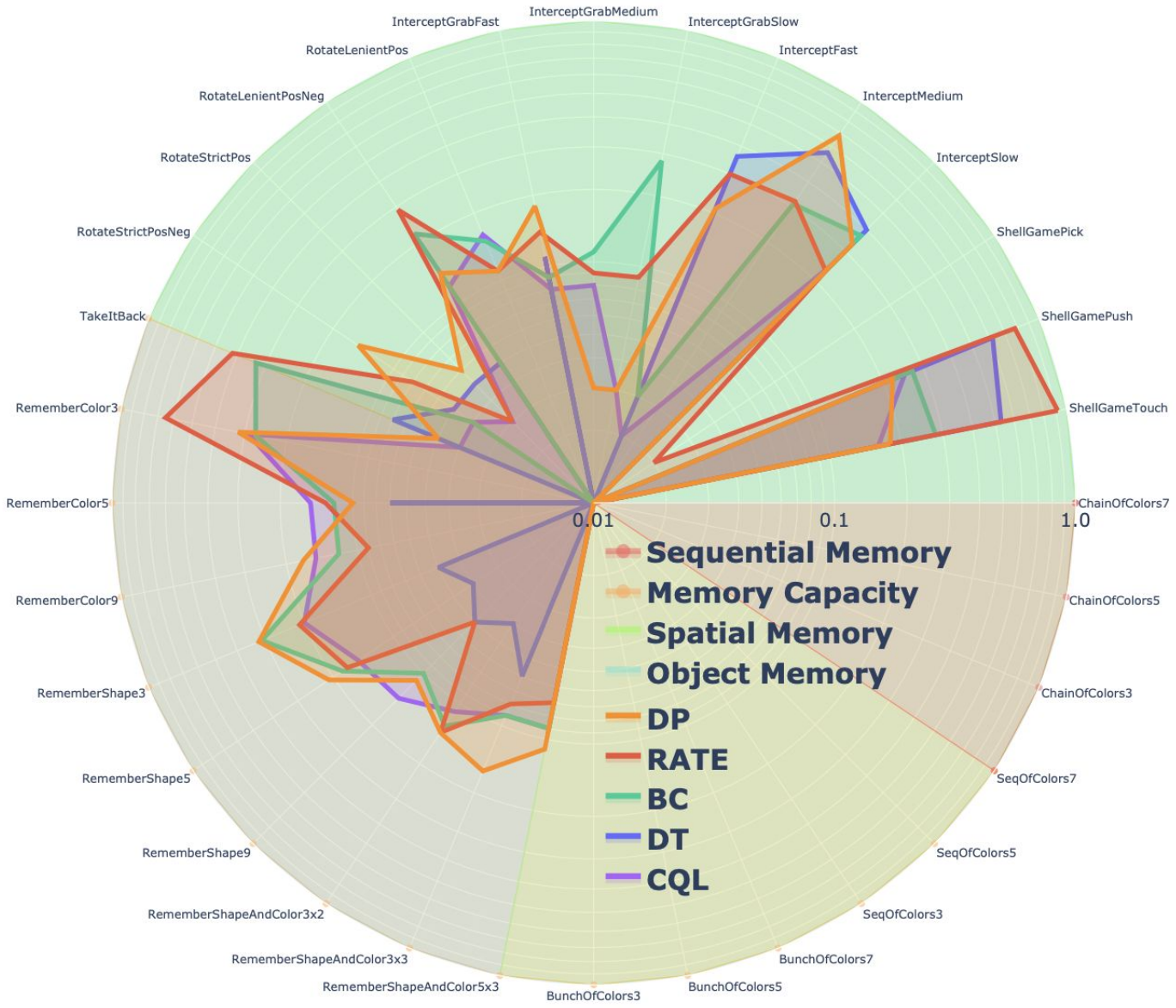
- **Easy to install:** pip install mikasa-robo-suite
- **Easy to customize!**

AIRI

MIKASA-Robo Memory online RL baselines

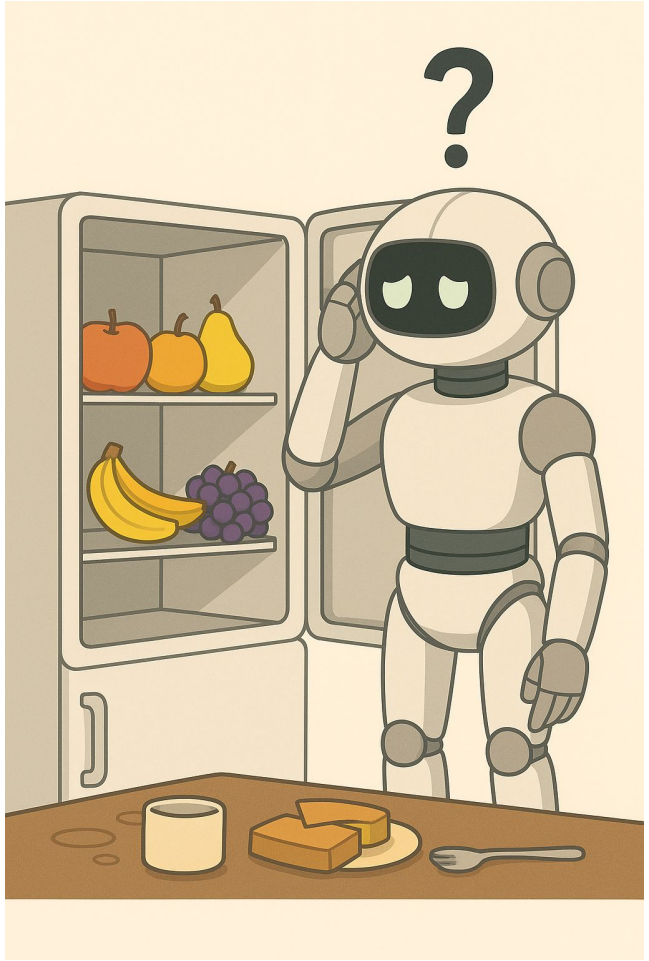


MIKASA-Robo Memory offline RL baselines



Problem 3: Current memory mechanisms are insufficient to solve truly complex, memory-intensive tasks

- Trained from RGB with sparse, binary rewards, none of DT, RATE, BC, CQL, or Diffusion Policy solve the majority of MIKASA-Robo tasks - capacity and sequential-memory tasks are especially unmet. Supplementary dense-reward runs improve some scores but still **reveal substantial gaps, reinforcing the benchmark's difficulty.**



MIKASA-Robo Memory VLA baselines

Model	ShellGameTouch	InterceptMedium	RememberColor3	RememberColor5	RememberColor9
Octo-small	0.46 ± 0.05	0.39 ± 0.04	0.45 ± 0.06	0.17 ± 0.03	0.11 ± 0.03
OpenVLA (K=4)	0.12 ± 0.05	0.06 ± 0.02	0.21 ± 0.00	0.09 ± 0.02	0.08 ± 0.02
OpenVLA (K=8)	0.47 ± 0.05	0.14 ± 0.03	0.59 ± 0.04	0.16 ± 0.03	0.06 ± 0.02
SpatialVLA (K=4)	0.23 ± 0.04	0.27 ± 0.04	0.27 ± 0.05	0.17 ± 0.03	0.11 ± 0.03
π_0 (K=4)	0.33 ± 0.05	0.42 ± 0.03	0.35 ± 0.04	0.22 ± 0.04	0.15 ± 0.02

- Our experiments highlights a **critical gap** in current VLA models: the **absence of effective memory mechanisms** leads to brittle performance on tasks demanding strong memory capabilities

Open source MIKASA repositories:

GitHub

Hugging Face

AIRI

Contacts

Kachaev Nikita

Research Engineer, AIRI

ttonyalpha@gmail.com