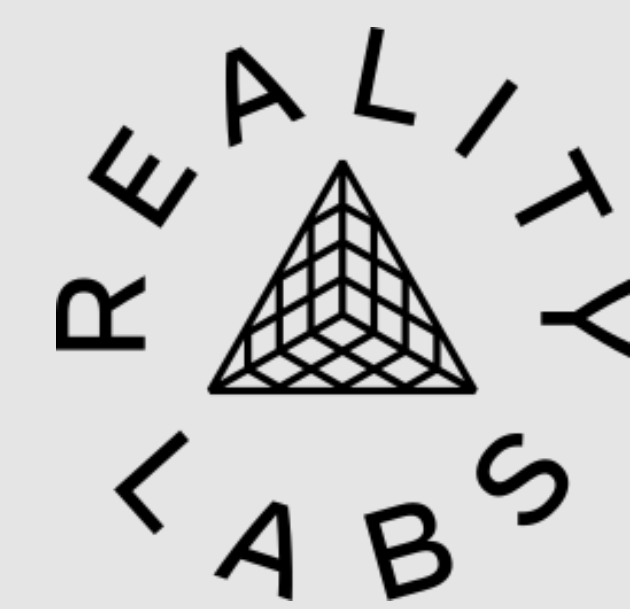


RDA: Reward Design Agent for Reinforcement Learning

Hoon Lee, Ajay Subramanian, Ben Abbatemateo, Vijay Veerabadran, Pedro Matias, Karl Ridgeway, Nitin Kamra



Why Reward Design is Difficult?

- Scalar reward must be aligned to the human’s intention.
- Task instructions often require long, multi-step behaviors.
- Reward function is difficult to assess without visually inspecting the resulting behavior.

Limitations of Existing Methods

- Hand-design: Slow, brittle, and error-prone.
- Eureka: Automates reward design, but numerical feedback cannot diagnose behavioral misalignment.

Instruction: “Walk towards the package, stand behind it, maintain contact, and push it to the goal location”

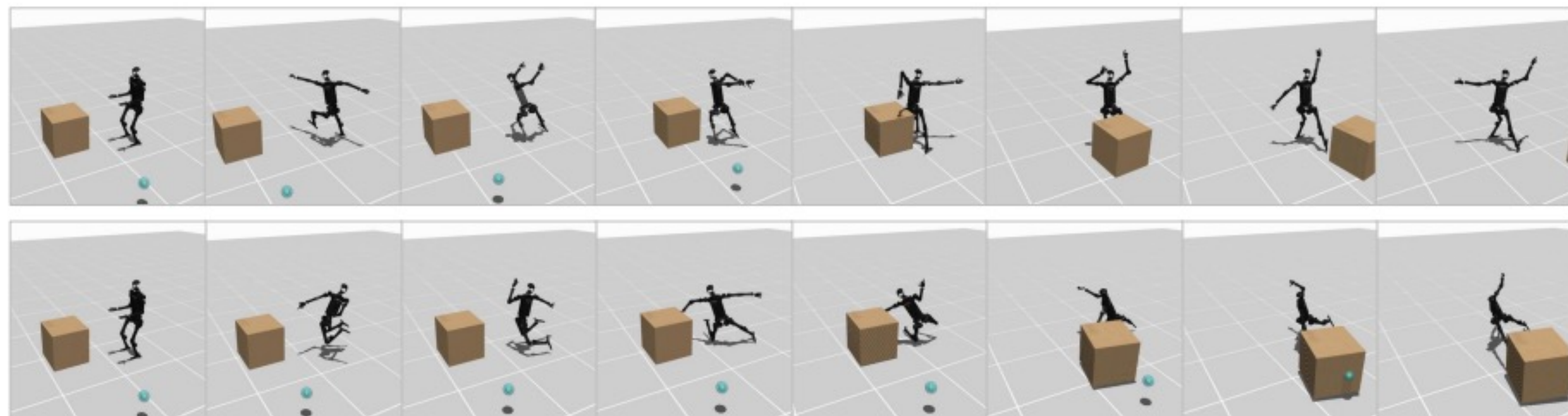
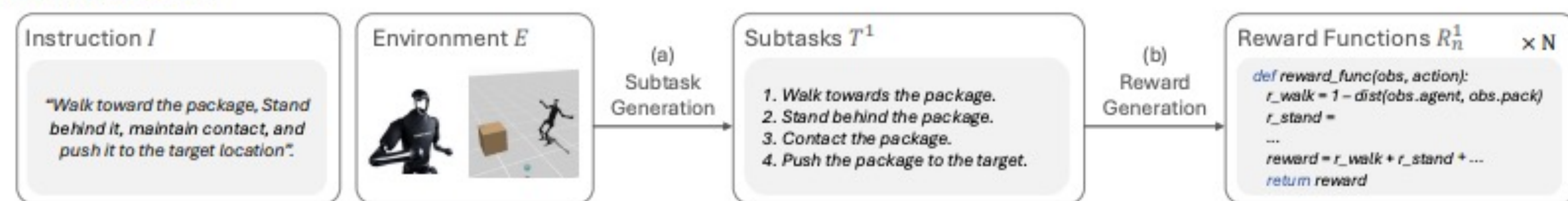


Figure 1: The intended behavior is to approach the package, stand behind it, maintain contact, and gently push it to the target. Rewards designed by both humans and Eureka achieve task success but induce misaligned behaviors, causing the humanoid to kick or throw the package rather than push it smoothly.

RDA: Reward Design Agent for Reinforcement Learning

I. Initialization



II. Evolutionary Search

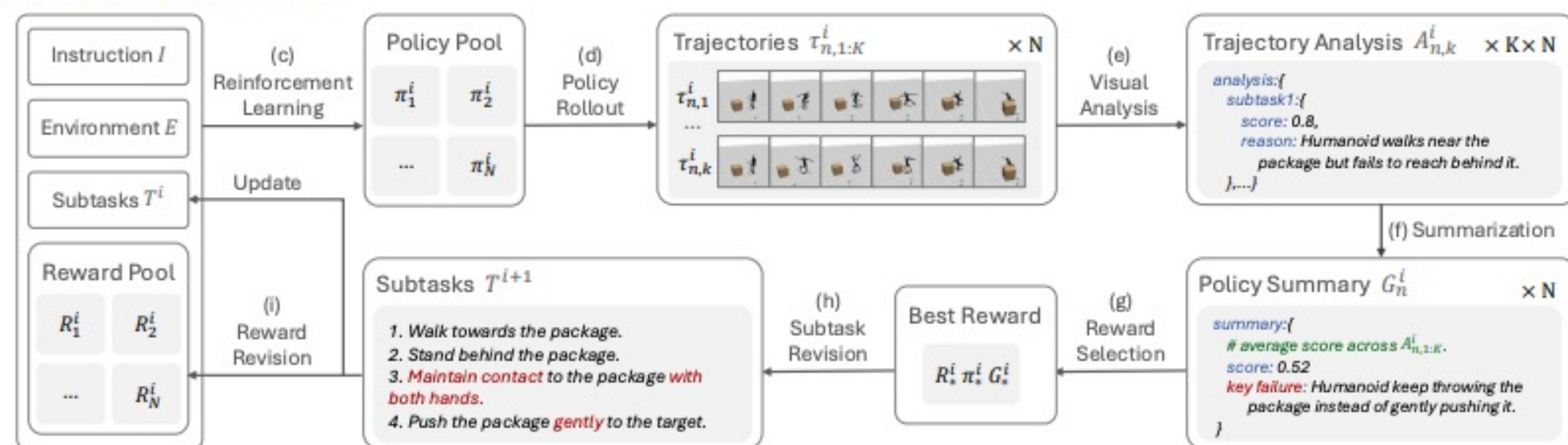


Figure 2: RDA mirrors humans’ reward designing process. Decompose → Generate → Evaluate → Reflect..

Experimental Setup

Environments

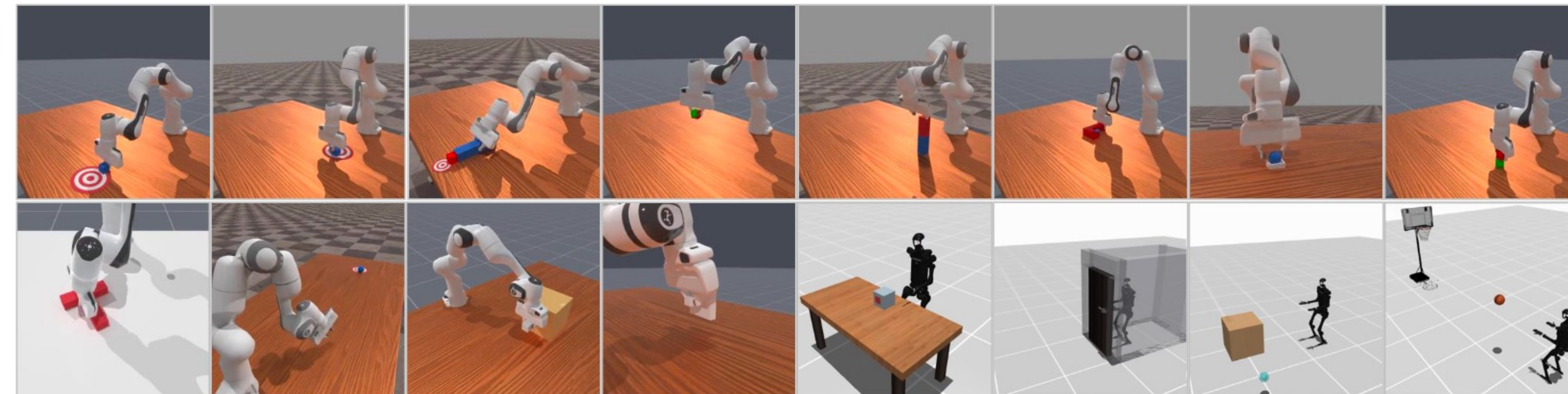


Figure 3: 12 tabletop manipulation environments from ManiSkill, 4 loco-manipulation tasks from HumanoidBench.

Training

- VLM Backbone: GPT-5.
- RL: SAC with Simba; 50M (ManiSkill), 10M (HumanoidBench) environment steps.
- Evolutionary Search: 4 candidates with 5 generations.

Evaluation

- Alignment Rate: GPT-4.1 evaluates behavioral alignment on a 5-point Likert scale.
- Success Rate: Measures task completion using the benchmark’s binary success indicator.

Experimental Results

- RDA generates rewards with strong intention alignment and task success.

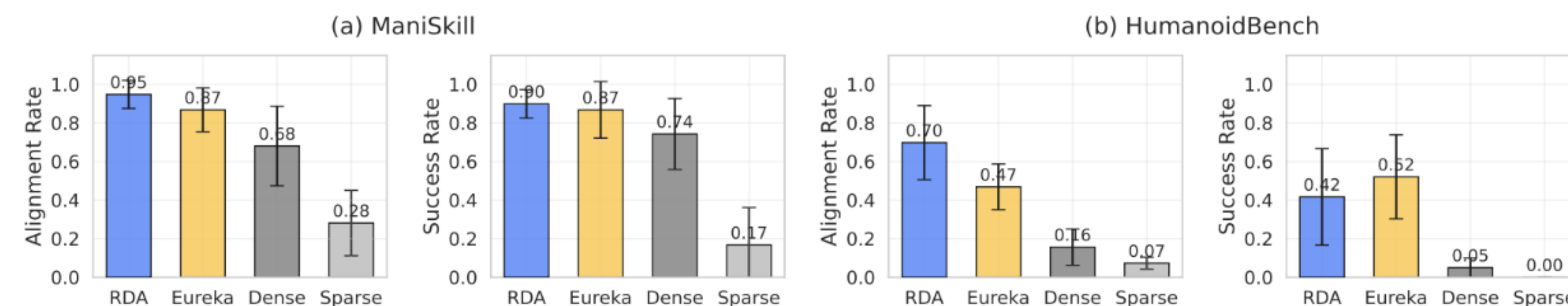


Figure 4: Results averaged across environments and 3 random seeds; the best candidate is selected by alignment.

Instruction: “Walk towards the package, stand behind it, maintain contact, and push it to the goal location”



Figure 5: RDA produces policy that faithfully follows the instructions

TL;DR: Living in the era of Vibe Coding, try RDA’s workflow in [claude.md](#).

