Rapidly Adapting Policies to the Real-World With Simulation Guided Fine-Tuning

Presenter: Tyler Westenbroek

Authors: Patrick Yin*, Tyler Westenbroek*, Simran Bagaria, Kevin Huang, Ching-an Cheng, Andrey Kobolov, Abhishek Gupta

UNIVERSITY of WASHINGTON

Microsoft Research

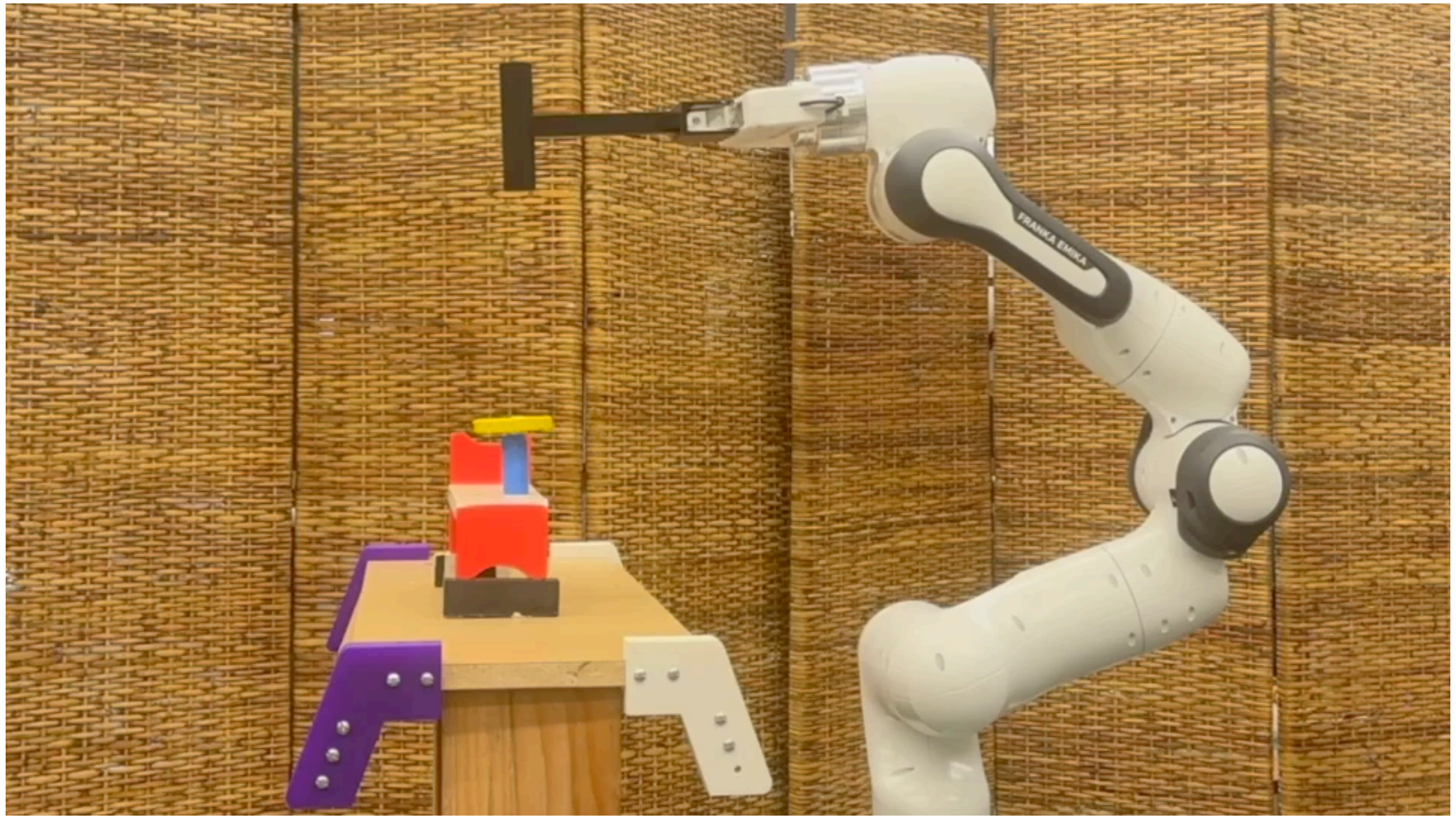
What’s missing for sim-to-real transfer?

Current approaches: use extensive domain randomization to make zero-shot transfer as robust as possible

- Dynamics randomization
- Visual randomization

Failure Mode: Simulators are wrong

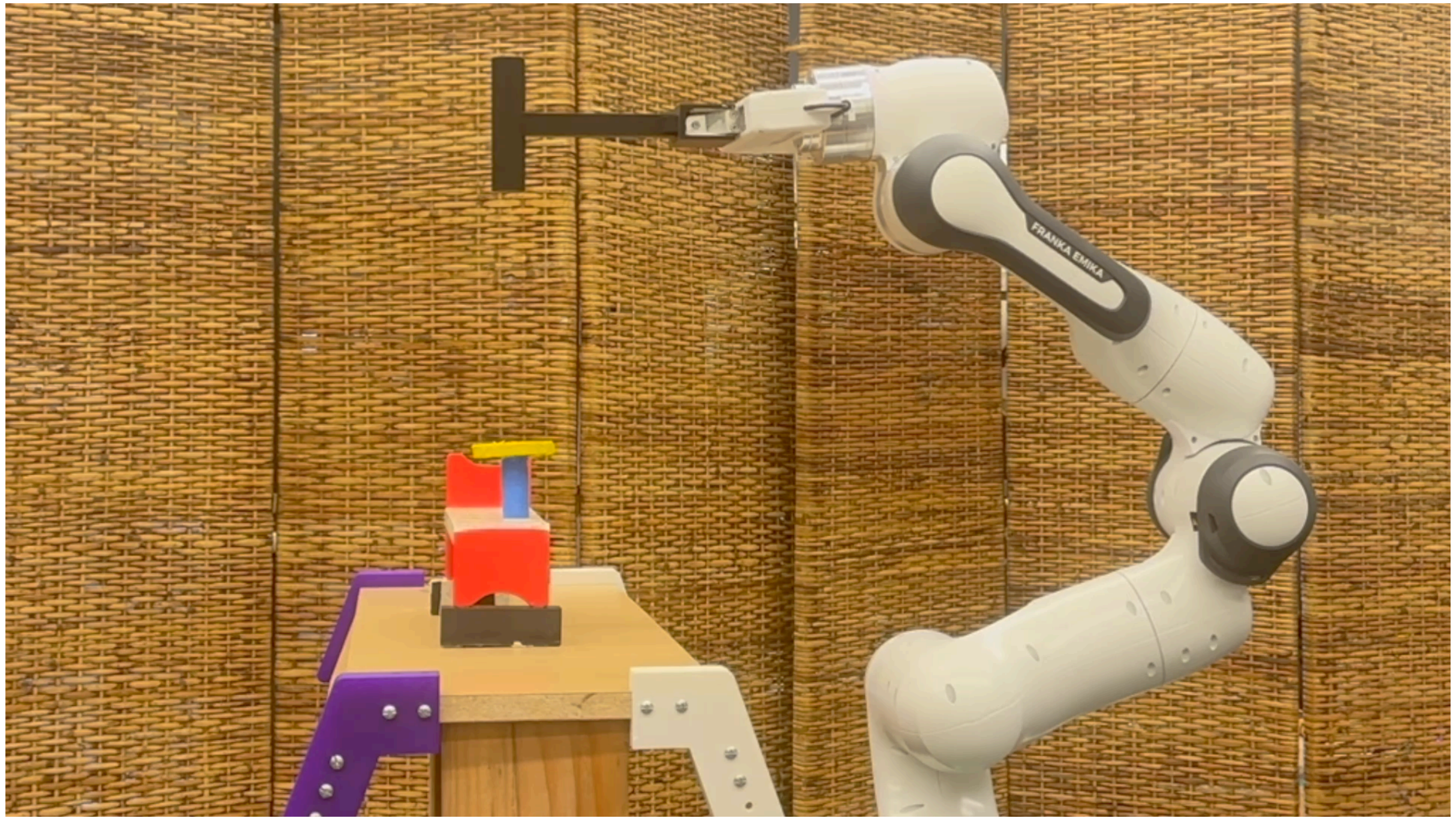
- Contact dynamics
- Deformables



Current Fine-Tuning Approaches

Current Approaches: Use off-domain data to <u>initialize</u> learning

- Pre-train policy
- Populate replay buffer



Failure Mode: Poor fine-tuning performance after transfer

- Policy performance initially collapses

SGFT: Simulation Guided Fine-Tuning

Key Question: Even though simulators are wrong, can we use them to learn effective exploration strategies for few-shot real world adaptation?

Key Idea: Use the value function learned in simulation V_{sim} to guide real-world exploration.

Key Insight: V_{sim} will have a maxima at goal states in both simulation and reality — we can follow V_{sim} to reach goals in the real-world!

SGFT: Simulation Guided Fine-Tuning

Value-Based Reward Shaping:

$$\bar{r}(s) = r(s) + \gamma V_{sim}(s') - V_{sim}(s)$$

Shortening Learning Horizon:

$$\sum_{t=0}^{\infty} r(s) \rightarrow \sum_{t=0}^{H} \bar{r}(s)$$

Modified Returns:

$$\underbrace{\gamma^H V_{sim}(s_H)}_{\text{bootstrapping}} + \underbrace{\sum_{t=0}^{H-1} r(s_t)}_{\text{original return}} - \underbrace{V_{sim}(s_0)}_{\text{baseline}}$$

- ✓ Tractable short horizon real returns
- ✓ Simulation bootstraps long-horizon behavior
- ✓ Strong learning signal reduces variance

SGFT: Simulation Guided Fine-Tuning

Value-Based Reward Shaping:

$$\bar{r}(s) = r(s) + \gamma V_{sim}(s') - V_{sim}(s)$$

Shortening Learning Horizon:

$$\sum^{\infty} r(s) \rightarrow \sum^{H} \bar{r}(s)$$

Generality: **We can apply this to any base RL algorithm!**

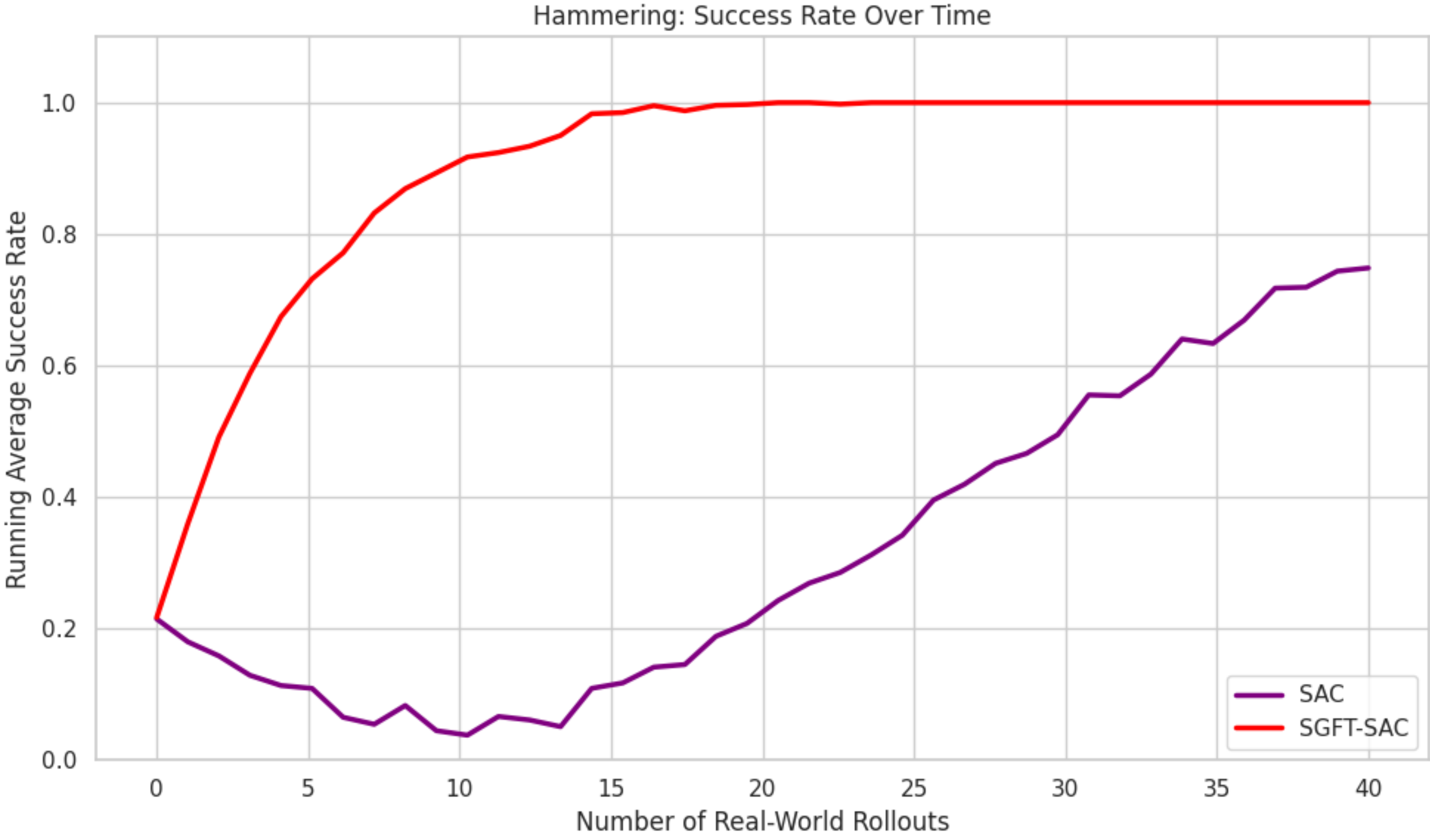
$$\gamma^H V_{sim}(s_H) + \sum_{t=0}^{H-1} r(s_t) - V_{sim}(s_0)$$

bootstrapping original return baseline

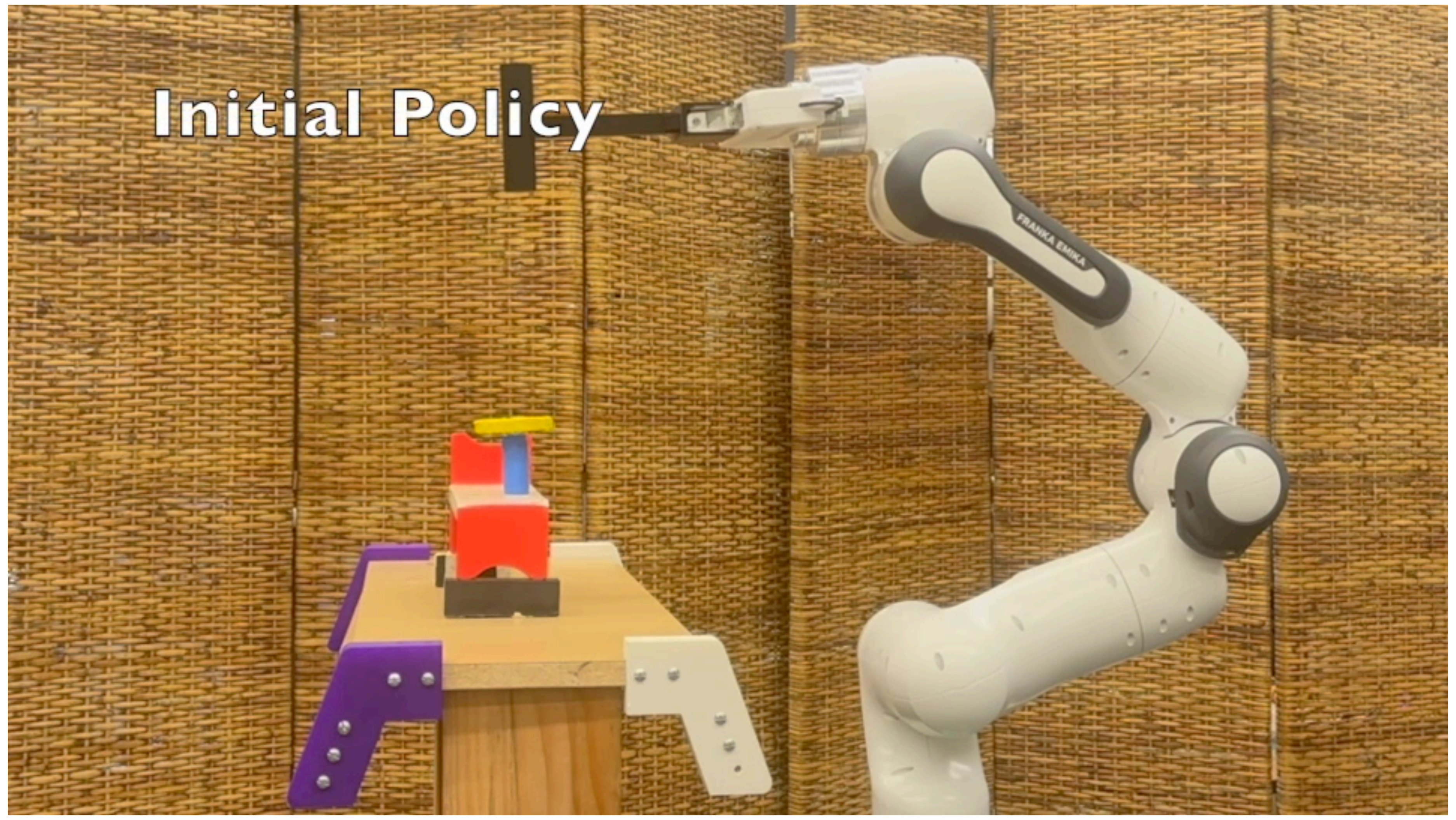
Optimizing short returns is tractable, but makes policy consistent with real dynamics

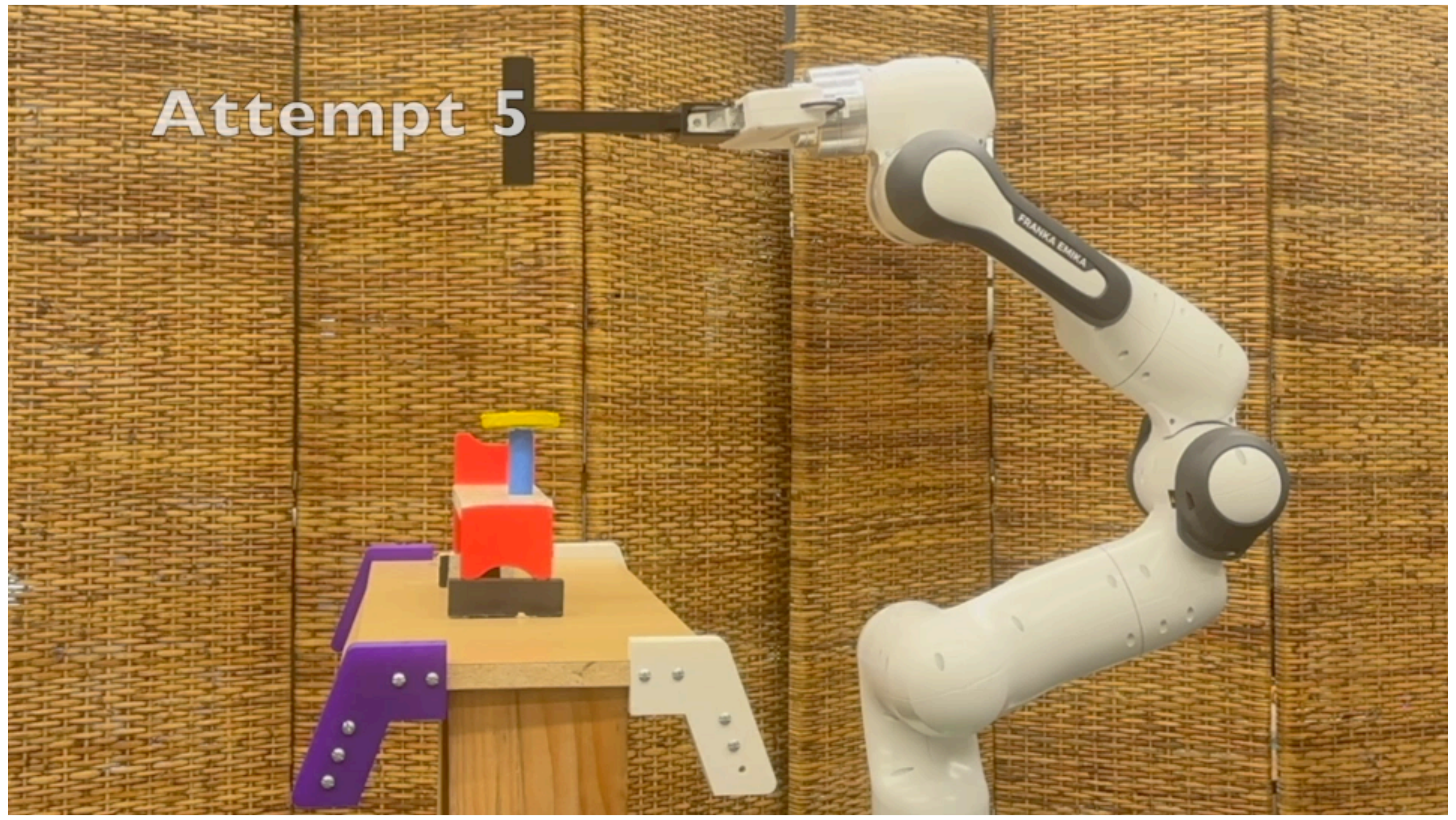
Small bias in learned policy

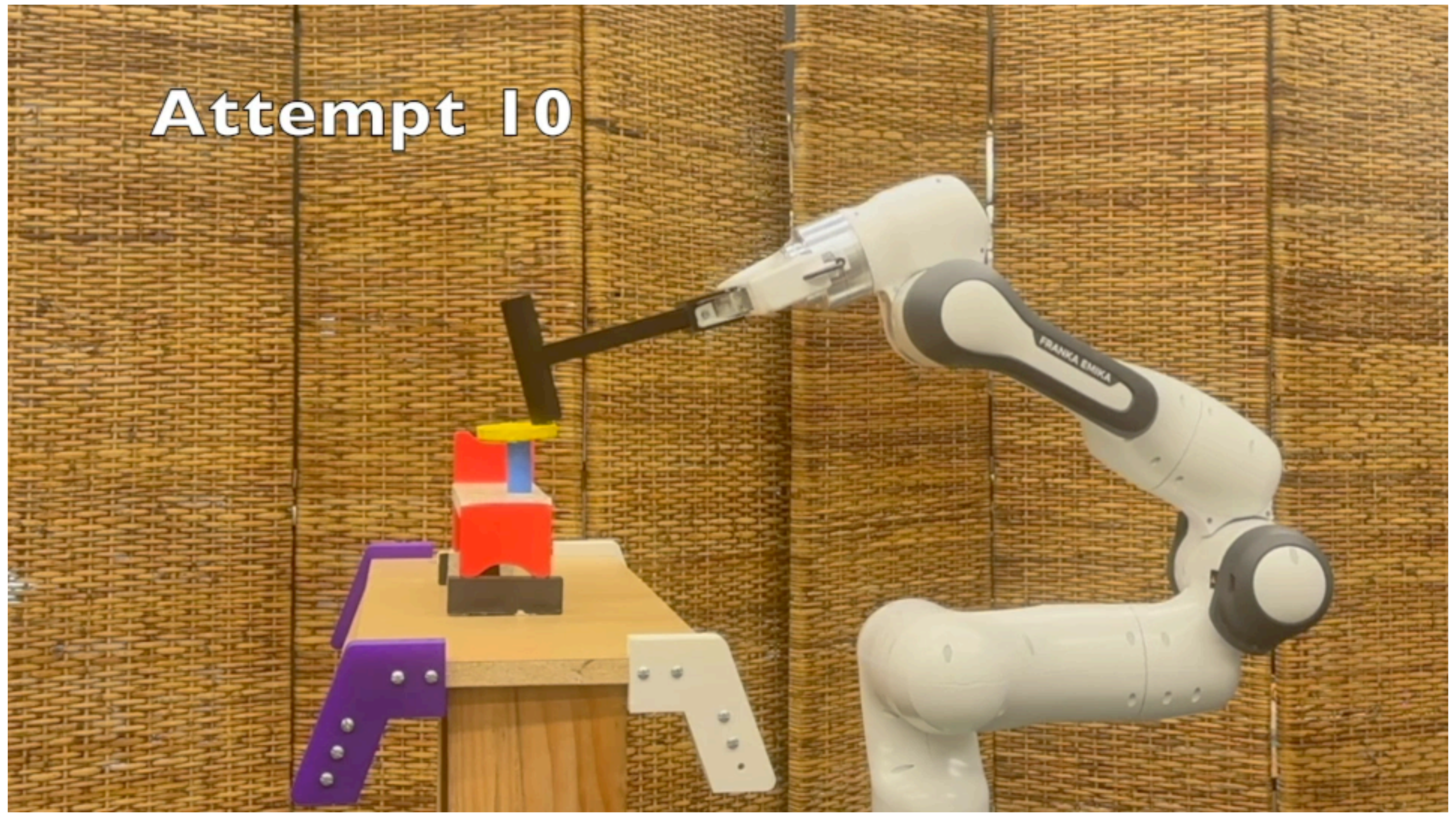
Consistent, Rapid Fine-tuning Progress

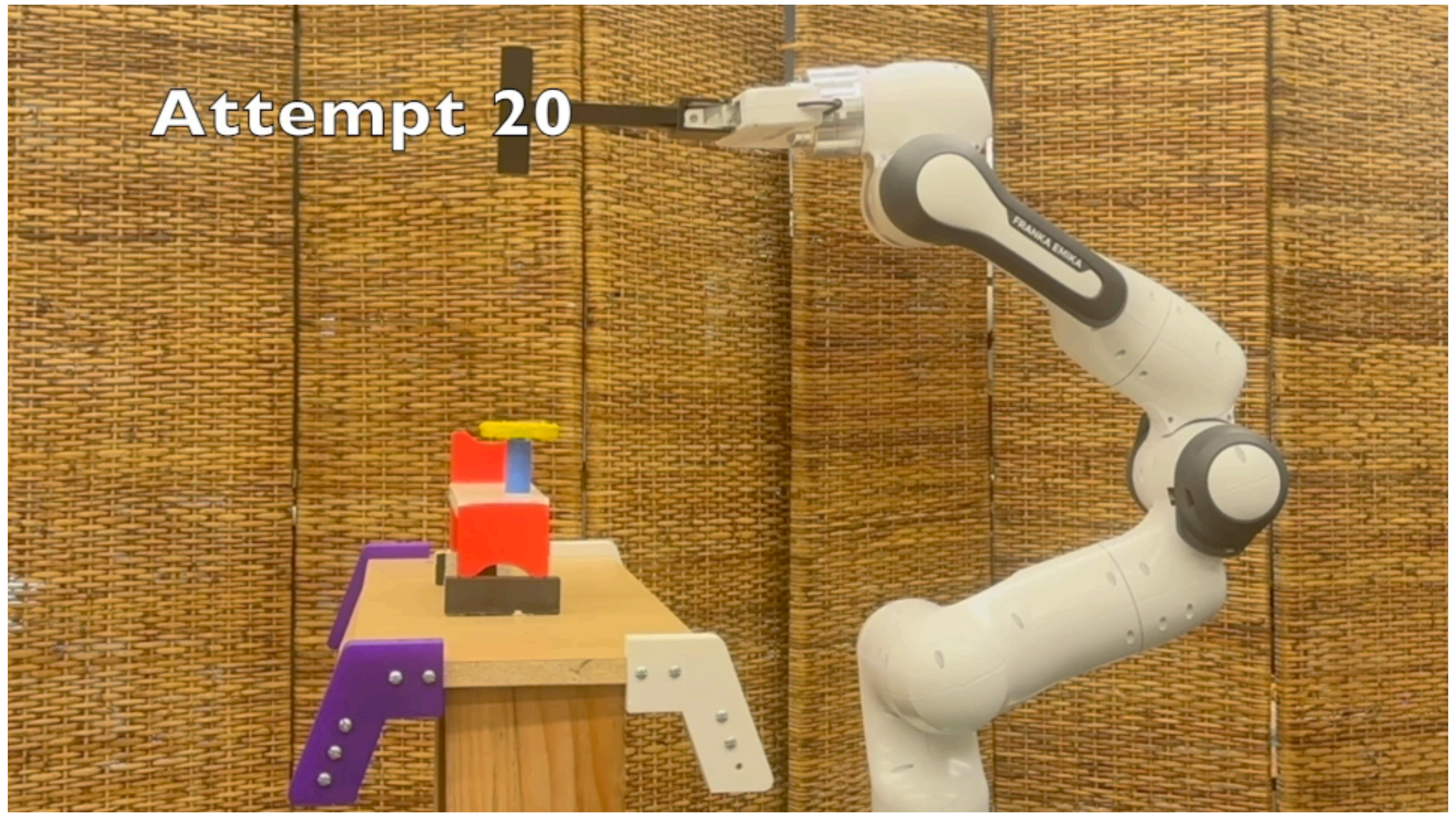


Consistent, Rapid Fine-tuning Progress









Substantial Improvement over Benchmarks