Time Your Rewards:

Learning Temporally Consistent Rewards from a Single Video Demonstration

Huaxiaoyue (Yuki) Wang, Will Huey*, Anne Wu, Yoav Artzi, Sanjiban Choudhury*
Cornell University

PORTAL

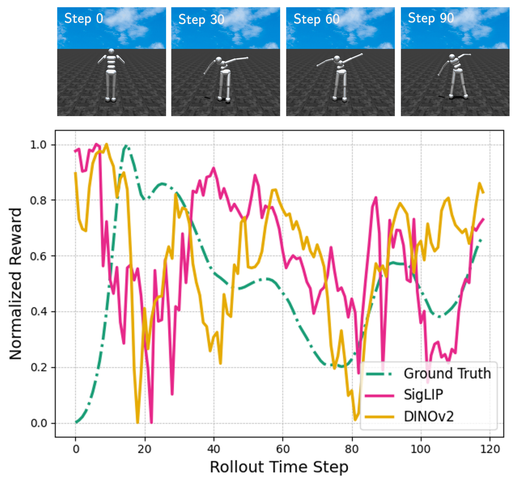
Learn the Task Specified in the Video Demo Through RL

↓

Define a class of reward function based on the video demo

Challenge 1:

Off-the-shelf vision models provide noisy rewards

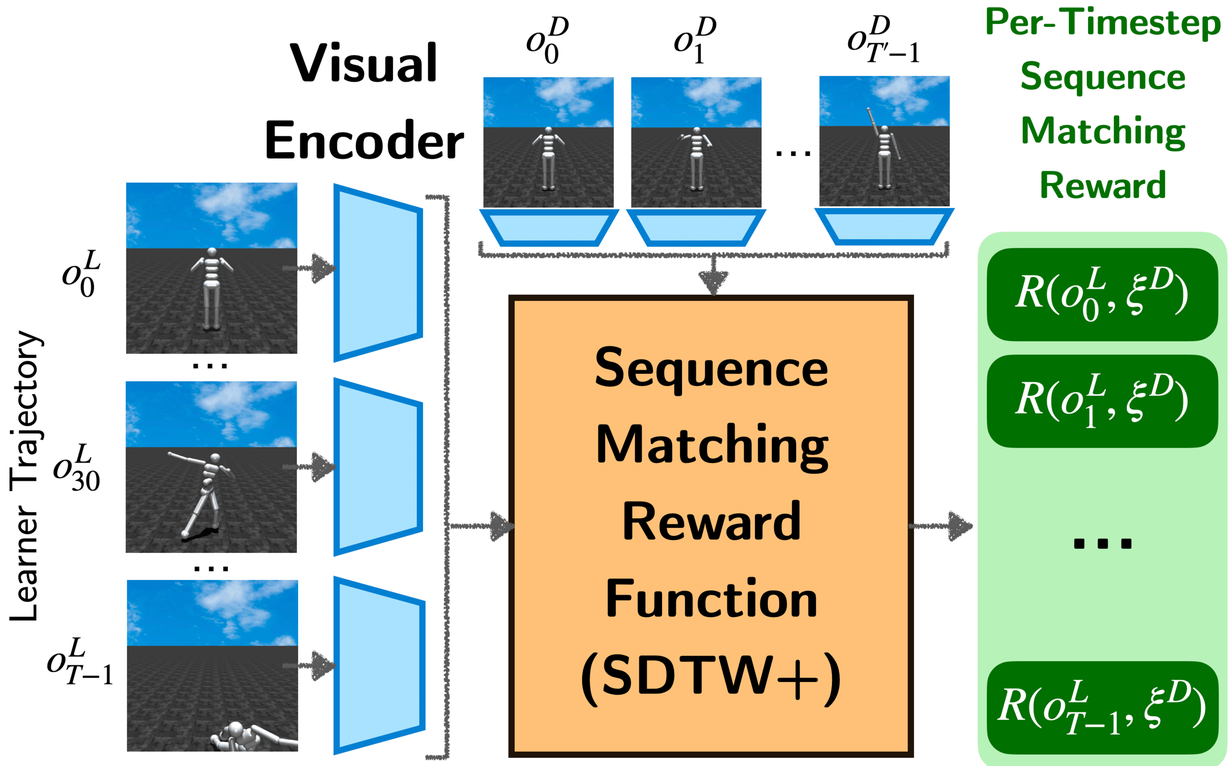


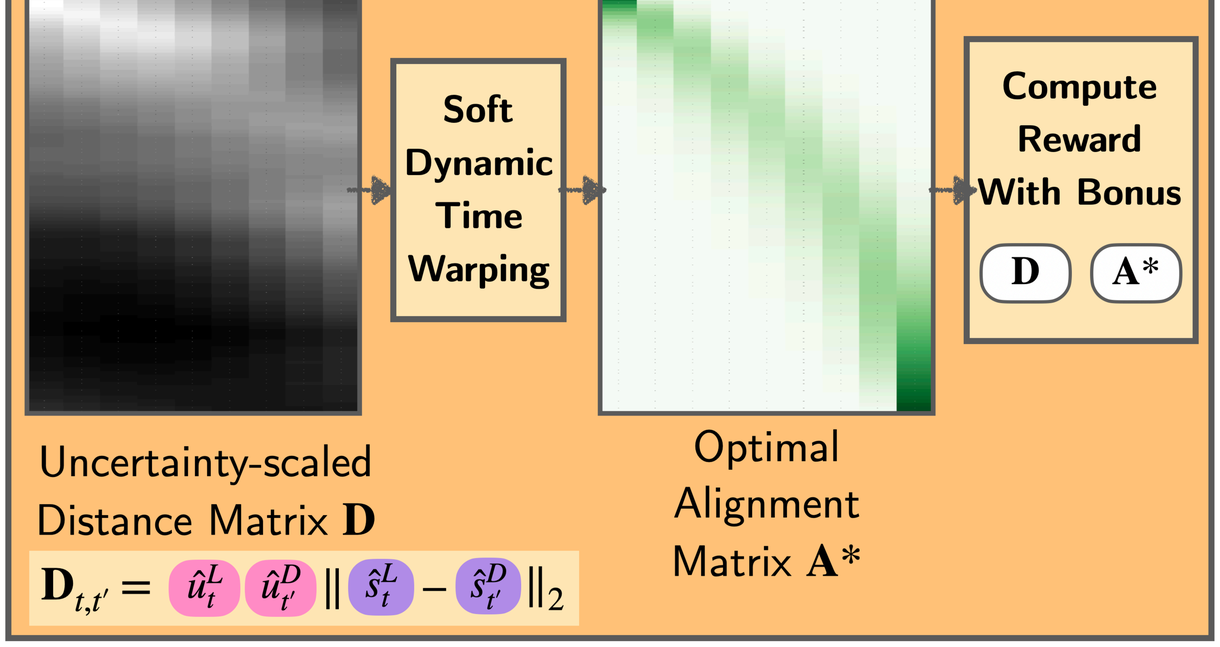
Challenge 2:

Certain tasks require completing subgoals in specific orders

A good reward function needs to:

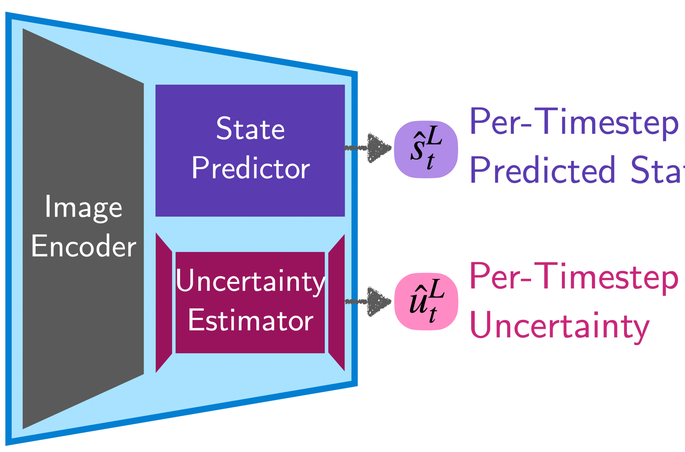
- ☐ **Enforce temporal constraints**
- ☐ **Avoid reward hacking**



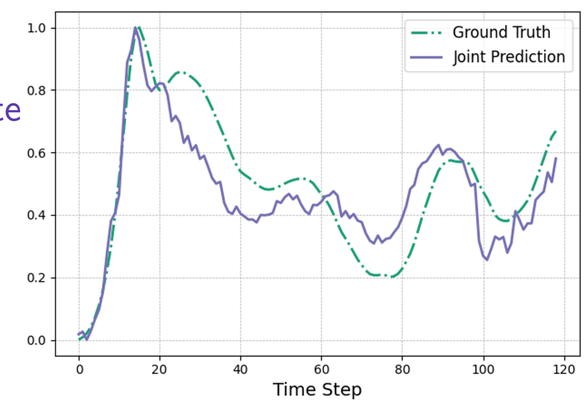


Visual Encoder

Finetuned on Privileged States + Uncertainty Estimation



More Reliable Signal



SDTW+ as the Sequence Matching Reward Fn

SDTW Instead of Optimal Transport → **Enforce Temporal Constraints**

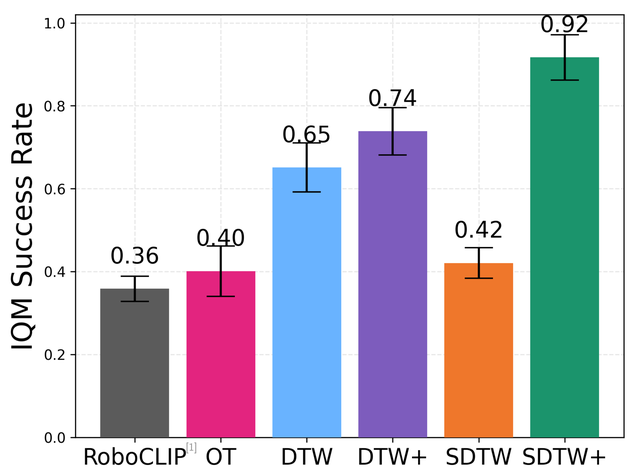
Cumulative Reward Bonus → **Reduce Reward Hacking**

$$\mathcal{R}_{SDTW+}(o_t^L, \xi^D) = \begin{cases} \mathcal{R}_{SDTW}(o_0^L, \xi^D) & \text{if } t = 0, \\ \mathcal{R}_{SDTW}(o_t^L, \xi^D) + \mathcal{B}(o_t^L, \xi^D) & \text{if } t > 0. \end{cases}$$

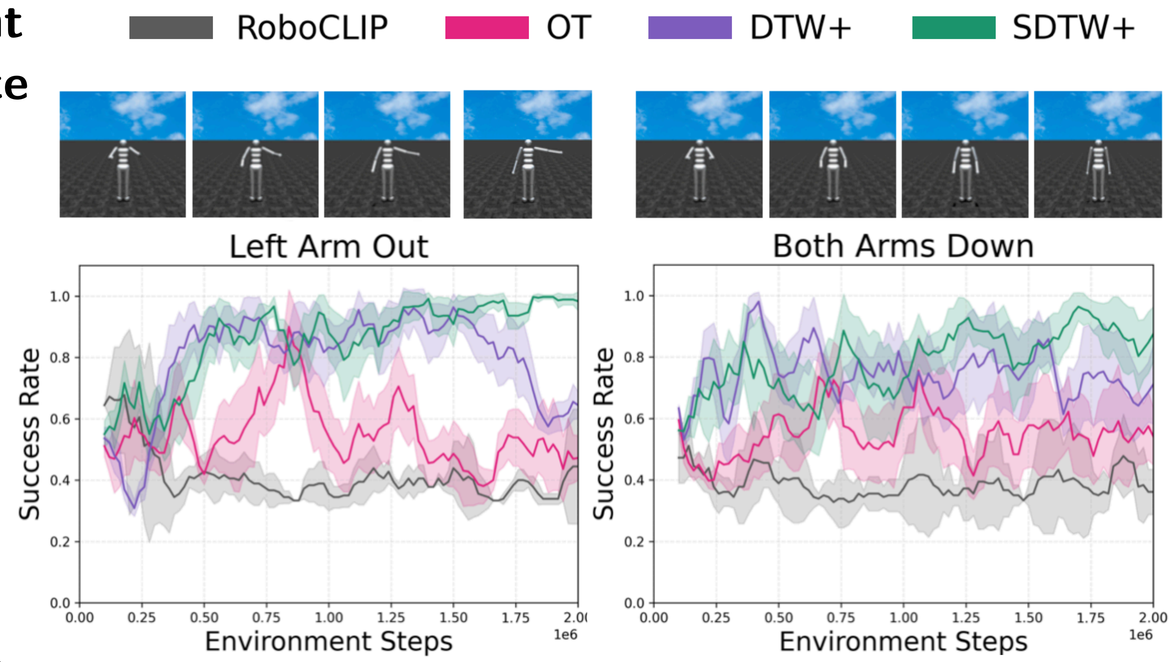
Based on the learner's progress at time t

SDTW+ reward helps agent

- **achieve higher success rate**
- **train stably**



OT: Optimal Transport DTW: Special Case of SDTW (smoothing param = 0)
DTW+: DTW with Cumulative Reward Bonus



Please check out our paper for more results.

Future Work

- Longer horizon, periodic tasks
  - More domains (e.g. cooking)
- Learning from human video demos

  - Vision encoder that bridges visual embodiment gap
- More theoretical analysis on reward hacking given different sequence matching rewards

[1] S. Sontakke, J. Zhang, S. Arnold, K. Pertsch, E. Bıyık, D. Sadigh, C. Finn, and L. Itti. Robo-clip: One demonstration is enough to learn robot policies. Advances in Neural Information Processing Systems, 36, 2024.