ICML
International Conference
On Machine Learning

PORTAL

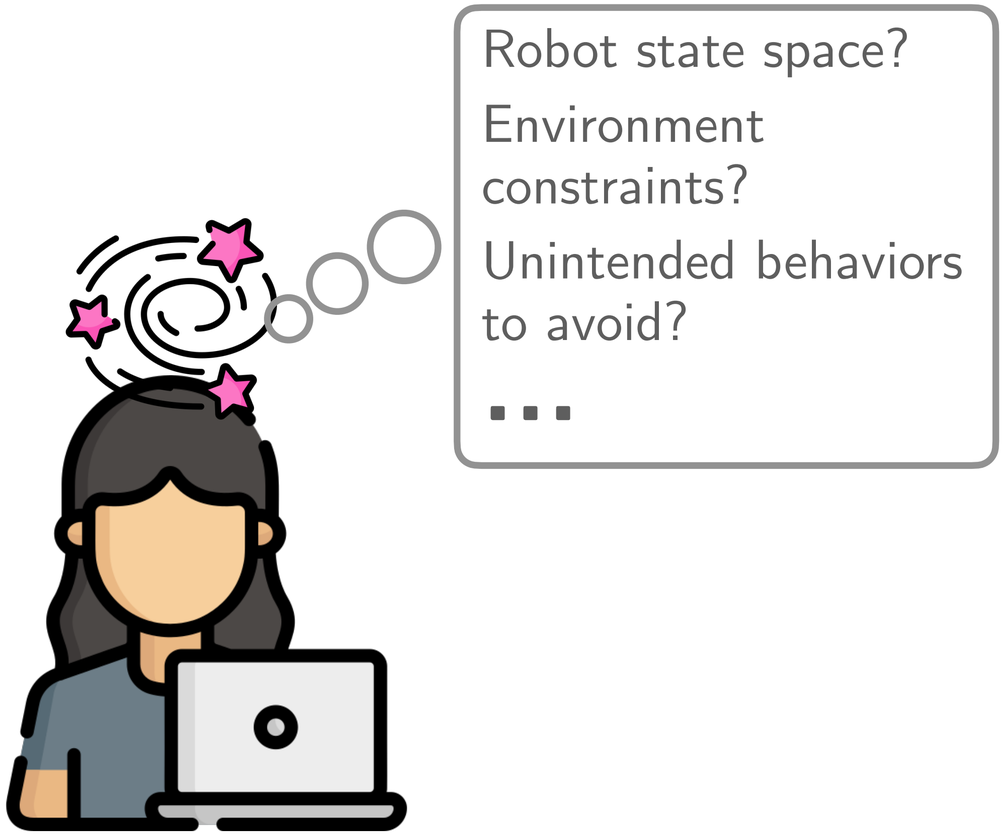
CORNELL UNIVERSITY
FOUNDED A.D. 1865

Imitation Learning from a Single Temporally Misaligned Video




William Huey*, Huaxiaoyue (Yuki) Wang*, Anne Wu, Yoav Artzi, Sanjiban Choudhury

Designing reward function is tedious



Designing reward function is tedious

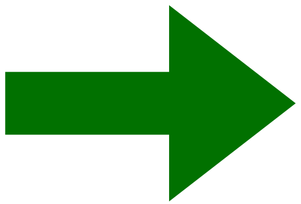
Instead, easier to provide a video demonstrating the task

Robot state space?

Environment constraints?

Unintended behaviors to avoid?

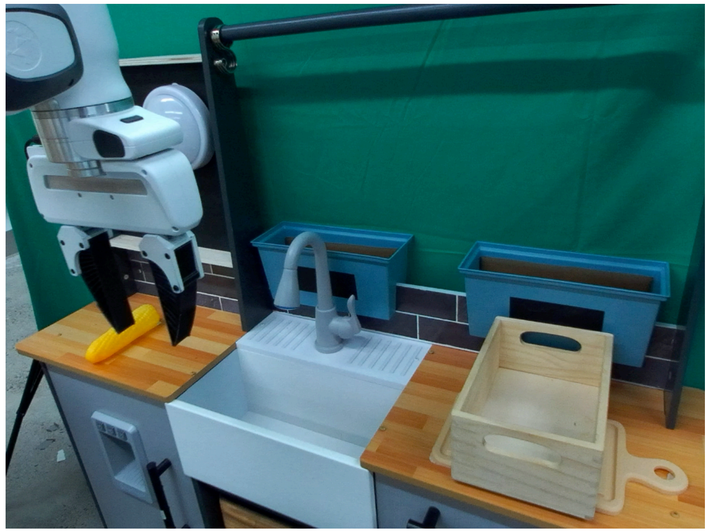
...



https://portal-cornell.github.io/X-Sim/

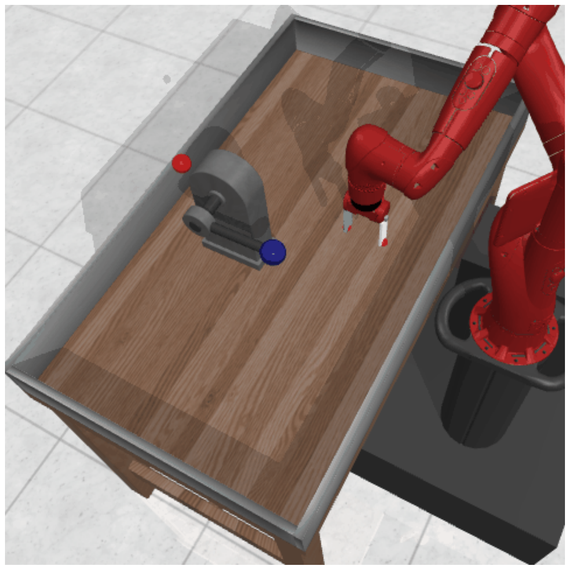
Demonstrations are often **temporally misaligned**

Different Speeds

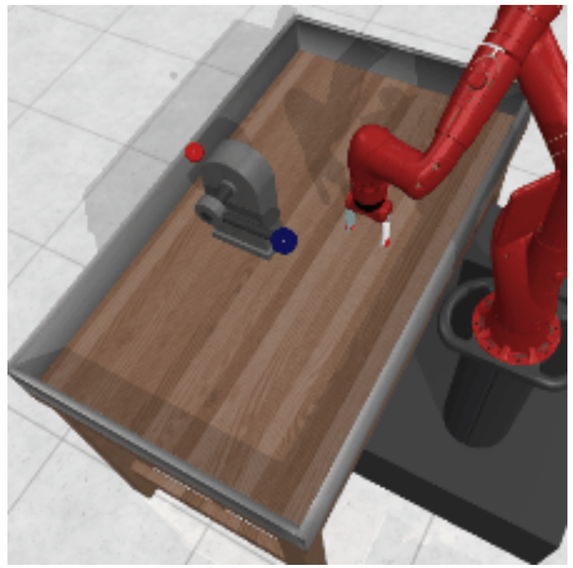


https://portal-cornell.github.io/X-Sim/

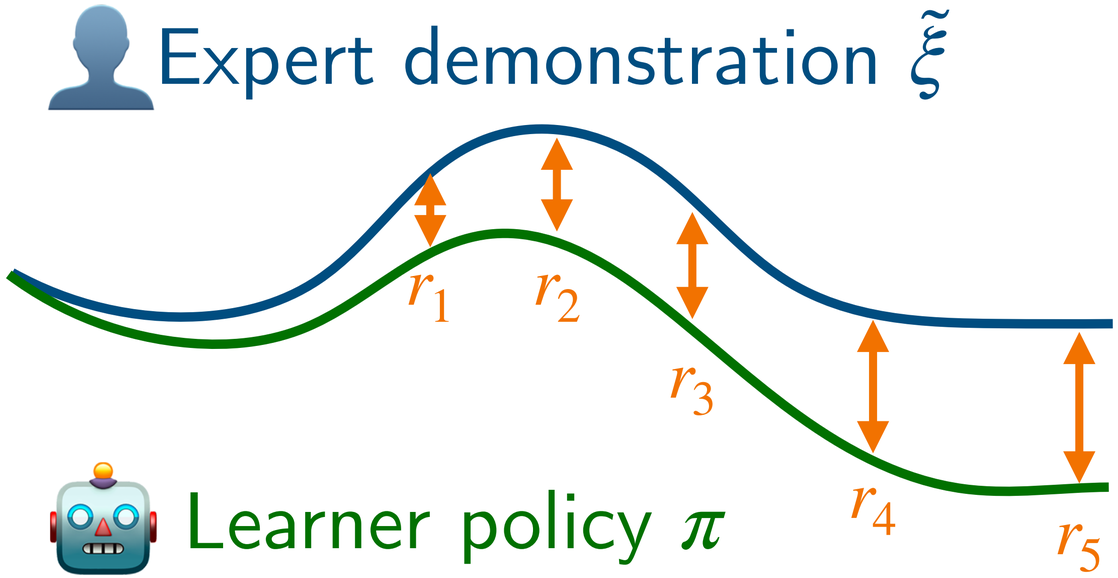
Demonstrations are often **temporally misaligned**



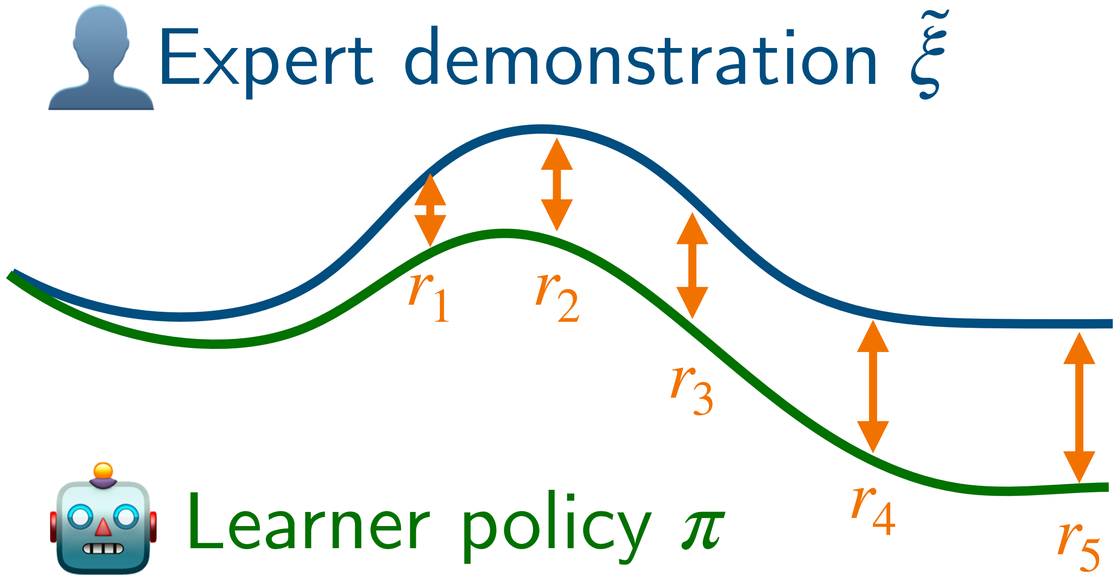
Different Speeds

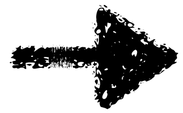


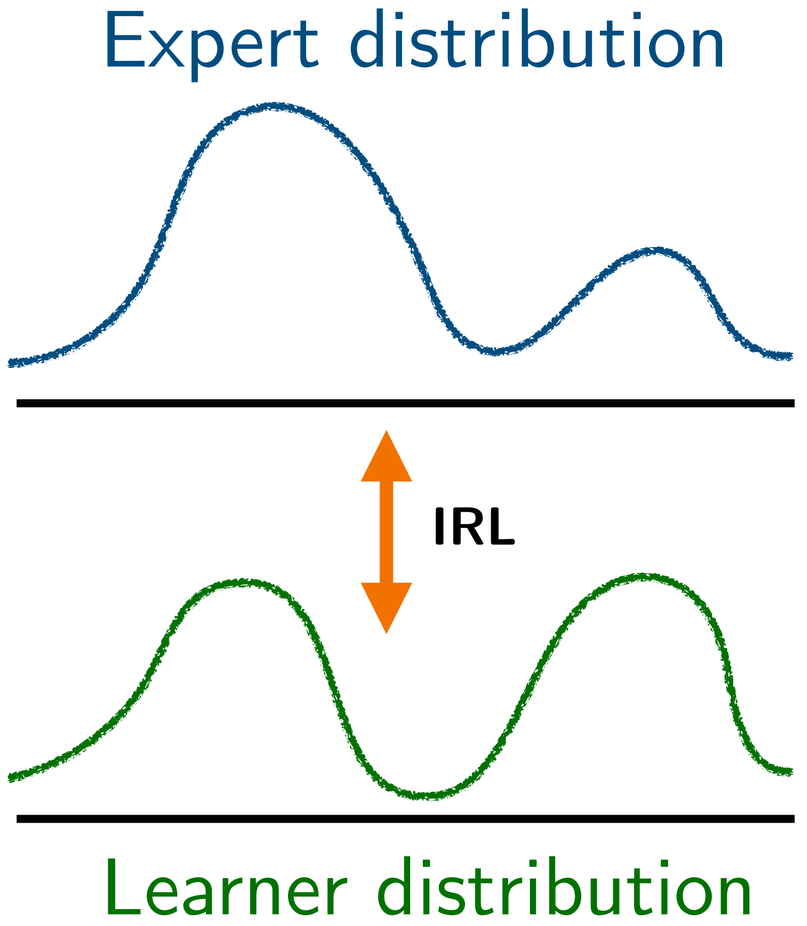
Existing approaches to learning from videos use **inverse reinforcement learning** (IRL)



IRL matches learner and expert distributions in expectation

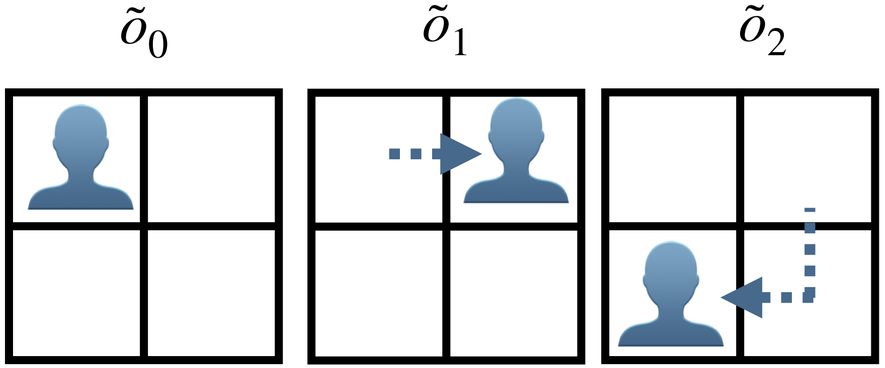




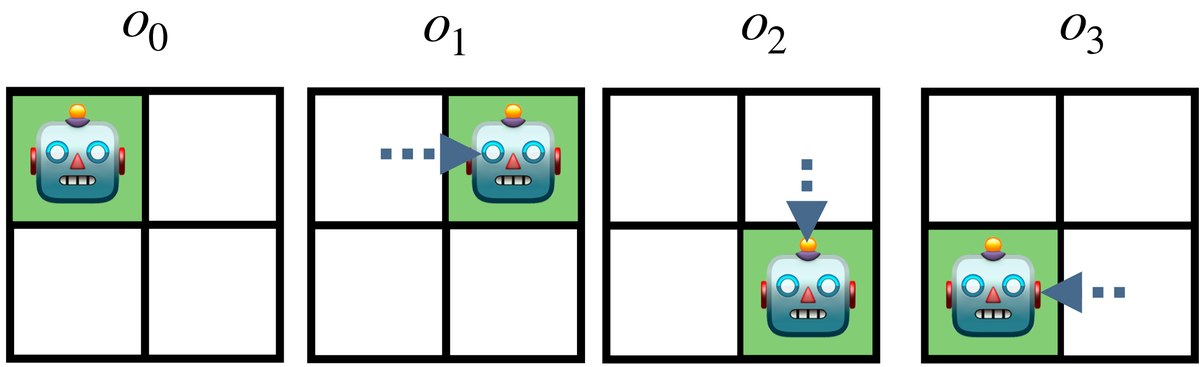


Swamy, G., Choudhury, S., Bagnell, J. A., and Wu, Z. S. Of moments and matching: A game-theoretic framework for closing the imitation gap. Proceedings of the 38th International Conference on Machine Learning, 2021.

Existing approaches to learning from videos match **distributions over video frames**



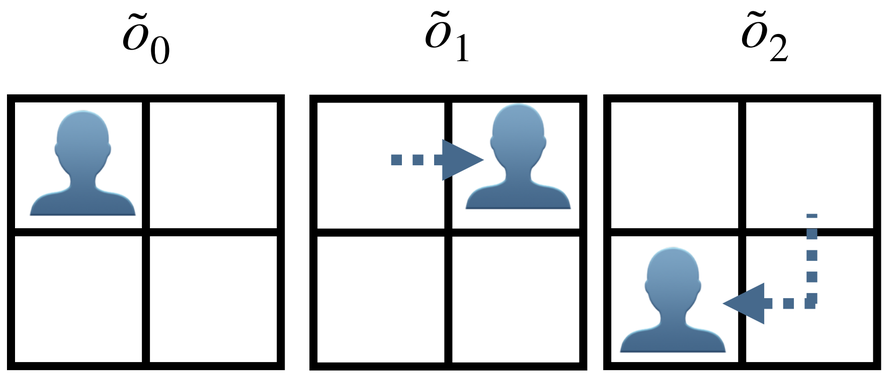
Demonstration video $\tilde{\xi}$



Learner video ξ

Tian, T., Xu, C., Tomizuka, M., Malik, J., and Bajcsy, A. What matters to you? towards visual representation alignment for robot learning. In The Twelfth International Conference on Learning Representations, 2024.

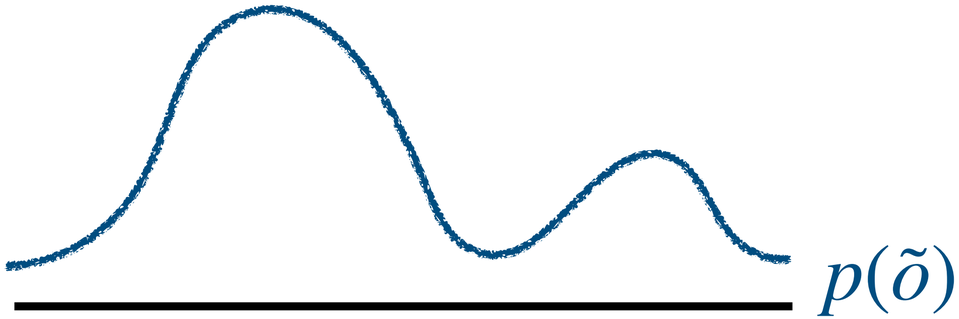
Existing approaches to learning from videos match **distributions over video frames**

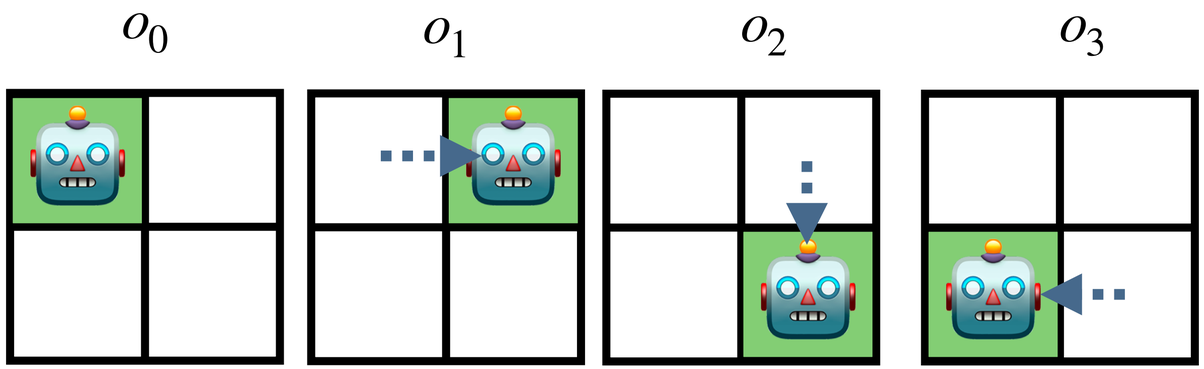


Demonstration video $\tilde{\xi}$

Frame encoder ϕ

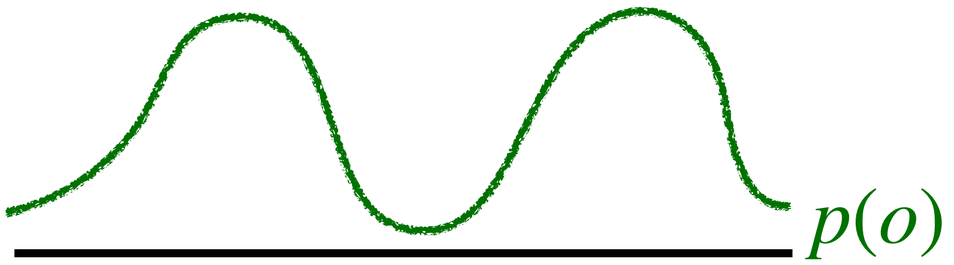
Expert frame distribution





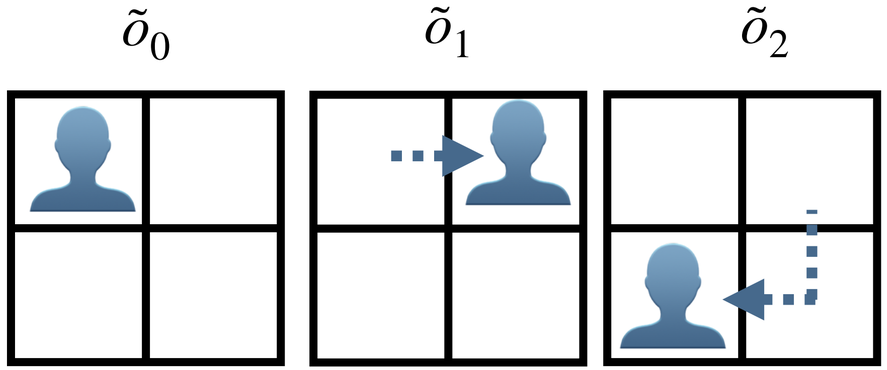
Learner video ξ

ϕ



Learner frame distribution

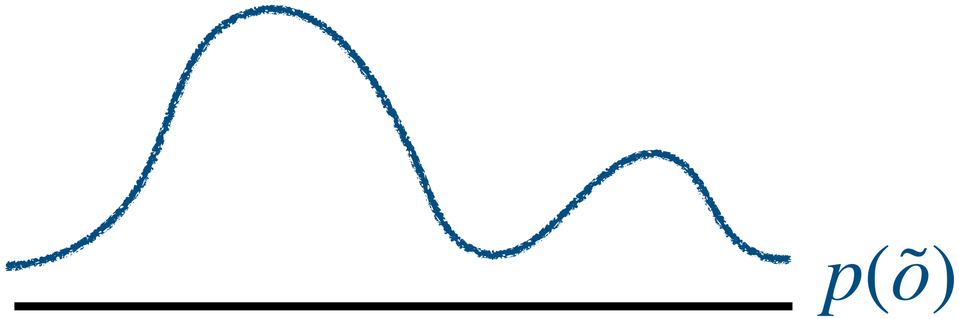
Existing approaches to learning from videos match **distributions over video frames**



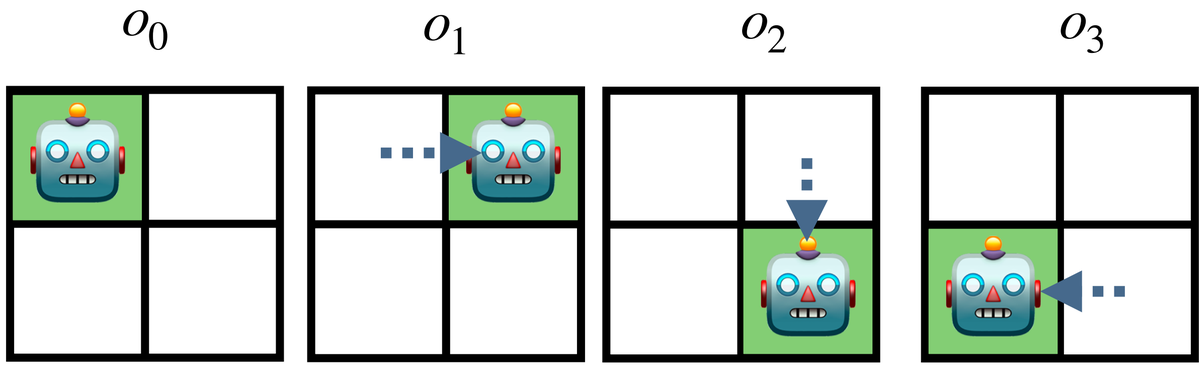
Demonstration video $\tilde{\xi}$

Frame encoder ϕ

Expert frame distribution

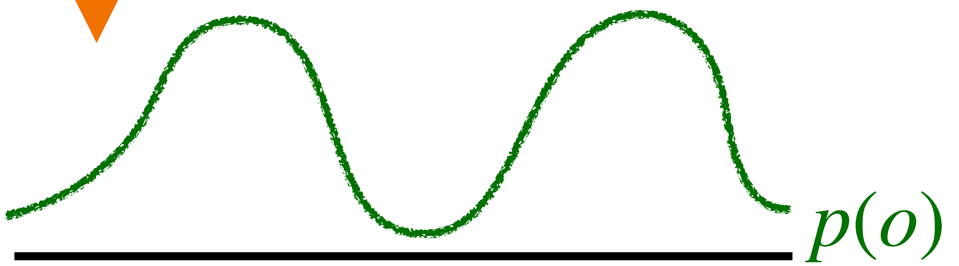


RL using integral probability metric (**optimal transport**) as reward



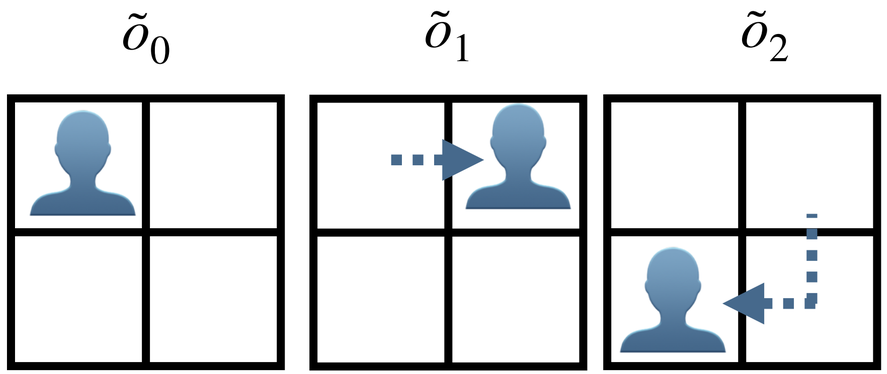
Learner video ξ

ϕ



Learner frame distribution

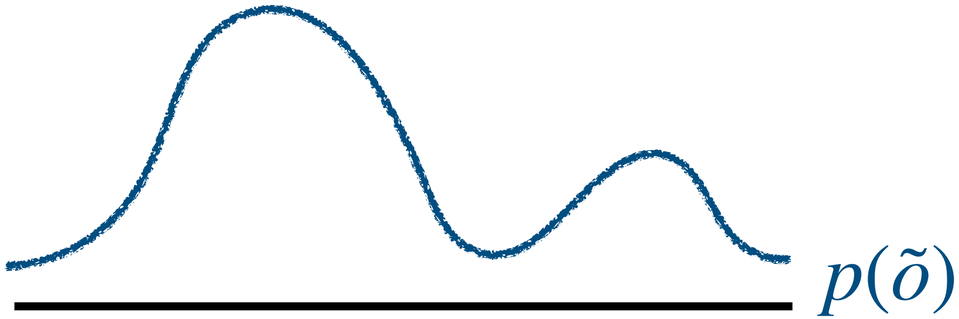
Existing approaches to learning from videos match **distributions over video frames**



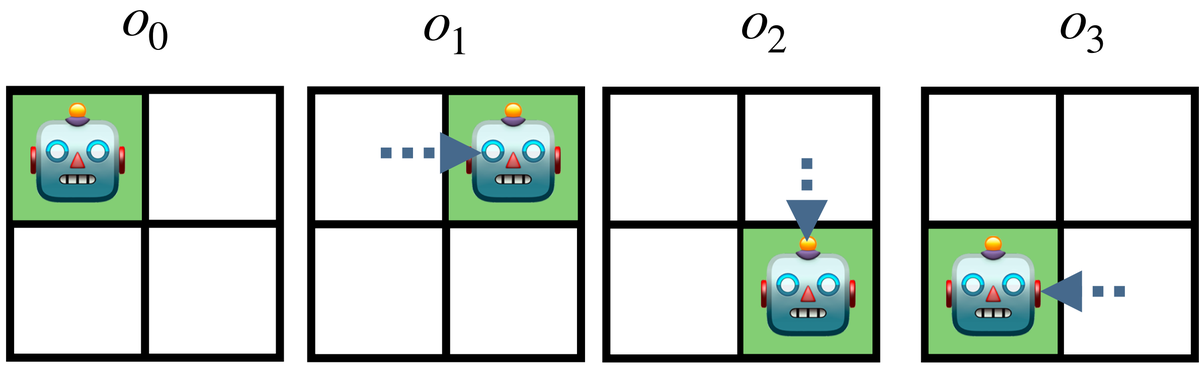
Demonstration video $\tilde{\xi}$

Frame encoder ϕ

Expert frame distribution

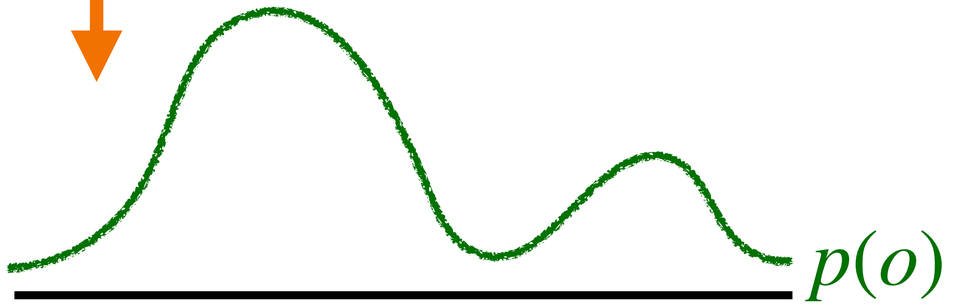


RL using integral probability metric (**optimal transport**) as reward



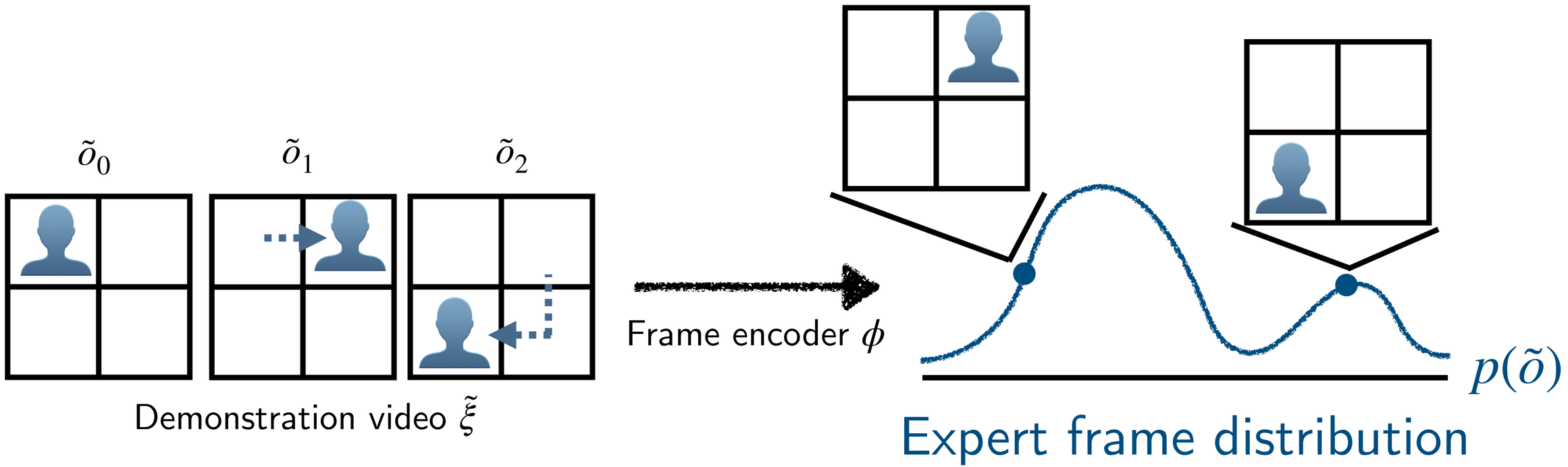
Learner video ξ

ϕ



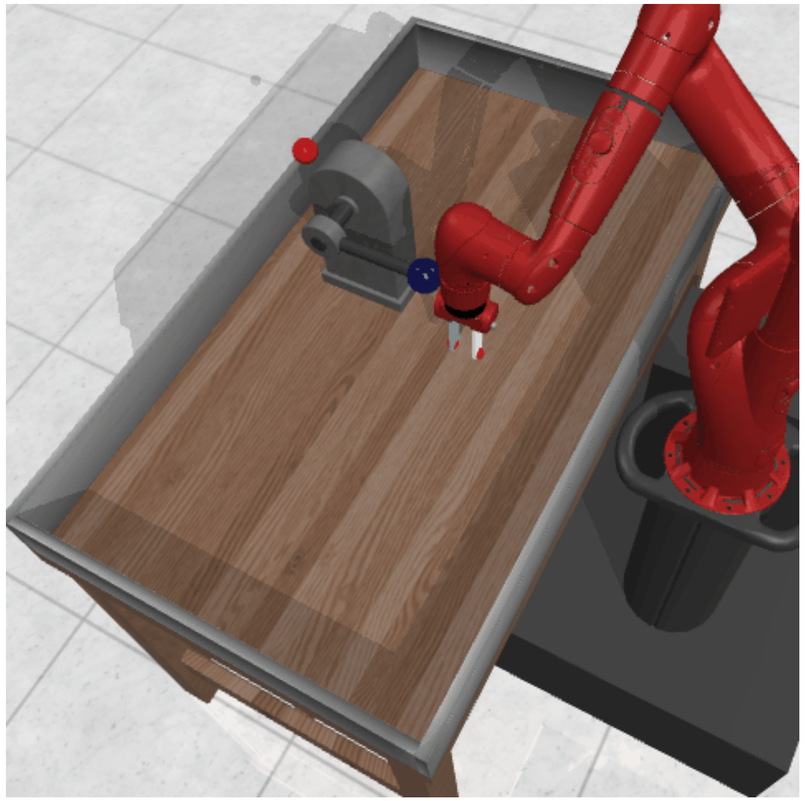
Learner frame distribution

Problem! Individual frames lack temporal information



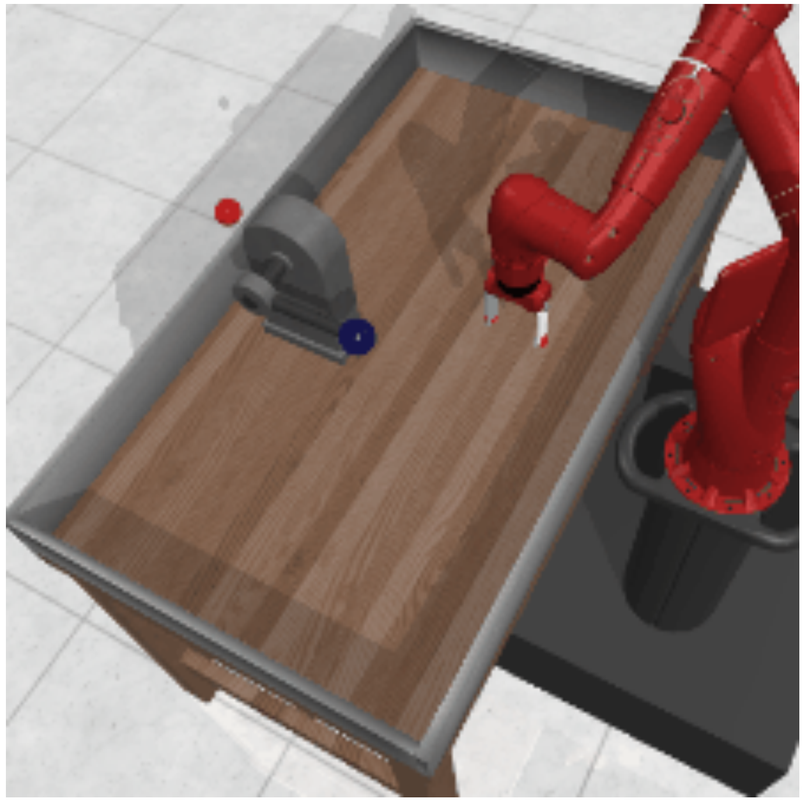
$p(\tilde{o})$ represents a collection of frames, not an ordered sequence

Experimental Example in Meta-world



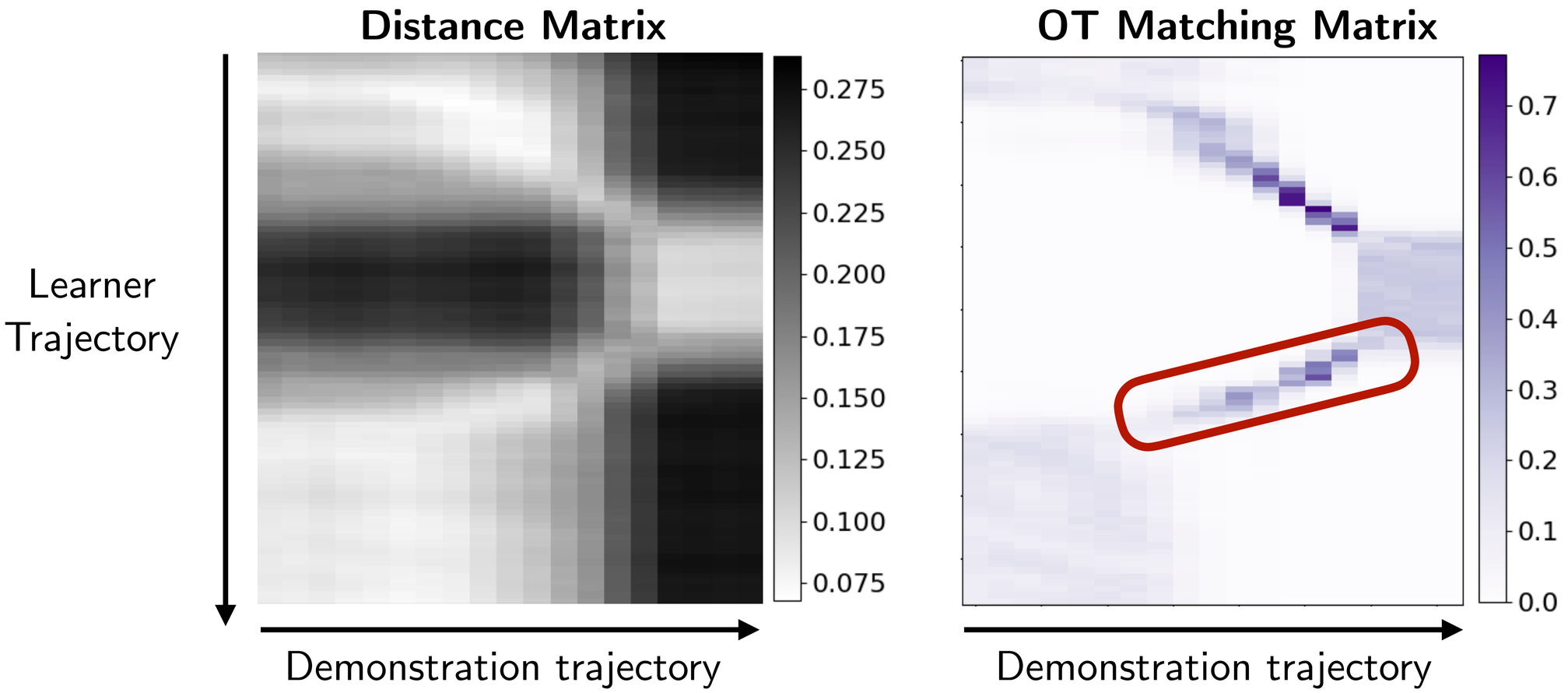
Demo

OT fails to enforce **temporal ordering**



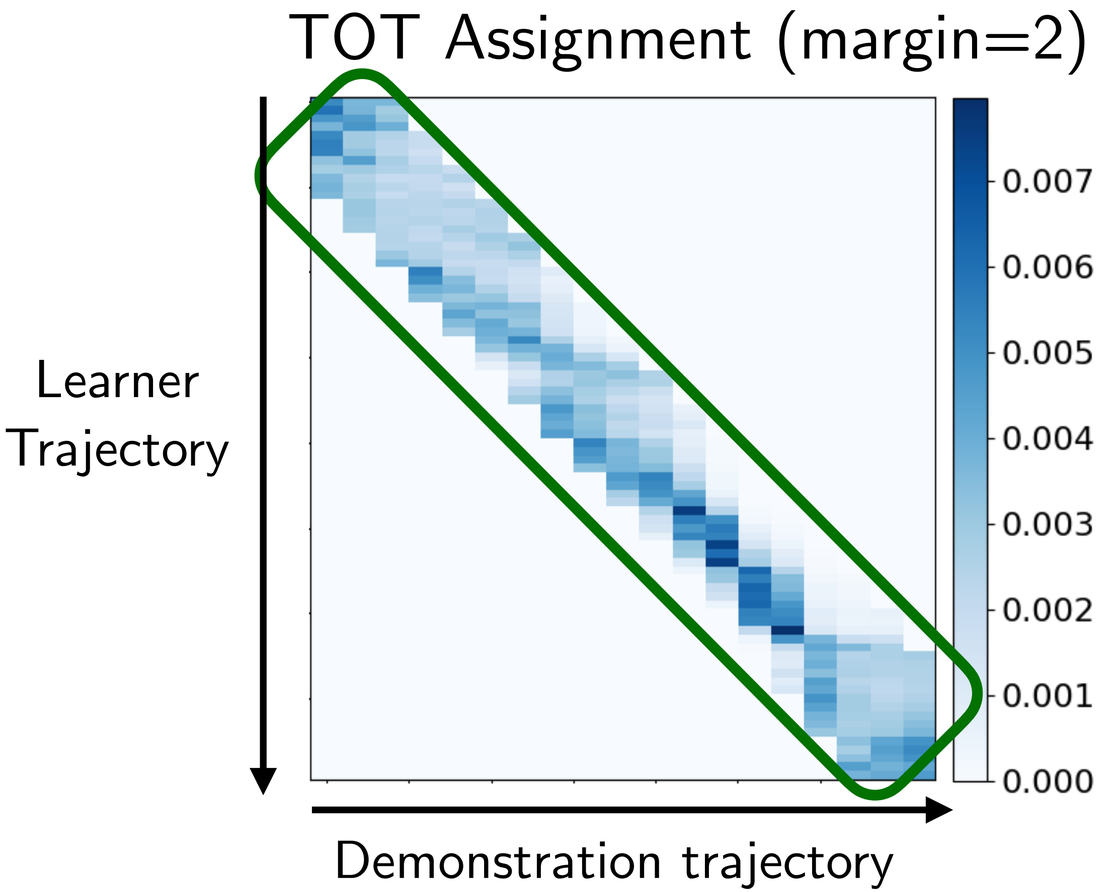
OT

OT fails to enforce **temporal ordering**



OT matches later learner frames to earlier subgoals.

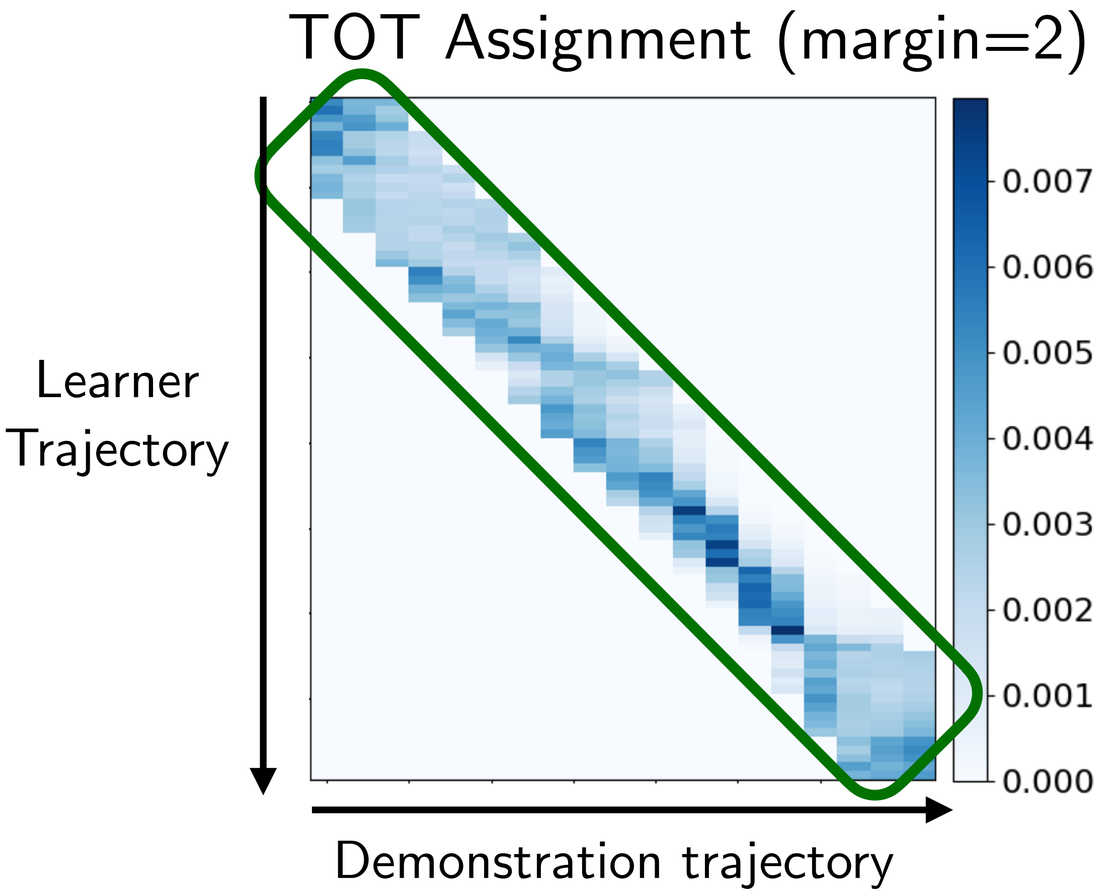
TemporalOT solves the ordering problem



Constrains OT to be near the diagonal

Fu, Y., Zhang, H., Wu, D., Xu, W., and Boulet, B. Robot policy learning with temporal optimal transport reward. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024

TemporalOT solves the ordering problem

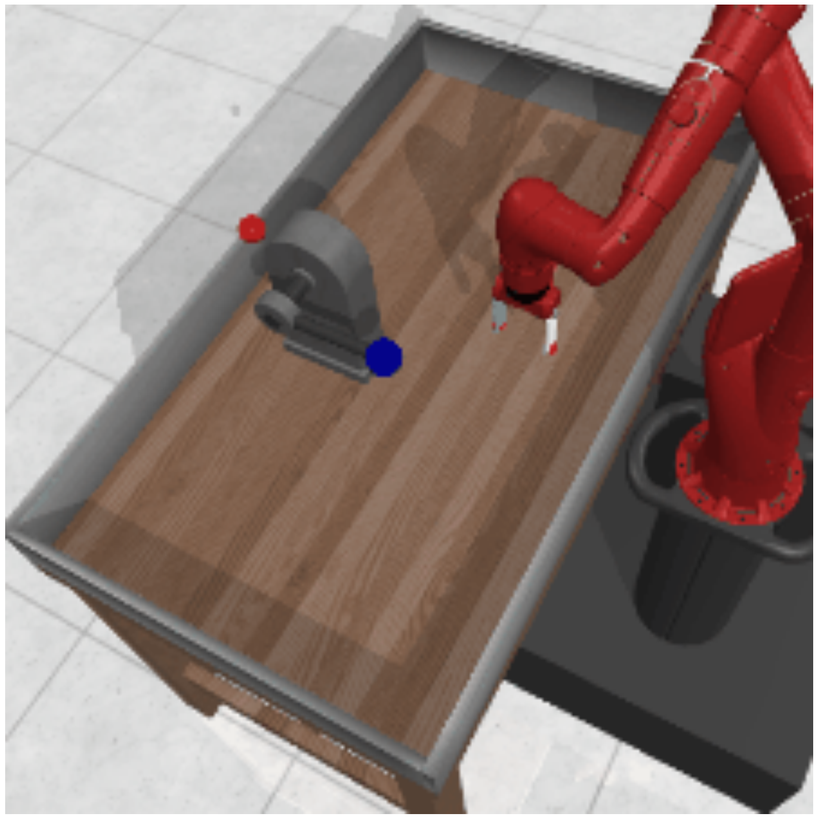


Constrains OT to be near the diagonal

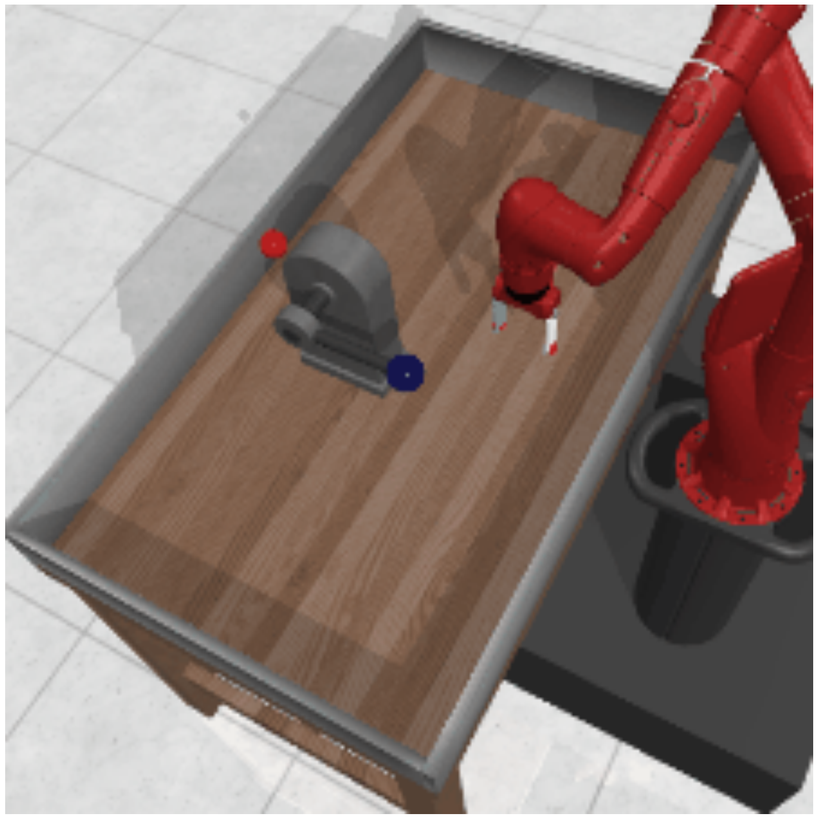
Problem: assumes learner and expert are *temporally aligned*

TemporalOT fails to enforce **subgoal coverage** under temporal misalignment

TOT (margin=10)

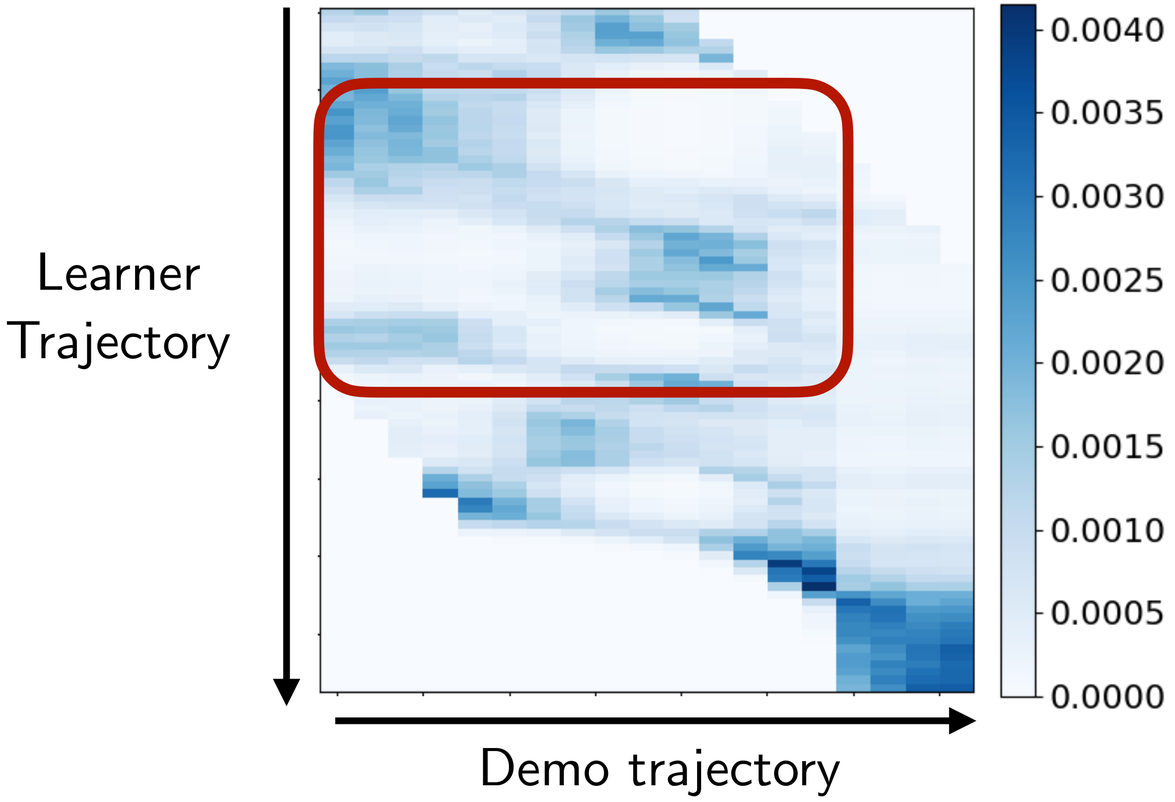


TOT (margin=2)



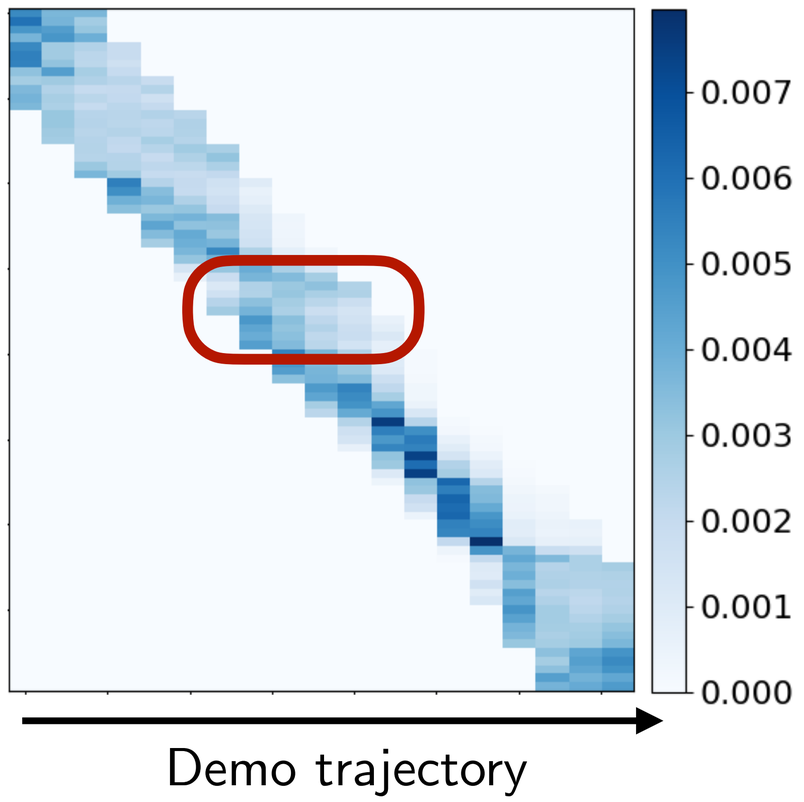
TemporalOT fails to enforce **subgoal coverage** under temporal misalignment

TOT Assignment (margin=10)



Fails to enforce temporal ordering due to the large mask window

TOT Assignment (margin=2)



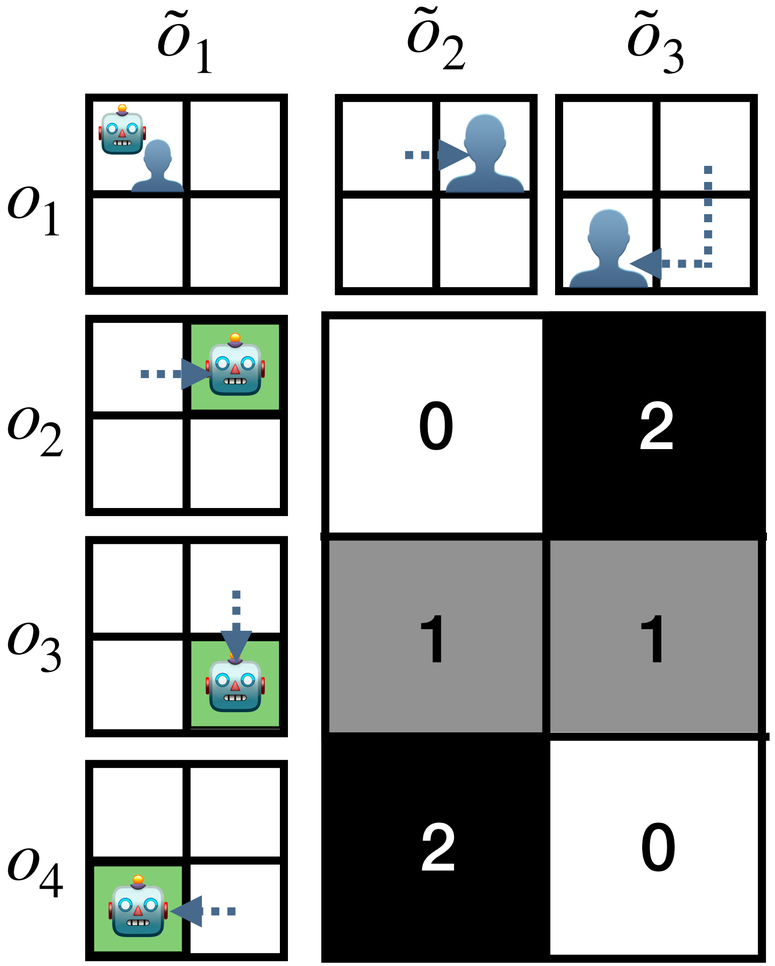
Over-constrains assignment, leading to slow and shaky motion

Instead of matching at
the frame-level,
we should match at the
sequence-level

Sequence-level matching should enforce:

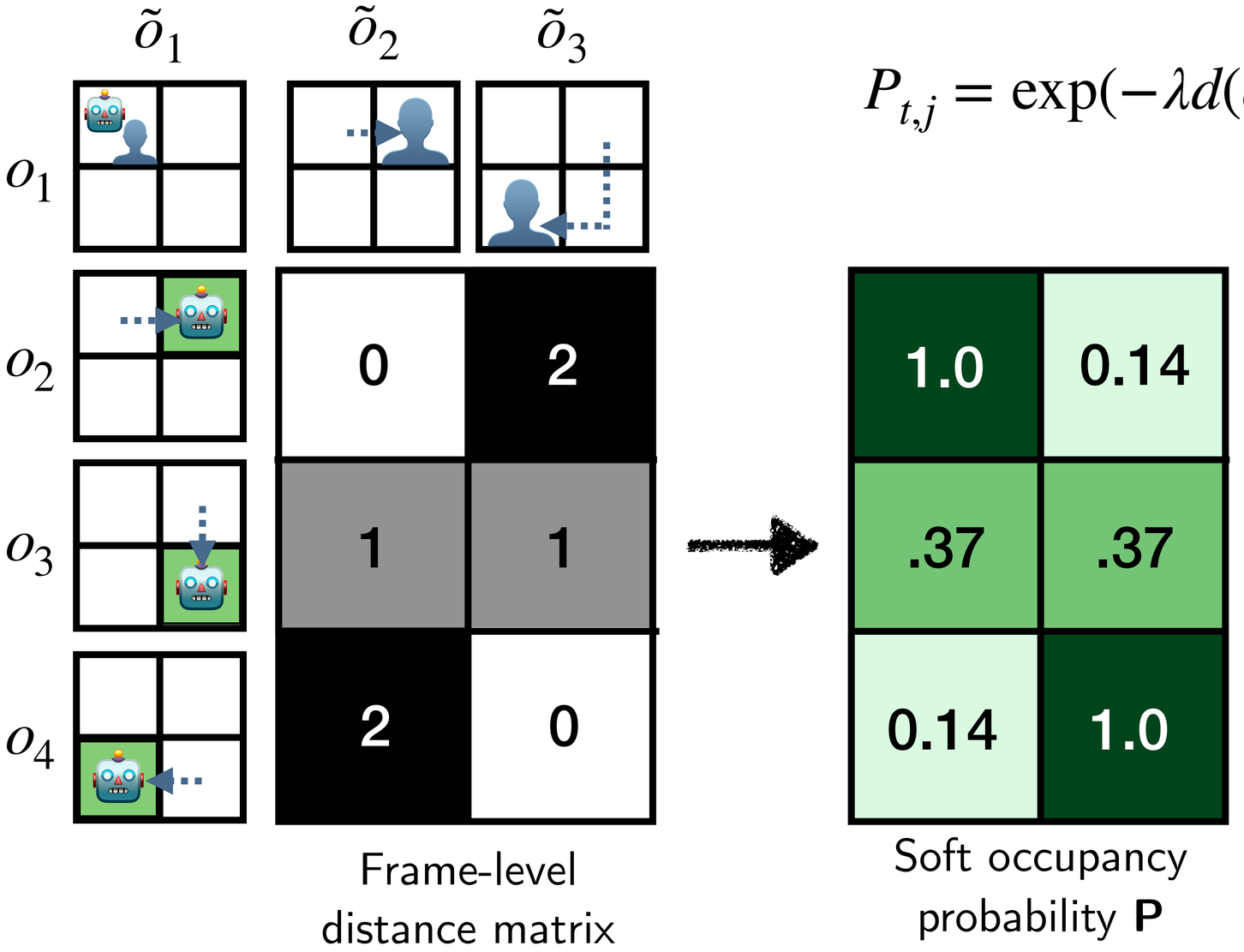
1. Subgoal **ordering**
2. Subgoal **coverage**

ORdered Coverage Alignment (**ORCA**)



Frame-level
distance matrix

ORdered Coverage Alignment (**ORCA**)

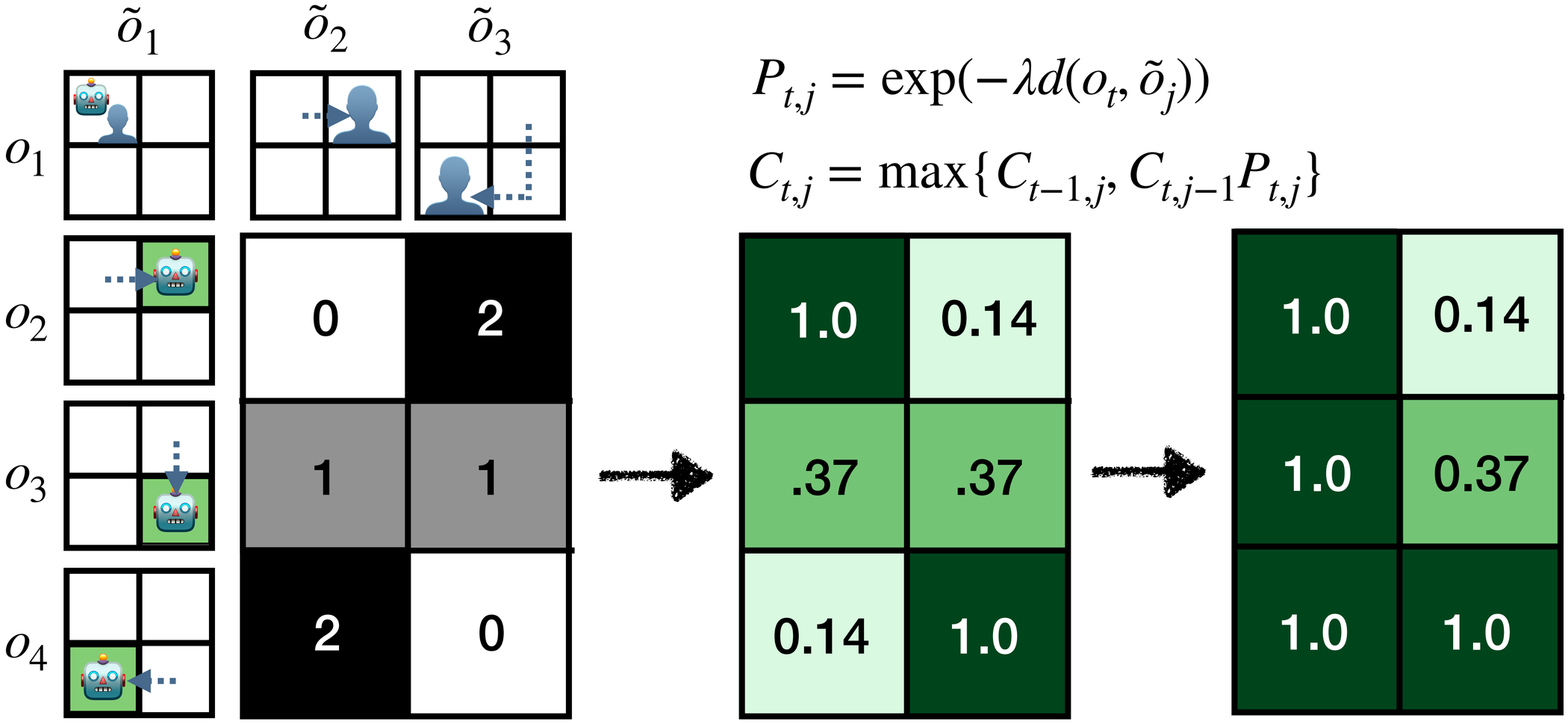


$$P_{t,j} = \exp(-\lambda d(o_t, \tilde{o}_j))$$

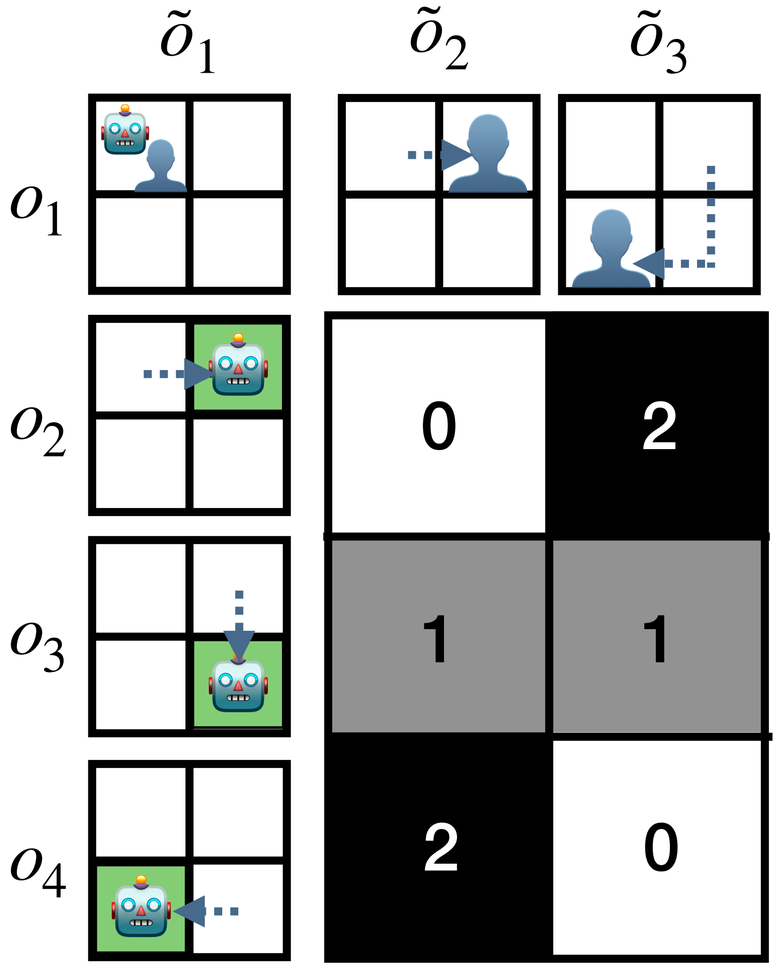
Frame-level distance matrix

Soft occupancy probability **P**

ORdered Coverage Alignment (**ORCA**)



ORdered Coverage Alignment (**ORCA**)

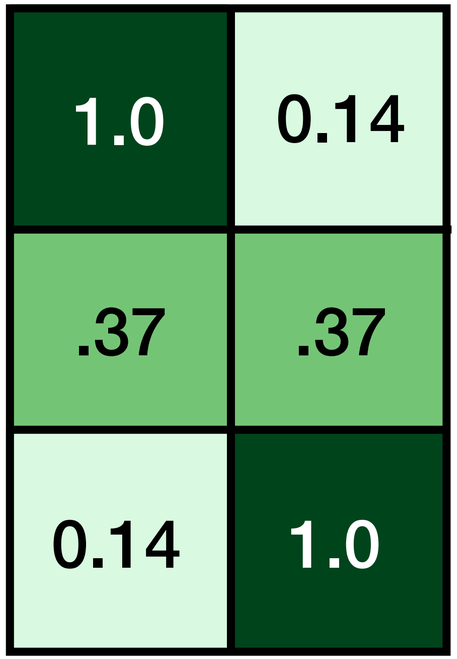


Frame-level distance matrix

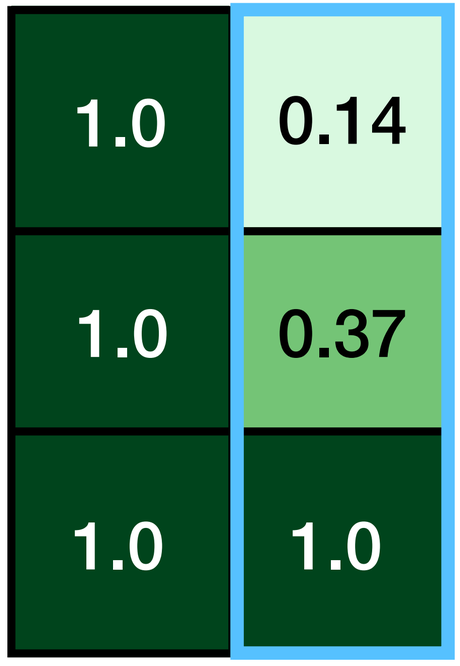
$$P_{t,j} = \exp(-\lambda d(o_t, \tilde{o}_j))$$

$$C_{t,j} = \max\{C_{t-1,j}, C_{t,j-1}P_{t,j}\}$$

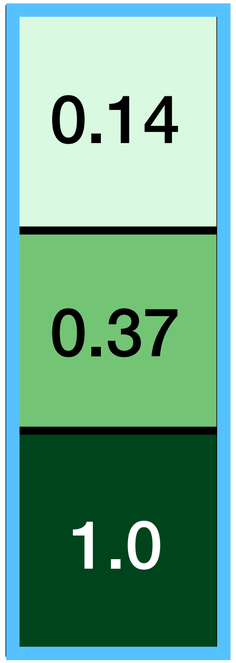
$$R_{ORCA} = C_{t,\tilde{T}-1}P_{t,\tilde{T}}$$



Soft occupancy probability **P**

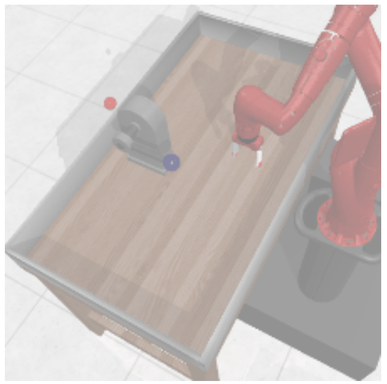


Ordered coverage matrix **C**

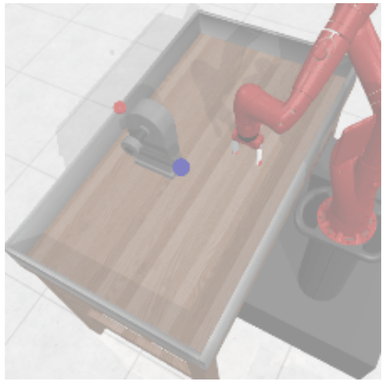


ORCA reward

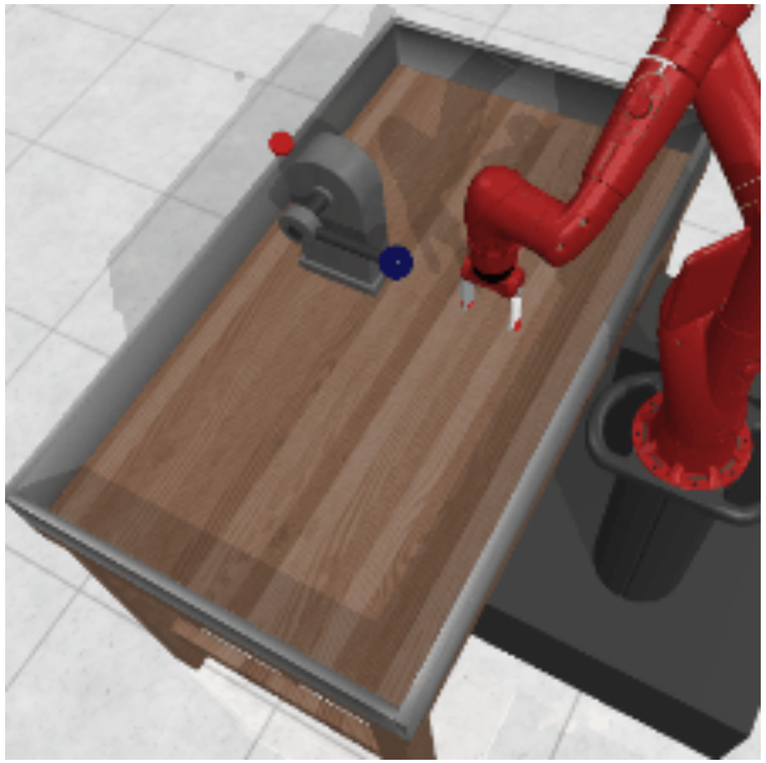
ORCA completes tasks efficiently and effectively



OT: ❌ subgoal order



TOT: ❌ subgoal coverage



ORCA: ✅ subgoal order ✅ subgoal coverage

*See our paper for proofs that ORCA enforces ordering and coverage

ORCA achieves better performance on temporally misaligned demonstrations

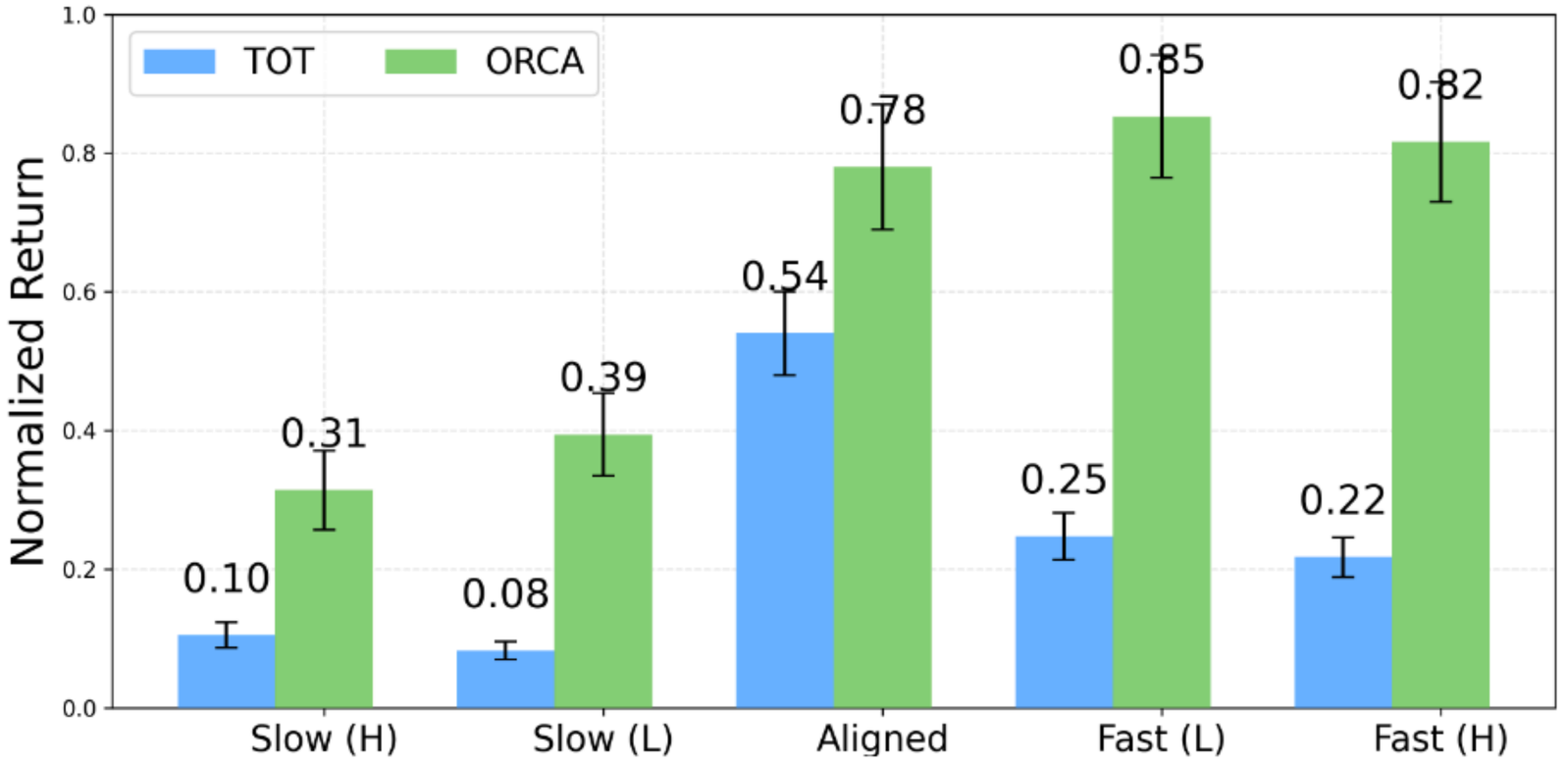
Category	Environment	Threshold	DTW	OT	TemporalOT	ORCA (NP)	**ORCA**
Easy	Button-press	0.30 (0.10)	0.00 (0.00)	0.00 (0.00)	0.10 (0.02)	0.45 (0.11)	**0.62 (0.11)**
	Door-close	0.34 (0.07)	0.00 (0.00)	0.00 (0.00)	0.19 (0.01)	0.86 (0.01)	**0.88 (0.01)**
Medium	Door-open	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.08 (0.01)	**1.60 (0.09)**	0.89 (0.13)
	Window-open	0.72 (0.14)	0.00 (0.00)	0.19 (0.06)	0.26 (0.05)	**0.86 (0.17)**	**0.85 (0.16)**
	Lever-pull	0.07 (0.02)	0.00 (0.00)	0.00 (0.00)	0.07 (0.03)	**0.27 (0.08)**	**0.28 (0.09)**
	Hand-insert	0.00 (0.00)	0.00 (0.00)	0.03 (0.02)	0.00 (0.00)	**0.08 (0.08)**	**0.04 (0.04)**
	Push	0.07 (0.05)	0.00 (0.00)	0.03 (0.01)	0.01 (0.01)	0.02 (0.02)	0.00 (0.00)
Hard	Basketball	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	**0.07 (0.03)**	0.01 (0.00)
	Stick-push	0.12 (0.04)	0.00 (0.00)	0.07 (0.02)	0.36 (0.00)	0.46 (0.13)	**1.25 (0.04)**
	Door-lock	0.00 (0.00)	0.05 (0.02)	0.04 (0.02)	0.00 (0.00)	**0.23 (0.09)**	**0.19 (0.08)**
	Average	0.16 (0.02)	0.01 (0.00)	0.04 (0.01)	0.11 (0.01)	**0.49 (0.04)**	**0.50 (0.04)**

(Top) Meta-world tasks

(Right) Humanoid tasks

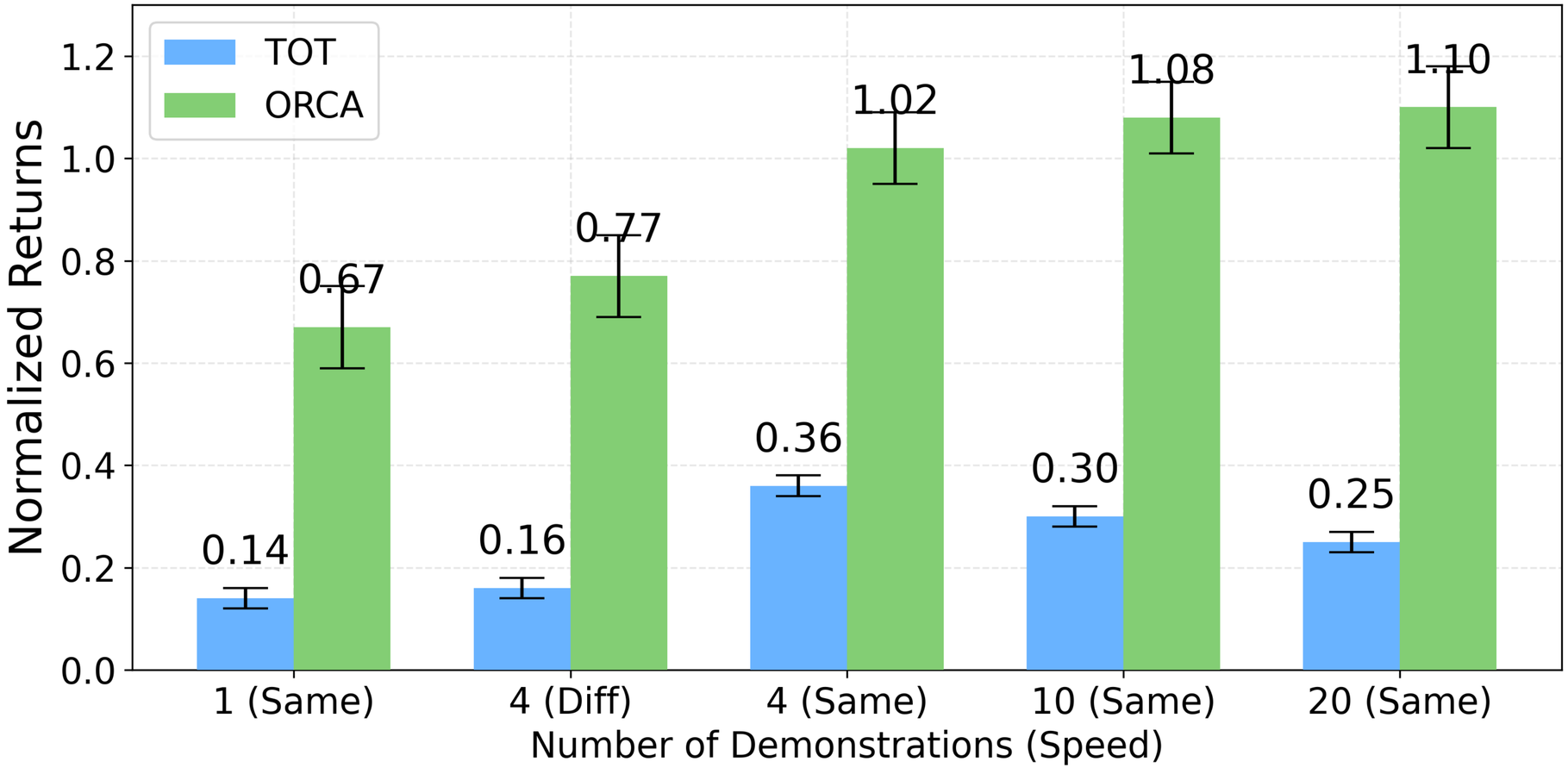
Task	TOT	ORCA (NP)	**ORCA**
Arm up (L)	5.29 (2.22)	65.9 (8.25)	**81.6 (3.65)**
Arm up (R)	7.67 (2.88)	**92.5 (4.71)**	49.6 (5.00)
Arms out	1.62 (0.75)	**72.7 (10.1)**	8.50 (2.60)
Arms down	11.6 (3.56)	19.7 (5.03)	**33.4 (7.20)**
Average	6.55 (2.35)	**62.9 (7.02)**	43.3 (4.61)

ORCA achieves better performance on temporally misaligned demonstrations



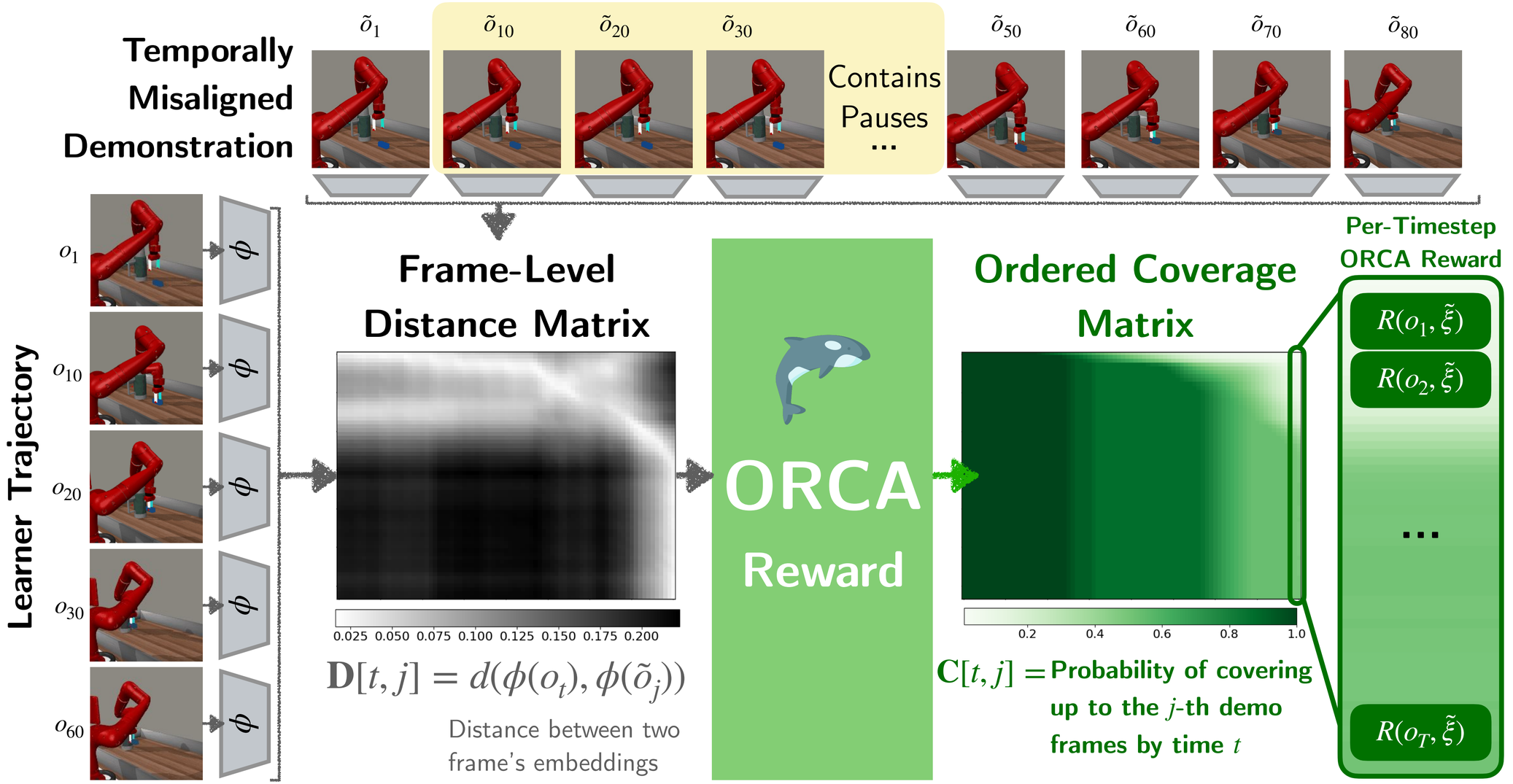
ORCA beats baselines given faster, slower, and same-speed demonstrations

ORCA scales with more demonstrations



ORCA beats baselines given multiple demonstrations, whether they are the same speed or different speeds

ORCA is a principled reward function for imitation learning from a **temporally misaligned video**



Thank You!




William Huey*, Huaxiaoyue (Yuki) Wang*, Anne Wu, Yoav Artzi, Sanjiban Choudhury

{wph52, hw575}@cornell.edu, willhuey.com, lunay0yuki.github.io