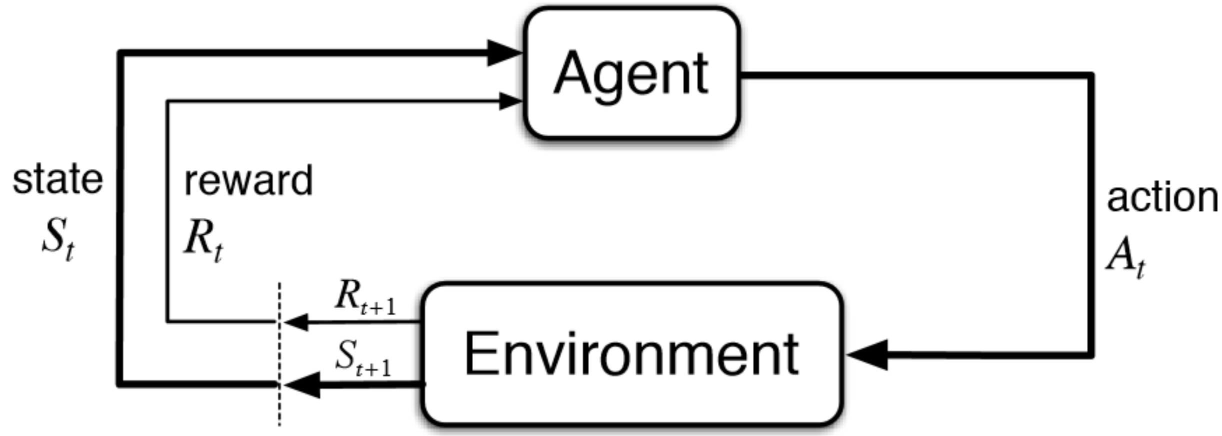
Dual RL: New Methods for Reinforcement and Imitation Learning

Amy Zhang

Assistant Professor, UT Austin

A Refresher on Reinforcement Learning



State-action/ State visitation distribution:

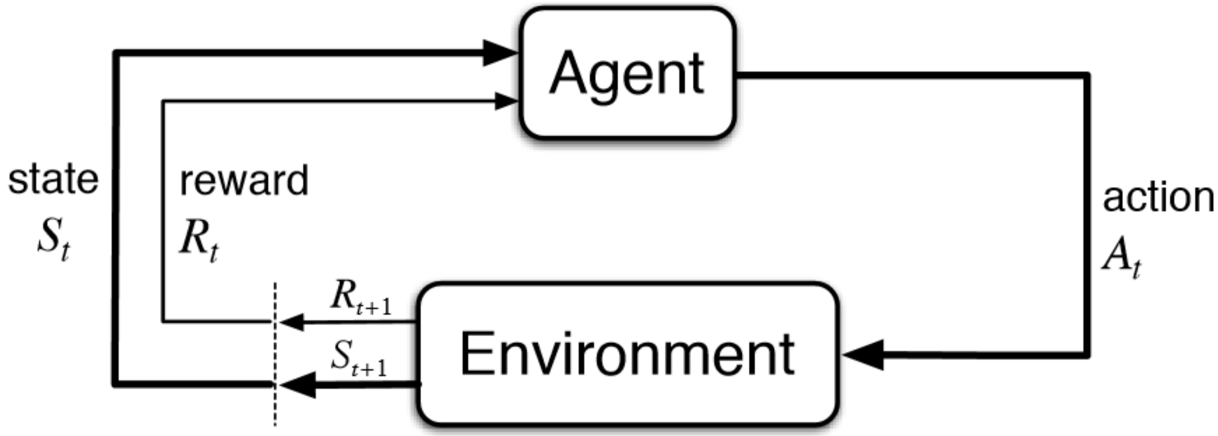
$$d^\pi(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s, a_t = a | s_0 \sim \rho_0, a_t \sim \pi(s_t), s_{t+1} \sim p(s_t, a_t)).$$

d^O : offline data visitation

Regularized optimization objective:

$$\max_\pi \mathbb{E}_{d^\pi(s,a)}[r(s,a)] - \alpha D_f(d^\pi(s,a) || d^o(s,a))$$

A Refresher on Reinforcement Learning



$$Q^*(s,a) = r(s,a) + \gamma \max_{a'} Q^*(s'(s,a),a')$$

$$V^*(s) = \max_{a} r(s,a) + \gamma V^*(s'(s,a))$$

Reinforcement Learning as a Linear Program

We can rewrite the problem as a convex optimization problem (CoP):

Primal-Q: $\max_{\pi, d \geq 0} \mathbb{E}_{d(s,a)}[r(s,a)] - D_f(d(s,a) || d^O(s,a))$

s.t. $d(s,a) = \underline{(1-\gamma) d_0(s) \pi(a|s) + \gamma \sum_{s',a'} d(s',a') p(s|s',a')} \pi(a|s)$

Constrains the visitation distribution *d* to be valid
Called *Bellman Flow constraint*

Manne, 1960

Lagrangian Dual without Constraints

- Define operator

$$T^{\pi_Q} Q(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim p, a_{t+1} \sim \pi_Q}[\gamma Q(s_{t+1}, a_{t+1})]$$

Dual-Q: $\max_{\pi} \min_{Q} (1-\gamma)\mathbb{E}_{d_0(s), \pi(a|s)}[Q(s,a)] + \alpha \mathbb{E}_{s,a \sim d^O}[f^*([T^{\pi} Q(s,a) - Q(s,a)]/\alpha)]$

where f^* is the convex conjugate of f

Dual-Q is overconstrained

Primal-V : $\max_{d \geq 0} \mathbb{E}_{d(s,a)}[r(s,a)] - \alpha D_f(d(s,a)||d^O(s,a))$

s.t. $\sum_{a \in \mathcal{A}} d(s,a) = (1-\gamma)d_0(s) + \gamma \sum_{s',a'} d(s',a')p(s|s',a')$

- Define operator:

$$TV(s_t, a_t) = r(s_t, a_t) + \mathbb{E}_{s_{t+1} \sim p(s_t, a_t)}[V(s_{t+1})]$$

dual-V: $\min_V (1-\gamma)\mathbb{E}_{d_0(s)}[V(s)] + \underline{\alpha \mathbb{E}_{s,a \sim d^O}[f^*([TV(s,a) - V(s)]/\alpha)]}$

Single-player non-adversarial optimization

Gives rise to RElaxed Coverage for Off-policy Imitation Learning (ReCOIL): Imitation from arbitrary experience

- Consider the f-divergence between the mixture distributions:

$$D_f(\underbrace{\beta d(s,a) + (1-\beta) d^S(s,a)}_{d^S_{mix}} \;||\; \underbrace{\beta d^E(s,a) + (1-\beta) d^S(s,a)}_{d^{E,S}_{mix}})$$

- Is a valid imitation learning objective: shares the same global minima as traditional objective ($d = d^E$)

- Avoids estimating $\frac{d^S(s,a)}{d^E(s,a)}$ which is ill-defined in state-action space with zero expert support.

d^E: expert data visitation, d^S suboptimal data visitation

ReCOIL: Imitation from arbitrary experience

- Primal-V with the mixture distributions:

Primal-V: $$\max_{d \geq 0} -D_f(d^S_{mix}(s,a) || d^{E,S}_{mix}(s,a))$$

s.t. $\sum_{a \in \mathcal{A}} d(s,a) = (1-\gamma) d_0(s) + \gamma \sum_{s',a'} d(s',a') p(s|s',a')$

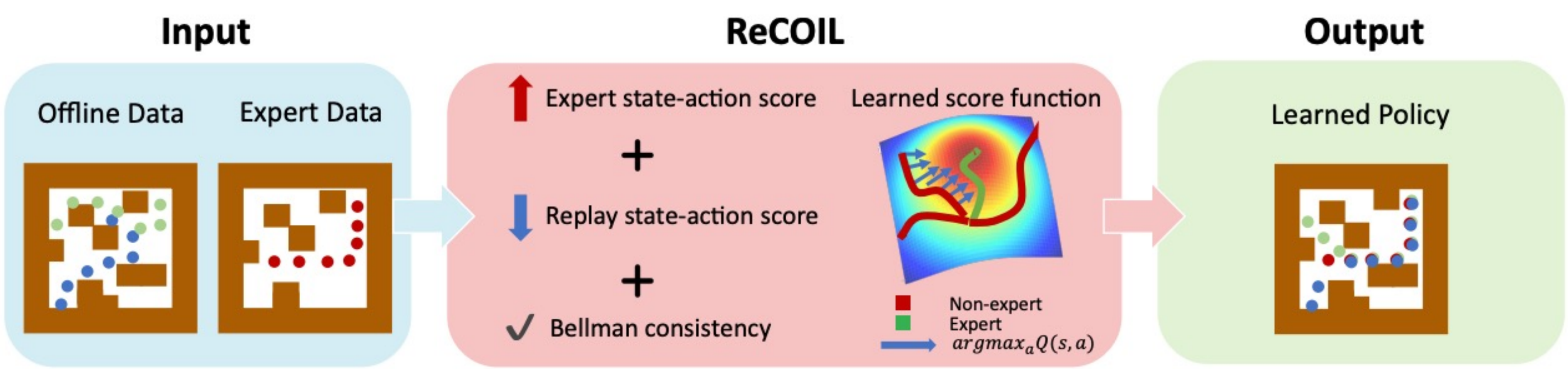
- Dual for Primal-V with mixture distributions:

ReCOIL-V

$$\min_V \beta(1-\gamma)\mathbb{E}_{d_0(s)}[V(s)] + \alpha \mathbb{E}_{s,a \sim d^{E,S}_{mix}}[f^*(TV(s,a) - V(s))] - (1-\beta)\,\mathbb{E}_{s,a\sim d^S}[TV(s,a) - V(s)]$$

What is ReCOIL doing behind the scenes?| Intuition

ReCOIL is just a Bellman-Consistent EBM.



ReCOIL: Imitation from arbitrary experience

ReCOIL-Q

$$\max_{\pi(a|s)} \min_{Q(s,a)} \beta(1-\gamma)\mathbb{E}_{d_0(s),\pi(a|s)}[Q(s,a)] + \mathbb{E}_{s,a\sim d^{E,R}_{mix}}\left[f_p^*(\mathcal{T}_0^\pi Q(s,a) - Q(s,a))\right]$$
$$- (1-\beta)\mathbb{E}_{s,a\sim d^R}[\mathcal{T}_0^\pi Q(s,a) - Q(s,a)]$$

ReCOIL-V

$$\min_{V(s)} \beta(1-\gamma)\mathbb{E}_{d_0(s)}[V(s)] + \mathbb{E}_{s,a\sim d^{E,R}_{mix}}\left[f_p^*(\mathcal{T}_0 V(s,a) - V(s))\right]$$
$$- (1-\beta)\mathbb{E}_{s,a\sim d^R}[\mathcal{T}_0 V(s,a) - V(s)]$$

Key features:

✓ Non-adversarial
✓ ReCOIL-V is a single player optimization instead of a game.
✓ Does not require learning a discriminator.
✓ Relaxes the coverage assumption
✓ Works for arbitrary f-divergence

Offline IL experiments: ReCOIL

Locomotion

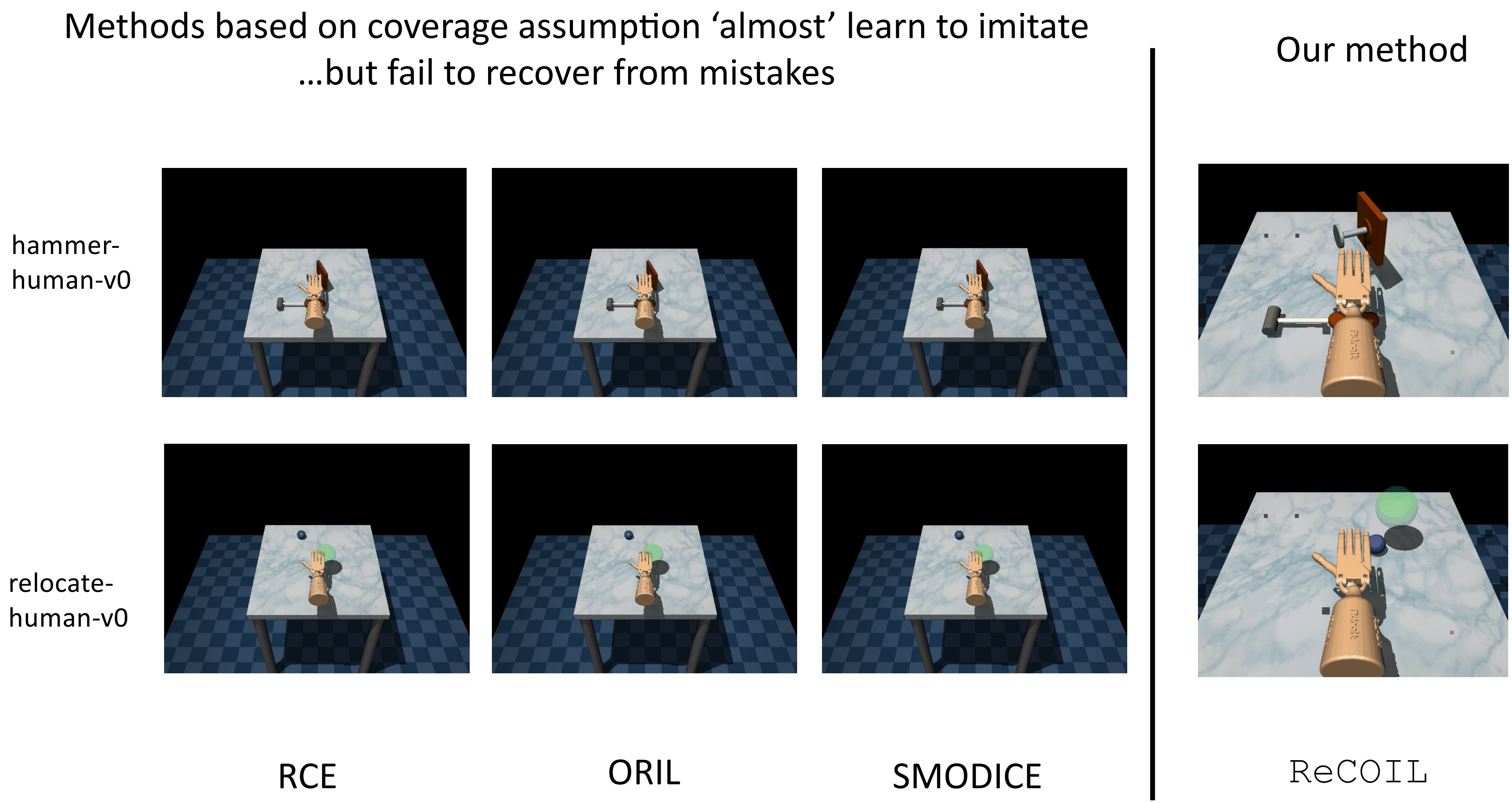
Manipulation

Suboptimal Dataset	Env	RCE	ORIL	SMODICE	ReCOIL
random+ expert	hopper	51.41±38.63	73.93±11.06	101.61±7.69	**108.18±3.28**
	halfcheetah	64.19±11.06	60.49±3.53	80.16±7.30	**80.20±6.61**
	walker2d	20.90±26.80	2.86±3.39	**105.86±3.47**	102.16±7.19
	ant	105.38±14.15	73.67±12.69	**126.78±5.12**	**126.74±4.63**
random+ few-expert	hopper	25.31±18.97	42.04±13.76	60.11±18.28	**97.85±17.89**
	halfcheetah	2.99±1.07	2.84±5.52	2.28±0.62	**76.92±7.53**
	walker2d	40.49±26.52	3.22±3.29	**107.18±1.87**	83.23±19.00
	ant	**67.62±15.81**	25.41 ± 8.58	-6.10±7.85	**67.14± 8.30**
medium+ expert	hopper	58.71±34.06	61.68±7.61	49.74±3.62	**88.51±16.73**
	halfcheetah	65.14±13.82	54.66±0.88	59.50±0.82	**81.15±2.84**
	walker2d	96.24±14.04	8.19±7.70	2.62±0.93	**108.54±1.81**
	ant	86.14±38.59	102.74±6.63	104.95±6.43	**120.36±7.67**
medium few-expert	hopper	66.15±35.16	17.40±15.15	47.61±7.08	**50.01±10.36**
	halfcheetah	61.14±18.31	43.24±0.75	46.45±3.12	**75.96±4.54**
	walker2d	85.28±34.90	6.81±6.76	6.00±6.69	**91.25±17.63**
	ant	67.95±36.78	81.53±8.618	81.53±8.618	**110.38±10.96**
cloned+expert	pen	19.60±11.40	-3.10±0.40	-3.36±0.71	**95.04±4.48**
	door	0.08± 0.15	-0.33±0.01	0.25± 0.54	**102.75±4.05**
	hammer	1.95±3.89	0.25± 0.01	0.15± 0.078	**95.77±17.90**
	relocate	-0.25±0.04	-0.29±0.01	1.75±3.85	**67.43±14.60**
human+expert	pen	17.81±5.91	-3.38±2.29	-2.20±2.40	**103.72±2.90**
	door	-0.05±0.05	-0.33±0.01	-0.20± 0.11	**104.70±0.55**
	hammer	5.00±5.64	1.89±0.70	-0.07±0.39	**125.19±3.29**
	relocate	0.02±0.10	-0.29±0.01	-0.16±0.04	**91.98± 2.89**
partial+expert	kitchen	6.875±9.24	0.00±0.00	39.16± 1.17	**60.0±5.70**
mixed+expert	kitchen	1.66±2.35	0.00±0.00	42.5±2.04	**52.0±1.0**

Table 2: The normalized return obtained by different offline IL methods trained on the D4RL suboptimal datasets with 1000 expert transitions.

Methods based on coverage assumption **fail when coverage is low** (few expert trajectories in dataset),
and
in **high dimensional tasks** where the Discriminator easily overfits.

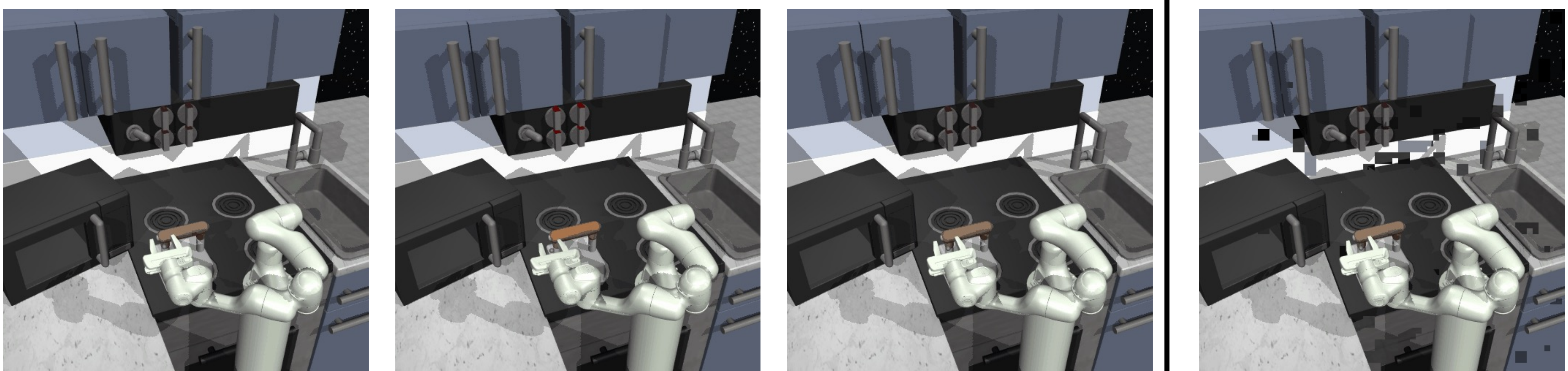
ReCOIL outperforms baselines by a large margin!


Methods based on coverage assumption 'almost' learn to imitate
...but fail to recover from mistakes
Our method
hammer-human-v0
relocate-human-v0
RCE
ORIL
SMODICE
ReCOIL

Methods based on coverage assumption ‘almost’ learn to imitate
…but fail to recover from mistakes

Our method

Environment: Kitchen-partial-v0



RCE ORIL SMODICE ReCOIL

Dual Formulation for Self-Supervised Pre-training

Key Idea: Learning from Human Videos as a *BIG Offline Goal-Conditioned RL* Problem

Offline Dataset:
Diverse Human Videos

$$\max_{\pi_H, \phi} \mathbb{E}_{\pi^H}\left[\sum_t \gamma^t r(o;g)\right] - D_{\mathrm{KL}}(d^{\pi_H}(o, a^H; g) \| d^D(o, \tilde{a}^H; g)),$$

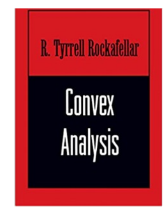
- Mathematically Sound
- What are human actions?
- Can’t be optimized in practice

Human videos are rich sources of goal-directed behavior!

Offline *Value* Learning on Human Videos

Offline Dataset:
Diverse Human Videos

$$\max_{\pi_H, \phi} \mathbb{E}_{\pi^H}\left[\sum_t \gamma^t r(o; g)\right] - D_{\mathrm{KL}}(d^{\pi_H}(o, a^H; g) \| d^D(o, \tilde{a}^H; g)),$$

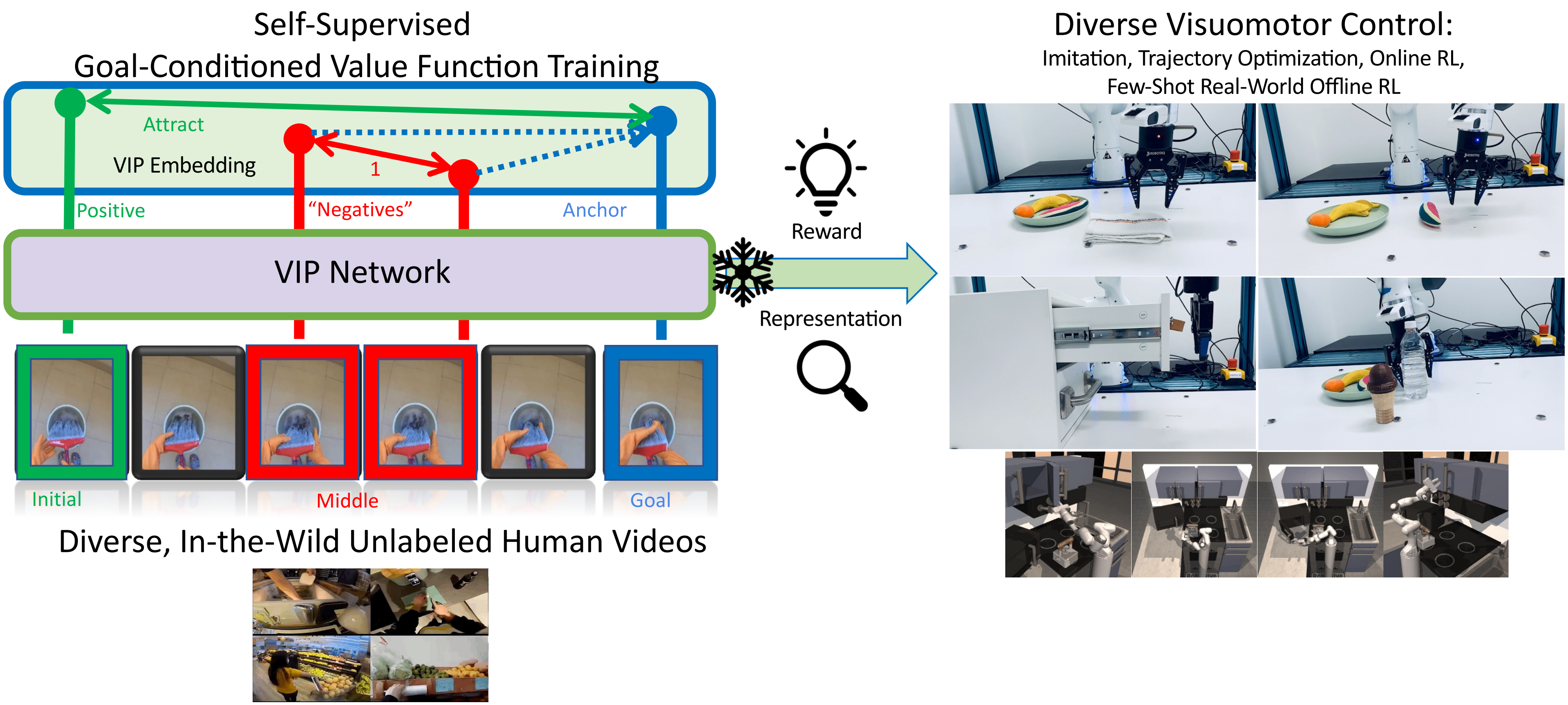


Dual Problem depends only on offline data! No dependence on actions!

$$\max_{\phi} \min_{V} \mathbb{E}_{p(g)}\left[(1-\gamma)\mathbb{E}_{\mu_0(o;g)}[V(\phi(o);\phi(g))] + \log \mathbb{E}_{(o,o';g)\sim D}\left[\exp\left(r(o,g) + \gamma V(\phi(o');\phi(g)) - V(\phi(o),\phi(g))\right)\right]\right]$$

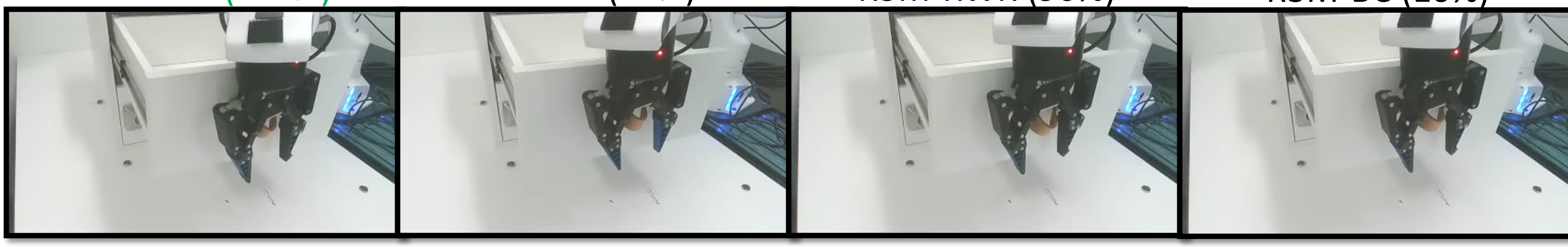
goal frame — initial frame — middle frame

VIP: Towards Universal Visual Reward and Representation Via **V**alue-**I**mplicit **P**re-Training

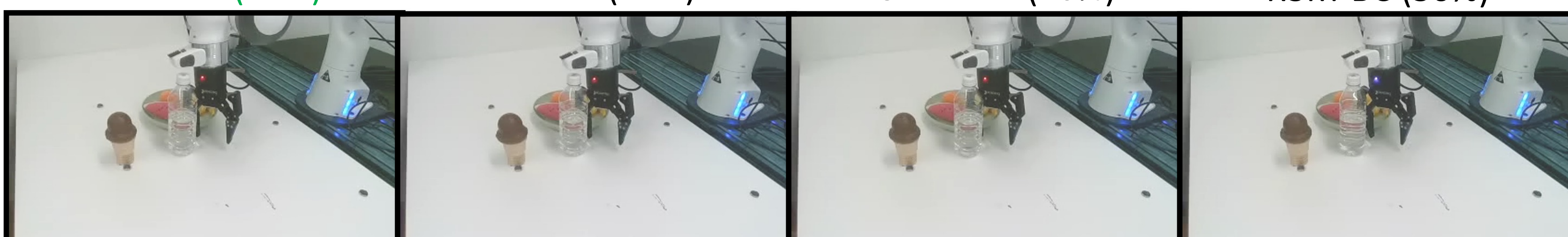


CloseDrawer & PushBottle

VIP-RWR (100%) | VIP-BC (50%) | R3M-RWR (90%) | R3M-BC (10%)

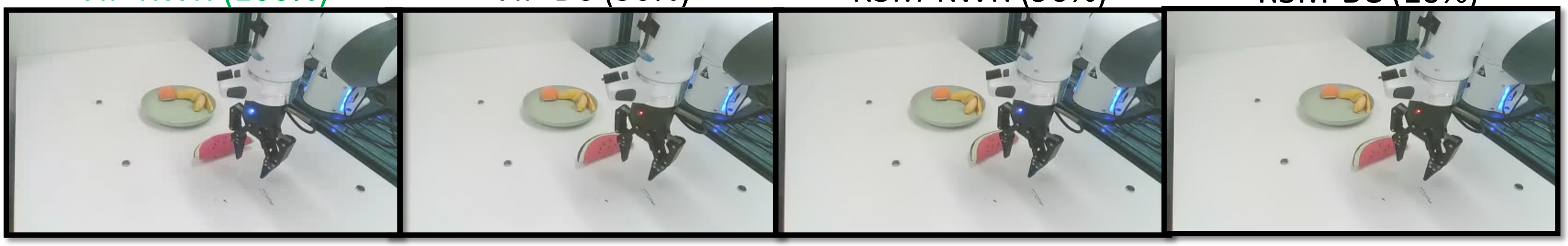


VIP-RWR (90%) | VIP-BC (50%) | R3M-RWR (70%) | R3M-BC (50%)

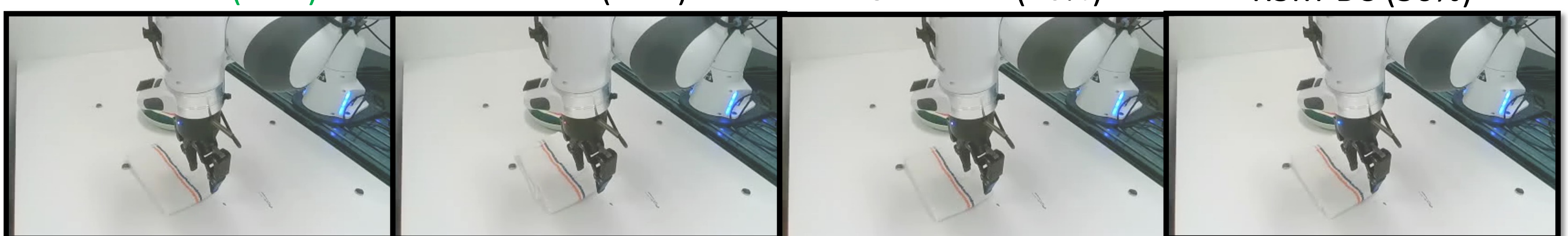


PickPlaceMelon & FoldTowel

VIP-RWR (100%) | VIP-BC (50%) | R3M-RWR (90%) | R3M-BC (10%)

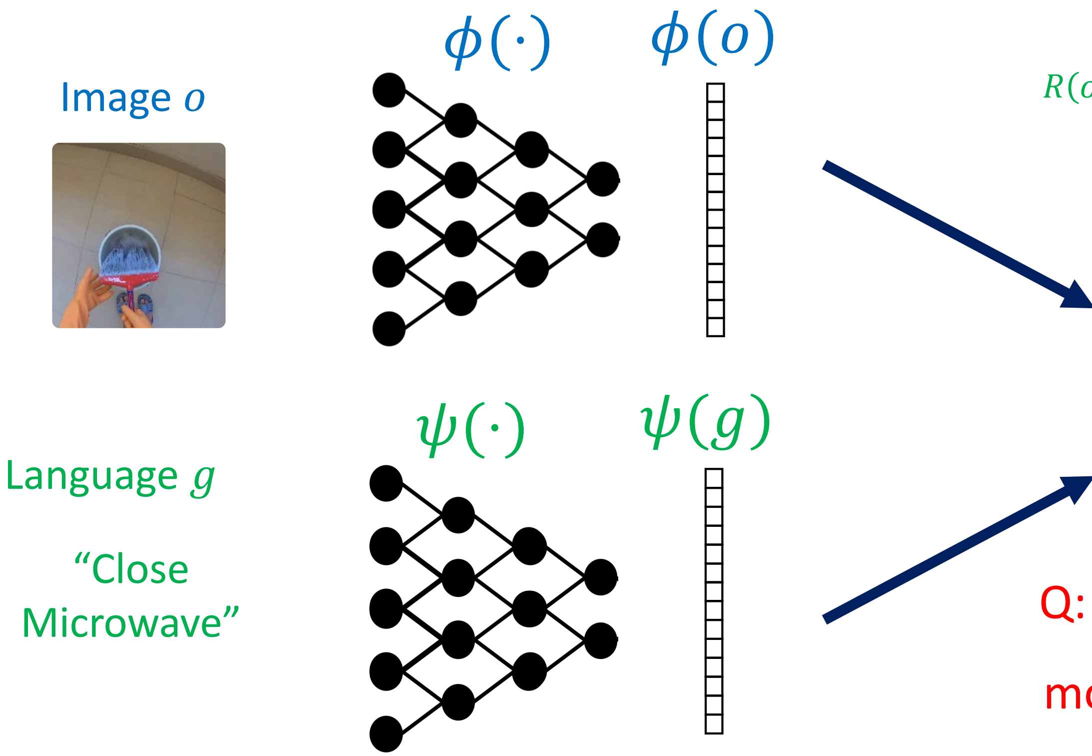


VIP-RWR (90%) | VIP-BC (50%) | R3M-RWR (70%) | R3M-BC (50%)



Extended to Multimodal Settings

This Work : Representations as Multi-Modal Value Functions



Shaped Dense Reward:
$R(o_{t+1}, o_t; g) \coloneqq V^*(\phi(o_{t+1}); \psi(g)) - V^*(\phi(o_t); \psi(g))$

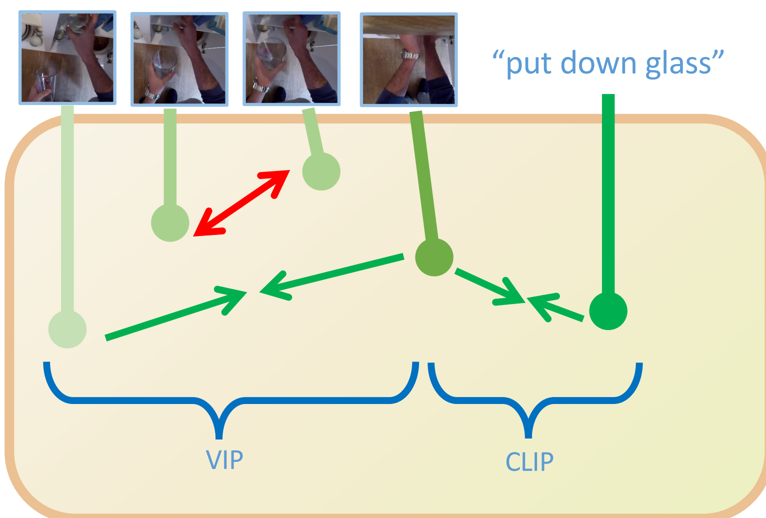
$V^*(\cdot; g)$	$V^*(o; g)$
Embedding Distance	0.6

Q: How can we effectively learn multi-modal implicit value representations?

Language-Image Value Learning (LIV)

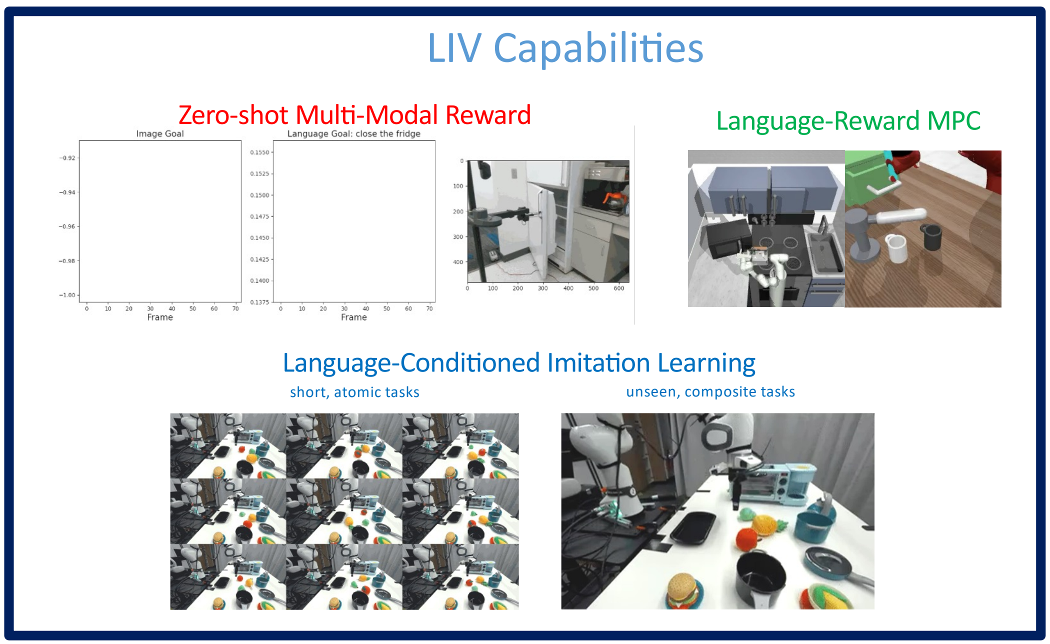
- Theory: Combining VIP and CLIP objectives amounts to learning a multi-modal value representation compatible with image and language goals

LIV Objective



Language-Image Value Learning (LIV) Applications

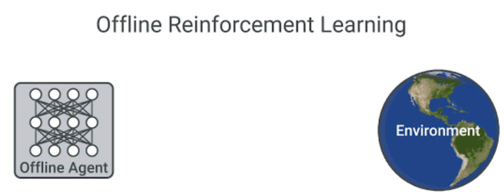
- SOTA results on pre-training, fine-tuning, and reward learning for language-conditioned robotic manipulation



LIV performs best on all RealRobot tasks

Closing: Unifying existing work with dual RL

- Most successful techniques for offline RL:



Policy Constraint

Ensure policy during training stays close to replay buffer policy

Pessimistic Value function

Ensure underestimation of Q for OOD actions in dataset

Implicit Maximization

$$Q(s,a) \coloneqq r(s,a) + \gamma \max_{a'} Q(s',a')$$

Ensure that the maximum is taken over in-distribution actions

We show all these classes of prior methods come from a unified dual perspective!

Closing: Unifying existing work with dual RL

We show a number of prior methods in IL and RL to be dual RL methods! Some surprising ones are CQL, Implicit Behavior Cloning, XQL, IQLearn.

	Dual RL Method	Gradient	Objective	dual-Q/V	Non-Adversarial?	Off-Policy Data	Coverage Assumption
RL	AlgaeDICE [56], GenDICE [81], CQL [43]	semi	reg. RL	Q	✗	Arbitrary	—
	OptiDICE [45]	full	reg. RL	V	✓	Arbitrary	—
	XQL [23], REPS [61], f-**DVL**	semi	reg. RL	V	✓	Arbitrary	—
	VIP [49], GoFAR [50]	full	reg. RL	V	✓	Arbitrary	—
	Logistic Q-learning [6]	full	reg. RL	QV[1]	✓	✗	—
IL	IQLearn [22], IBC [15]	semi	$D_f(\rho^\pi \Vert \rho^E)$	Q	✓	Expert-only	✗
	IVLearn	semi	$D_f(\rho^\pi \Vert \rho^E)$	V	✓	Expert-only	✗
	OPOLO [82], OPIRL [32]	semi	$D_{rkl}(\rho^\pi \Vert \rho^E)$	Q	✗	Arbitrary	✓
	ValueDICE [40]	semi	$D_{rkl}(\rho^\pi \Vert \rho^E)$	Q	✗	Arbitrary	✓
	SMODICE [48]	full	$D_{rkl}(\rho^\pi \Vert \rho^E)$	V	✓	Arbitrary	✓
	DemoDICE [38], LobsDICE [37]	full	$D_{rkl}(\rho^\pi \Vert \rho^E) + \alpha D_{rkl}(\rho^\pi \Vert \rho^R)$	V	✓	Arbitrary	✓
	P^{2}IL [79]	full	$D_C(\rho^\pi \Vert \rho^E)$[1]	QV[1]	✗	✗	✗
	ReCOIL-Q	full	$D_f(\rho^\pi_{mix} \Vert \rho^{E,R}_{mix})$	Q	✗	Arbitrary	✗
	ReCOIL-V	full	$D_f(\rho^\pi_{mix} \Vert \rho^{E,R}_{mix})$	V	✓	Arbitrary	✗

To Summarize

- Dual RL provides a unifying perspective on imitation learning and regularized reinforcement learning
- Gives rise to new IL and RL algorithms
- Can also be leveraged for pre-training representations and reward functions in vision and multimodal settings

1. Dual RL: Unification and New Methods for Reinforcement and Imitation Learning. Harshit Sikchi, Qinqing Zheng, AZ, Scott Niekum. *In submission.*
2. VIP: Towards Universal Visual Reward and Representation via Value-Implicit Pre-Training, Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar*, AZ*. *ICLR 2023.*
3. LIV: Language-Image Representations and Rewards for Robotic Control. Yecheng Jason Ma · Vikash Kumar · AZ· Osbert Bastani · Dinesh Jayaraman. *ICML 2023.*