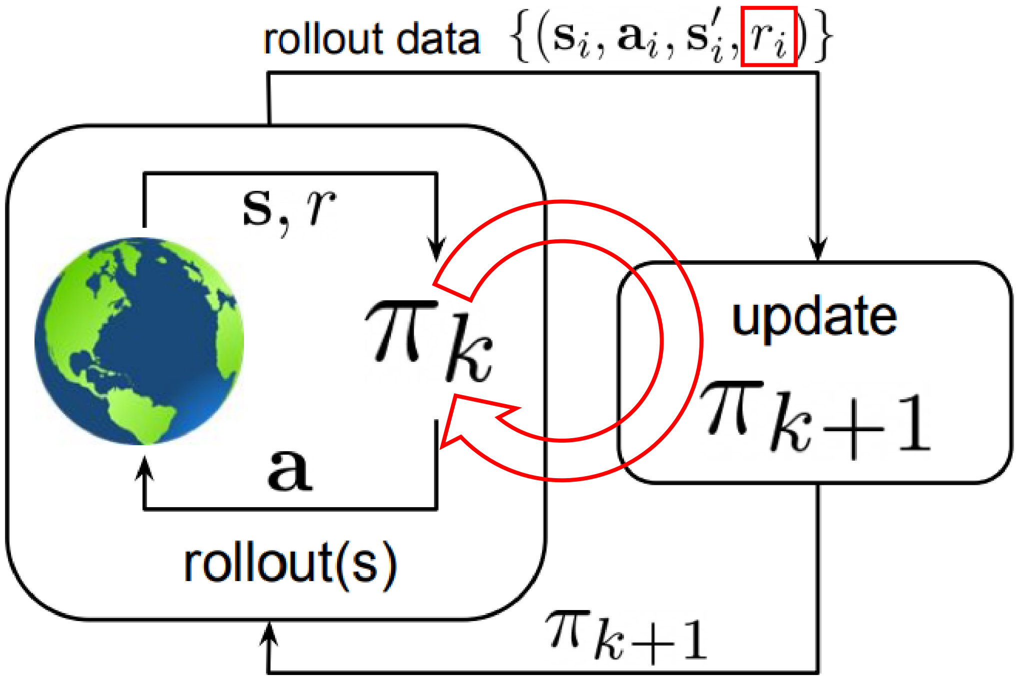
Beyond Reward: Offline Preference-guided Policy Optimization

Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, Donglin Wang
Machine Intelligence Lab (MiLAB), Westlake University & Zhejiang University

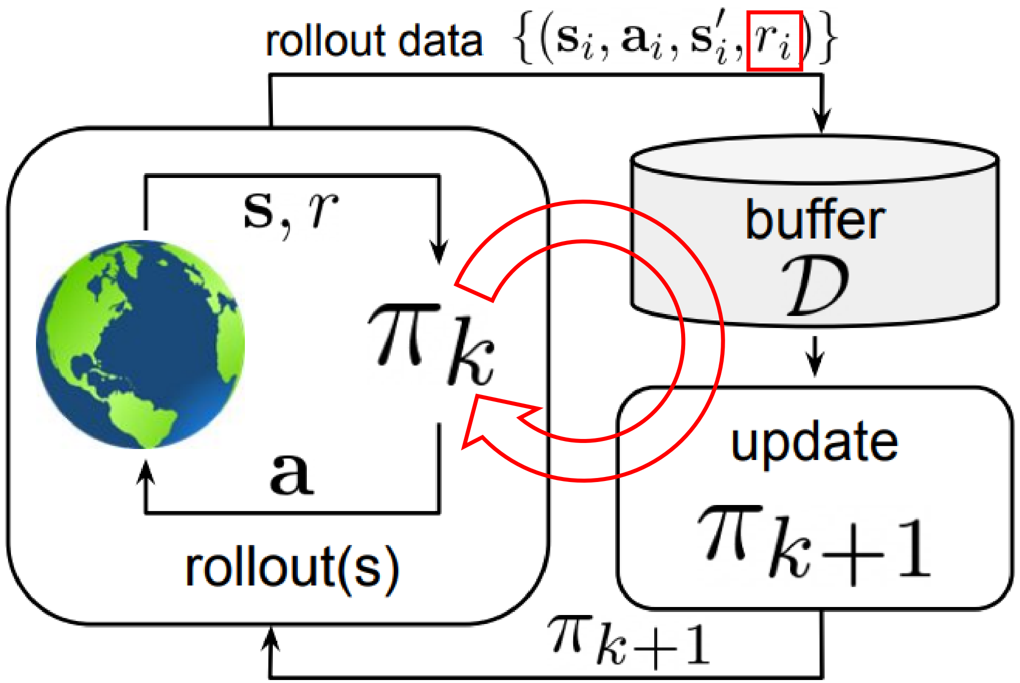
ICML 40 Years
International Conference On Machine Learning

Background

Online Reinforcement Learning



On-policy

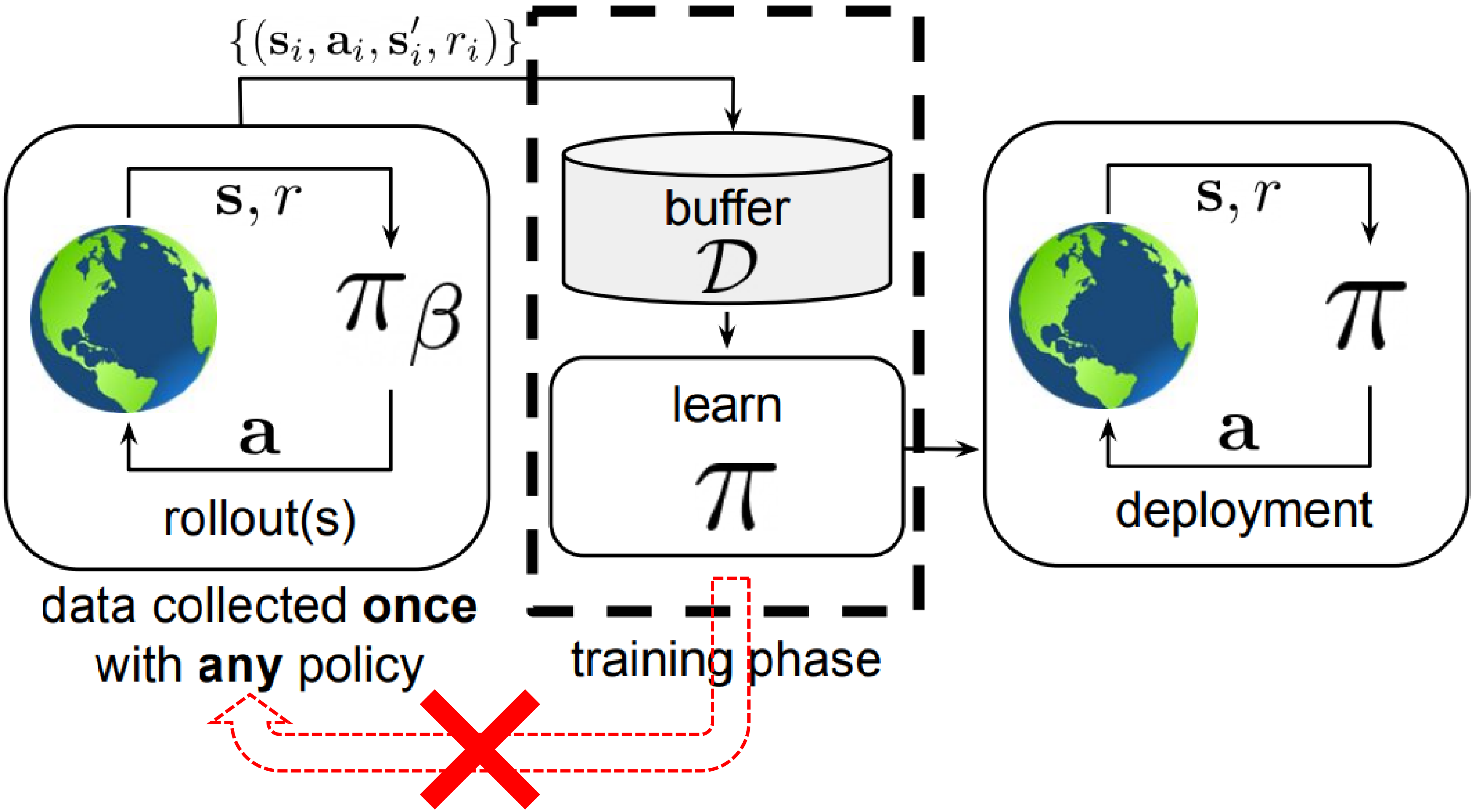


Off-policy

Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." arXiv preprint arXiv:2005.01643 (2020).

Background

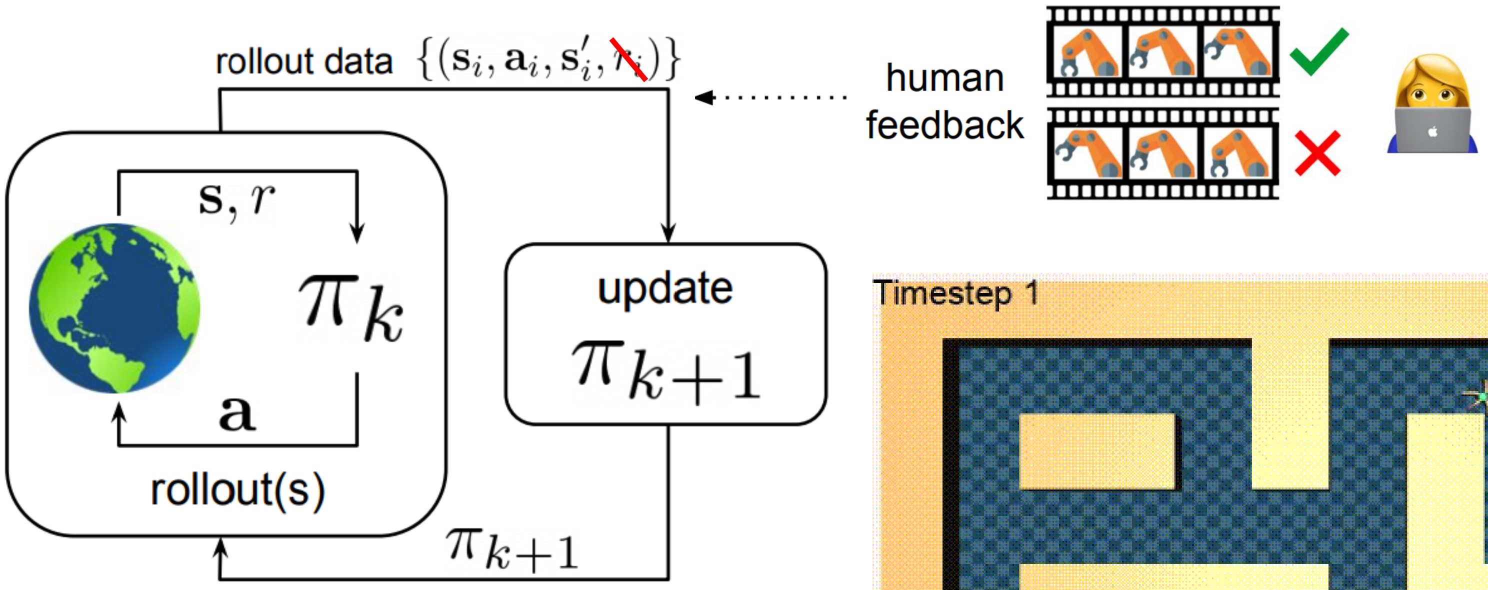
Offline Reinforcement Learning

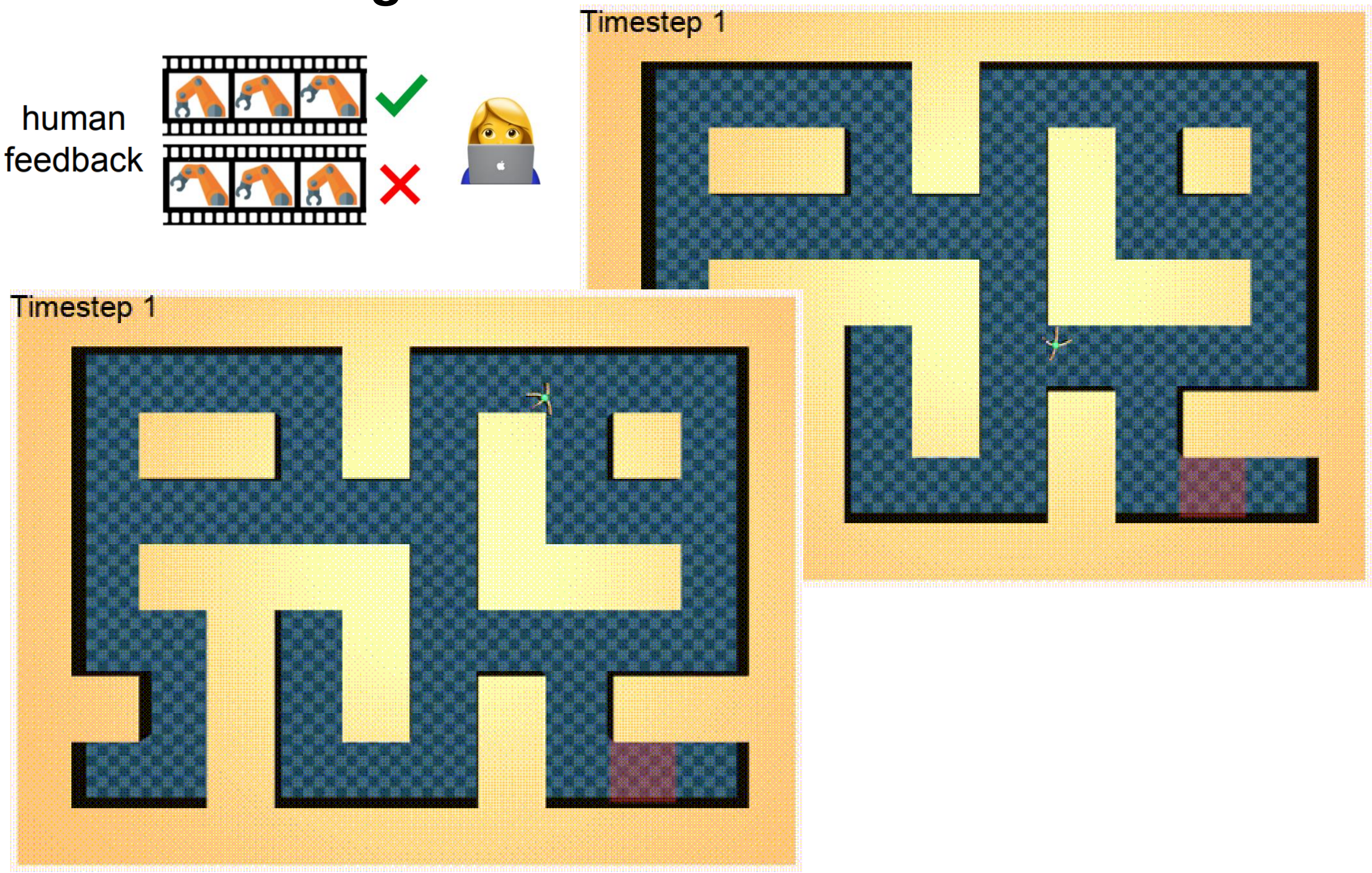


Levine, Sergey, et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems." arXiv preprint arXiv:2005.01643 (2020).

Background

Preference-based Reinforcement Learning





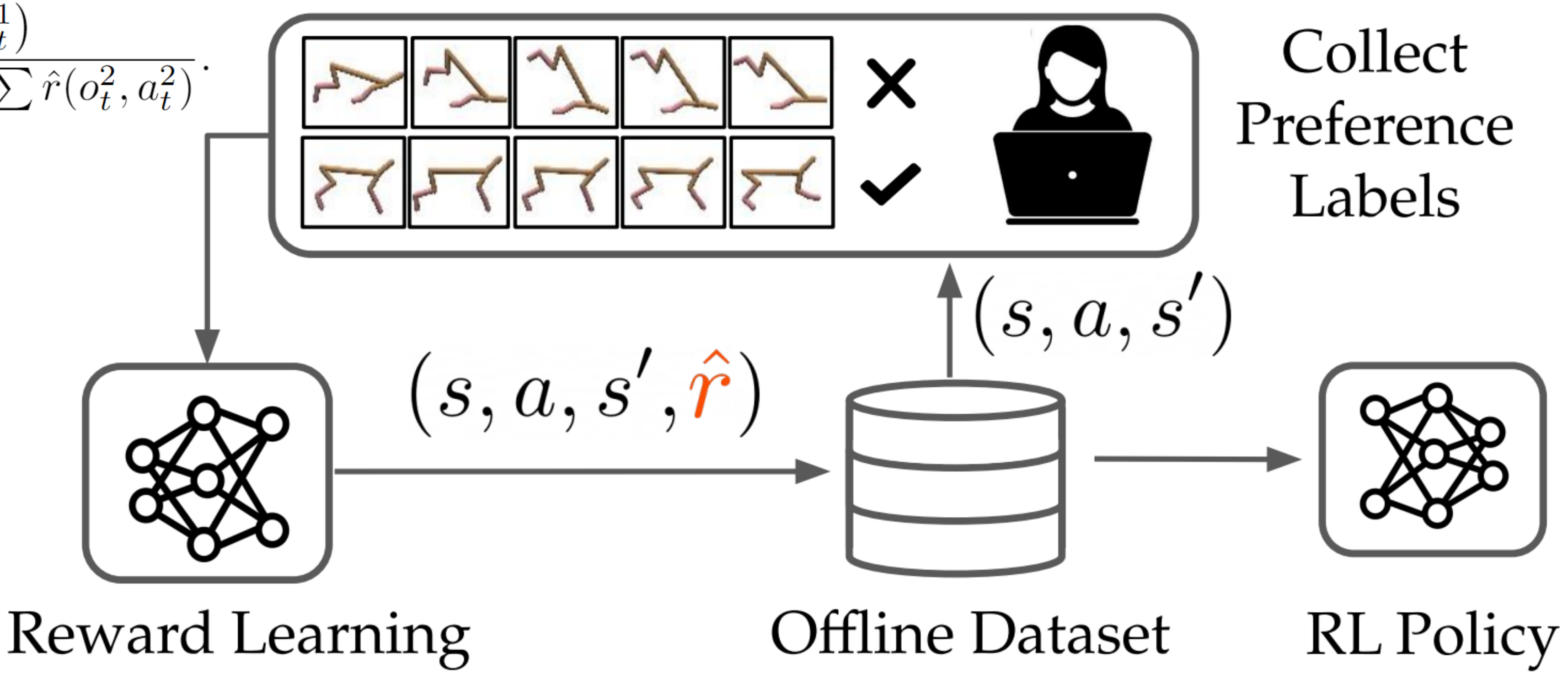
Christiano, Paul F., et al. "Deep reinforcement learning from human preferences." Advances in neural information processing systems 30 (2017).
Kim, Changyeon, et al. "Preference Transformer: Modeling Human Preferences using Transformers for RL." arXiv preprint arXiv:2303.00957 (2023).

Background

Offline Preference-based Reinforcement Learning

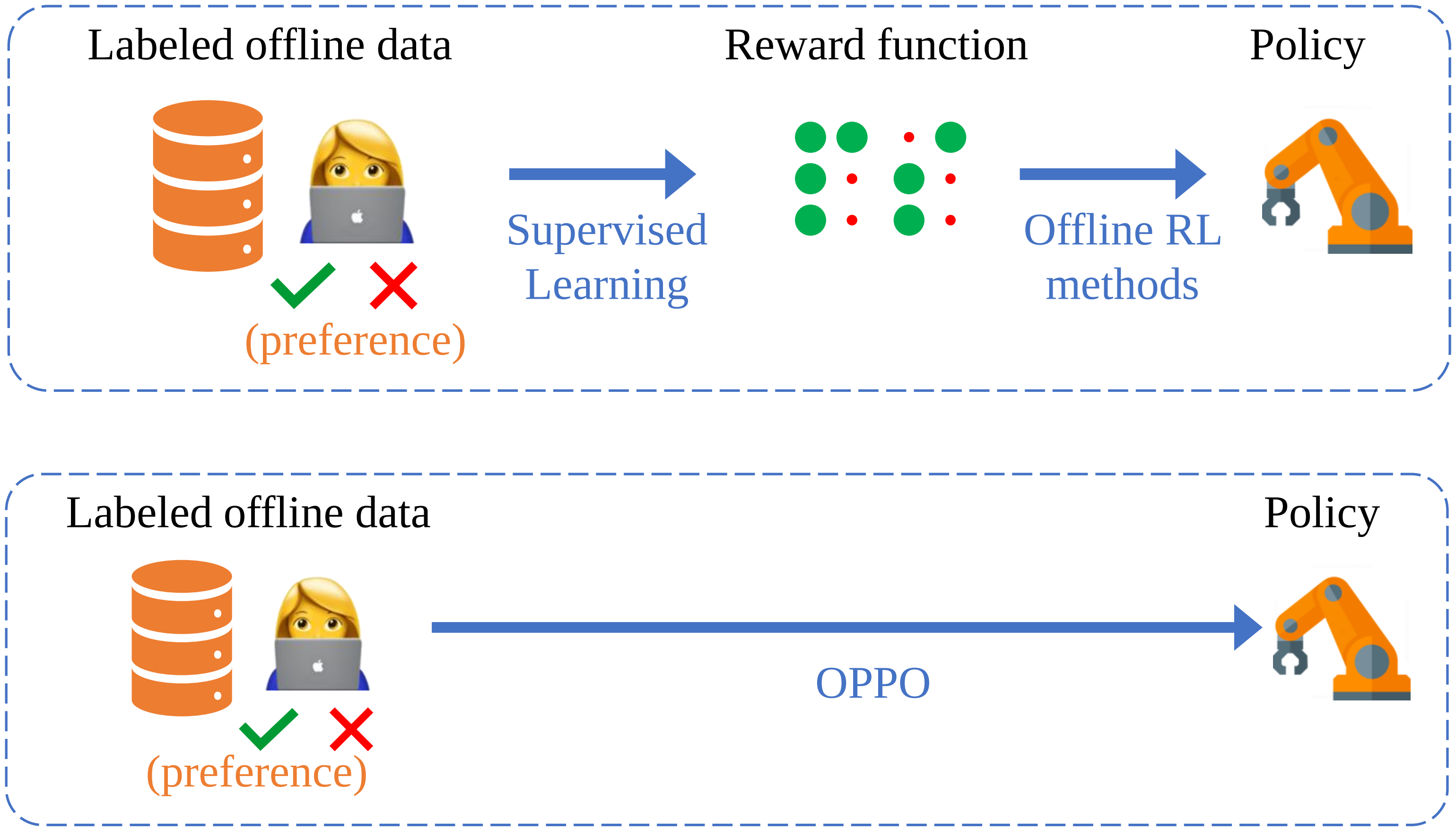
$$\text{loss}(\hat{r}) = -\sum_{(\sigma^1,\sigma^2,\mu)\in\mathcal{D}} \mu(1)\log \hat{P}[\sigma^1 \succ \sigma^2] + \mu(2)\log \hat{P}[\sigma^2 \succ \sigma^1].$$

$$\hat{P}[\sigma^1 \succ \sigma^2] = \frac{\exp \sum \hat{r}(o_t^1, a_t^1)}{\exp \sum \hat{r}(o_t^1, a_t^1) + \exp \sum \hat{r}(o_t^2, a_t^2)}.$$



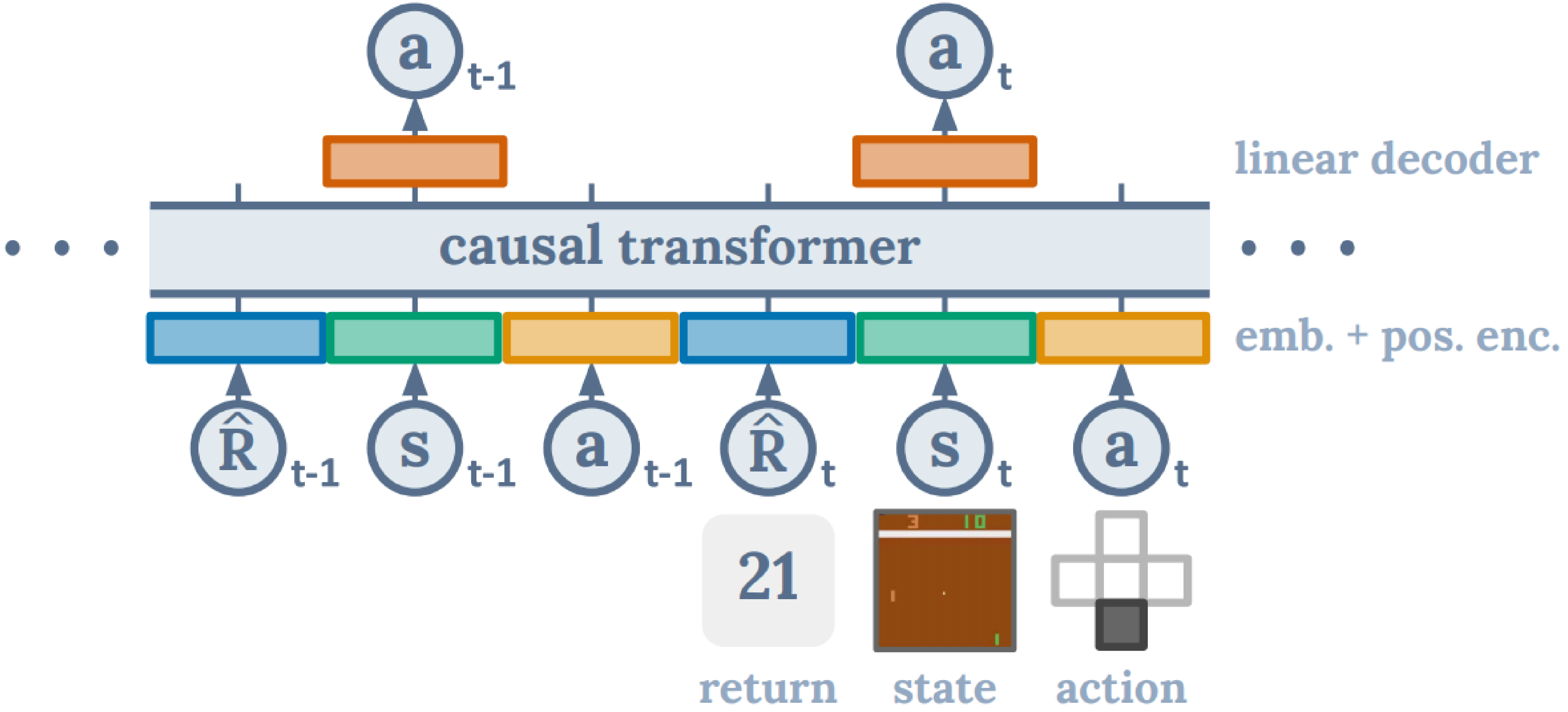
Shin, Daniel, Daniel S. Brown, and Anca D. Dragan. "Offline preference-based apprenticeship learning." arXiv preprint arXiv:2107.09251 (2021).

Offline Preference-guided Policy Optimization (OPPO)



Hindsight Information Matching

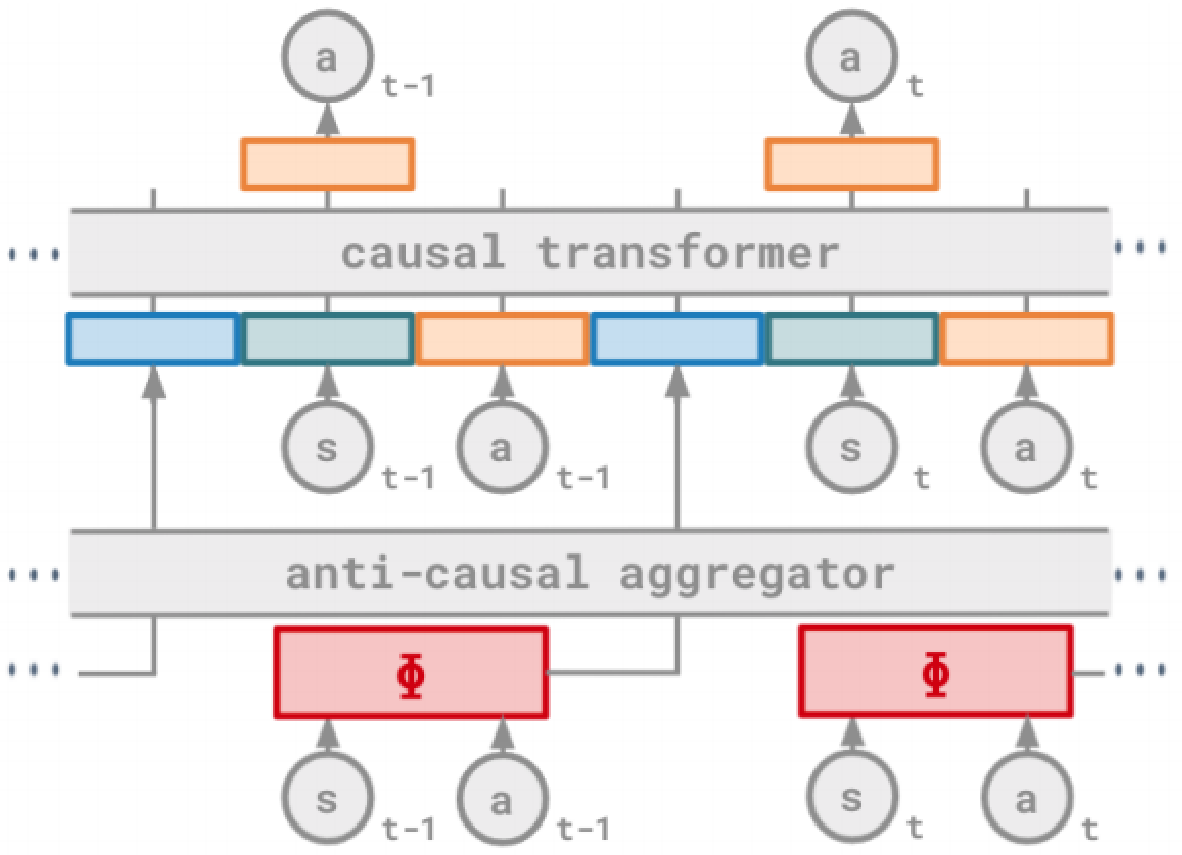
Decision transformer



Chen, Lili, et al. "Decision transformer: Reinforcement learning via sequence modeling." Advances in neural information processing systems 34 (2021): 15084-15097.

Hindsight Information Matching

Generalized Decision transformer



$$\min_{\pi} \mathbb{E}_{z\sim p(z),\tau\sim\rho_z^{\pi}(\tau)}\left[D(I^{\Phi}(\tau), z)\right]$$

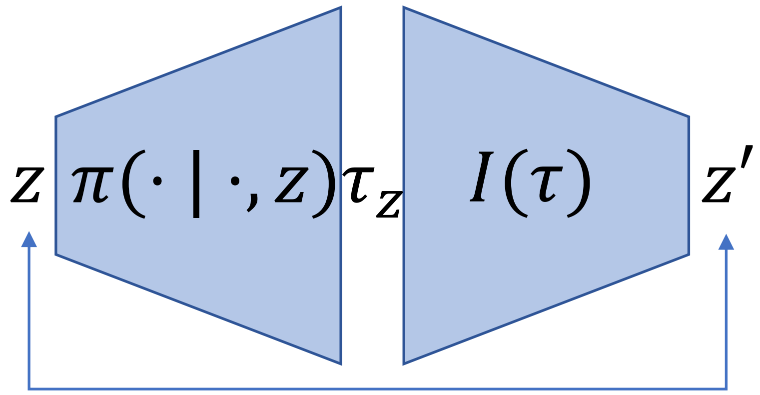
Method	$\Phi(\mathbf{s}, \mathbf{a})$	Aggregator
DT (Chen et al., 2021a)	$r(s, a)$	Summation
DT-X (Section 5.3)	Learned	Summation
CDT (Section 5.2)	$r(s, a)$ or **any**	Binning
BDT (Section 5.4)	Learned	Transformer

Furuta, Hiroki, Yutaka Matsuo, and Shixiang Shane Gu. "Generalized Decision Transformer for Offline Hindsight Information Matching." ICLR22.

Hindsight Information Matching

Information Matching

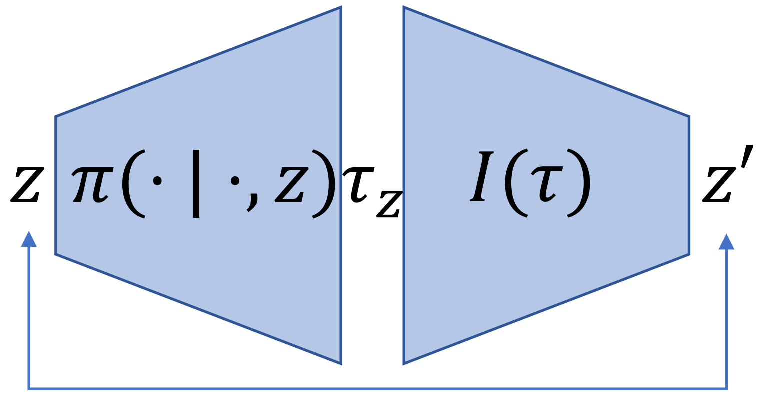
$$\min_{\pi} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z}), \tau_{\mathbf{z}} \sim \pi(\tau_{\mathbf{z}})} \left[\ell \left(\mathbf{z}, I(\tau_{\mathbf{z}}) \right) \right]$$



Hindsight Information Matching

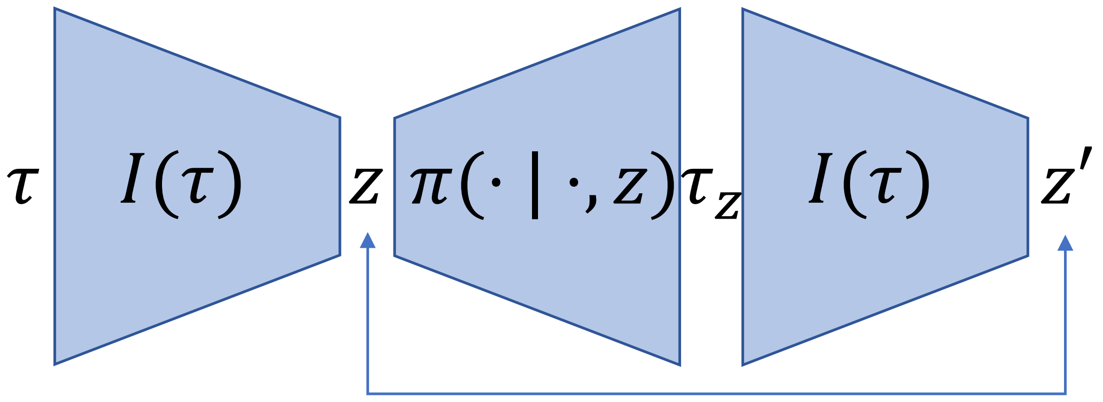
Information Matching

$$\min_{\pi} \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}),\tau_{\mathbf{z}}\sim\pi(\tau_{\mathbf{z}})} \left[\ell\left(\mathbf{z}, I(\tau_{\mathbf{z}})\right)\right]$$



Hindsight Information Matching

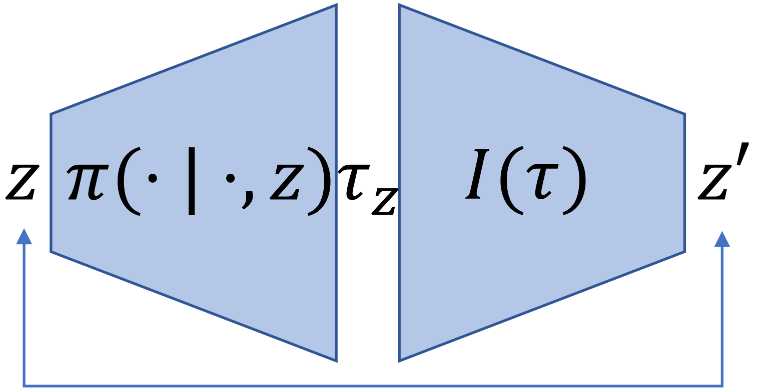
$$\min_{\pi} \mathbb{E}_{\tau\sim\mathcal{D}(\tau),\tau_{\mathbf{z}}\sim\pi(\tau_{\mathbf{z}})} \left[\ell\left(I(\tau), I(\tau_{\mathbf{z}})\right)\right]$$



Hindsight Information Matching

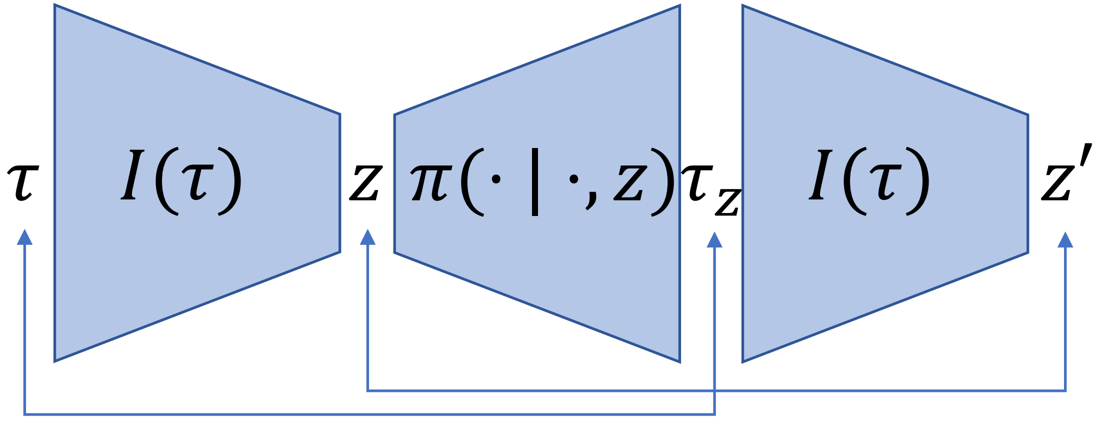
Information Matching

$$\min_{\pi} \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}),\tau_{\mathbf{z}}\sim\pi(\tau_{\mathbf{z}})}\left[\ell\left(\mathbf{z}, I(\tau_{\mathbf{z}})\right)\right]$$



Hindsight Information Matching

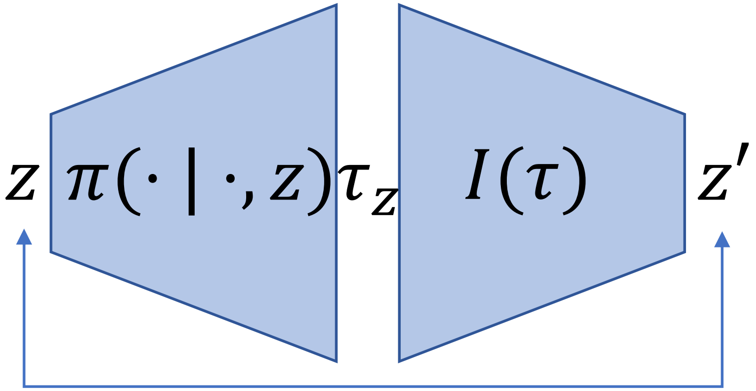
$$\min_{\pi} \mathbb{E}_{\tau\sim\mathcal{D}(\tau),\tau_{\mathbf{z}}\sim\pi(\tau_{\mathbf{z}})}\left[\ell\left(I(\tau), I(\tau_{\mathbf{z}})\right) + \ell\left(\tau, \tau_{\mathbf{z}}\right)\right]$$



Hindsight Information Matching

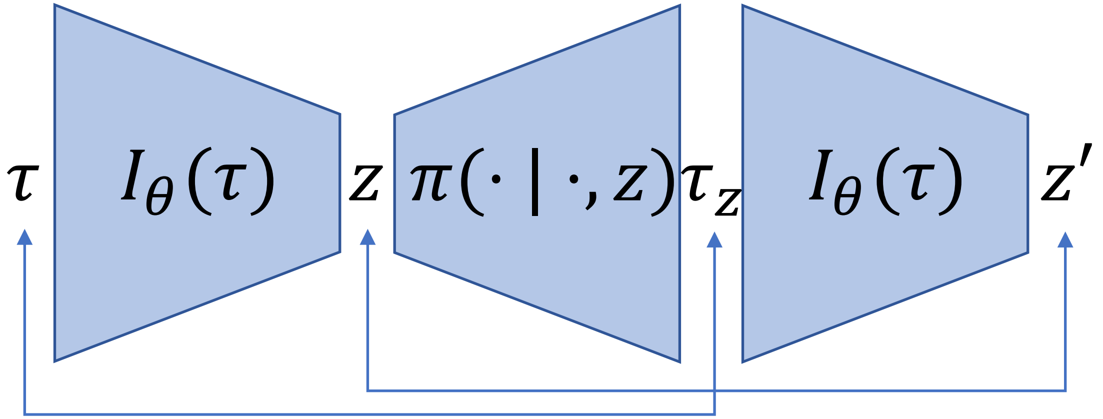
Information Matching

$$\min_{\pi} \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}),\tau_{\mathbf{z}}\sim\pi(\tau_{\mathbf{z}})}\left[\ell\left(\mathbf{z}, I(\tau_{\mathbf{z}})\right)\right]$$

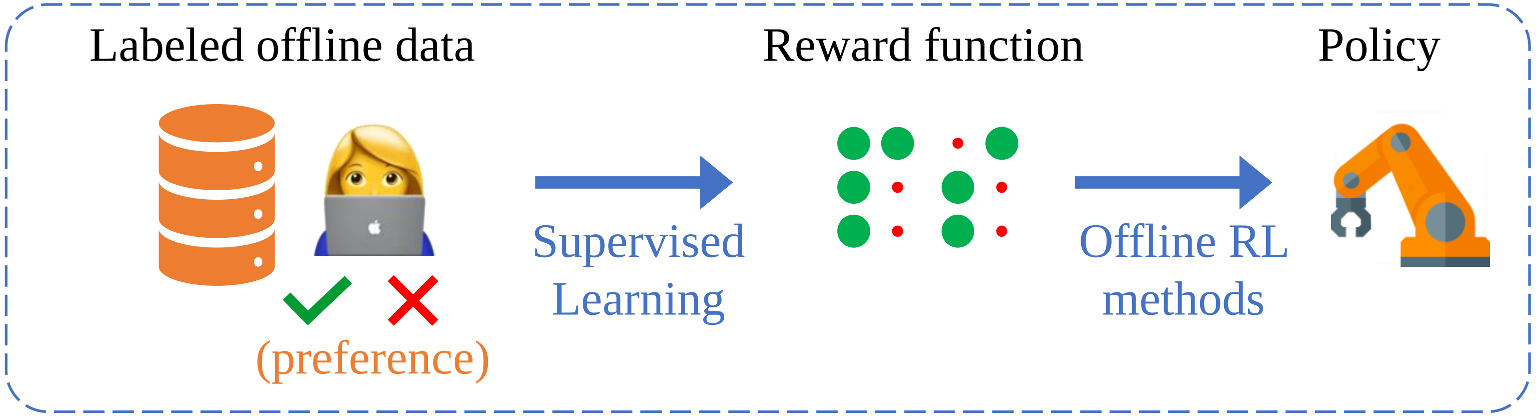


Hindsight Information Matching

$$\min_{\pi, I_\theta} \mathcal{L}_{\text{HIM}} := \mathbb{E}_{\tau\sim\mathcal{D}(\tau),\tau_{\mathbf{z}}\sim\pi(\tau_{\mathbf{z}})}\left[\ell\left(I_\theta(\tau), I_\theta(\tau_{\mathbf{z}})\right)+\ell\left(\tau, \tau_{\mathbf{z}}\right)\right]$$



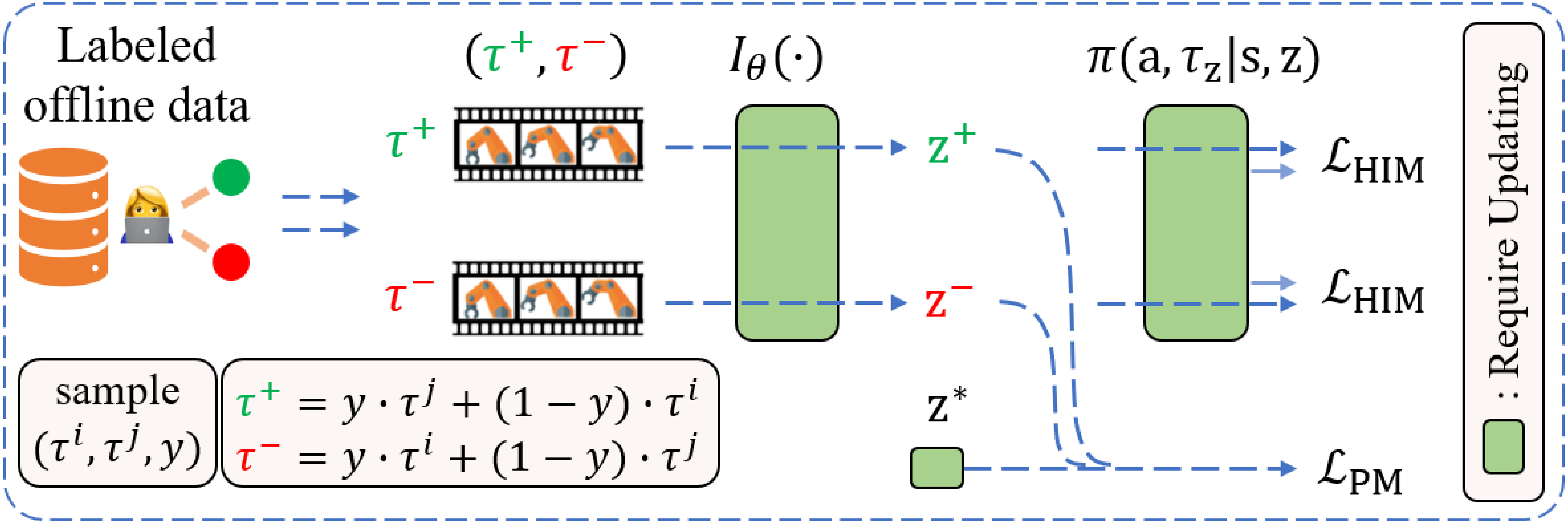
Offline Preference-guided Policy Optimization (OPPO)



$$\pi(\cdot \mid \cdot, z)$$

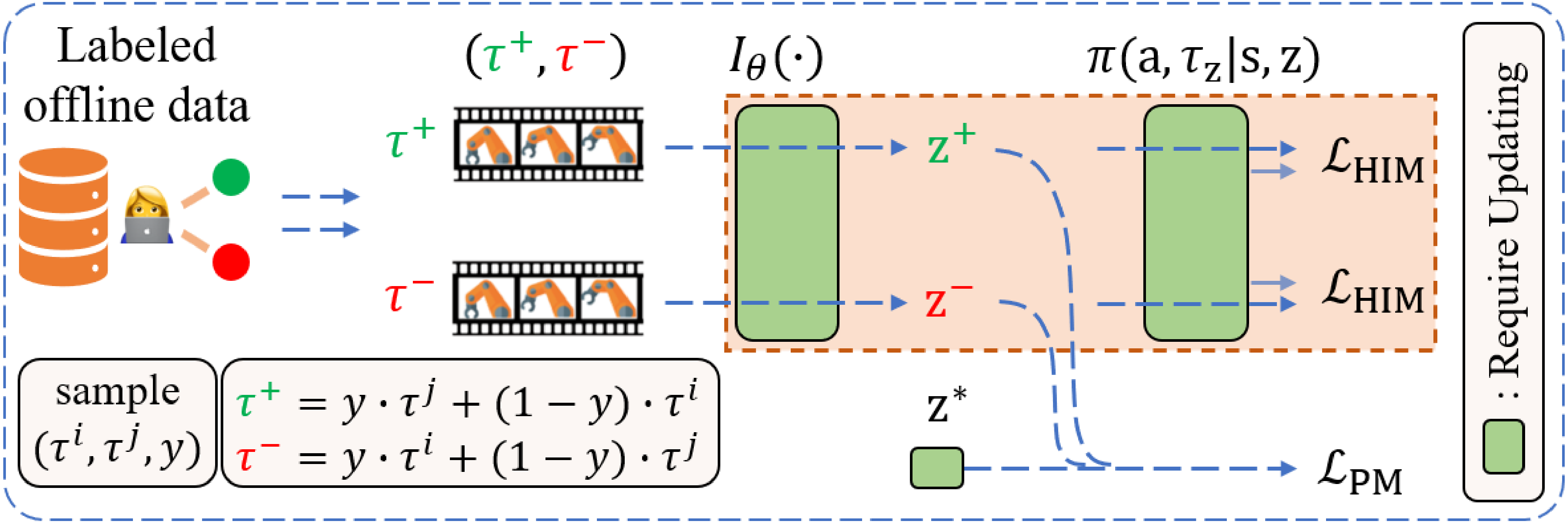
$$z^*$$

Method



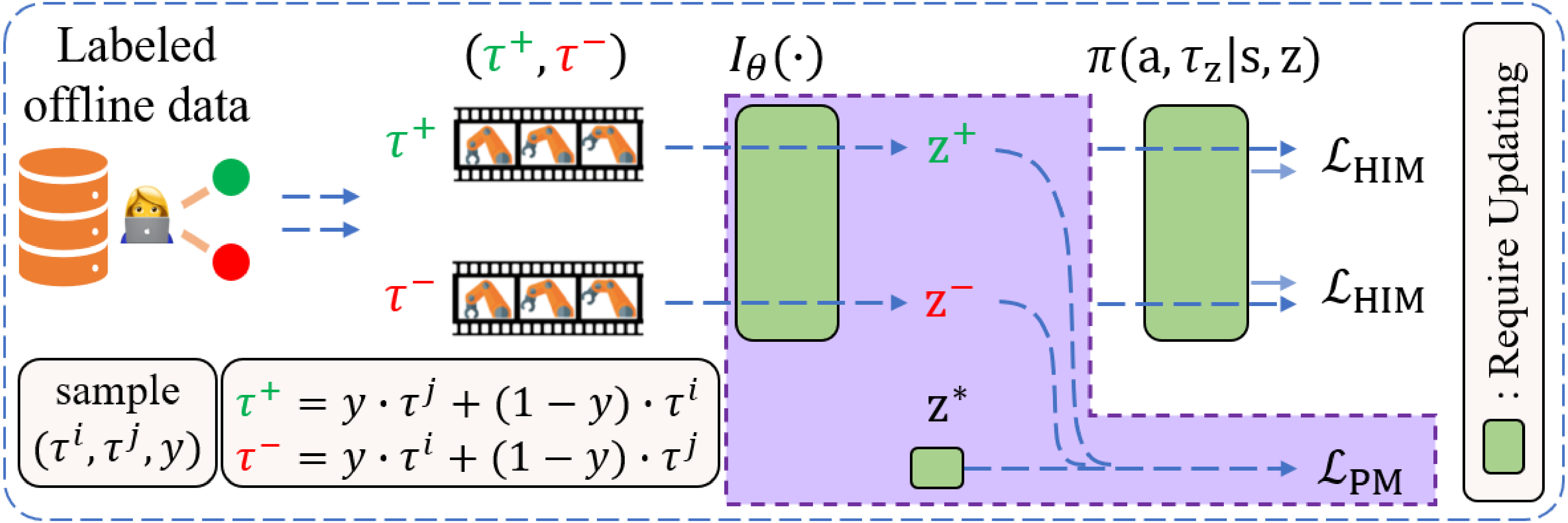
Labeled offline data
(τ^+, τ^-)
$I_\theta(\cdot)$
$\pi(a, \tau_z | s, z)$
τ^+
τ^-
z^+
z^-
z^*
$\mathcal{L}_{HIM}$
$\mathcal{L}_{HIM}$
$\mathcal{L}_{PM}$
: Require Updating
sample (τ^i, τ^j, y)
$\tau^+ = y \cdot \tau^j + (1 - y) \cdot \tau^i$
$\tau^- = y \cdot \tau^i + (1 - y) \cdot \tau^j$

Method



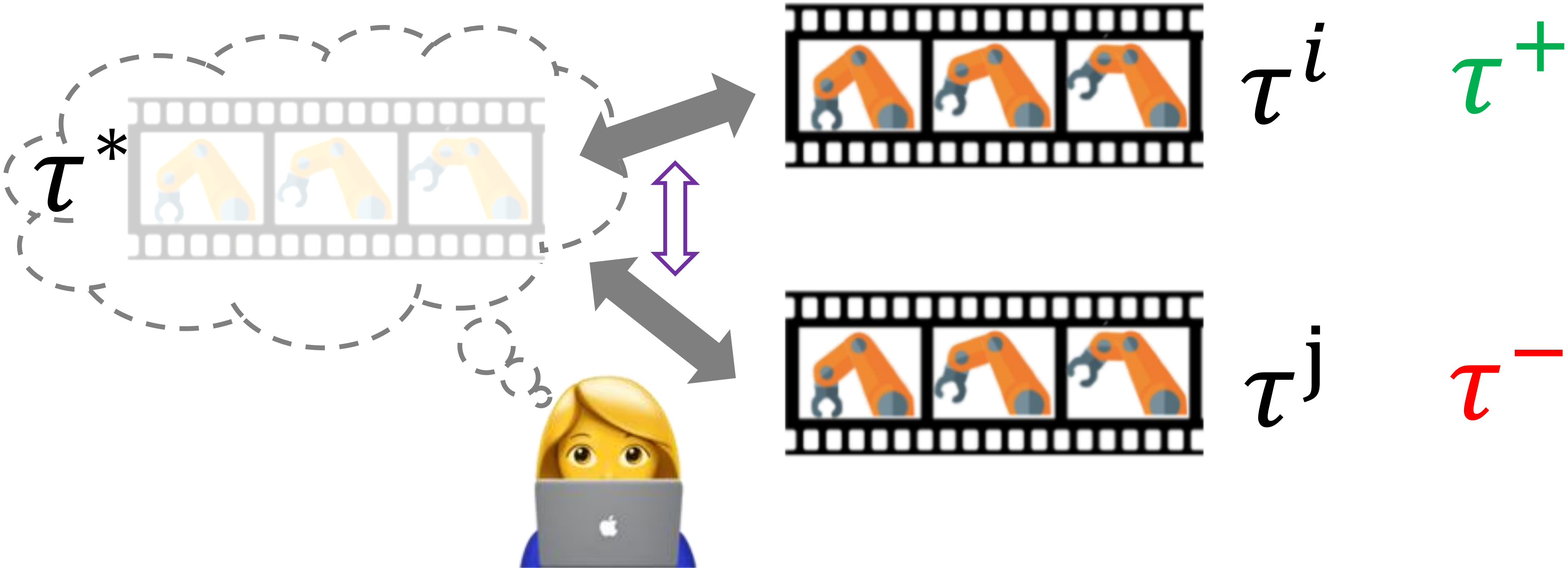
$$\min_{\pi, I_\theta} \mathcal{L}_{\mathrm{HIM}} := \mathbb{E}_{\tau\sim\mathcal{D}(\tau), \tau_\mathrm{z}\sim\pi(\mathrm{z})}\left[\ell\left(I_\theta(\tau), I_\theta(\tau_\mathrm{z})\right) + \ell(\tau, \tau_\mathrm{z})\right]$$

Method



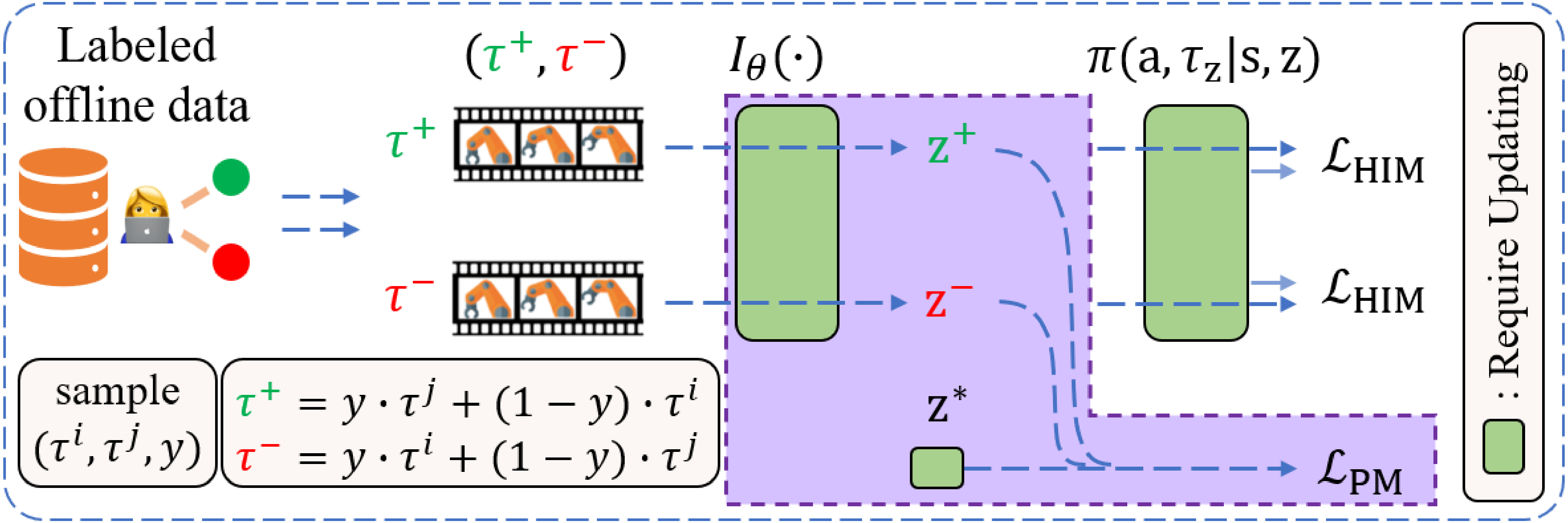
$$\min_{\mathbf{z}^*, I_\theta} \mathcal{L}_{\text{PM}} := \mathbb{E}_{(\tau^i,\tau^j,y)\sim\mathcal{D}_>}[max(\ell(\mathbf{z}^*, \mathbf{z}^+) - \ell(\mathbf{z}^*, \mathbf{z}^-) + \text{margin}, 0)]$$

Method



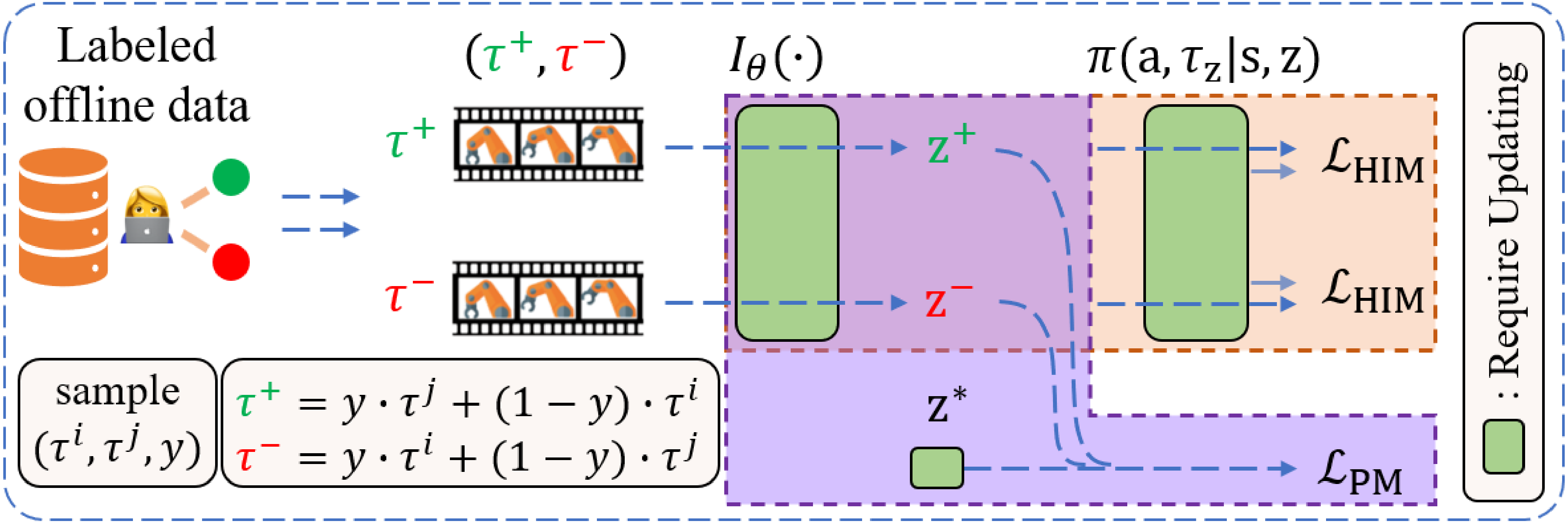
$$\min_{\mathbf{z}^*, I_\theta} \mathcal{L}_{\mathrm{PM}} := \mathbb{E}_{(\tau^i,\tau^j,y)\sim\mathcal{D}_>}\left[max\left(\ell(\mathbf{z}^*, \mathbf{z}^+) - \ell(\mathbf{z}^*, \mathbf{z}^-) + \mathrm{margin}, 0\right)\right]$$

Method



$$\min_{\mathbf{z}^*, I_\theta} \mathcal{L}_{\text{PM}} := \mathbb{E}_{(\tau^i, \tau^j, y) \sim \mathcal{D}_>} [max(\ell(\mathbf{z}^*, \mathbf{z}^+) - \ell(\mathbf{z}^*, \mathbf{z}^-) + \text{margin}, 0)]$$

Method



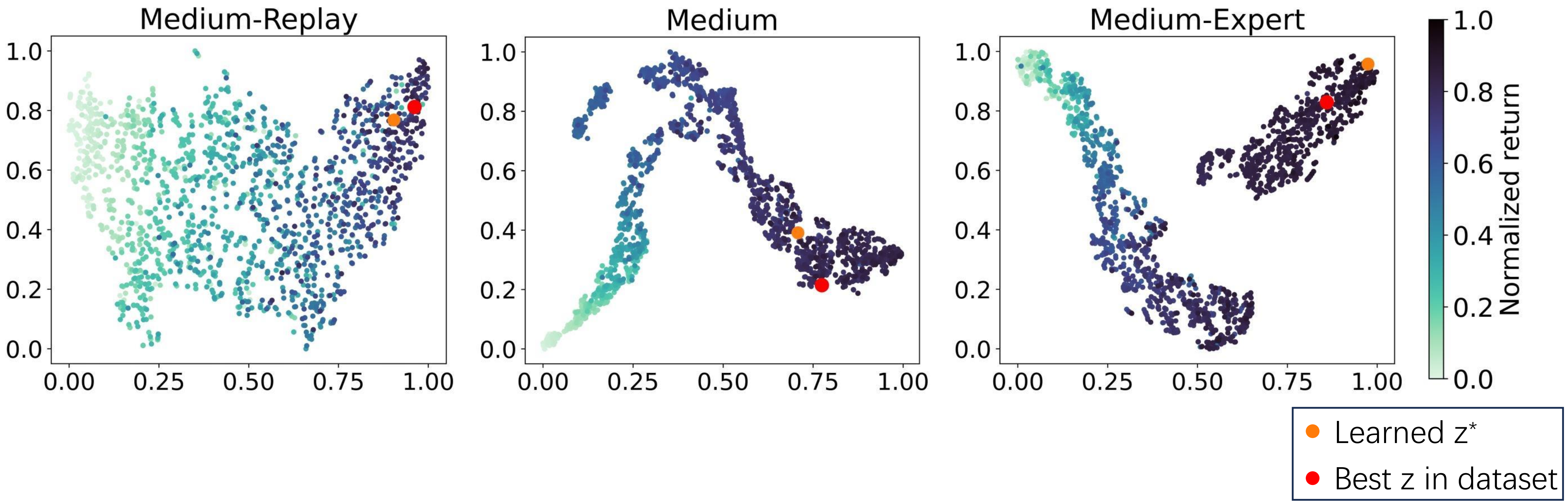
$$\min_{\pi, I_\theta} \mathcal{L}_{HIM} := \mathbb{E}_{\tau \sim \mathcal{D}(\tau), \tau_z \sim \pi(z)} \left[\ell\left(I_\theta(\tau), I_\theta(\tau_z)\right) + \ell(\tau, \tau_z) \right]$$

$$\min_{\mathbf{z}^*, I_\theta} \mathcal{L}_{PM} := \mathbb{E}_{(\tau^i, \tau^j, y) \sim \mathcal{D}_>} \left[max(\ell(\mathbf{z}^*, \mathbf{z}^+) - \ell(\mathbf{z}^*, \mathbf{z}^-) + \text{margin}, 0) \right]$$

$$\mathcal{L}_{total} := \boxed{\mathcal{L}_{HIM}} + \alpha \boxed{\mathcal{L}_{PM}} + \beta \mathcal{L}_{norm}$$

Experiments

Does the learned z-space (encoded by the learned $I_\theta(\cdot)$) align with the given preference?



Experiments

Can the learned optimal contextual policy $\pi(\cdot \mid \cdot, z^*)$ outperform the policy $\pi(\cdot \mid \cdot, z)$ that is conditioned on any other context $z \in \{I_\theta(\tau) | \tau \in D\}$?

Environment	**Dataset**	$\mathbf{z}^*$	$\mathbf{z}_{\text{high}}$	$\mathbf{z}_{\text{low}}$
Hopper	Medium-Expert	**108.0** ± **5.1**	94.2 ± 24.3	79.1 ± 28.8
	Medium	**86.3** ± **3.2**	55.8 ± 7.9	51.6 ± 13.8
	Medium-Replay	**88.9** ± **2.3**	78.6 ± 26.3	26.6 ± 15.2
Walker	Medium-Expert	**105.0** ± **2.4**	**106.5** ± **9.1**	93.4 ± 7.4
	Medium	**85.0** ± **2.9**	64.9 ± 24.9	72.6 ± 10.6
	Medium-Replay	**71.7** ± **4.4**	55.7 ± 24.8	6.8 ± 1.7
Halfcheetah	Medium-Expert	**89.6** ± **0.8**	48.3 ± 14.4	42.6 ± 2.6
	Medium	**43.4** ± **0.2**	42.5 ± 3.9	42.4 ± 3.2
	Medium-Replay	**39.8** ± **0.2**	35.6 ± 8.5	33.9 ± 9.2
Sum		**717.7**	581.9	448.9

Experiments

Can OPPO achieve the competitive performance compared with other offline baselines?

Environment	Dataset	Ours	DT+r	DT+r_ψ	CQL+r	IQL+r	BC
Hopper	Medium-Expert	**108.0 ± 5.1**	**111.0 ± 0.5**	95.6 ± 27.3	**111.0**	91.5	79.6
	Medium	**86.3 ± 3.2**	76.6 ± 3.9	73.3 ± 3.0	58.0	66.3	63.9
	Medium-Replay	**88.9 ± 2.3**	**87.8 ± 4.7**	72.5 ± 22.2	48.6	**94.7**	27.6
Walker	Medium-Expert	105.0 ± 2.4	**109.2 ± 0.3**	**109.7 ± 0.1**	98.7	**109.6**	36.6
	Medium	**85.0 ± 2.9**	80.9 ± 3.1	81.1 ± 2.1	79.2	78.3	77.3
	Medium-Replay	71.7 ± 4.4	**79.6 ± 3.1**	**80.4 ± 4.4**	26.7	73.9	36.9
HalfCheetah	Medium-Expert	**89.6 ± 0.8**	86.8 ± 1.3	**88.4 ± 0.7**	62.4	86.7	59.9
	Medium	43.4 ± 0.2	43.4 ± 0.1	43.2 ± 0.2	44.4	**47.4**	43.1
	Medium-Replay	39.8 ± 0.2	39.2 ± 0.3	38.8 ± 0.3	**46.2**	44.2	4.3
Sum		**717.7**	714.5	683.0	575.2	692.4	429.2

Thank you for listening!

Beyond Reward: Offline Preference-guided Policy Optimization

Yachen Kang, Diyuan Shi, Jinxin Liu, Li He, Donglin Wang

arXiv **Project page** **Github**