Meta-Reinforcement Learning with Adaptation from Human Feedback via Preference-Order-Preserving Task Embedding

Siyuan Xu & Minghui Zhu

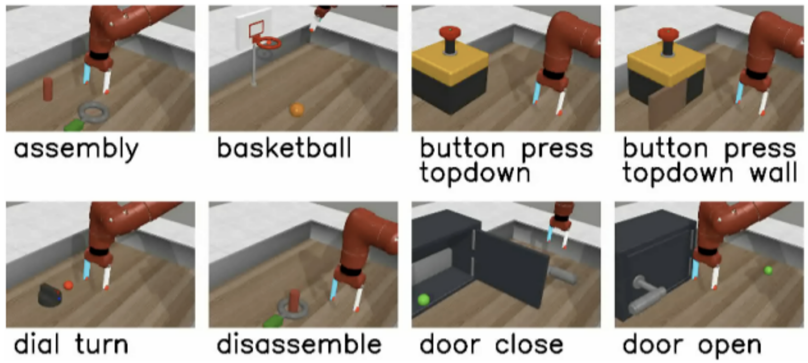
School of Electrical Engineering and Computer Science
The Pennsylvania State University

International Conference on Machine Learning
July, 2025

Meta-RL with adaptation from human feedback

Meta-RL:

Train a meta policy π_ϕ (meta-traning) such that π_ϕ can be adapted to task T given a small dataset $\boldsymbol{D_t}$ of the task T (meta-test)



Meta-RL v.s. Meta-RL with Adaptation from Human Feedback

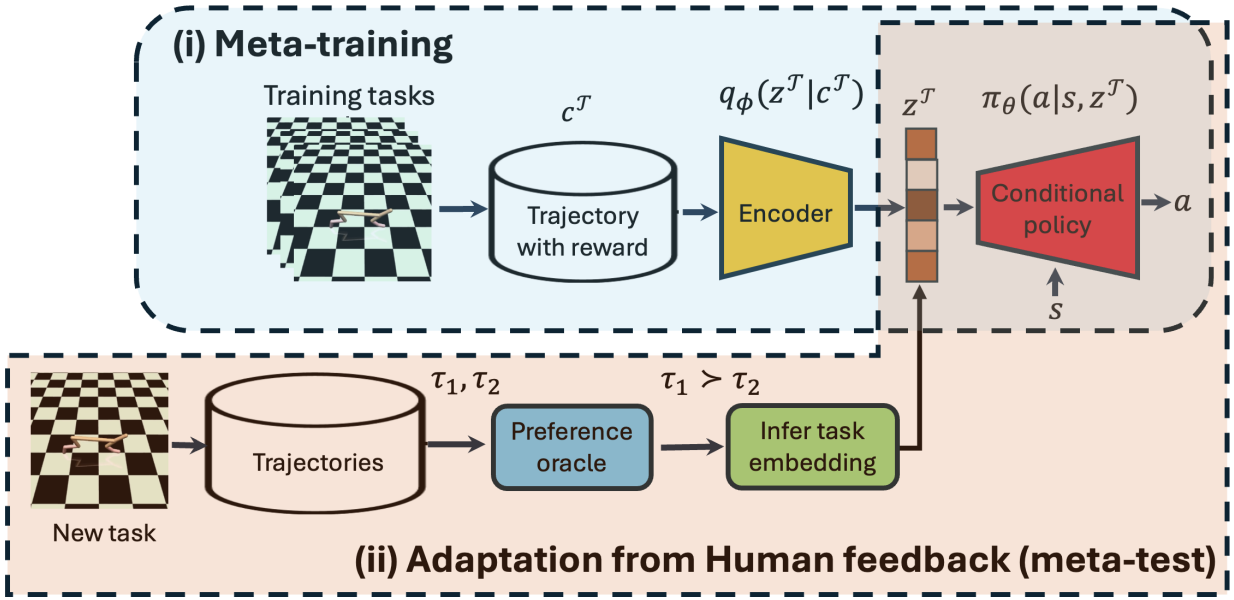
Adaptation (meta-test) data $\boldsymbol{D_t}$ in Meta-RL:

- Trajectories with reward signals $\{(s_t, a_t, s_{t+1}, r_t)\}_H$

Adaptation (meta-test) data $\boldsymbol{D_t}$ in Meta-RL with Adaptation from Human Feedback:

- Trajectories without reward signals: $\tau = \{(s_t, a_t, s_{t+1})\}_H$
- Human preference feedback: $\{\tau_i \succ \tau_j\}_K$

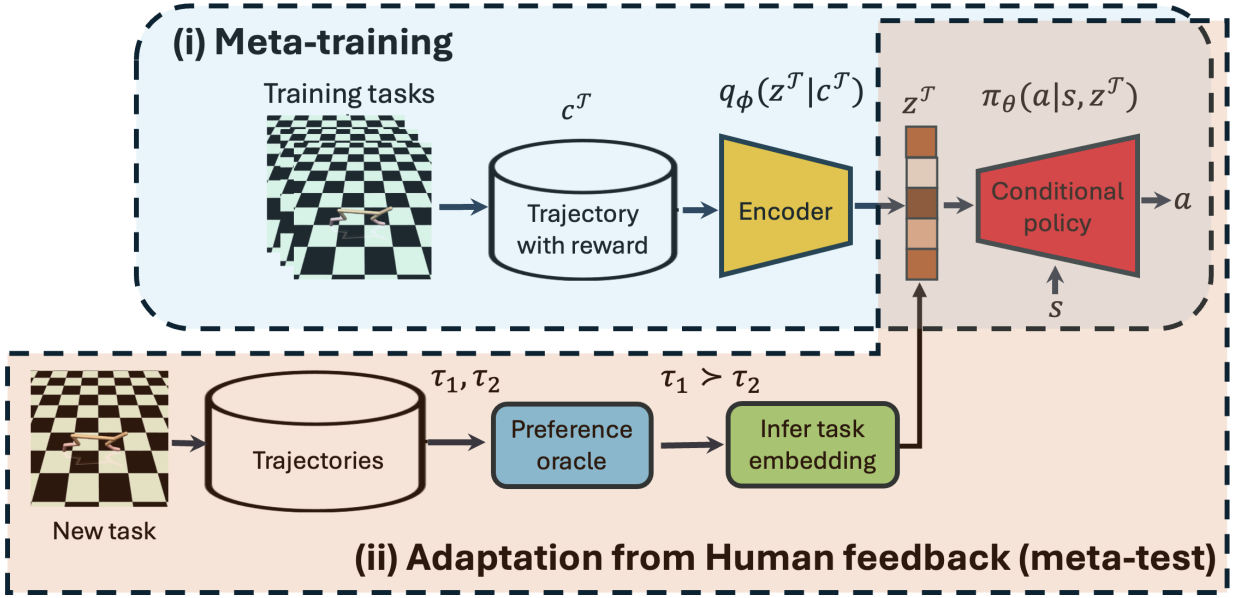
Adaptation via preference-order-preserving embedding (POEM)



Overview:

- Meta-training: learn a task encoder $q_\phi(z^{\mathcal{T}}|c^{\mathcal{T}})$ encodes tasks into embeddings and a policy decoder to decode the policy
- Meta-test: infer the task embedding from human preferences

Adaptation via preference-order-preserving embedding (POEM)

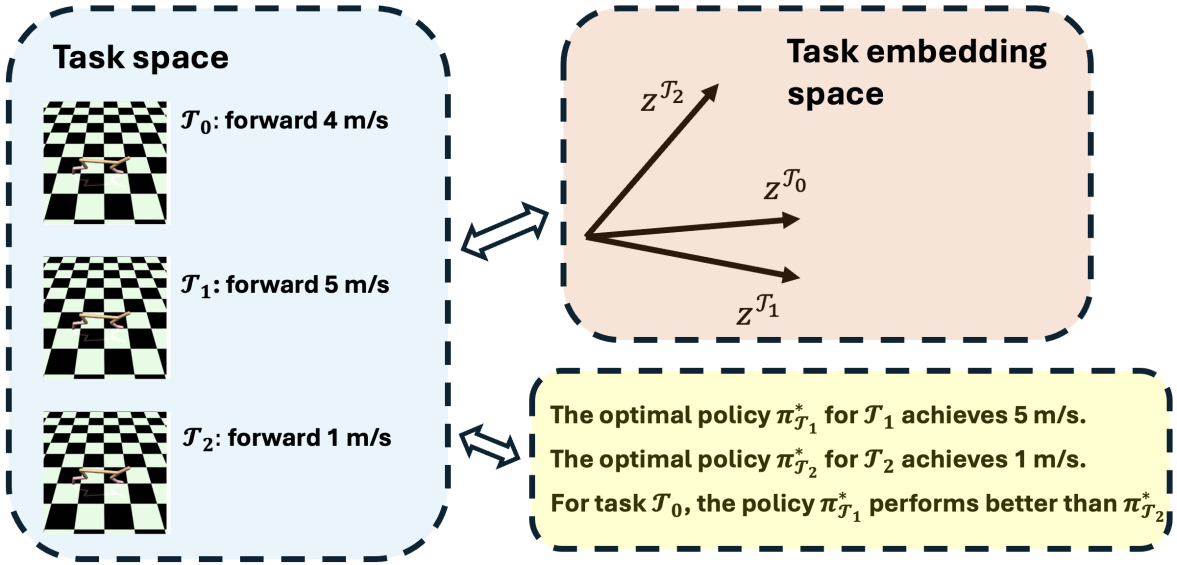


Challenge:

- In the meta-test, it cannot directly encode the human preference data into the task embedding space

Preference-Order-Preserving Embedding Encoder

Insights for preference-order-preserving embedding encoder $q_\phi(z^{\mathcal{T}}|c^{\mathcal{T}})$:

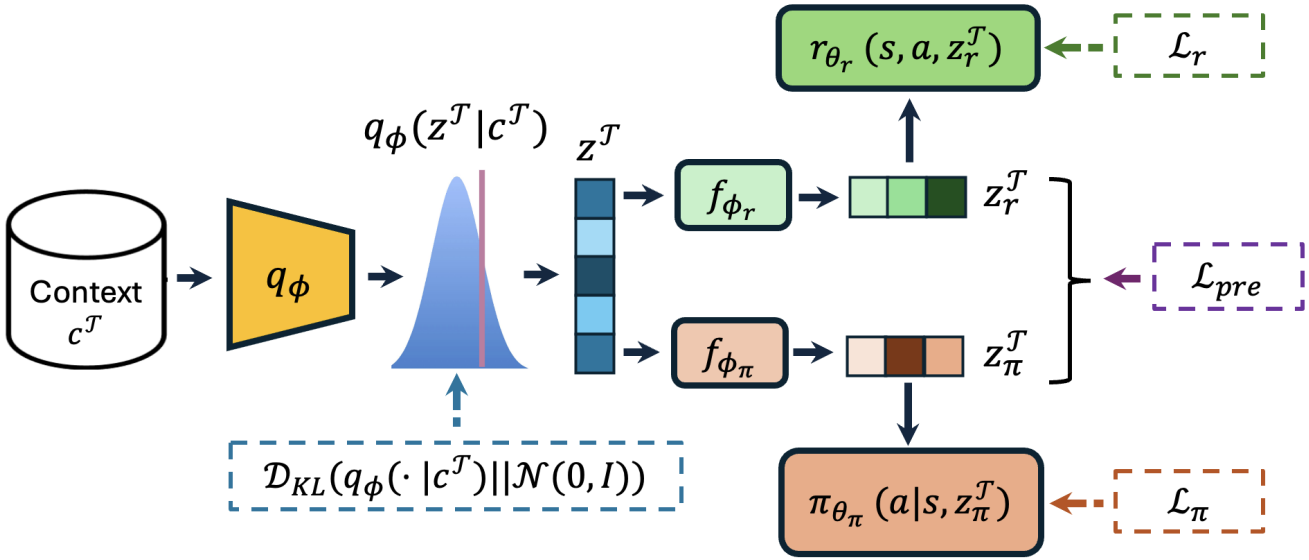


- Connect the relations over the task embeddings with human preference feedback:

$$S(z^{\mathcal{T}_0}, z^{\mathcal{T}_1}) \geq S(z^{\mathcal{T}_0}, z^{\mathcal{T}_2}) \iff \tau^{\mathcal{T}_0}(\pi^*_{\mathcal{T}_1}) \succ \tau^{\mathcal{T}_0}(\pi^*_{\mathcal{T}_2}) \text{ under } \mathcal{T}_0,$$

- We prove that a preference-order-preserving embedding encoder exists.

Meta-training with Preference-Order-Preserving Task Embedding



Loss functions imposed on the encoder-decoder network:

- KL divergence loss $\mathcal{D}_{KL}$ for enforcing the posterior $q_\phi(\cdot|c^{\mathcal{T}})$ to $\mathcal{N}(0, I)$
- Reward reconstruction loss $\mathcal{L}_r$ to recover the true reward $r_{\mathcal{T}}$ by $r_{\theta_r}(\cdot, z^{\mathcal{T}_r})$
- Policy loss $\mathcal{L}_\pi$ for recovering optimal policies by $\pi_{\theta_\pi}(\cdot, z^{\mathcal{T}_\pi})$
- Preference loss $\mathcal{L}_{pre}$ for enforcing $z^{\mathcal{T}}$ to be preference-order-preserving

Meta-training with Preference-Order-Preserving Task Embedding

The preference loss $\mathcal{L}_\pi$:

$$\mathcal{L}_{pre}(\Phi, \mathcal{T}, \mathcal{T}_1, \mathcal{T}_2) \triangleq \mathbb{E}_{z^{\mathcal{T}}, z^{\mathcal{T}_1}, z^{\mathcal{T}_2} \sim q_\phi(\cdot|c^{\mathcal{T}}), q_\phi(\cdot|c^{\mathcal{T}_1}), q_\phi(\cdot|c^{\mathcal{T}_2})}$$
$$\left[D_{KL} \left(\mathbb{I}[\tau^{\mathcal{T}}(\pi^*_{\mathcal{T}_1}) \succ_{\mathcal{T}} \tau^{\mathcal{T}}(\pi^*_{\mathcal{T}_2})] \parallel \Pr[S(z^{\mathcal{T}}, z^{\mathcal{T}_1}) > S(z^{\mathcal{T}}, z^{\mathcal{T}_2})] \right) \right],$$

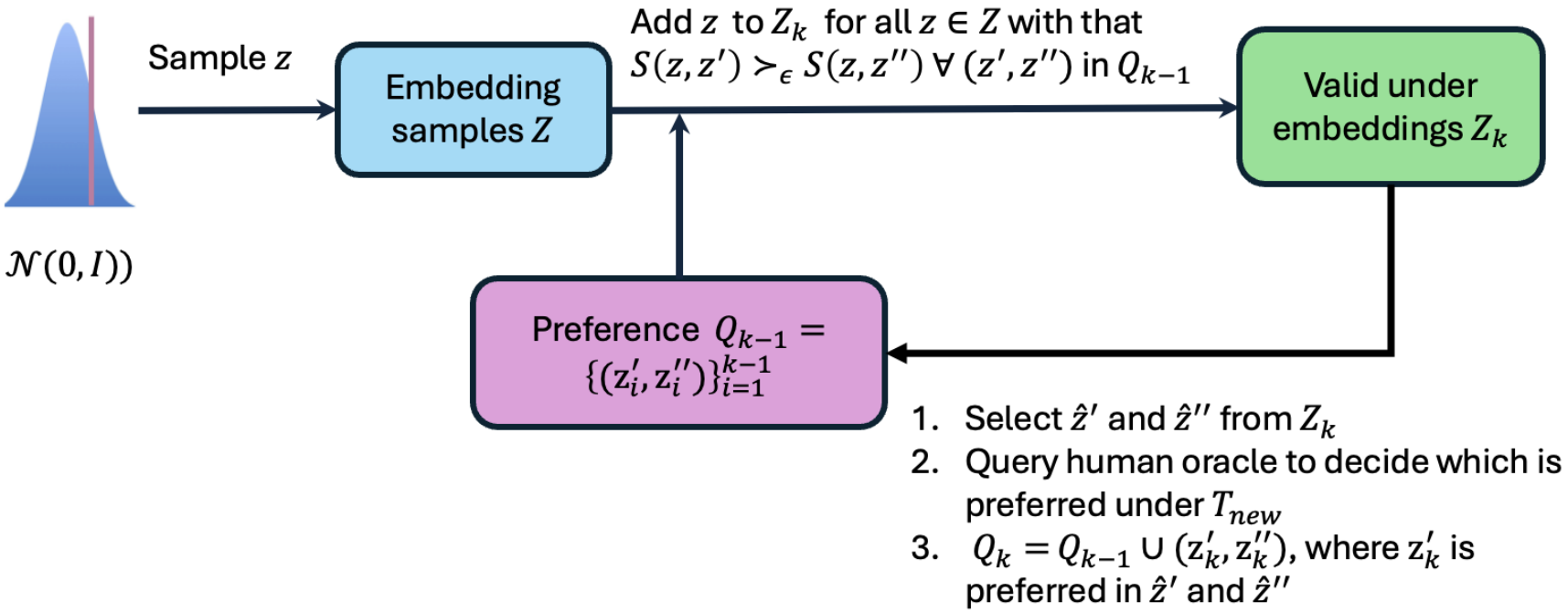
where

- $\tau^{\mathcal{T}}(\pi^*_{\mathcal{T}_1})$ and $\tau^{\mathcal{T}}(\pi^*_{\mathcal{T}_2})$ are trajectories generated by $\pi^*_{\mathcal{T}_1}$ and $\pi^*_{\mathcal{T}_2}$ on $\mathcal{T}$.
- $\mathbb{I}[\tau^{\mathcal{T}}(\pi^*_{\mathcal{T}_1}) \succ_{\mathcal{T}} \tau^{\mathcal{T}}(\pi^*_{\mathcal{T}_2})]$ is the ground-truth preference.
- $\Pr[S(z^{\mathcal{T}}, z^{\mathcal{T}_1}) > S(z^{\mathcal{T}}, z^{\mathcal{T}_2})]$ is the preference prediction based on the preference-order-preserving embedding. Following the Bradley-Terry model, we use the predicted probability

$$\Pr[S(z^{\mathcal{T}}, z^{\mathcal{T}_1}) > S(z^{\mathcal{T}}, z^{\mathcal{T}_2})] = \frac{\exp\left(S(z^{\mathcal{T}}, z^{\mathcal{T}_1})\right)}{\exp\left(S(z^{\mathcal{T}}, z^{\mathcal{T}_1})\right) + \exp\left(S(z^{\mathcal{T}}, z^{\mathcal{T}_2})\right)}.$$

Adaptation from Human Preference by Task Embedding Inference

Given task T_{new}, in iteration k:



Select (z'_k, z''_k) from $\mathcal{Z}_k$ by maximum candidate elimination:

$$(z'_k, z''_k) = \arg\min_{z',z'' \in \mathcal{Z}_k} (\max\{|\mathcal{Z}^{(1)}|, |\mathcal{Z}^{(2)}|\}), \text{ where}$$

$\mathcal{Z}^{(1)} = \{z \in \mathcal{Z}_k : S(z,z') >_\epsilon S(z,z'')\}$ and $\mathcal{Z}^{(2)} = \{z \in \mathcal{Z}_k : S(z,z'') >_\epsilon S(z,z')\}$.

Theoretical guarantee

Convergence in distribution

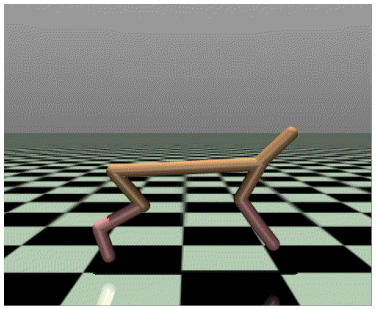
Assume that (i) the task encoder $f : \Gamma \to \mathbb{R}^d$ holds that, for $\mathcal{T} \sim \mathbb{P}(\Gamma)$, $z^{\mathcal{T}} = f(\mathcal{T})$ follows the normal distribution $\mathcal{N}(0, I)$; (ii) f_r and $f_\pi : \mathbb{R}^d \to \mathbb{R}^d$ are mappings such that the preference-order-preserving property is satisfied; (iii) the policy network with θ_π is optimal, i.e., $\pi_{\theta_\pi}(\cdot|z_\pi^{\mathcal{T}}) = \pi_{\mathcal{T}}^*$. Suppose the preference oracle on task $\mathcal{T}_{new}$ holds the error at most ϵ. Under certain mild conditions on the mappings f_r and f_π,

(a) The probability density function (PDF) of Z_k at $z^{\mathcal{T}_{new}}$ has
$$P(Z_k = z^{\mathcal{T}_{new}}) \geq \frac{P(Z=z^{\mathcal{T}_{new}})}{C_1 \cdot \left(\frac{1}{2}\right)^k + C_2 \log\left(\frac{1+2\epsilon}{1-2\epsilon}\right)};$$

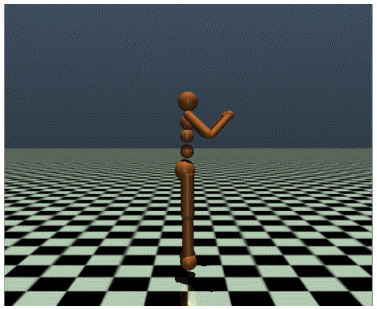
(b) When $\epsilon = 0$, $Z_k \xrightarrow{d} z^{\mathcal{T}_{new}}$, i.e., Z_k converges to $z^{\mathcal{T}_{new}}$ in distribution.

Here, C_1 and C_2 are the constants, and $z^{\mathcal{T}_{new}}$ is the embedding of $\mathcal{T}_{new}$ such that $J_{\mathcal{T}_{new}}(\pi_{\theta_\pi}(\cdot|z_\pi^{\mathcal{T}_{new}})) = J_{\mathcal{T}_{new}}(\pi_{\mathcal{T}}^*)$.

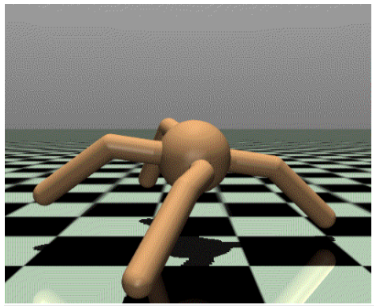
Experiments setting



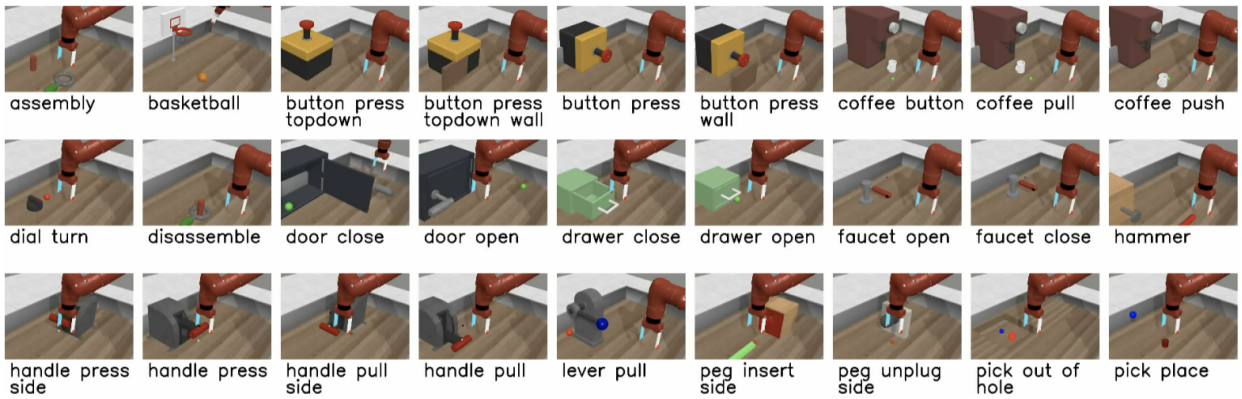
(a) Half-Cheetah



(b) Humanoid



(c) Ant



(d) MetaWorld

Figure: Visualization of robotic locomotion environments, including Half-Cheetah, Humanoid, and Ant, simulated by Mujoco and the MetaWorld environments.

Experiments results

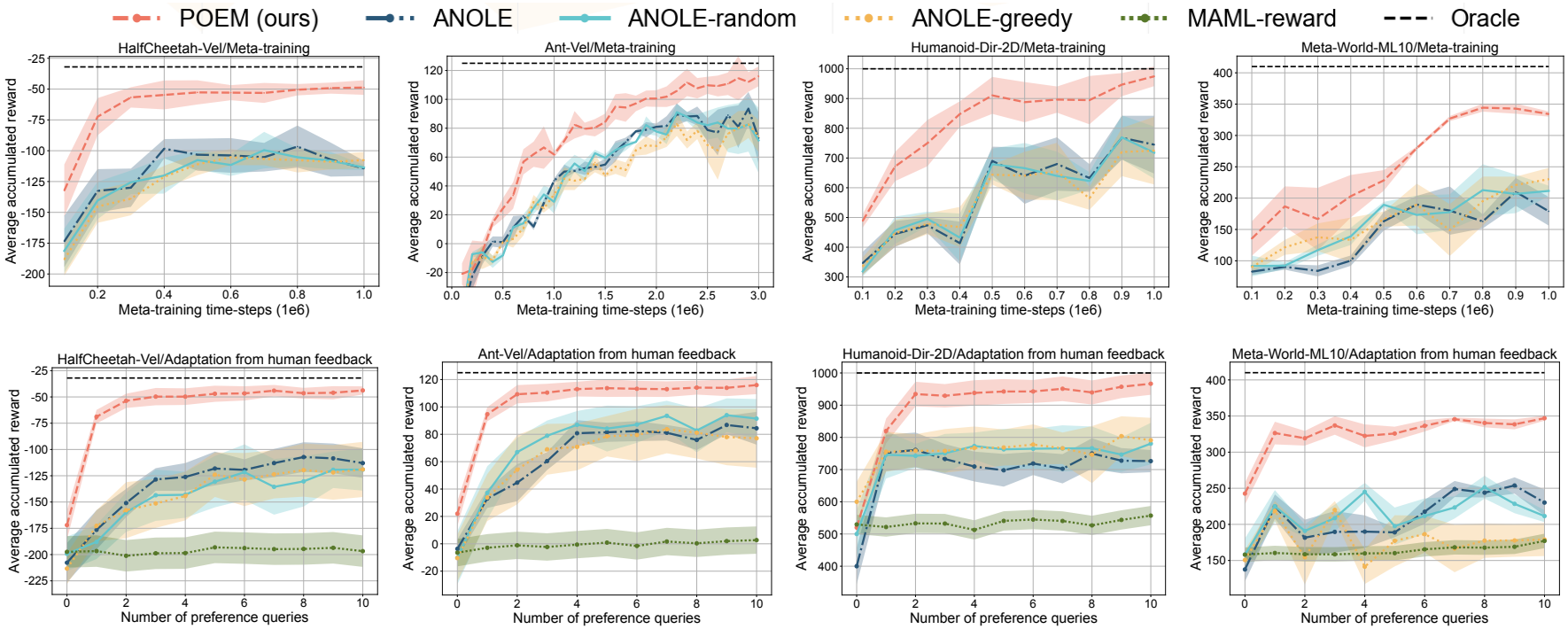


Figure: Performance on HalfCheetah-Vel, Ant-Vel, Humanoid-Dir, and MetaWorld-ML10. Average accumulated reward on test tasks v.s. samples collected during meta-training (**Top**) and during adaptation from human feedback (**Bottom**).

Conclusion

- Propose adaptation via preference-order-preserving embedding (POEM) for meta-RL with adaptation from human feedback.
- Theoretically guarantee the convergence in distribution of the proposed method.
- Experimentally validate the effectiveness of the algorithm in continuous control environments on Mujoco and MetaWorld.