Reflect-then-Correct: Rebalancing Task Optimization for Generalizable Meta-Reinforcement Learning via Distributional Value Error Reduction

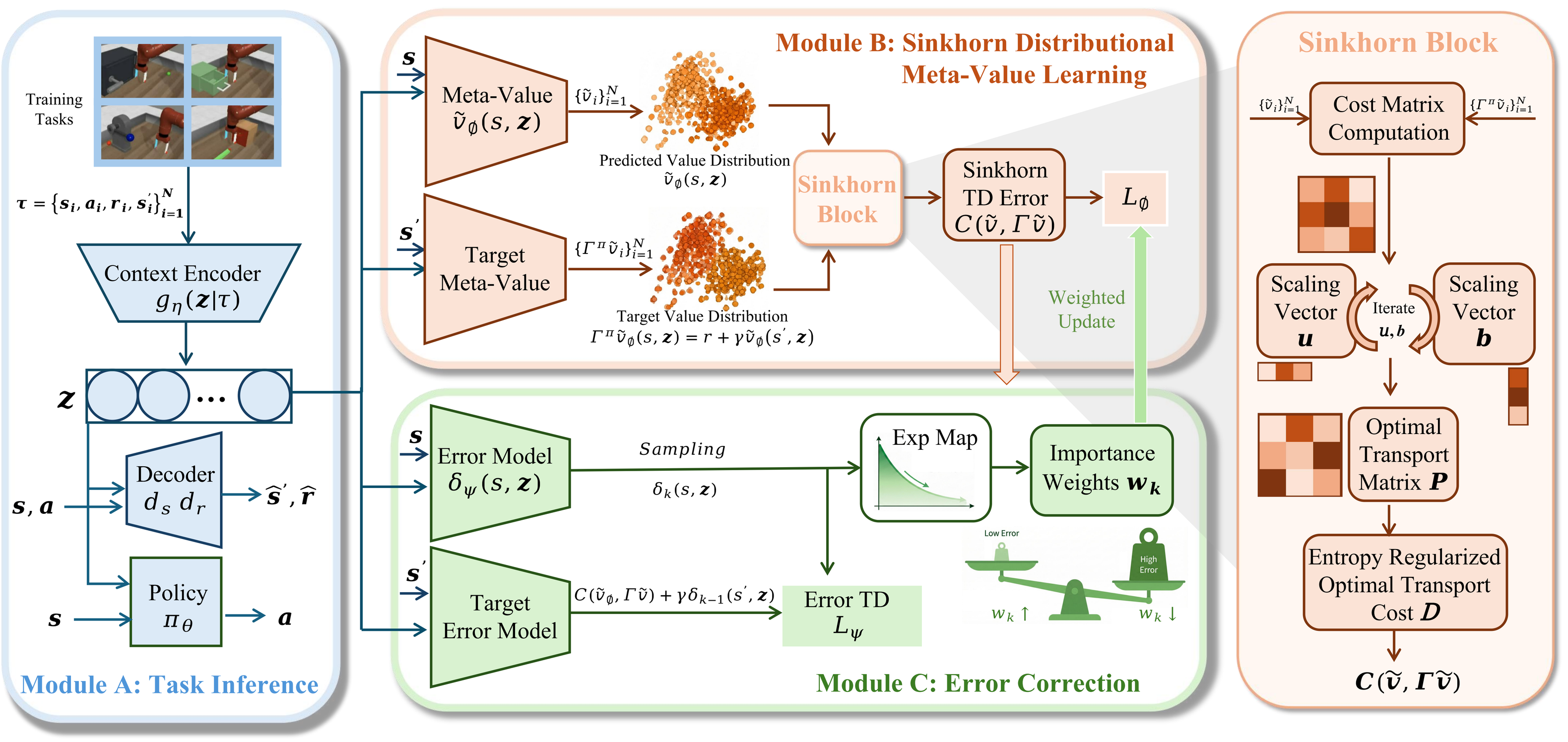
Min Wang[1], Xin Li[†1], Ye He[1], Mingzhong Wang[2], Yonggang Zhang[3]

[1]Beijing Institute of Technology, China [2]University of the Sunshine Coast, Australia [3]Jilin University, China [†] Corresponding author

Background

Despite notable advances in improving adaptability, prior meta-Reinforcement Learning (meta-RL) studies have primarily focused on parametric tasks, where the tasks differ solely in underlying parameters. Meta-RL faces significant challenges in non-parametric settings, where vastly different return scales across diverse tasks cause severe gradient interference. Existing categorical solutions attempt to normalize these scales but often fail due to rigid discretization and quantization errors.

Method



Sinkhorn Distributional Meta-Value

Entropy regularized optimal transport cost:

$$\mathcal{D}(\tilde{v}, \Gamma^{\pi}\tilde{v}) = \min_{\Pi \in \mathbf{\Pi}(\tilde{v}, \Gamma^{\pi}\tilde{v})} \left\{ \int c(x,y)\,\mathrm{d}\Pi(x,y) + \lambda \mathcal{H}(\Pi, \tilde{v} \otimes \Gamma^{\pi}\tilde{v}) \right\}. \tag{1}$$

Sinkhorn TD error of meta-value:

$$\mathcal{C}(\tilde{v}, \Gamma^{\pi}\tilde{v}) = \mathcal{D}(\tilde{v}, \Gamma^{\pi}\tilde{v}) - \frac{1}{2}\Big(\mathcal{D}(\tilde{v}, \tilde{v}) + \mathcal{D}(\Gamma^{\pi}\tilde{v}, \Gamma^{\pi}\tilde{v})\Big),$$

$$\text{where} \quad \mathcal{D} = \min_{P \in \mathbb{R}_{+}^{N\times N}} \{\langle P, c\rangle; P\mathbf{1}_N = \mathbf{1}_N, P^{\top}\mathbf{1}_N = \mathbf{1}_N\} + \lambda \textstyle\sum_{ij} P_{ij}(\log P_{ij} - 1) + \lambda, \tag{2}$$

The optimal transport plan:

$$\exists u^* \in \mathbb{R}_{+}^{n}, b^* \in \mathbb{R}_{+}^{n} \text{ s.t. } P^* = \operatorname{diag}(u^*)\,\mathcal{K}\operatorname{diag}(b^*), \tag{3}$$

$$u^{(l+1)} \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}b^{(l)}} \quad \text{and} \quad b^{(l+1)} \leftarrow \frac{\mathbf{1}_N}{\mathcal{K}^{\top}u^{(l+1)}}. \tag{4}$$

Meta-Value Error Correction

Optimization Problem of Meta-Value

$$\min_{p_k} \mathbb{E}_{d^{\pi_k}}\left[\mathcal{C}(\tilde{v}_k, \tilde{v}^*)\right] \quad \text{s.t.} \sum_{s,\mathbf{z}} p_k(s,\mathbf{z}) = 1,$$
$$\tilde{v}_k = \arg\min_{\tilde{v}} \mathbb{E}_{p_k}\left[\mathcal{C}(\tilde{v}, \Gamma^{*}\tilde{v}_{k-1})\right]. \tag{5}$$

Theorem 1 *Let $p_k(s,\mathbf{z})$ denote the solution to a relaxed version of the optimization problem in Eq. 5. Then we have*

$$p_k \propto \frac{\exp\left(-\mathcal{C}(\tilde{v}_k, \tilde{v}^*)\right)}{\alpha^*}, \tag{6}$$

and the corresponding importance weights are given by

$$w_k(s,\mathbf{z}) = \frac{p_k}{\mu_k} \propto \exp\left(-\frac{\gamma\left[\mathcal{P}^{\pi_{k-1}}\delta_{k-1}\right](s,\mathbf{z})}{\beta}\right), \tag{7}$$

where μ_k denotes the sampling distribution, $\alpha^* \in \mathbb{R}^+$ is the Lagrange multiplier with the constraint $\sum_{s,\mathbf{z}} p_k(s,\mathbf{z}) = 1$, and $\beta > 0$ controls the sharpness of reweighting.

Experimental Results

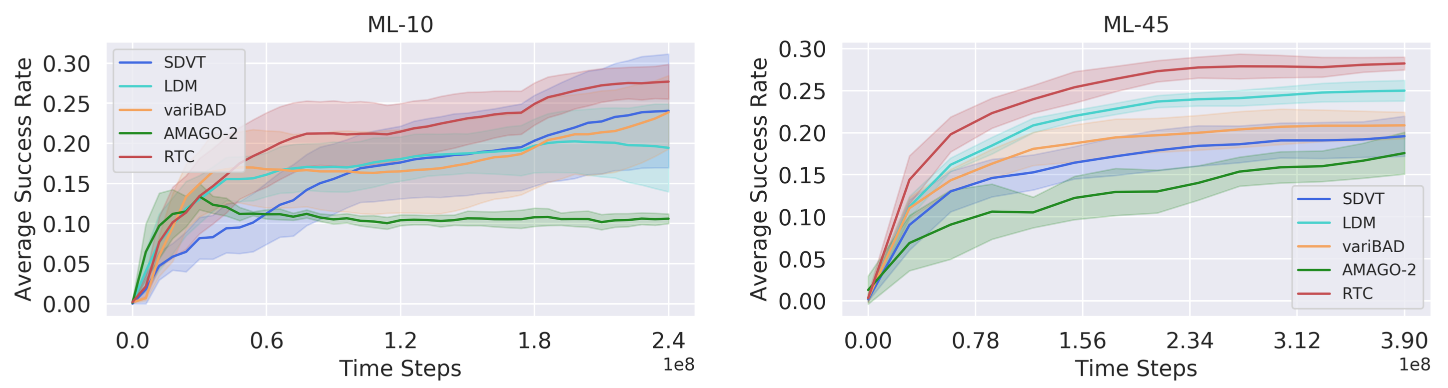
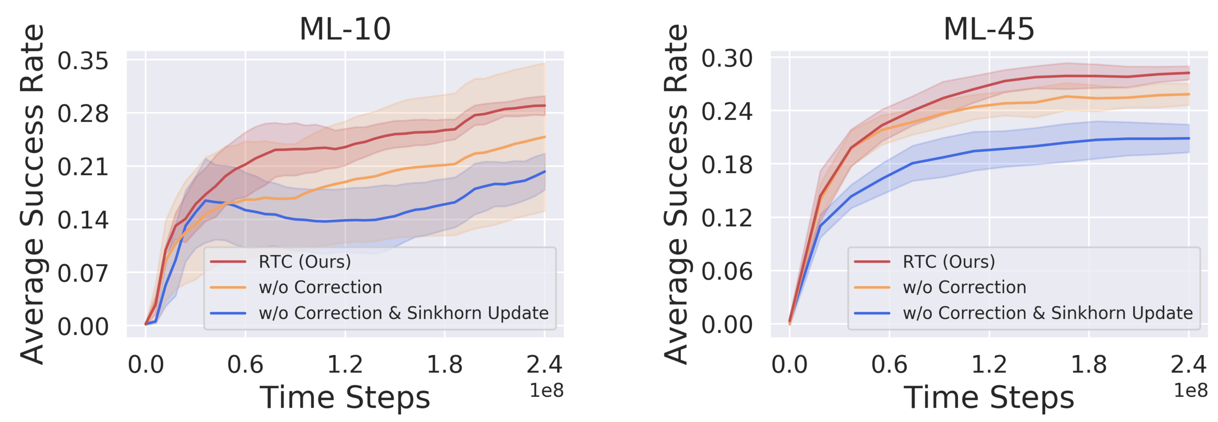
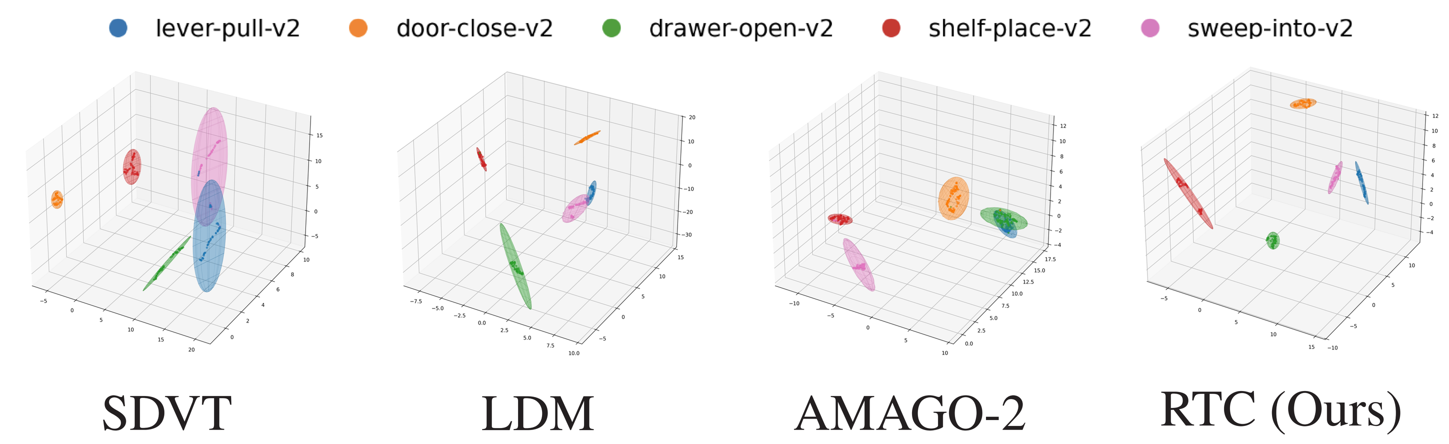


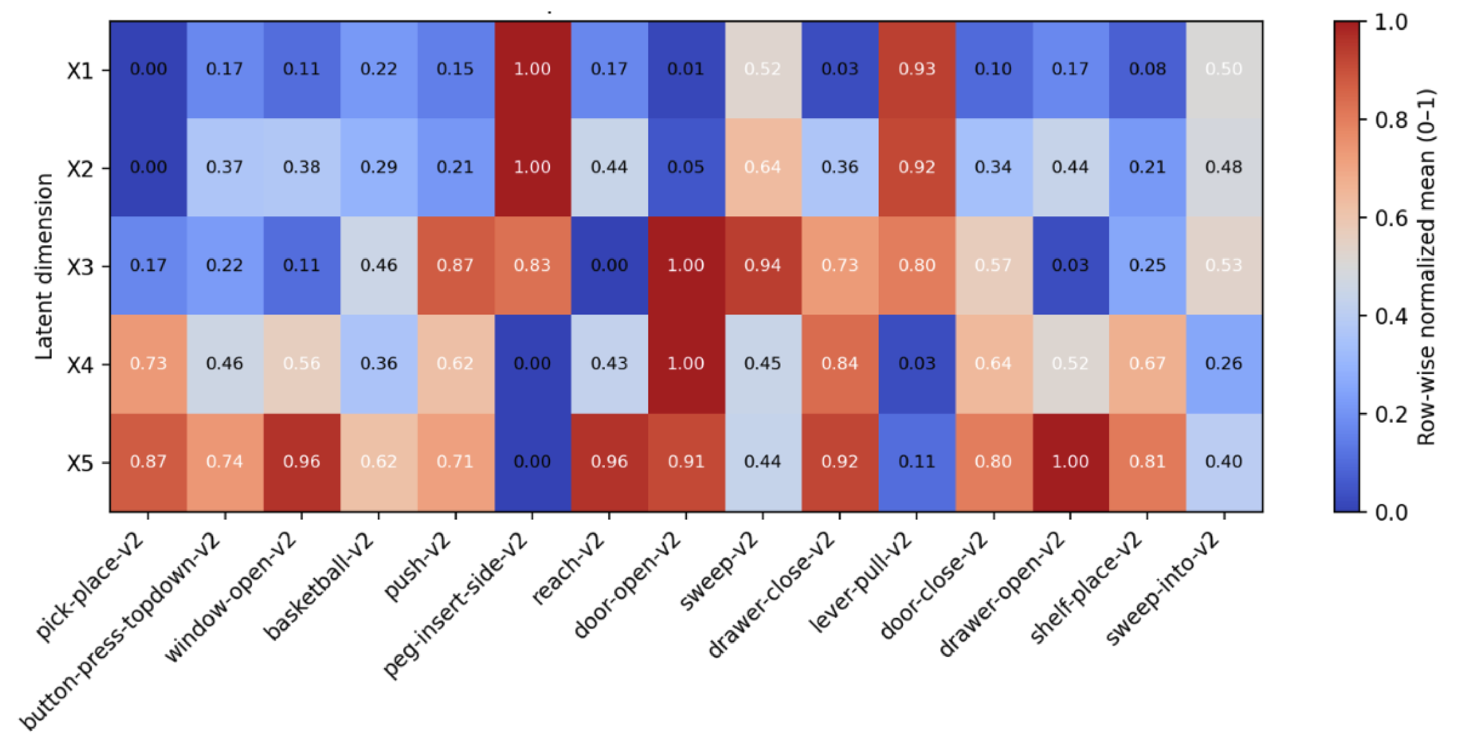
Table 1: Converged average train/test success rate ± standard error (%) and test return on ML-10 and ML-45.

	Success Rate (%)				Return	
	ML-10		ML-45		ML-10	ML-45
Methods	Train	Test	Train	Test	Test	Test
SDVT	60.61 ± 11.66	24.16 ± 11.18	34.69 ± 9.33	21.19 ± 2.17	978.93 ± 170.25	615.40 ± 61.94
LDM	57.34 ± 10.54	17.92 ± 11.09	34.41 ± 2.40	<u>26.57</u> ± 3.67	895.35 ± 199.53	<u>712.63</u> ± 45.53
variBAD	68.95 ± 10.30	<u>28.39</u> ± 3.76	41.24 ± 1.40	23.27 ± 3.57	<u>1075.50</u> ± 207.92	655.70 ± 42.16
AMAGO-2	**98.44** ± 0.83	10.89 ± 2.11	**79.91** ± 3.62	24.07 ± 5.16	841.25 ± 133.86	549.50 ± 73.38
RTC (Ours)	<u>79.53</u> ± 2.11	**30.42** ± 3.10	<u>62.62</u> ± 1.64	**28.66** ± 2.02	**1394.75** ± 74.93	**757.95** ± 42.82

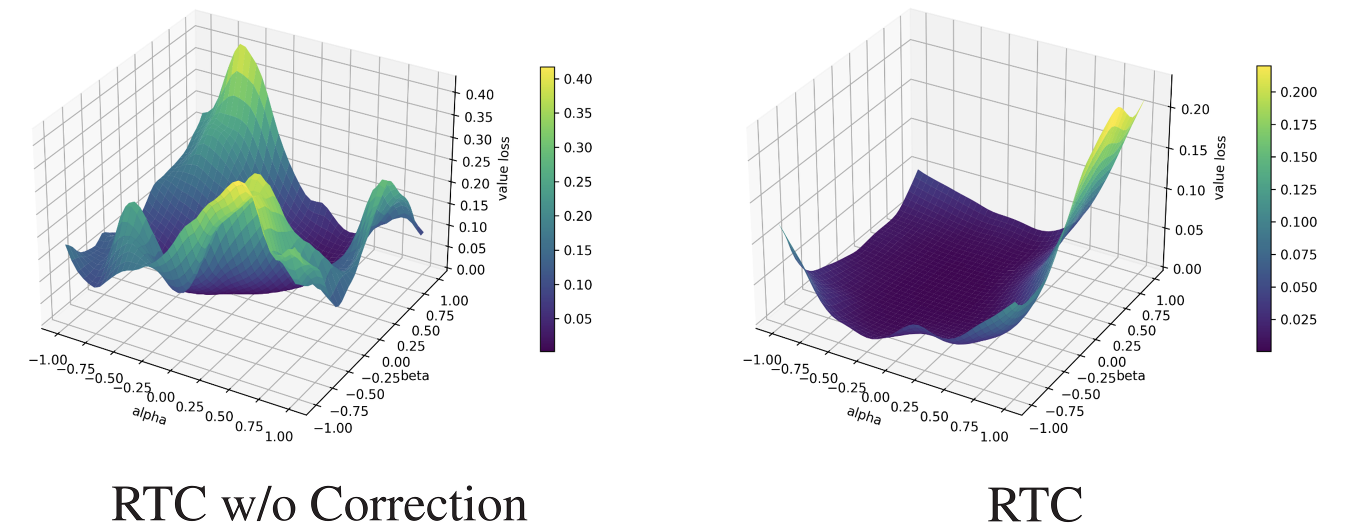


Visualization of Task Representation





Landscape of the Meta-Value Network



Conclusion

We propose RTC, which leverages the Sinkhorn divergence to achieve geometry-aware alignment across diverse tasks while explicitly correcting statistical bias. By modeling the temporal accumulation of Bellman inconsistencies through a recursive error model and reweighting transitions accordingly, RTC combines flexible distributional representations with error-aware optimization.

Paper Code