EC-DIFFUSER: MULTI-OBJECT MANIPULATION VIA ENTITY-CENTRIC BEHAVIOR GENERATION (ICLR 2025)

Carl Qi, Dan Haramati, Tal Daniel, Aviv Tamar, Amy Zhang

PRESENTED BY CARL QI
Ph.D. Student, The University of Texas at Austin

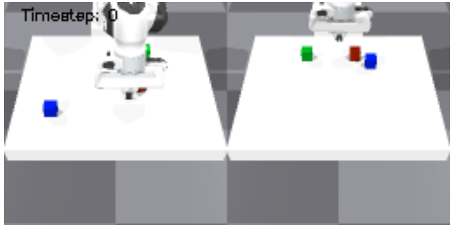
TEXAS
The University of Texas at Austin

Object Manipulation in the Real World

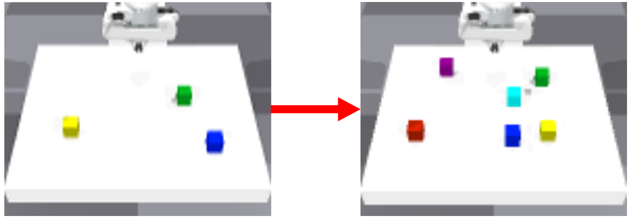
High-dimensional observations (pixels)
Long-horizon planning
Account for entity-entity relations and interactions

The Goal-Conditioned Multi-Object Manipulation Problem

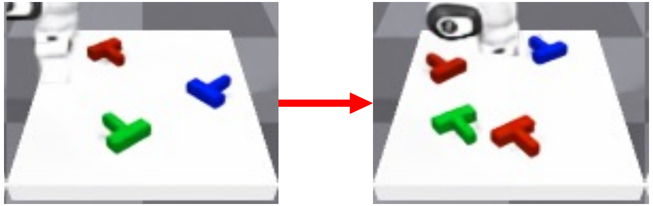
Given an observation and a goal, the policy outputs manipulation actions.
Access to an offline dataset of demonstrations.



We want to achieve **Compositional Generalization.**



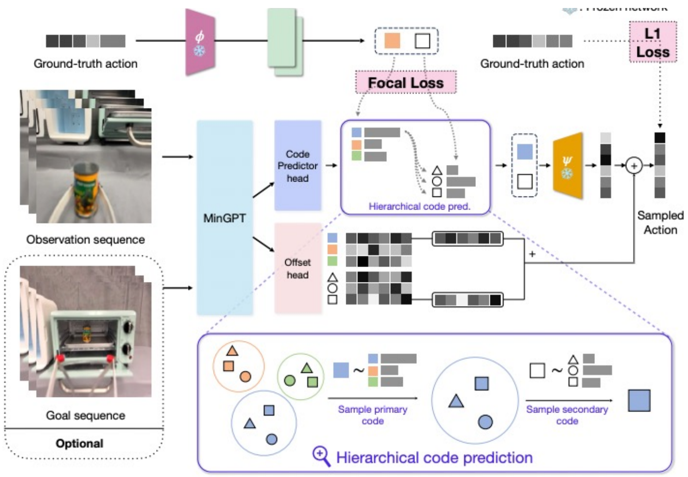
Train Test



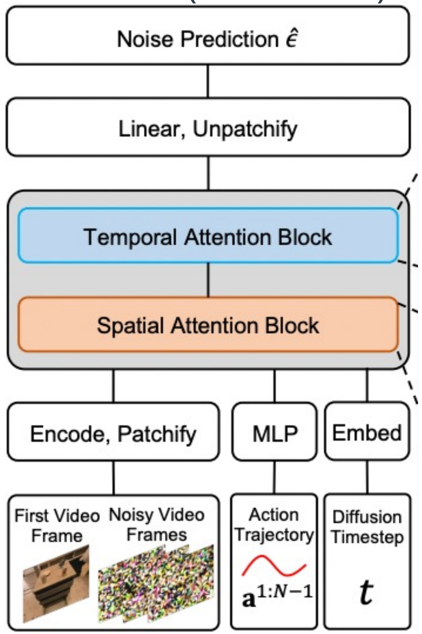
Train Test

Prior Works on Learning Behavioral Cloning Polices for Multi-Object Manipulation

VQ-BeT(Lee et al.)



IRASim (Zhu et al.)

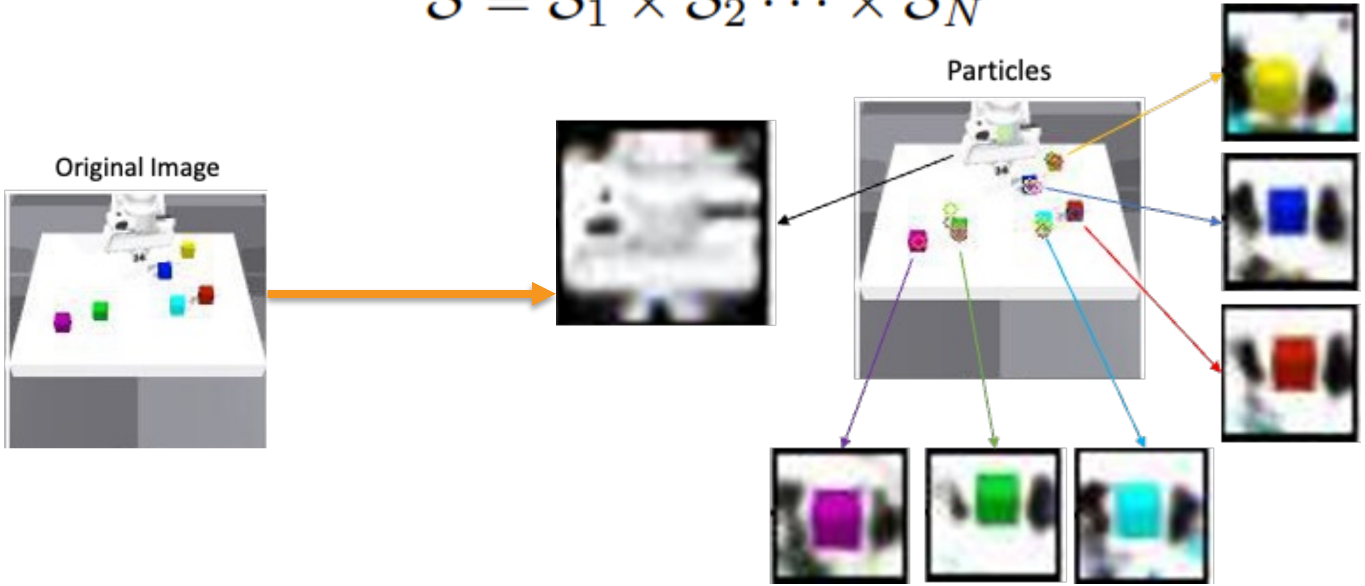


Zhu et al. IRASim: Learning Interactive Real-Robot Action Simulators, 2024.
Lee et al. Behavior Generation with Latent Actions, 2024.

Structures in Multi-Object Manipulation

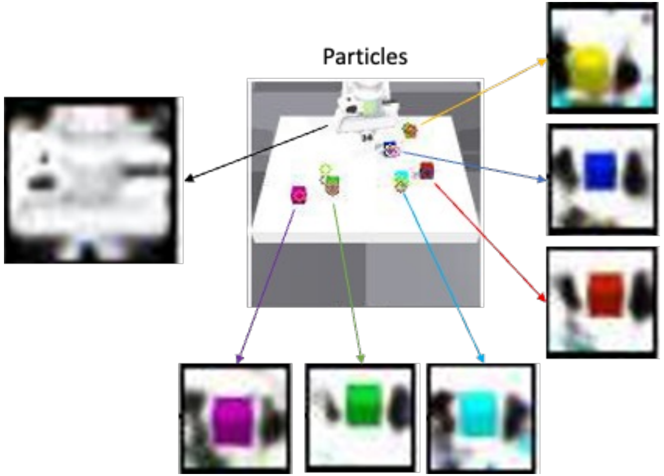
We consider consider a **factorized** “state” space

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \cdots \times \mathcal{S}_N$$

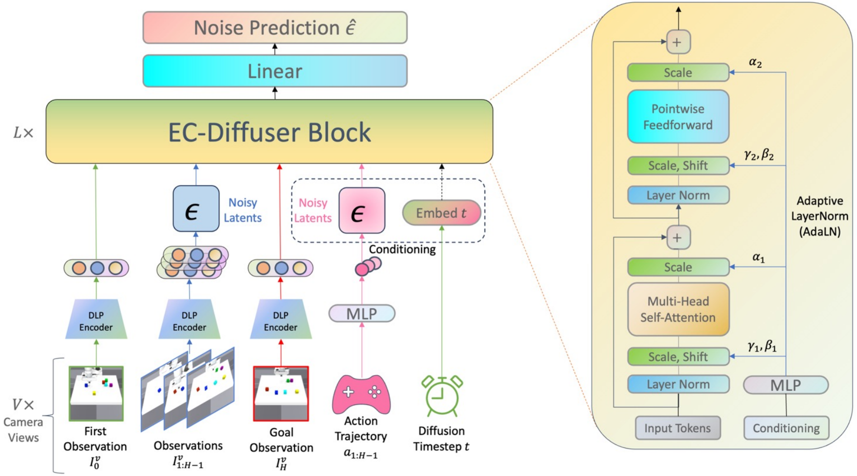


Method – Our Overall Pipeline

Learning an Object-Centric Representation

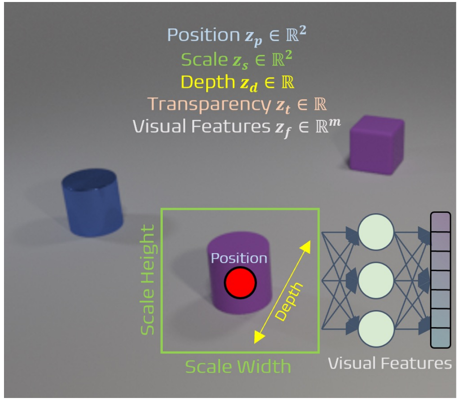


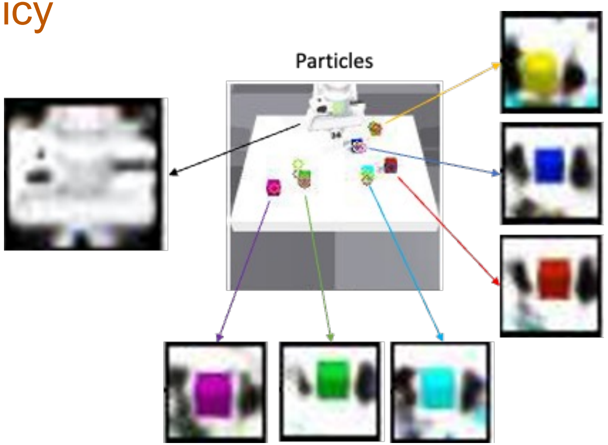
Learning an Entity-Centric Policy



Step 1 - Object-Centric Representation of Images

States, goals acquired from DLP encoder
Multi-view perception
Pre-trained on data collected with a random policy





$$z = \left\{ [z_{position}, z_{scale}, z_{depth}, z_{transparency}, z_{visual}]_i \right\}_{i=1}^{K} \in \mathbb{R}^{K \times (6+l)}$$

Daniel et al. Unsupervised Image Representation Learning with Deep Latent Particles, 2022.

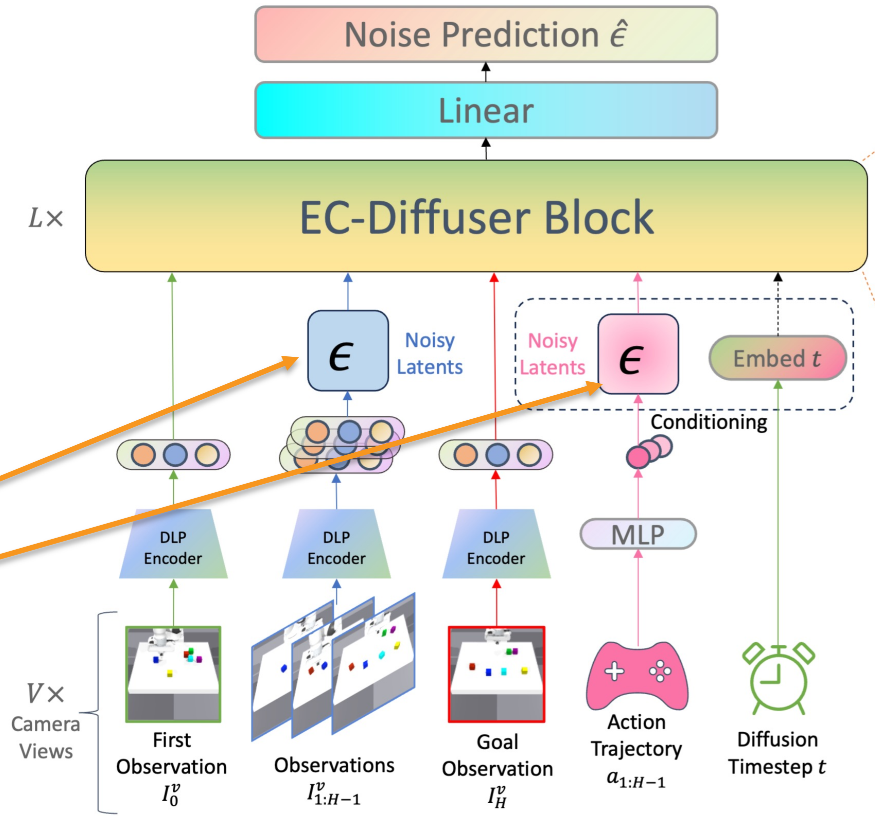
Step 2 – Entity-Centric Diffuser

Permutation-Equivariant transformer model that takes in particles and actions as input

Diffusion over particles + actions jointly

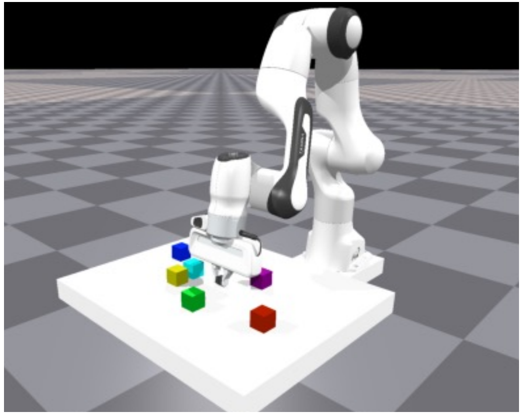
Trained with l1 loss on unordered particles

Diffused latents

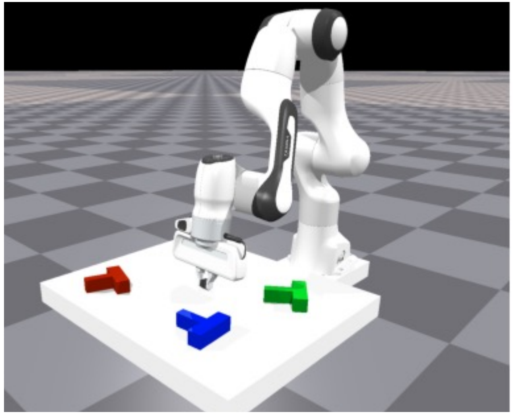


Experiment Results

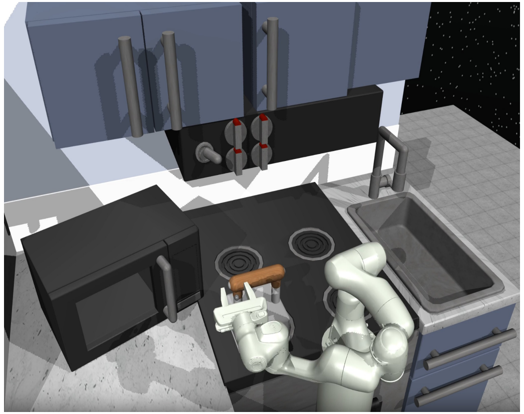
PushCube - Trained on 1-3 Cubes
PushT – Trained on 1-3 T-blocks
FrankaKitchen – trained on 4 objects



PushCube



PushT



FrankaKitchen

Experiment Results

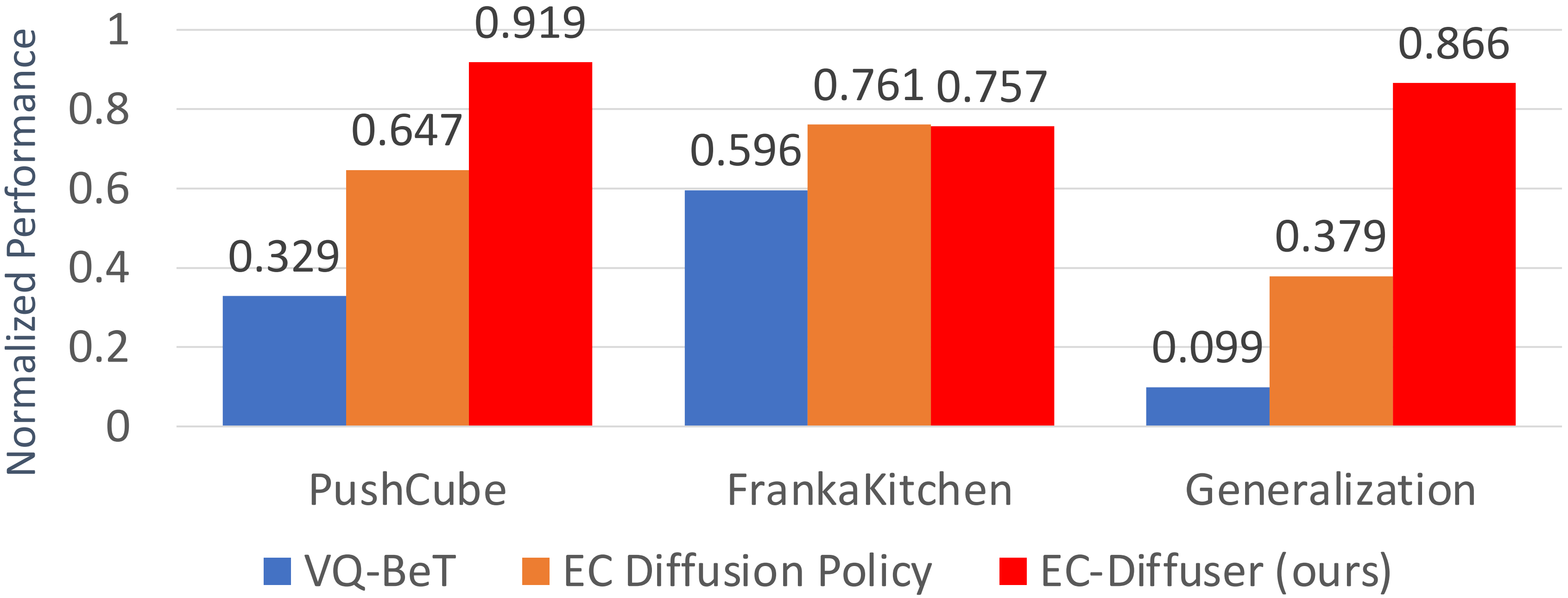
w/o Diffusion

Env (Metric)	# Obj	VQ-BeT	Diffuser	EIT+BC (DLP)	EC Diffusion Policy (DLP)	EC-Diffuser (DLP)
PushCube (Success Rate ↑)	1	**0.929 ± 0.032**	0.367 ± 0.027	0.890 ± 0.019	0.887 ± 0.031	**0.948 ± 0.015**
	2	0.052 ± 0.010	0.013 ± 0.011	0.146 ± 0.125	0.388 ± 0.106	**0.917 ± 0.030**
	3	0.006 ± 0.001	0.002 ± 0.004	0.141 ± 0.164	0.668 ± 0.169	**0.894 ± 0.025**
PushT (Avg. Radian Diff. ↓)	1	1.227 ± 0.066	1.522 ± 0.159	0.835 ± 0.081	0.493 ± 0.068	**0.263 ± 0.022**
	2	1.520 ± 0.056	1.540 ± 0.050	1.465 ± 0.034	1.214 ± 0.147	**0.452 ± 0.068**
	3	1.541 ± 0.045	1.542 ± 0.045	1.526 ± 0.047	1.538 ± 0.040	**0.805 ± 0.256**
FrankaKitchen (Goals Reached ↑)	-	2.384* ± 0.123	0.846 ± 0.101	2.360 ± 0.088	**3.046 ± 0.156**	**3.031 ± 0.087**

w/o Object-Centric structure

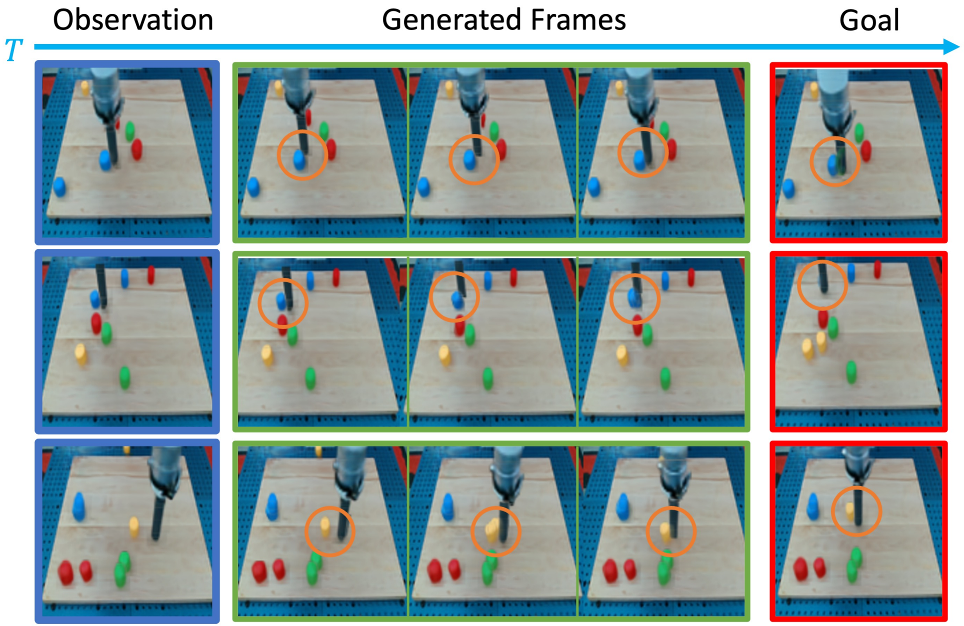
Generalization Results

Generalization Results



EC-Diffuser for Real World Problems

LanguageTable is a real-world dataset of object manipulation
EC-Diffuser can generate good future observations



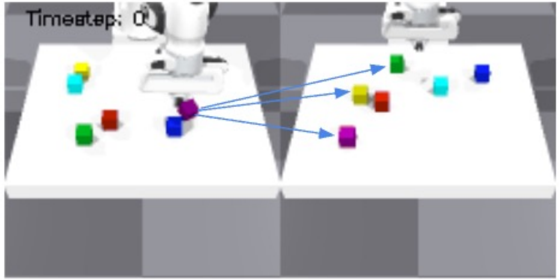
Future Research

Goal-Conditioned Reinforcement Learning with Sparse Rewards

$$J(\pi) = \mathbb{E}_{\substack{a_t \sim \pi(\cdot|s_t,g), g\sim p_g \\ s_{t+1}\sim \mathcal{T}(\cdot|s_t,a_t)}} \left[\sum_t \gamma^t r(s_t, a_t, g) \right]$$

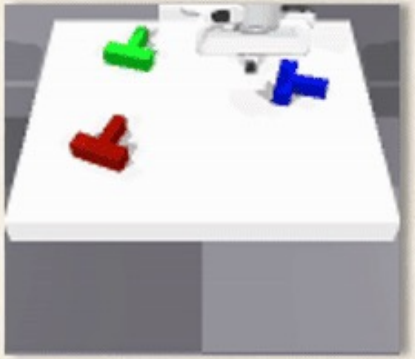
$$r_g(s_t, a_t, g) = \mathbb{1}(\text{the goal is reached})$$

$$\text{chamfer}(P_1, P_2) = \frac{1}{2n}\sum_{i=1}^{n} |x_i - \text{NN}(x_i, P_2)| + \frac{1}{2m}\sum_{j=1}^{n} |x_j - \text{NN}(x_j, P_1)|$$



Can we learn a structured reward model that generalizes across tasks?

EC-Diffuser



Train