XDex: Learning Cross-Embodiment Dexterous Grasping with 1000 Hands

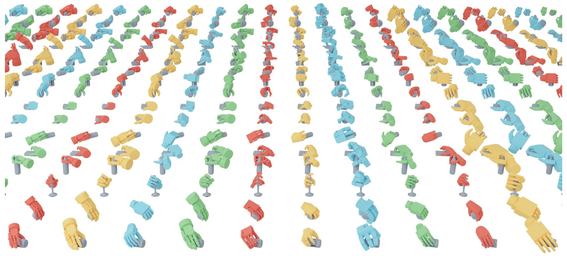
Keyi Wang* **Jianglong Ye*** **Chengjing Yuan** **Xiaolong Wang**

UC San Diego * Equal contribution keyiw.github.io/xdex

Why Cross-Embodiment Grasping?

- Learned grasp synthesis usually targets **one hand**: each new hand needs new data and a new model.
- **Retargeting** ignores the object, so transferred poses often penetrate or miss it.

Hypothesis: scaling the **number and diversity of embodiments** drives generalization to unseen hands.

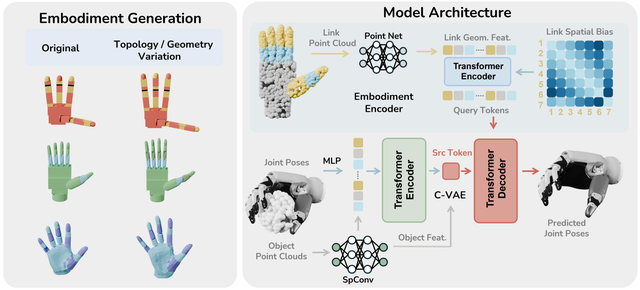


Rows: one grasp on one object, transferred across hands. Columns: one hand (color) on different objects.

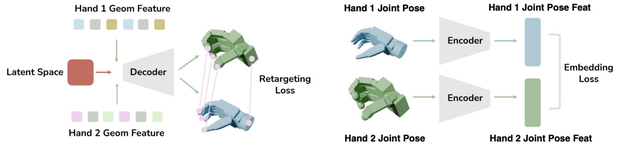
1,000	7k	200k
robot & human hands	object scenes	source grasps

- **Disentangles** **what** to grasp (object), **how** (pose latent z) and **with what** (hand tokens).
- **Embodiment encoder** over per-link geometry and URDF topology.
- **Largest cross-embodiment grasp dataset**, with grasps paired across hands.
- One model for **grasp synthesis** and **object-aware retargeting**.

1,000 Hands & Embodiment-Aware Model

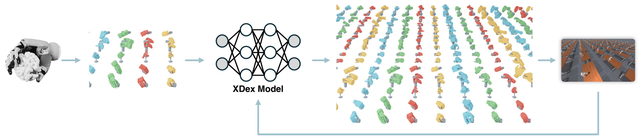


- **1,000 hands:** 8 robot + MANO bases with varied geometry & topology → **200 MANO + 800 robot**.
- **Embodiment encoder:** per-link PointNet + URDF-aware Transformer → link tokens.
- **CVAE:** link tokens decode joint poses from latent z + local object feature.



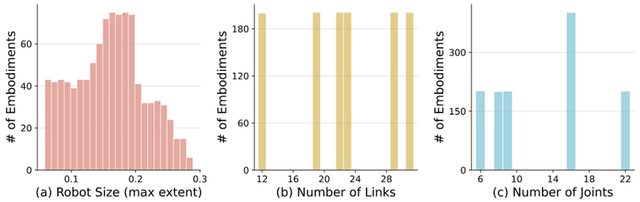
Consistency losses. Left: retargeting loss. Right: embedding loss.

Self-Improving Data Loop



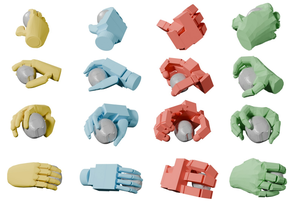
① Optimize **200k grasps** on one source hand (Inspire), 7k objects.
② Retarget to **6 sampled hands** per object → seed pairs.
③ Train an initial XDex model on the seed set.
④ Infer on **all 1,000 hands**, filter by IK + simulation, retrain.

Embodiment Distribution



The 1,000 generated hands: size (max extent), number of links, number of joints.

Grasp Synthesis Results



Disentangled grasp poses

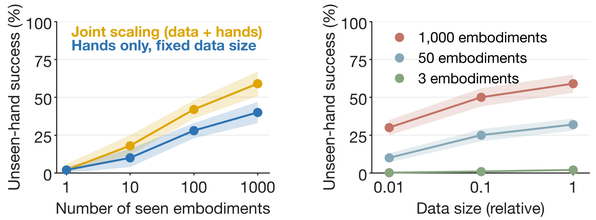
- Each **row**: same object point p and latent z, decoded on **different hands** → consistent grasps.
- Each **column**: different z on one hand → **diverse poses**.

Method	SR ↑	Pen ↓	Pos Diff ↓	Vec Diff ↓	H mean ↑
Retargeting	0.52 / 0.54	0.41 / 0.72	1.06 / 1.22	4.12 / 2.97	4.02 / 5.35
D(R,O)*	0.34 / 0.10	0.89 / 1.51	4.34 / 5.42	5.24 / 5.66	4.32 / 4.07
XDex (ours)	**0.77 / 0.59**	**0.36 / 0.53**	**1.21 / 1.84**	**3.02 / 3.04**	**5.09 / 5.50**
w/o paired data	0.73 / 0.50	0.37 / 0.55	1.28 / 2.72	3.03 / 3.87	5.06 / 5.26
w/o emb. encoder	0.74 / 0.57	0.39 / 0.59	1.23 / 1.91	3.11 / 3.15	5.09 / 5.13
w/o consist. loss	0.76 / 0.58	0.38 / 0.53	1.28 / 1.96	3.04 / 3.26	5.08 / 5.23

Seen / unseen hands. SR: success rate. Pen: penetration (cm). Pos/Vec Diff: cross-hand consistency (cm). H: joint entropy (bits).

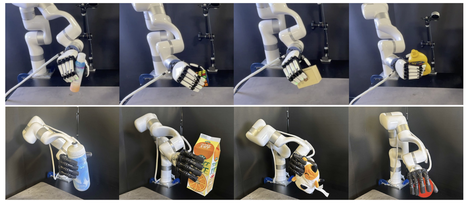
Best success on **seen (0.77)** and **unseen (0.59)** hands, and the most consistent of the **learned** methods. Removing paired data hurts unseen hands most.

Embodiment Scaling Law



Scaling data alone is **insufficient**: with 3 hands, unseen success stays near 0%. **Embodiment diversity** gets it to 59%.

Real-World Deployment



87.5% XHand (seen)

77.5% Ability (unseen)

xArm, single camera, open-loop; 8 objects × 5 trials per hand.

Object-Aware Neural Retargeting



- Human hand → robot hand, **conditioned on the object**.
- Real-world test with twin bottles.
- **Top:** keypoint-only retargeting fails.
- **Bottom:** **XDex succeeds**.