General Flow as Foundation Affordance for Scalable Robot Learning

Chengbo Yuan, Chuan Wen, Tong Zhang, Yang Gao

IIIS, Tsinghua University

Shanghai Qi Zhi Institute, Shanghai Artificial Intelligence Laboratory

Introduction

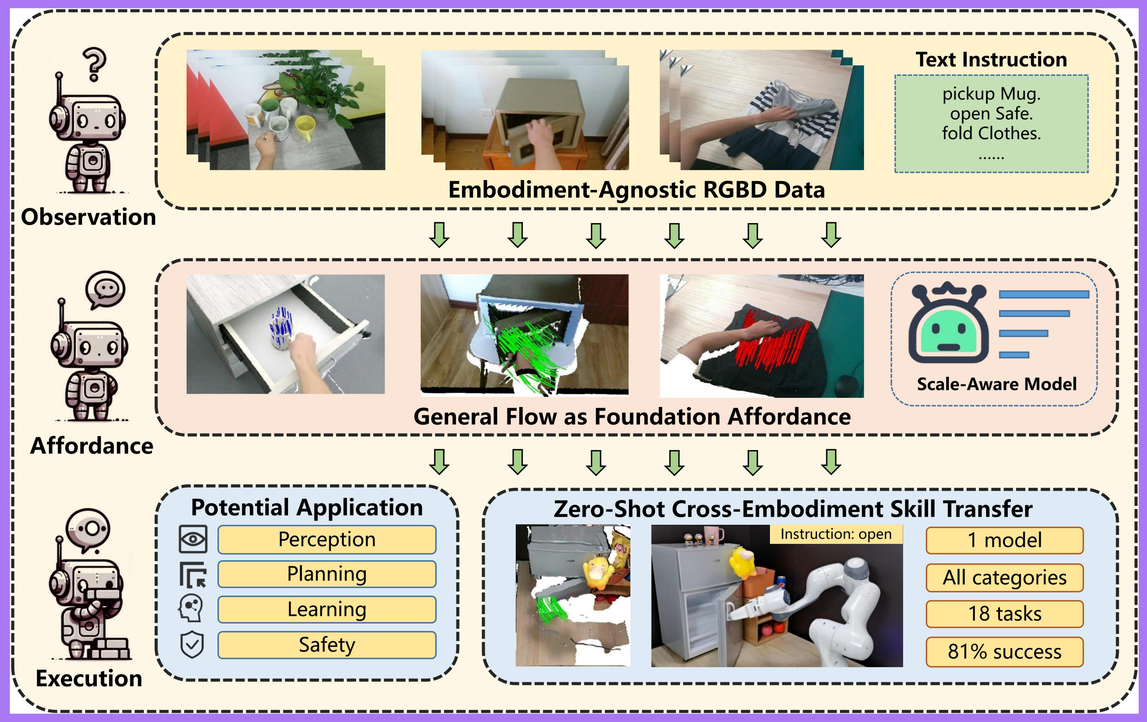
We address the challenge of acquiring real-world manipulation skills with a scalable framework.

We hold the belief that **identifying an appropriate prediction target** the is crucial for achieving efficient and universal learning. It should satiesfy

(1) scalability: could leverage cross-embodiment data resources.

(2) wide application: suitable for multiple object categories, including rigid, articulated, and soft bodies.

(3) stable skill transfer: providing actionable guidance with a small inference domain-gap.



Generalizable Multi-Task Robot Policy Derived Solely from Human Datasets

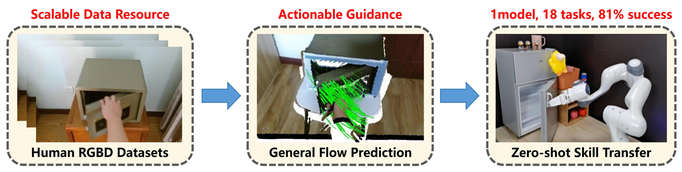
By leveraging 3D flow representation, we address the most challenging scenario for cross-embodiment robot learning:

(1) Only RGBD human video datasets

(2) Multi-task policy with only one model

(3) No robot data, zero-shot skill transfer

(4) Novel scene, novel object, novel view

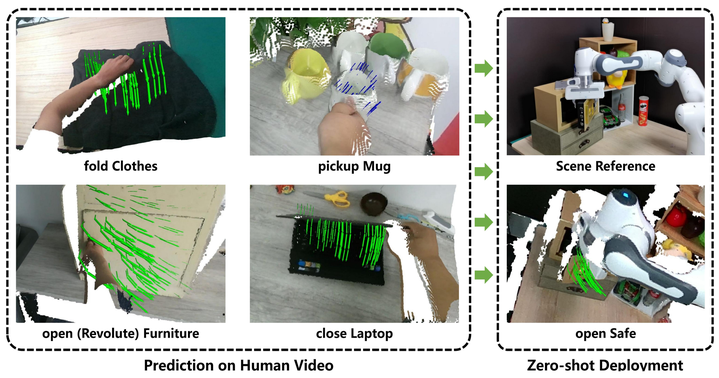


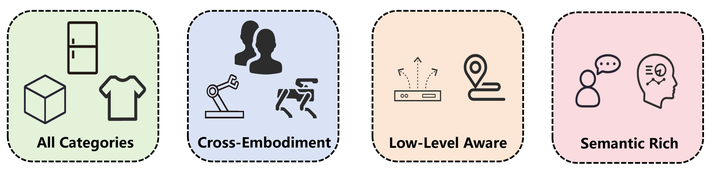
3D Flow as Affordance

We identify **3D flow** representation as foundation affordance and prediction target to achieve all of these demands.

We named it "general flow" because, in theory, it can be applied universally across any embodiment, any object category, and any task with instruction control.

Definition of General Flow Given a perception observation S (from any embodiment) and a task instruction I, for N_q 3D query points $Q \in R^{N_q \times 3}$ in space, the general flow $F \in R^{N_q \times T \times 3}$ represents the trajectories of these points over T future timestamps.





Application: Scalable Human-to-Robot Skill Transfer

To illustrate the power of 3D flow representation as a foundational affordance, we apply it under the most challenging condition: training exclusively **on a large-scale human dataset**, then deriving a **multi-task** policy in a **zero-shot** fashion.

Experiment: 18 tasks, 6 scenes, zero-shot human-to-robot skill transfer

For real-world human-to-robot skill transfer experiments, we select 8 objects (covering rigid, articulated and soft bodies) across 6 scenes, encompassing 18 manipulation tasks.

The experiment contain **novel objects, novel background, novel view, novel embodiment (human v.s. robot)**. Our method achieves 81% success rate in such challenge setting.

Object	Action-1	SR-1	Action-2	SR-2	Action-3	SR-3
Mug	pickup	10/10	putdown	9/10	-	-
Toy Car	pickup	10/10	putdown	10/10	push	5/10
Clothes	fold	8/10	-	-	-	-
Safe	open	9/10	close	10/10	-	-
Box	open	10/10	close	10/10	-	-
Drawer	open (pull)	4/10	open (grasp)	3/10	close	10/10
Refrigerator	open (pull)	7/10	open (grasp)	9/10	close	10/10
Laptop	open	5/10	close	7/10	-	-
Average Success Rate			**81% (146 / 180)**			

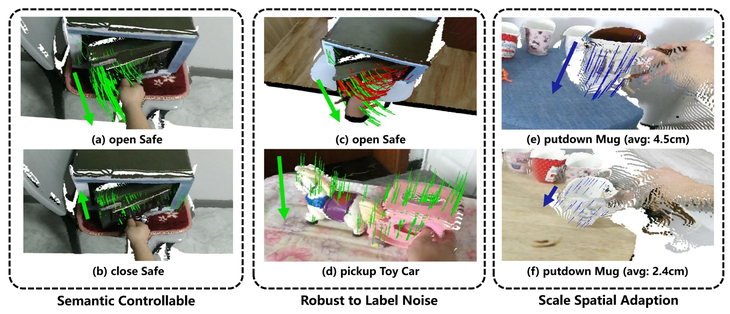
Emergent Properties from General Flow Training

We extract training data from HOI4D and self-collected datasets, and train a flow prediction model based on PointNeXt and CLIP backbone. During training, the model shows some emergent properties:

(1) Semantic controllable: predict varied flows for identical input point clouds when provided with different instructions.

(2) Resilience to label noise: despite some severe label noise our model accurately predicts the correct trend.

(3) Scale spatial adaption: dynamically adjusts its prediction scale in response to the spatial relationships of objects



Future Works

(1) Training a model for fast 4D reconstruction of monocular HOI videos (**will release soon 🚀🚀**).

(2) Scaling up 3D flow prediction model. Training in larger scale with RGB HOI videos.