General Flow as Foundation Affordance for Scalable Robot Learning

Chengbo Yuan[2345], Chuan Wen[123], Tong Zhang[123], Yang Gao[123†]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Shanghai Artificial Intelligence Laboratory, [3]Shanghai Qi Zhi Institute
[4]School of Computer Science, Wuhan University, [5]Hubei Luojia Laboratory
[†]Corresponding Author
michael.yuan.cb@whu.edu.cn, {cwen20, zhangton20}mails.tsinghua.edu.cn, gaoyangiiis@mail.tsinghua.edu.cn

https://general-flow.github.io/

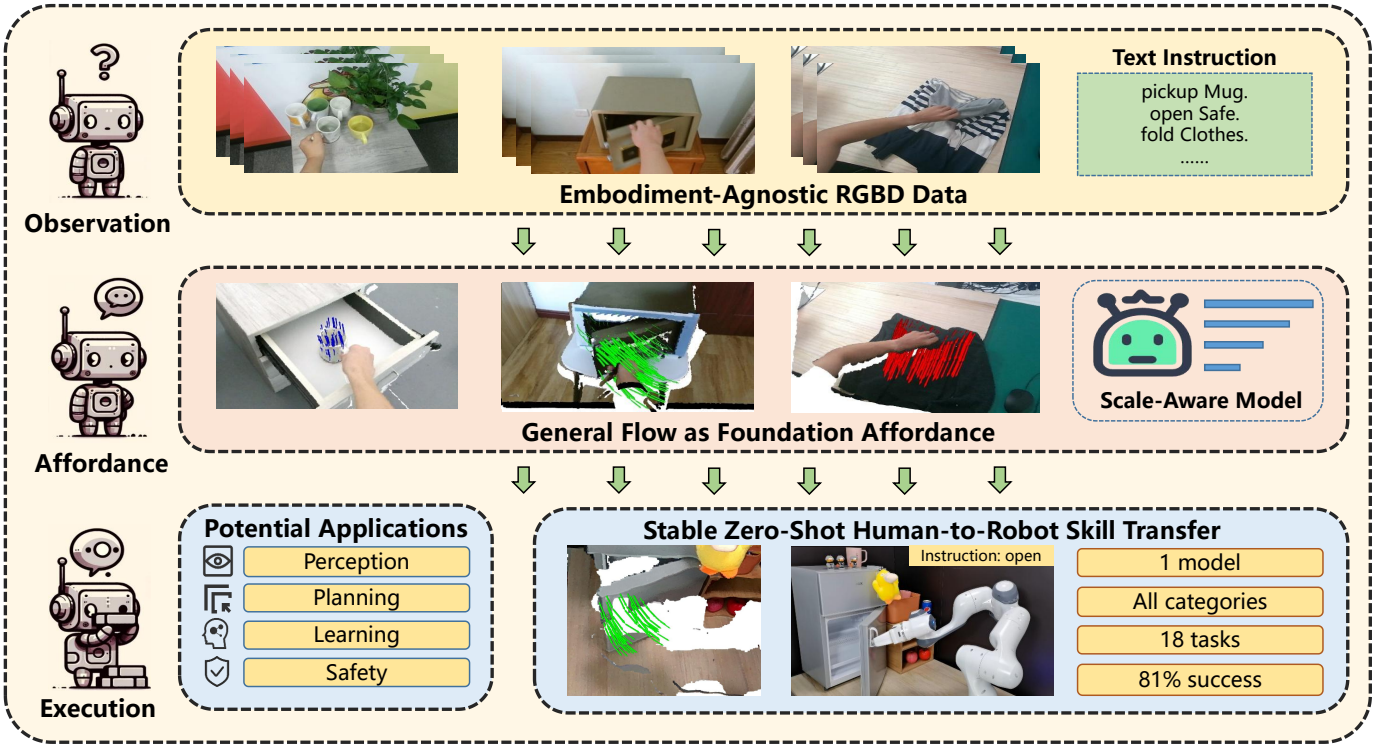


Fig. 1: We propose General Flow as Foundation Affordance. Its properties and applications are analyzed to reveal its great power. We design a scale-aware algorithm for general flow prediction and achieve stable zero-shot cross-embodiment skill transfer in the real world. These findings highlight the transformative potential of general flow in spearheading scalable general robot learning.

***Abstract*—We address the challenge of acquiring real-world manipulation skills with a scalable framework. Inspired by the success of large-scale auto-regressive prediction in Large Language Models (LLMs), we hold the belief that identifying an appropriate prediction target capable of leveraging large-scale datasets is crucial for achieving efficient and universal learning. Therefore, we propose to utilize flow, which represents the future trajectories of 3D points on objects of interest, as an ideal prediction target in robot learning. To exploit scalable data resources, we turn our attention to cross-embodiment datasets. We develop, for the first time, a language-conditioned prediction model directly from large-scale RGBD human video datasets. Our predicted flow offers actionable geometric and physics guidance, thus facilitating stable zero-shot skill transfer in real-world scenarios. We deploy our method with a policy based on closed-loop flow prediction. Remarkably, without any additional training, our method achieves an impressive 81% success rate in human-to-robot skill transfer, covering 18 tasks in 6 scenes. Our framework features the following benefits: (1) scalability: leveraging cross-embodiment data resources; (2) universality: multiple object categories, including rigid, articulated, and soft bodies; (3) stable skill transfer: providing actionable guidance with a small inference domain-gap. These lead to a new pathway towards scalable general robot learning. Data, code, and model weights will be made publicly available.**

I. Introduction

We aim to reveal a potential pathway for replicating the success of Large Language Models (LLMs) in the domain of robot learning. Specifically, we are interested in developing a new framework that enables scalable learning for robot manipulation. With more data and larger model training in the future, this framework has the potential to progressively enhance the capabilities of robots, i.e., the scaling law that has been observed in LLMs [82]. Inspired by the LLMs training paradigm [14], we believe that two key elements contribute to their strong generalization abilities: (1) a vast training dataset with a small inference domain gap, such as all texts from the internet in LLMs, and (2) a foundational prediction task with appropriate supervision signals, such as text-token prediction in LLMs. How can we translate these elements into robot learning?

Confronted with the challenges of collecting real-world robot data [46, 67], we pivot towards large-scale human datasets. These data resources guarantee scalability and a small inference domain-gap (no simulation-to-reality problem), key ingredients for effective generalization. Moreover, human manipulation data provides a vast, real-world dataset rich in diverse physics interactions and dynamic behaviors that closely align with robot manipulation. The next step is to identify a foundational prediction target conducive to widespread downstream tasks. We propose affordance for this role. Rooted in Gibson's theory [31], affordance concentrates on the potential actions associated with an object, remaining neutral to specific manipulators. This characteristic positions affordance as a cornerstone in cross-embodiment robot learning.

What affordance format will lead to a foundation prediction target that is universal for object categories and provides actionable geometric and physics guidance for downstream applications? In this paper, we propose **General Flow as Foundation Affordance** (as shown in Figure 1) to achieve this goal. This affordance elucidates the future trajectories of 3D points on the object of interest. Our key observation is that predicting keypoint motion is an efficient way to express geometric and physics information. Take the task of 'open Safe' as an instance (in the middle of Figure 1): the general flow represents future positions of points on the safe. Then, a robot can infer the safe's articulation by noting a static flow on the body and a moving flow on the door. It can also gain a resilient motion primitive for the opening skill by following the door's flow. We term our affordance "general flow" due to its capability for universal robot learning: **(1) scalability**: leveraging different embodiment data resources, e.g., humans and different robots; **(2) universality**: multiple object categories, including rigid, articulated, and soft bodies, with a large number of potential applications (see Section II-B). **(3) stable skill transfer**: providing actionable geometric and physics information with a small inference domain-gap, even sufficient for zero-shot execution.

In this paper, a novel framework is proposed to leverage general flow as the training target for foundational affordance learning. We first develop pipelines to extract 3D flow labels directly from RGBD human video datasets. We find prediction of dense flow in real-world scene point clouds remains a formidable challenge, primarily due to the variability of trajectory scales and the need to enhance robustness in zero-shot scenarios. To address these issues, we employ scale-aware strategies in both the data and model aspects, complemented by augmentation techniques that focus on embodiment occlusion (human hand and robot arm) and query point sampling (3D points on objects of interest), thereby boosting zero-shot stability. Remarkably, our model (named *"ScaleFlow"*), with fewer parameters, surpasses existing methods, setting a strong baseline for future research. Moreover, our system, trained at scale, demonstrates notable competencies such as language-driven semantic control, resilience to label noise, and spatial commonsense understanding.

To showcase the full potential of general flow affordance, we elect to tackle one of its most challenging applications: zero-shot cross-embodiment execution. Implementing a straightforward heuristic policy derived from closed-loop flow prediction, we evaluate our approach on a Franka-Emika robot in a real-world setup. Distinct from prior methodologies [4], without any additional training, our system accomplishes **stable zero-shot human-to-robot skill transfer**. Despite the simplicity of the derived policy, our approach, fueled by the rich actionable geometric and physics guidance of general flow affordance, notches **an impressive 81% average success rate in 18 diverse tasks across 6 scenes, covering multiple categories of object types like rigid, articulated, and soft bodies**. It is also noteworthy that our approach capably **handles complex environmental challenges,** such as robot segmentation errors, novel categories, and various robot configurations (initial gripper positions, grasp states, and manipulation directions), to a considerable extent. These findings highlight the transformative potential of general flow in spearheading scalable general robot learning.

In summary, our contributions can be concluded as follows:

- We introduce the framework of General Flow as Foundation Affordance, substantiating its feasibility and effectiveness, which is a new pathway towards scalable robot learning.
- We propose a robust, scale-aware algorithm that utilizes 3D flow labels derived from RGBD human video datasets, achieving remarkable accuracy in predicting complex real-world flow scenarios.
- We apply a simple heuristic policy, based on our model, to a Franka-Emika robot, successfully enabling stable zero-shot human-to-robot skill transfer across various object categories. This results in an impressive 81% success rate in 18 tasks across 6 scenes, marking a significant milestone in the real-world application of flow-based methods.

II. Related Work

A. Universal Robot Learning for General Manipulation

Real-World General Robot Learning Research in general-purpose robotic manipulation in real-world settings is constantly evolving, with a focus on integrating Large Language Models (LLMs) for high-level planning [13, 44, 37, 43, 19, 41] and exploring direct actionable guidance through LLMs [45], albeit with challenges due to overlooked physics dynamics. The development of large models for direct low-level control [11, 12, 9, 66] faces scalability issues due to intensive data requirements [46, 67]. This highlights the need for a training framework that achieves a balance between actionable output and scalability. In this work, we achieve this through a universal robot learning approach based on general flow prediction.

Embodiment-Agnostic Framework for Robot Learning To leverage large-scale, cross-embodiment data resources [62, 28, 16, 55, 27], multiple embodiment-agnostic frameworks [32] are proposed for robot learning. Prior works [65, 56, 57, 90] employ large-scale visual pre-training to develop embodied-aware pretrained representations, but these demonstrate limited generalization [58, 38, 42]. Alternative approaches seek to derive action signals from image or video generation [20, 35, 25, 23, 21, 50, 7], but these methods are resource-intensive and often yield redundant information. Affordances extracted from simulators [88, 63, 81, 30, 89, 29] are another focus, yet they struggle with the significant sim-to-real domain-gap, particularly in 3D environments. Recent efforts [3, 4, 61] attempt to directly acquire geometric-aware structured information in human video but require burdensome in-lab training. Bharadhwaj et al. [8] use hand poses as a bridge for guiding robot manipulation, but the limited geometric information extraction leads to unstable performance. Instead, we leverage 3D flow-based affordance to achieve reliable solutions. This leads to a stable skill transfer with an impressive 81% success rate in the real world.

Keypoints and Flow for Robot Learning Systems Utilizing keypoints and flow efficiently reveals the fine-grained geometric properties of the physical world, prompting researchers to harness their power in robot learning systems. Previous studies [59, 26, 24, 78, 91, 79, 60, 68] use keypoints as state descriptors for robot learning. The potential of flow prediction has also been noted [86, 76, 17, 2, 80]. However, these approaches either depend on embodiment-specific data, limiting their scalability [69, 74], or are simulation-based, facing significant sim-to-real domain-gap due to imperfect real-world RGBD point cloud generation [22, 93]. In this paper, we extend the work of Seita et al. [74], Eisner et al. [22] into a more general version (termed "general flow") both in data resources and downstream applications. We acquire a large-scale 3D flow prediction model directly from RGBD human video datasets and achieve stable zero-shot skill transfer. Besides, we propose a novel scale-aware architecture design and robust augmentation techniques to address the challenges of real-world flow prediction. Compared with previous models [70, 71, 72, 22, 74, 69], our methods achieve a significant performance enhancement.

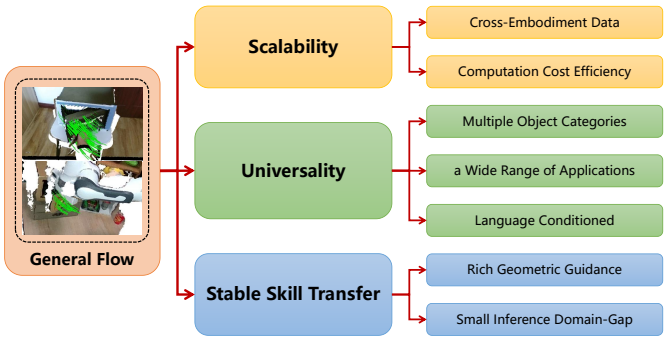


Fig. 2: General flow affordance offers scalability, universality, and stable skill transfer simultaneously, paving a new way for scalable general robot learning.

B. Potential Application for General Flow Prediction

General flow prediction (formally defined in the next section) can provide a foundational affordance prior [31, 15] for robot manipulation. Here, we outline some representative potential downstream applications. In **perception**, using algorithms like clustering, general flow can facilitate coarse part-level segmentation of objects [64] and serve as a prior for pose prediction and tracking [95, 83, 84]. For **planning**, robot motion planning can be executed [59, 26, 97, 22] based on flow prediction. Additionally, general flow can act as a strong policy prior [92] for **robot learning**, applicable in both imitation [74, 85] and reinforcement learning [4, 3]. In **safety** applications, the robot can detect anomalies by tracking keypoints and comparing them with predicted flow [36]. Finally, general flow can also provide semantic priors for other embodied tasks, such as Human-Object Interaction (HOI) synthesis [52].

Verifying all these applications is beyond the scope of this paper. Instead, we focus on the most challenging task: **stable real-world cross-embodiment zero-shot skill transfer**, a task not yet achieved by any existing flow-based work.

III. General Flow as Foundation Affordance

A. General Flow Affordance

Manipulation tasks typically consist of functional grasp and subsequent motion [1, 47]. In this paper, we mainly focus on the affordance [31] of later. We introduce "general flow" as an affordance that provides comprehensive, actionable guidance in terms of geometry and physics for downstream manipulation tasks:

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.*

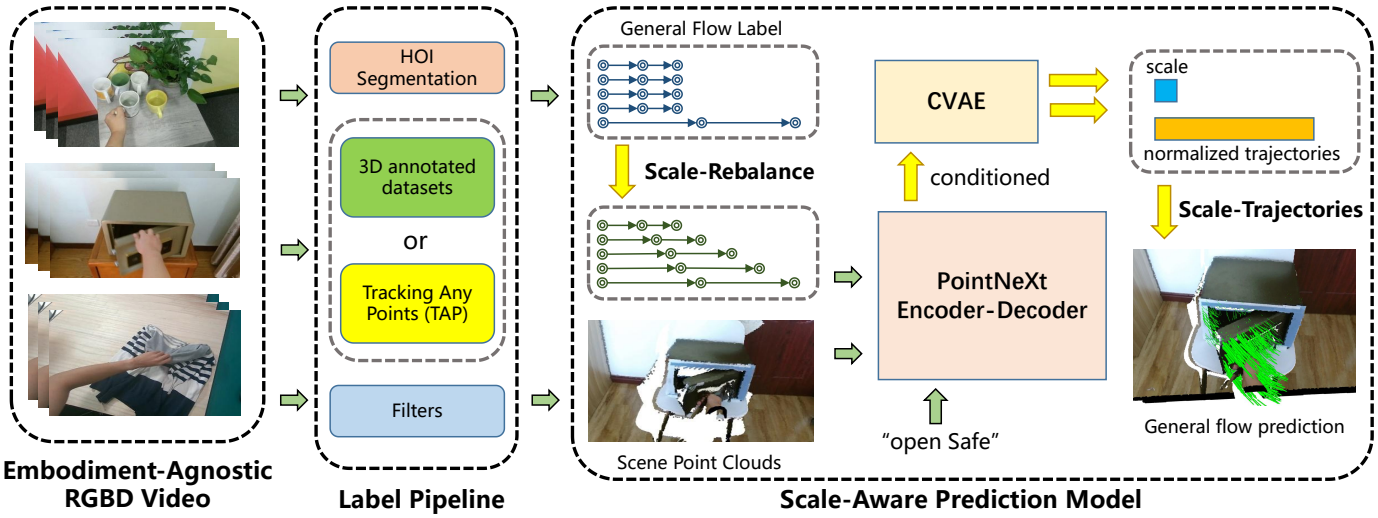


Fig. 3: The framework of our prediction model. We build pipelines to extract general flow labels from both RGBD human video datasets. Then, multiple key design elements are utilized to enhance the scale-awareness and robustness of the prediction model.

In this work, we use point clouds from real-world RGBD camera streams as our perception state S, eliminating sim-to-real transfer domain-gap concerns. We term our affordance "general flow" to emphasize its broad applicability across different embodiments, object categories, and downstream applications. Next, we delve into its distinct properties to highlight its potential and effectiveness.

B. Properties of General Flow

We explore the advantageous properties of the general flow affordance (see Figure 2) to demonstrate its capability as a bridge to a scalable, general robot learning framework. It provides several key benefits:

- **Scalability:** General flow allows for direct utilization of cross-embodiment data, such as human video datasets, circumventing the challenge of accessing large volumes of real robot data [74, 66]. Moreover, representing physics as future trajectories is a resource-efficient abstraction of motion dynamics, especially compared with full video generation [20, 50].
- **Universality:** General flow represents a unified abstraction of physical dynamics across multiple object categories, e.g., rigid, articulated, and soft bodies [29]. It also provides support for a wide range of applications (Section II-B). Additionally, its predictions are contingent on language instructions, enabling the execution of various behaviors within a single scene.
- **Stable Skill Transfer:** This benefit arises from two aspects. First, general flow offers richer geometric and physical guidance, especially compared to pretrained representations [65, 56] and coarse motion trends [4, 8]. Second, its reliance on real-world data eliminates any sim-to-real domain-gap issue. [22, 88].

Considering these points, we posit that general flow offers a scalable prediction target for foundation robot learning, similar to "text token" in Large Language Models (LLMs). Given that scaling up has led to strong generalization and emergent phenomena in LLMs [82, 14], we expect similar progressive enhancements in robot capabilities through larger-scale training. In the future, we aim to achieve this by harnessing larger RGBD datasets [34] recently released or by combining RGB video resources [16, 33] with depth estimation techniques [10, 50].

IV. Embodiment-Agnostic and Scale-Aware General Flow Prediction

In this section, we propose a framework for general flow prediction that is agnostic to specific embodiments, which is outlined in Figure 3. We first design pipelines to extract flow labels from RGBD human video datasets. To manage the variable scales of trajectories and account for real-world noise, we integrate key designs that enhance the model's scale-awareness and robustness in predictions. For more details on the label pipeline and training insights, please refer to Appendix B,C,D.

A. General Flow Label Acquisition

We introduce methods for acquiring general flow labels from two types of cross-embodiment datasets. All tools and pipelines will be open-source to benefit future research.

From 3D Annotated Datasets [51, 28, 55]: Utilizing the detailed 3D labels from these datasets, we first randomly sample points within the active object and then calculate its future position using ground-truth pose and camera parameters.

From Annotation-Free RGBD Videos [27, 34]: We first perform Human-Object-Interaction (HOI) segmentation [75, 96]

to obtain the active object mask. Points are then sampled within this mask, and the future 2D trajectory is tracked using the Tracking Any Point (TAP) tools [48]. The 3D label of the general flow is determined through back projection in both the spatial and temporal dimensions.

To reduce the effect of noise in the annotations and the pipeline, multiple techniques and filters are employed. Additionally, we retain the hand mask for potential use in subsequent training augmentations.

B. Scale-Aware Prediction Model

Our model processes natural language instructions I, scene point cloud features $P_s \in R^{N_s \times 6}$ (comprising N_s points with XYZ+RGB attributes), and N_q spatial query points $Q \in R^{N_q \times 3}$ (comprising N_q points with XYZ attributes). The aim is to predict a trajectory set, or "flow", denoted as $F \in R^{N_q \times T \times 3}$. For the i-th query point $p^i \in R^3$, its trajectory is defined as $F^i \in R^{T \times 3}$, with the absolute position at time t represented as $F_t^i \in R^3$ for $t = 1, 2, \cdots, T$. Initially, F_0^i is set to the input position of the query point p^i. We observe enhanced performance when predicting relative displacements rather than absolute positions. Thus, our refined goal is to predict $\Delta p_t^i = F_t^i - F_{t-1}^i$ for $t = 1, 2, \cdots, T$. The trajectory length for each query point p^i is defined as $Len(F^i) = \sum_{t=1}^{T} ||\Delta p_t^i||$.

A primary challenge in real-world flow prediction is the significant variance in trajectory lengths across different query points. For instance, in the "open Safe" task, the trajectories of points on the door are substantially longer than those on the safe body. To address this, we apply Total Length Normalization (TLN) to uniformly rescale trajectories. For the original prediction target $\{\Delta p_t^i \mid t = 1..T\}$, we define the scale L^i and normalized target $\{\Delta n_t^i\}$ as:

$$\Delta n_t^i = \frac{\Delta p_t^i}{L^i} \quad where \quad L^i = Len(F^i) \tag{1}$$

Our ablation study demonstrates that TLN yields the best performance compared with other normalization methods (Appendix F), leading to its adoption in subsequent experiments. The original prediction target F^i can be easily reconstructed from the predicted values of Δn_t^i and L^i.

Next, we describe our model's architecture (Figure 4). To facilitate multimodal control, we integrate instruction semantics early in the process. Instructions are converted into semantic features using a CLIP [73] encoder, then their dimensions are reduced via MLPs (to d_I) to align with point features. Features of scene point clouds P_s include 3D positions and RGB values, while query point clouds P_q substitute RGB values with a learnable embedding $E \in R^3$ (it serves as a "query identifier" and remains the same for all query points). We first concatenate aligned text features with point cloud features, then concatenate the features of scene points and query points, forming merged point cloud features $P_M \in R^{(N_s+N_q)\times(3+3+d_I)}$. The merged features are processed through a PointNeXt [72] backbone with a segmentation head to extract geometric features. The features of query points serve as condition variables for a conditional VAE [49, 40], generating the final predictions $\Delta\hat{n}_t$ and $\hat{L}$. For consistent scene semantics, a single latent variable z is sampled for all query points within the same scene during inference.

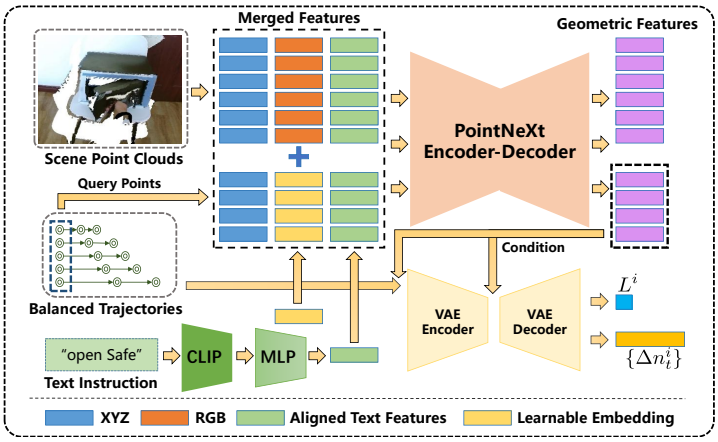


Fig. 4: Design architecture of our model.

C. Loss Calculation with Scale Rebalance

For skills such as 'open safe', where most query points are static (e.g., points on the safe's body), direct model training leads to a strong bias towards predicting stationary trajectories. This results from a scale imbalance in our datasets. To mitigate this, we implement scale rebalance across the dataset. First, we employ the K-Means algorithm to cluster each data point's general flow by scale L^i. As a result, we obtain N_r clusters of 3D points. We represent the original point ratios of each cluster as $\{r_i \mid i = 1..N_r\}$. Except for the cluster with the largest number of points, we perform resampling for all other clusters. The resampled distribution is given by:

$$\tilde{r}_i = \frac{e^{r_i/\tau}}{\sum_{i=1}^{N_r} e^{r_i/\tau}} \tag{2}$$

which is smoother than the original distribution. By default, we set τ to 1.

The final loss function comprises trajectory reconstruction loss $\mathcal{L}_{recon} = \frac{1}{N_q}\sum_{i,t} ||\Delta\hat{n}_t^i - \Delta n_t^i||^2$, scale regression loss $\mathcal{L}_{scale} = \frac{1}{N_q}\sum_{i,t} ||\hat{L}^i - L^i||^2$, and VAE KL-divergence loss $\mathcal{L}_{KL}$. To minimize cumulative error, we also incorporate an MSE loss $\mathcal{L}_{acc} = \frac{1}{N_q}\sum_{i,t} ||\hat{F}_t^i - F_t^i||^2$ for the recovered accumulative shift. Thus, the total loss is expressed as:

$$\mathcal{L} = \mathcal{L}_{recon} + \beta_1\mathcal{L}_{scale} + \beta_2\mathcal{L}_{KL} + \beta_3\mathcal{L}_{acc} \tag{3}$$

Our experiments underscore the importance of adequately weighting $\mathcal{L}_{scale}$ by setting $\beta_1 = 25$ and $\beta_2, \beta_3 = 1$, which is crucial for optimal performance.

D. Augmentations for Zero-Shot Robustness

In light of the complex environmental challenges encountered in zero-shot real-world deployments, we propose two technical augmentations to boost zero-shot generalization robustness:

- **Hand Mask (HM) Augmentation:** We encounter occlusions from human hands in our training data while facing occlusions from robot arms during deployments. Therefore, it is crucial to enhance the model's resilience to embodiment occlusions. To achieve this, we manipulate the presence of points on the hand in the input scene point clouds. We choose one of three rules, with probabilities p_{h1}=0.5, p_{h2}=0.2, and p_{h3}=0.3: (1) deleting all hand points; (2) keeping all hand points; and (3) sampling a random anchor point on the hand and retaining only points with a distance from the anchor greater than 12cm.
- **Query Points Sampling (QPS) Augmentation:** Different downstream applications may require varying query point sampling methods. Consequently, our model must be adaptable to various query point distributions. We achieve this by augmenting the training process. In each training iteration, we select a subset of available query points using one of two rules, based on probabilities p_{s1}=0.7, p_{s2}=0.3: (1) complete random sampling; (2) randomly selecting an anchor query point and then choosing a specific number of points closest to the anchor.

Our ablation studies validate that these augmentations, combined with scale rebalance, significantly improve zero-shot performance without adversely affecting in-domain prediction results (Appendix F and Table I,III).

V. General Flow Prediction Experiment

A. Experimental Setting

Dataset: For rigid and articulated objects, we utilize the HOI4D dataset [55] to train our general flow prediction model. This extensive RGBD video dataset includes 16 categories and 800 objects, encompassing 44.4 hours of recording. It provides comprehensive 3D labels, such as active object segmentation, 3D pose, and camera parameters. To further explore general flow in soft object manipulation, we collect RGBD videos for the "fold clothes" task using the RealSense D455 camera, comprising 6 garments, 30 rollouts, and 605 extracted clips.

Baseline: Given the absence of an identical problem setting in previous work, we adapt three types of relevant work to our setting:

- **2D Models:** To investigate the importance of 3D geometry information, we employ pretrained **ResNet** [39] and **Vision Transformer (VIT)** [18] models from the timm [87] library as feature extractors. We finetune these models, combining their 2D visual features with aligned text features and processing them through an MLP for direct flow regression. We also evaluate the performance of the **R3M representation** [65]. Both finetuning and frozen modes are considered for R3M.
- **VAT-MART [88]:** This model, originally designed for predicting affordance with single contact points, is adapted to our setting. We only utilize the 3D trajectory prediction branch of VAT-MART, replacing its task indentifier with aligned text features while keeping the rest of the model unchanged.
- **3D Backbones: FlowBot3D** [22] and **ToolFlowNet** [74] share a similar problem setting with ours. They originally used plain **PointNet++** [71] for flow prediction in simulation without language supervision. For fair comparison, we implement an improved version, replacing PointNet++ with the stronger **PointNeXt** [72] backbone as a geometric feature extractor. The extracted features, combined with aligned text features, are then processed through an MLP for general flow regression.

TABLE I: Results of general flow prediction, with the best outcomes highlighted in **bold** and the second-best outcome underlined. "ADE-H" and "FDE-H" denote evaluations that include hand points in the model's input. Even with fewer parameters, our model's performance is still significantly better than that of competitors.

Model	ADE↓	FDE↓	ADE-H↓	FDE-H↓	Param(M)
ResNet18	0.0754	0.1071	/	/	13.160
R3M(frozen)	0.0755	0.1056	/	/	11.861
R3M(finetune)	0.0754	0.1069	/	/	11.861
VAT-MART	0.0716	0.1220	0.0717	0.1220	1.577
VIT-B-224	0.0681	0.0948	/	/	86.614
PointNeXt-B	0.0396	0.0537	0.0392	0.0529	4.134
PointNeXt-L	0.0383	0.0516	0.0380	0.0512	15.583
ScaleFlow-S	0.0374	0.0501	0.0372	0.0498	0.906
ScaleFlow-B	0.0358	0.0477	0.0356	0.0474	5.622
ScaleFlow-L	**0.0355**	**0.0470**	**0.0352**	**0.0467**	17.088

We refer to our model as "**ScaleFlow**" in subsequent discussions. For 3D backbones and our model, we train multiple versions with varying model sizes. More details are available in Table I and Appendix C. To ensure a fair comparison, scale rebalance, HM augmentation and QPS augmentation are applied to all baselines.

Training Details: We utilize 1.5s video clips as our training data and set the time steps of general flow to 3 for all data sources. The dataset is divided into training, validation, and test sets in an 80%, 10%, 10% ratio, resulting in 51693, 6950, and 6835 clips respectively, with no identical object instances across sets. Each sample consists of 2048 scene points sampled in an $80\times80\times80$ cm^3 cube space around the center of the flow start points using the furthest point sampling (FPS) algorithm. During training, we randomly sample 128 query points, while for validation, 512 points are randomly sampled. Given that our datasets include ground-truth labels for object parts, we

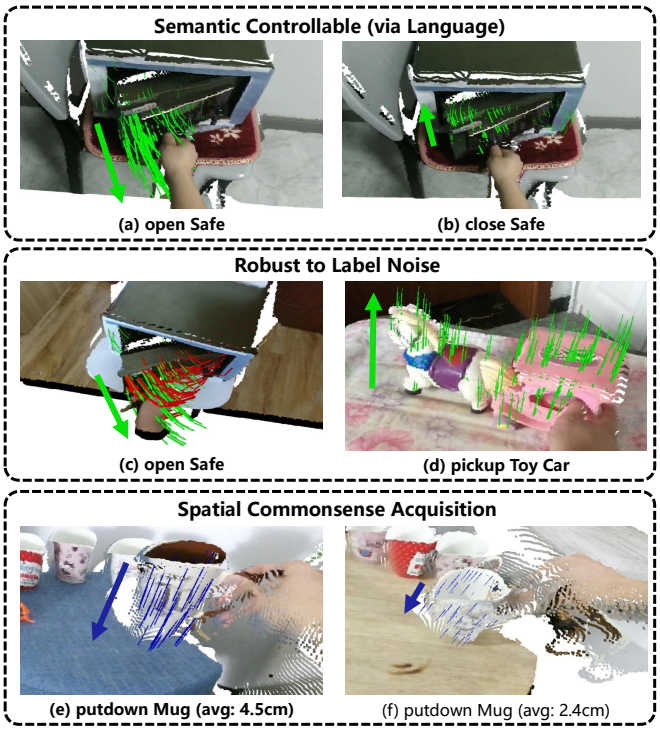


Fig. 5: Emergent properties of general flow prediction are demonstrated. The arrow indicates the coarse direction of the predicted flow. In images (a) and (b), the same input is used, differing only in the text instruction. For (c) and (d), the color red represents the extracted label, while green denotes the model's prediction. In (e) and (f), "avg" signifies the average trajectory lengths of all query points.

distribute 512 query points across each part equally during testing to enhance evaluation effectiveness. It's important to note that we do not use any part labels in model validation and real-word testing.

We train all the models with a batch size of 128, utilizing an AdamW optimizer and a learning rate of 0.001. For validation and testing, we set the batch sizes to 256. All models are trained for 200 epochs. Notably, the base version of ScaleFlow (ScaleFlow-B), which has 5.62M parameters, completes its training in 10 hours on a single NVIDIA GeForce RTX 3090 GPU.

Evaluation Metrics: We use 3D Average Displacement Error (ADE) and Final Displacement Error (FDE) in meters [5, 53] as evaluation metrics. For VAE-dependent models, metrics are averaged over 10 samplings. We also test robustness to hand occupancy in inputs across all 3D models, which are denoted as ADE-H and FDE-H.

B. Prediction Results

The results in Table I demonstrate that our model achieves superior performance on all metrics, even with fewer parameters. Overall, 3D models are superior to 2D models, indicating the importance of 3D geometry. Notably, with appropriate augmentations, the model is robust for hand occupancy. We show visualization of flow prediction in Appendix E. Furthermore, our **ablation study** (Appendix F) results quantitatively illustrate the effectiveness of all designs.

Through large-scale training, our model not only **captures rich semantic information** but also becomes adeptly **controllable through language modality**. As depicted in Figure 5(a)(b), our model demonstrates the capability to predict varied flows for identical input point clouds when provided with different instructions. Furthermore, it is remarkably **robust to label noise**. Figure 5(c)(d) showcases two instances of this resilience: despite severe label noise (notable deviation in "open Safe" and near-static in "pickup Toy Car"), our model accurately predicts the correct trend. Additionally, our model **gains spatial commonsense** through scalable training. It dynamically adjusts its prediction scale in response to the spatial relationships of objects, such as ending on the table and scaling up for longer distances, as seen in Figure 5(e)(f). All these emerging phenomena reveal the benefits of large-scale training.

Algorithm 1 Heuristic Close-Loop Policy from General Flow

Require: Task instruction I, camera stream $\mathcal{C}$, pretrained FastSAM model $\mathcal{M}_{seg}$, pretrained general flow predictor $\mathcal{M}_{flow}$, operation space controller $\mathcal{M}_{control}$.
$p_{base} \leftarrow$ 2D position of Franka-Emika base
$p_{extra} \leftarrow$ user interface (optional)
repeat
  $O_{rgbd} \leftarrow \mathcal{C}$
  $O_{seg} \leftarrow \mathcal{M}_{seg}(O_{rgbd}, \text{prompts}=[p_{base}, p_{extra}])$
  Recover Point Clouds $P_{scene} \leftarrow BackProject(O_{seg})$
  Gripper Position $g \leftarrow \mathcal{M}_{control}$
  Query Points $Q \leftarrow Radius(P_{scene}, g, 10\text{cm})$
  General Flow $F \leftarrow \mathcal{M}_{flow}(P_{scene}, Q, I)$
  SE(3) Transformation $\mathcal{T} \leftarrow$ SVD-Alignment(F)
  Execution: $\mathcal{M}_{control}(\mathcal{T})$
until Task Finished or Failed

VI. Zero-shot Real World Manipulation

In this section, we address one of the most challenging tasks: **stable zero-shot human-to-robot skill transfer in real-world scenarios**, demonstrating the foundational capabilities of general flow prediction. Utilizing **only a single prediction model** paired with a straightforward heuristic policy derived from closed-loop prediction, we achieve an impressive 81% average success rate. This success spans categories including rigid, articulated, and soft bodies and covers 18 tasks across 6 distinct scenes. Deriving more effective policies through approaches like few-shot imitation learning [85] or reinforcement learning [4] is left for future work.

A. Heuristic Policy with General Flow

Here we present our heuristic policy based on close-loop flow prediction (Algorithm 1). The fundamental idea is to treat the cluster of scene points near the gripper as a miniature rigid

TABLE II: Result of real-world manipulation with **one model for all tasks**. For the "open" task of "Storage Furniture", "pull" means execution with an opened gripper in a pulling manner, while "grasp" is with a closed gripper on the handle. Our method could achieve stable skill transfer among rigid, articulated, and soft bodies in most tasks.

Object	Action	Success Rate
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
Prismatic Furniture (Drawer)	open (pull)	4 / 10
	open (grasp)	3 / 10
	close	10 / 10
Revolute Furniture (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)**

body and forecast its future movements. Then we can derive a corresponding transformation for robot execution.

We use a RealSense D455 RGBD camera positioned behind the Franka-Emika Arm to capture an ego-view stream. The robot's static base during manipulation acts as a reference for the FastSAM model [96] to segment the robot. More prompt points or customized models [32] can also be employed. Post-segmentation, we reconstruct 3D scene point clouds and select query points near the gripper. These points, together with the scene point clouds and the text instruction, are fed into our ScaleFlow-B model to predict the general flow. The SVD algorithm [6] is used to align the robot arm's movement with the predicted flow. The Deoxys library [98] serves as an impedance controller for operation space control.

B. Real World Experiment Setting

For real-world experiments (Figure 6 and Appendix G), we select 8 objects (covering rigid, articulated and soft bodies) across 6 scenes, encompassing 18 manipulation tasks. The rigid category includes Mug and Toy Car. Articulated objects are Safe, Box (which can be approximated as an atypical design of Safe), Laptop, Refrigerator, and Drawer, while the soft category includes Clothes. We perform "pickup" and "putdown" actions for rigid objects, with an additional "push" action for the Toy Car. Articulated objects undergo "open" and "close" tasks, and the soft object is subjected to "fold" action. (See Figure 9 in Appendix G for visualization).

As general flow affordance guides post-grasp motion, we manually position the robotic arm for task initiation. This can be replaced with automatic methods, as demonstrated in Ko et al. [50]. For storage furniture with handles, we evaluate the performance of both opened and closed grippers. Each task undergoes 10 trials, and the success rates are recorded. Discussions about the quantitative success criteria of each task can be found in Appendix G. We also discuss the real-world baseline model [4] in Appendix H.

C. Results and Analysis

For result analysis, we focus on the following keywords and ask the following questions:

- **Transfer Ability:** Does general flow facilitate stable zero-shot human-to-robot skill transfer?
- **Segmentation Error:** How robust is the system against segmentation errors and robot occupancy?
- **Novel Category:** Can this model generalize to the shapes of new categories that are significantly different from the training instances?
- **Grasp Manner:** Is this object-centric system robust to variations in grasp state and gripper position?
- **Diverse Setting:** How well can general flow adapt to diverse scenes and manipulation directions?
- **Augmentation Effectiveness:** What impact do the designed technical augmentations have on enhancing the system's zero-shot capabilities?

Figure 6 presents a comprehensive overview of our analysis distribution. Following this, we delve into a detailed quantitative and qualitative examination to address these questions.

Stable Zero-Shot Skill Transfer Our results, presented in Table II, demonstrate that using a general flow as a bridge enables our framework to achieve stable zero-shot human-to-robot skill transfer. An impressive 81% success rate in such challenging settings underscores the strong transfer ability of general flow in cross-embodiment robot learning. To our knowledge, this is the first flow-based work to reach such a level of zero-shot transfer performance in real-world experiments. For tasks with success rates below 60%, we meticulously analyze the reasons and propose feasible future solutions in Appendix L.

Robustness to Segmentation Error Our findings reveal that random hand mask augmentation during training significantly enhances the model's robustness to errors in the segmentation maps of FastSAM [96]. Figure 6(a) illustrates this advantage with two examples. Notably, even with almost failed robot segmentation (as in the "open safe" task), our method still predicts meaningful flow to facilitate task completion in a closed-loop manner. Coupled with the with-hand prediction evaluation results (Table I), this evidences the strong adaptability of general flow to both human hand and incomplete Franka-Emika robot body occupancy. These findings lay the

Fig. 6: We achieve stable zero-shot human-to-robot skill transfer in the real world, encompassing 18 tasks with rigid, articulated, and soft objects across 6 scenes. It also demonstrates several key strengths of the general flow, such as **robustness to segmentation errors, adaptability to novel categories, and versatility in grasping manners and manipulation directions**. For additional insights, including execution videos and visualizations of general flow, please refer to Appendix E,G and project website.

groundwork for extending general flow to a broader spectrum of robot and human data.

Generalization to Novel Categories To probe the boundary of general flow's generalization capabilities, we experiment with a "Box" category, which can be approximated as an atypical design of "Safe", referred to "Box (Atypical Safe)". For comparison, we also test a conventional "Safe". Figure 6(b) presents these instances. Surprisingly, the success rate for manipulating "Box" is even higher than that of the ordinary one (100% vs. 90% for the "open" task, Table II), attributed to the structure of "Box" allowing more trajectory deviation without losing grip on the door. This underscores the strong generalization capacity of general flow methods, likely due to our model's training on real-world datasets, which avoids any simulation-to-reality domain-gap and focuses the model on the geometric and physics features of point clouds.

Robustness to Grasp Position and Manner As general flow is an embodiment-agnostic and object-centric method, it is expected to be resilient to variations in gripper position and grasp manner. To test this, we conduct manipulations using two storage pieces (a Refrigerator and a Drawer), leveraging their handles for different grasp methods. Figure 6(c) displays these different execution manners. Our model successfully completes tasks regardless of the gripper's state. The impact of the gripper being open or closed varies with the task; for instance, a closed gripper yields a higher success rate in "open Refrigerator". The closed gripper imposes constraints that prevent it from slipping off the door, and as a result, the impedance controller is then able to correct minor deviations in the action. Conversely, for "open Drawer", performance decreases due to the gripper slipping on the soft leather handle. We also ensure a diverse range of gripper positions, observing minimal impact on prediction and execution efficacy. The object-centric nature of general flow lays a solid foundation for leveraging diverse policy behaviors during execution.

Handling Diverse Scenes and Directions We examine the extent to which general flow can handle changes in scene and direction. Diverging from previous works' settings [22] (demonstrated in Appendix G), we distribute our tasks across six diverse scenes and perform scene-based prediction, eliminating the need for clean segmentation of the manipulated object. Despite these challenging conditions, our model proves robust against interference from environmentally irrelevant items. We also vary the direction of movable objects during our experiments. Figure 6(d) showcases the most challenging

TABLE III: Result of the ablation study in the real world. We take the success rate as the evaluation metric.

	open Safe	close Drawer	Avg
full	9 / 10	10 / 10	95%
w/o Scale Rebalance	7 / 10	8 / 10	75%
w/o HM Augmentation	8 / 10	6 / 10	70%
w/o QPS Augmentation	5 / 10	4 / 10	45%

example in this regard. We find that our heuristic policy successfully pushes a Toy Car in different directions to a certain extent. With more data training, we anticipate the model will achieve a deeper understanding of 3D visual mechanisms and semantics.

Role of Technical Augmentation Finally, we investigate the role of designed augmentations in enhancing the robustness of zero-shot transfer. We select two representative tasks ("open Safe" and "close Drawer") for evaluation. The ablation comparisons included:

- **w/o Scale Rebalance:** Utilizing the original flow label without rebalancing based on scale cluster results.
- **w/o HM Augmentation:** Setting $p_{h1} = 1.0$, which means erasing all points on the hand throughout the entire training process.
- **w/o QPS Augmentation:** Setting $p_{s1} = 1.0$, which means relying solely on random sampling for training query point selection.

The results in Table III indicate that each augmentation significantly contributes to robust zero-shot execution. Notably, hand augmentation markedly affects tasks with substantial embodiment occupancy and occlusion, such as "close Drawer". Query point sampling augmentation emerges as particularly influential. Currently, the PointNeXt [72] architecture inherently couples the extraction of query point features, leading to a reliance on query point sampling augmentation in our framework. We anticipate that future advancements in disentanglement architecture within 3D learning will address this issue thoroughly.

VII. Conclusion

In this paper, we introduce General Flow as a Foundation Affordance for scalable robot learning. For the first time, we develop a flow prediction model directly from large-scale RGBD human video datasets and successfully deploy it with a heuristic policy for stable zero-shot human-to-robot skill transfer. Our framework marks a stride in achieving scalability, universality, and stable skill transfer concurrently. We believe our work paves the way for innovative research in scalable general robot learning. In the future, we plan to further extend general flow learning to RGB datasets [16, 33], utilize depth estimation techniques [10, 50], and leverage larger RGBD datasets [34] that emerged concurrently with our work.

Acknowledgments

This work is supported by the Ministry of Science and Technology of the People´s Republic of China, the 2030 Innovation Megaprojects "Program on New Generation Artificial Intelligence" (Grant No. 2021AAA0150000). This work is also supported by the National Key R&D Program of China (2022ZD0161700). We would also like to thank Xianfan Gu for participating in the discussions of this project.

References

[1] Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. In *Conference on Robot Learning*, pages 3453–3467. PMLR, 2023.

[2] Artemij Amiranashvili, Alexey Dosovitskiy, Vladlen Koltun, and Thomas Brox. Motion perception in reinforcement learning with dynamic objects. In *Conference on Robot Learning*, pages 156–168. PMLR, 2018.

[3] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[4] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790, 2023.

[5] Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13702–13711, 2023.

[6] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992.

[7] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. *arXiv preprint arXiv:2312.00775*, 2023.

[8] Homanga Bharadhwaj, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Zero-shot robot manipulation from passive human videos. *arXiv preprint arXiv:2302.02011*, 2023.

[9] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. *arXiv preprint arXiv:2309.01918*, 2023.

[10] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

[11] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine

Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[12] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[13] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318. PMLR, 2023.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Dongpan Chen, Dehui Kong, Jinghua Li, Shaofan Wang, and Baocai Yin. A survey of visual affordance recognition based on deep learning. *IEEE Transactions on Big Data*, 2023.

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[17] Siyuan Dong, Devesh K Jha, Diego Romeres, Sangwoon Kim, Daniel Nikovski, and Alberto Rodriguez. Tactile-rl for insertion: Generalization to objects of unknown geometry. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6437–6443. IEEE, 2021.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[19] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[20] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *arXiv preprint arXiv:2302.00111*, 2023.

[21] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

[22] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *arXiv preprint arXiv:2205.04382*, 2022.

[23] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *arXiv preprint arXiv:2305.14343*, 2023.

[24] Peter Florence, Lucas Manuelli, and Russ Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5(2):492–499, 2019.

[25] Jialu Gao, Kaizhe Hu, Guowei Xu, and Huazhe Xu. Can pre-trained text-to-image models generate visual goals for reinforcement learning? *arXiv preprint arXiv:2307.07837*, 2023.

[26] Wei Gao and Russ Tedrake. kpam 2.0: Feedback control for category-level robotic manipulation. *IEEE Robotics and Automation Letters*, 6(2):2962–2969, 2021.

[27] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[28] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018.

[29] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023.

[30] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. Rlafford: End-to-end affordance learning for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5880–5886. IEEE, 2023.

[31] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[32] Kevin Gmelin, Shikhar Bahl, Russell Mendonca, and Deepak Pathak. Efficient rl via disentangled environment and agent representations. *arXiv preprint arXiv:2309.02435*, 2023.

[33] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[34] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. *arXiv preprint arXiv:2311.18259*, 2023.

[35] Xianfan Gu, Chuan Wen, Jiaming Song, and Yang Gao. Seer: Language instructed video prediction with latent diffusion models. *arXiv preprint arXiv:2303.14897*, 2023.

[36] Pinyao Guo, Hunmin Kim, Nurali Virani, Jun Xu, Minghui Zhu, and Peng Liu. Roboads: Anomaly detection against sensor and actuator misbehaviors in mobile robots. In *2018 48th Annual IEEE/IFIP international conference on dependable systems and networks (DSN)*, pages 574–585. IEEE, 2018.

[37] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *Conference on Robot Learning*, pages 3766–3777. PMLR, 2023.

[38] Nicklas Hansen, Zhecheng Yuan, Yanjie Ze, Tongzhou Mu, Aravind Rajeswaran, Hao Su, Huazhe Xu, and Xiaolong Wang. On pre-training for visuo-motor control: Revisiting a learning-from-scratch baseline. *arXiv preprint arXiv:2212.05749*, 2022.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[40] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

[41] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning. *arXiv preprint arXiv:2311.17842*, 2023.

[42] Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. *arXiv preprint arXiv:2304.04591*, 2023.

[43] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[44] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

[45] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[46] Dmitry Kalashnikov, Jacob Varley, Yevgen Chebotar, Benjamin Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. *arXiv preprint arXiv:2104.08212*, 2021.

[47] Aditya Kannan, Kenneth Shaw, Shikhar Bahl, Pragna Mannam, and Deepak Pathak. Deft: Dexterous fine-tuning for real-world hand policies. *arXiv preprint arXiv:2310.19797*, 2023.

[48] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023.

[49] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[50] Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B Tenenbaum. Learning to Act from Actionless Video through Dense Correspondences. *arXiv:2310.08576*, 2023.

[51] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021.

[52] Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913*, 2023.

[53] Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292, 2022.

[54] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[55] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022.

[56] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[57] Yecheng Jason Ma, William Liang, Vaidehi Som, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. *arXiv preprint arXiv:2306.00958*, 2023.

[58] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Pieter Abbeel, Jitendra Malik, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *arXiv preprint arXiv:2303.18240*, 2023.

[59] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019.

[60] Lucas Manuelli, Yunzhu Li, Pete Florence, and Russ Tedrake. Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning. *arXiv preprint arXiv:2009.05085*, 2020.

[61] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint arXiv:2308.10901*, 2023.

[62] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2019.

[63] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on Robot Learning*, pages 1666–1677. PMLR, 2022.

[64] Shujon Naha, Qingyang Xiao, Prianka Banik, Md Alimoor Reza, and David J Crandall. Part segmentation of unseen objects using keypoint guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1742–1750, 2021.

[65] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[66] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. https://octo-models.github.io, 2023.

[67] Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.

[68] Chuer Pan, Brian Okorn, Harry Zhang, Ben Eisner, and David Held. Tax-pose: Task-specific cross-pose estimation for robot manipulation. In *Conference on Robot Learning*, pages 1783–1792. PMLR, 2023.

[69] Carl Qi, Sarthak Shetty, Xingyu Lin, and David Held. Learning generalizable tool-use skills through trajectory generation. *arXiv preprint arXiv:2310.00156*, 2023.

[70] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[71] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[72] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in Neural Information Processing Systems*, 35:23192–23204, 2022.

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[74] Daniel Seita, Yufei Wang, Sarthak J Shetty, Edward Yao Li, Zackory Erickson, and David Held. Toolflownet: Robotic manipulation with tools via predicting tool flow from point clouds. In *Conference on Robot Learning*, pages 1038–1049. PMLR, 2023.

[75] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[76] Bokui Shen, Zhenyu Jiang, Christopher Choy, Leonidas J Guibas, Silvio Savarese, Anima Anandkumar, and Yuke Zhu. Acid: Action-conditional implicit visual dynamics for deformable object manipulation. *arXiv preprint arXiv:2203.06856*, 2022.

[77] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[78] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022.

[79] Yin-Tung Albert Sun, Hsin-Chang Lin, Po-Yen Wu, and Jung-Tang Huang. Learning by watching via keypoint extraction and imitation learning. *Machines*, 10(11): 1049, 2022.

[80] Mel Vecerik, Carl Doersch, Yi Yang, Todor Davchev, Yusuf Aytar, Guangyao Zhou, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. *arXiv preprint arXiv:2308.15975*, 2023.

[81] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas J Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d artic-

ulated objects via few-shot interactions. In *European Conference on Computer Vision*, pages 90–107. Springer, 2022.

[82] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[83] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074. IEEE, 2021.

[84] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 606–617, 2023.

[85] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

[86] Thomas Weng, Sujay Man Bajracharya, Yufei Wang, Khush Agrawal, and David Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In *Conference on Robot Learning*, pages 192–202. PMLR, 2022.

[87] Ross Wightman. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[88] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects. *arXiv preprint arXiv:2106.14440*, 2021.

[89] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. *arXiv preprint arXiv:2303.11057*, 2023.

[90] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[91] Ruinian Xu, Fu-Jen Chu, Chao Tang, Weiyu Liu, and Patricio A Vela. An affordance keypoint detection network for robot manipulation. *IEEE Robotics and Automation Letters*, 6(2):2870–2877, 2021.

[92] Weirui Ye, Yunsheng Zhang, Mengchen Wang, Shengjie Wang, Xianfan Gu, Pieter Abbeel, and Yang Gao. Foundation reinforcement learning: towards embodied generalist agents with foundation prior assistance. *arXiv preprint arXiv:2310.02635*, 2023.

[93] Harry Zhang, Ben Eisner, and David Held. Flowbot++: Learning generalized articulated objects manipulation via articulation projection. *arXiv preprint arXiv:2306.12893*, 2023.

[94] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022.

[95] Shaobo Zhang, Wanqing Zhao, Ziyu Guan, Xianlin Peng, and Jinye Peng. Keypoint-graph-driven learning framework for object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1065–1073, 2021.

[96] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[97] Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling Wang, Pengfei Li, Guyue Zhou, Chao Yang, et al. 3d implicit transporter for temporally consistent keypoint discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3869–3880, 2023.

[98] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. *arXiv preprint arXiv:2210.11339*, 2022. doi: 10.48550/arXiv.2210.11339.

APPENDIX

A. Overview of Appendix

In this appendix, we offer additional implementation details and a discussion of general flow. *The label pipeline, code and model weights will be released in the future*. Readers are also welcome to check out more details with them. This appendix is structured as follows:

- **Label Extraction:** we delve into the pipeline specifics utilized for extracting general flow labels from RGBD human video datasets in App. B.
- **Model Architecture and Training:** we provide an in-depth look at our model's architecture (App. C) and the details of its training process (App. D).
- **Flow Visualizations:** We present visualizations of general flow predictions in both in-domain data and zero-shot real-world executions in App. E.
- **Ablation Study:** A comprehensive quantitative ablation study is conducted in App. F, rigorously testing the effectiveness of our algorithmic design.
- **Real-World Experiment:** This section is dedicated to an expansive elucidation of real-world experiments, encompassing the experimental setup (App. G), baseline comparisons (App. H), robot system development (App. I), policy derivation strategies (App. J), inference latency measurements (App. K), and analysis of failure cases (App. L).
- **Limitations and Future Directions** Insights into the current limitations of our approach and prospective directions for improvement are shared in App. M.
- **Codebases:** Acknowledgements are extended to the multiple codebases that have been instrumental in supporting this project in App. N.

Additional videos and flow visualizations are included on project website. For guidance on viewing these materials, please refer to the *'README.pdf'*.

B. Label Extraction Pipeline

General flow labels can be directly extracted from 3D human datasets or RGBD videos. Figure 7 displays some data resources we utilize.

B.1 From 3D Annotated Datasets

We select the HOI4D dataset [55] as our primary resource due to its relatively large scale. This dataset offers comprehensive 3D labels, which are crucial for supporting 4D (point clouds + timestamps) Human-Object-Interaction (HOI) research. The labels we employ include RGBD images, camera parameters, object pose labels, scene segmentation masks, and action labels.

For effective closed-loop control, we divide the original action clips into multiple 1.5-second sub-clips, spaced at 0.15-second intervals, with a total of 3 time steps. For sub-clips that contain non-contact prefix actions (such as moving hands towards objects), we create 4 extended sub-clips with start timestamps within the semantic-less prefix.

Fig. 7: Examples of our cross-embodiment data resource

The model's input comes from the first image of each sub-clip. We identify and match key elements (objects and hands involved in the manipulation) from the instructions with their corresponding masks, considering the remainder as a background mask. Each mask is converted into a point cloud from RGBD values and down-sampled to one point per 0.02cm voxel. To adjust for noise in the HOI4D masks, we expand the hand mask by 8 pixels and shrink the object masks by 2 pixels.

We then proceed to extract general flow labels. Initial query points are selected within the masks of the objects of interest. Addressing segmentation noise, we maintain only the overlapping masks from the previous, current, and subsequent frames, using a homography matrix for projection onto the current frames. These points are chosen randomly, and their future trajectories are calculated based on ground-truth poses. We project all data back to the initial frame using the camera parameter labels. To correct for camera shake in the extrinsic parameter labels, we identify trajectories with shifts under 0.02 cm, compute their average as the camera shake, and deduct this from all points.

B.2 From RGBD Human Videos

Given the constraints of current RGBD Human-Object Interaction (HOI) datasets, which are either small in scale [28] or lack semantic richness [27] (mainly limited to pick & place actions), and considering the notable scarcity of resources for soft objects, we opt to collect our own RGBD videos. Coincidentally, Grauman et al. [34] releases a large-scale, semantically rich 4D HOI dataset during the same period as our project, representing a promising resource for future work. Our collection of RGBD videos for the "fold Clothes" task, captured with a D455 depth camera, includes 30 rollouts of 6 different types of clothing, resulting in 605 extracted clips. We plan to release this dataset in the future.

Echoing the process used in HOI4D, we maintain a duration of 1.5 seconds for each clip, with intervals of 0.15 seconds. Initially, we apply HOI segmentation [94] to acquire masks for hands and active objects. Utilizing HOI detection results from [75], we input bounding box outputs into FastSAM [96]

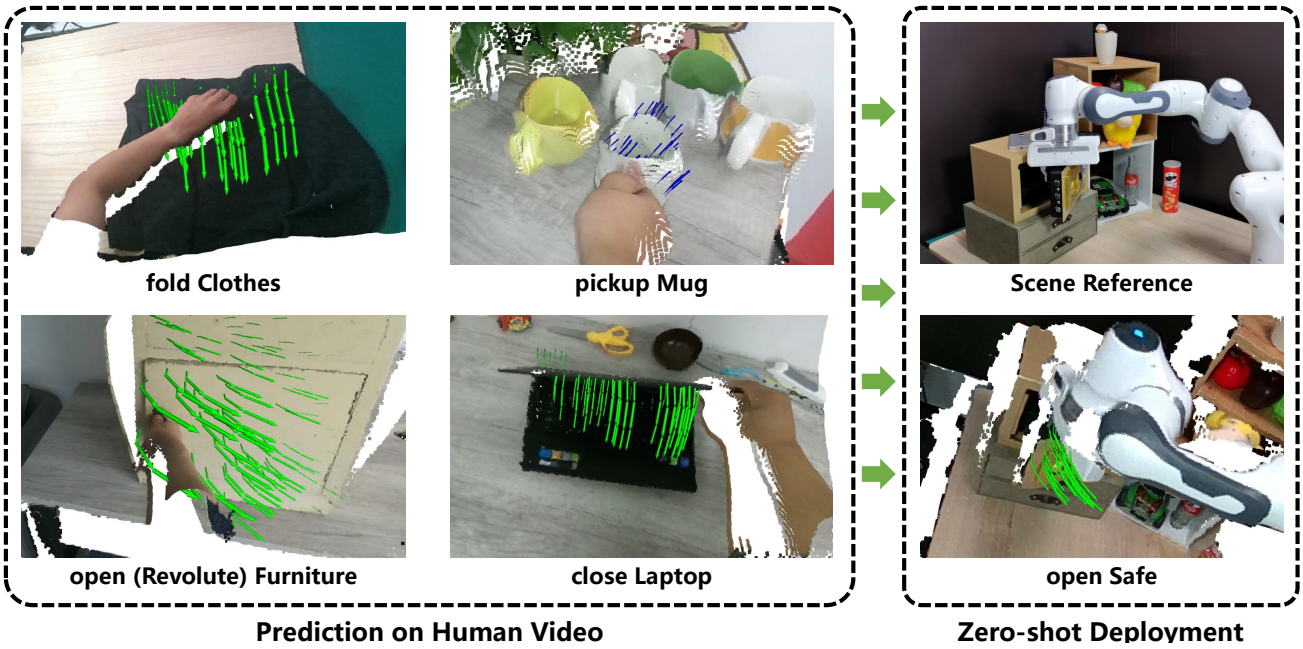


Fig. 8: Visualization of flow prediction. For additional insights, including more human video predictions and zero-shot results for *all 18 tasks*, please refer to project website.

for refined results. We retain only the masks with a confidence level above 0.5 for subsequent processing. After segmentation, we randomly sample 1024 points on the active object and employ co-tracker [48] to track their future positions in 2D pixel space. We exclude trajectories affected by occlusion, disappearance, or breakdown of depth values midway, and project the remaining trajectories back to 3D space and the first frame of the clip to derive the final general flow labels.

Our pipeline functions automatically, without the need for manual intervention. As general flow captures the geometric dynamics of the physical world, extending beyond mere object-centric interactions, our system effectively manages noise factors such as segmentation and point sampling errors (e.g., selecting query points on non-target objects due to segmentation mistakes), particularly during large-scale training.

C. Model Architecture

ScaleFlow: While the main body of our paper covers the bulk of our design, we offer further details in this section. The alignment width for CLIP [73] text features is set at 6, aligning with the dimension of the original point cloud features (RGB+XYZ). In the conditional Variational Autoencoder (VAE) [49] segment, we utilize a 2-layer Multilayer Perceptron (MLP) to encode the latent variable. This is followed by another 2-layer MLP that functions as the VAE decoder. We employ separate 2-layer MLPs for scale and normalized-trajectory prediction, each featuring a hidden dimension of 512. For ScaleFlow-B, our backbone configuration mirrors that of PointNeXt-B [72]. In ScaleFlow-L, we increase the backbone width from 32 to 64. Conversely, for ScaleFlow-S, PointNeXt-B is substituted with PointNeXt-S, and the CVAE utilizes a simpler 1-layer encoder and decoder, each with a hidden dimension of 384. The loss function parameters are configured as $\beta_1 = 25$ and $\beta_2, \beta_3 = 1$, with a focus on enhancing scale prediction. We maintain the latent variable dimension at 16. For more detailed information, please consult the configuration files in our code repository.

Baseline: We preserve the architecture from the original repository, with the sole modification of incorporating text features before the prediction MLP to transition the model into a multimodal version. We derive aligned text features from the original CLIP features. The dimension of these text features is set to 32. For ResNet [39] and Vision Transformer [18], we utilize the standard 'ResNet18' and 'VIT-B-224' versions, respectively. The default pretrained weights are loaded using the Timm library [87]. For R3M [65], its 'ResNet18' version is employed. In our architecture, all MLPs dedicated to the final flow prediction consist of 2 layers, featuring hidden dimensions of 512 and 256, respectively.

D. Training Details

This section outlines additional training details not covered in the main text. For each data point, we start by calculating the center of the initial points of the general flow. We then create a cubic space centered on this point, with each side measuring 80 cm (a 40 cm perception range suffices for most tasks). To standardize point numbers for batch training, we use sampling (with the Furthest Points Sampling Algorithm) or re-sampling to obtain 2048 scene points. Standard augmentation techniques for point cloud prediction [72], including random rotation, shifting, scaling, coordinate normalization, color jittering, and feature dropping, are applied.

In each training iteration, we randomly select 128 trajectories from the available flow labels. Additionally, to boost robustness in downstream zero-shot prediction, we implement technical augmentations as described in the main paper. For scale rebalance, we set the number of clusters to 4 and the default temperature to 1. The training process utilizes the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.0001. We incorporate 10 warmup epochs, followed by a cosine scheduler for 200 epochs. The training for ScaleFlow-B can be completed within 10 hours using an Intel(R) Xeon(R) Gold 5220R CPU and a single NVIDIA GeForce RTX 3090 GPU.

E. Flow Visualization

We showcase the capabilities of our model, ScaleFlow-B, through visualizations of its predictions. These demonstrations include applications on original human videos as well as zero-shot real-world executions, as depicted in Figure 8. The visualizations highlight our model's proficiency in predicting semantically and geometrically meaningful flows, even amidst the challenges posed by noisy real-world point clouds and significant embodiment occlusions. For a more comprehensive view, including zero-shot predictions for all 18 tasks, please refer to project website.

F. Quantitative Ablation Study

Experiment Setting: We conduct an ablation study to evaluate the key design elements of our methods. The variants tested include:

- **w/o Text EarlyFusion:** aligned text features are concatenated with PointNeXt features (having a dimension of 32) instead of the original point clouds.
- **w/o Scale Normalization:** the Conditional Variational AutoEncoder (CVAE) predicts general flow without scale normalization. We explore two versions: one with absolute position prediction and another with relative displacement prediction.
- **w TDN Scale Normalization:** this approach employs normalization to adjust the length of absolute displacement to 1, rather than the total length.
- **w SDN Scale Normalization:** normalization is used to set the length of each step to 1.
- **w $\beta_1 = 1$ (weight of scale-loss):** this test is designed to assess the importance of adequately weighting scale prediction in the loss function.
- **w/o central crop:** all scene point clouds in a 2m operation space are fed into the model without cube space cropping. The active object points average only about 2% in this setup.
- **w/o robustness augmentation:** these variants omit three types of technical augmentations (Scale Rebalance, Hand Mask Augmentation, Query Point Sampling Augmentation) to determine their impact on the model's prediction accuracy in our benchmark.

TABLE IV: Results of the ablation study on general flow prediction, with the best results highlighted in **bold** and the second-best results underlined.

	Test-ADE (w/o hand)	**Test-FDE (w/o hand)**
full	**0.0358**	**0.0477**
w/o Text EarlyFusion	0.0370	0.0495
w/o Scale Normalization (relative)	0.0376	0.0504
w/o Scale Normalization (absolute)	0.0381	0.0512
w TDN Scale Normalization	0.0374	0.0500
w SDN Scale Normalization	0.0377	0.0510
w $\beta_1 = 1$ (weight of scale-loss)	0.0368	0.0493
w/o central crop	0.0399	0.0538
w/o Scale Rebalance	0.0359	0.0478
w/o HM Augmentation	0.0366	0.0488
w/o QPS Augmentation	**0.0358**	**0.0477**

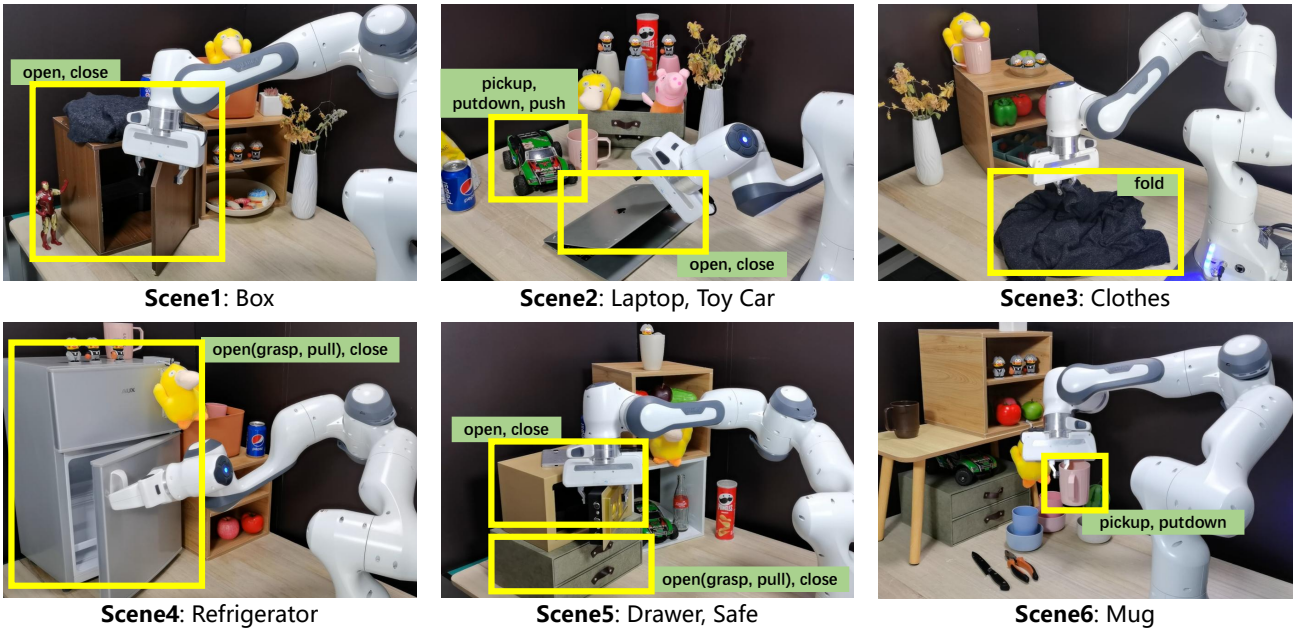
Experiment Result: The results of the ablation study are summarized in Table IV. For all variants except those without robustness augmentation, there is a noticeable degradation in model performance. In regards to the ablation of the three technical augmentations, it is evident that they do not detrimentally affect benchmark performance. Notably, the hand mask augmentation even significantly enhances in-domain prediction, which is an interesting observation.

G. Real-World Environment Setting

In our real-world experiment (Figure 9), we select 8 objects, including rigid, articulated, and soft bodies, as featured in our human video resources. We manually define multiple tasks and their corresponding success conditions for each object, resulting in a total of 18 distinct tasks. For a complete listing of these tasks, please refer to the content in the main paper (Section VI-B). For "Refrigerator" and "Drawer", we refer to them as "Storage Furniture" in the instructions. "Box" is also referred to as "Safe" since it can be seen as an approximation of an atypical design of "Safe".

These objects are arranged into 6 scenes, as depicted in Figure 9. For objects that are movable, such as "Mug" and "Toy Car", we randomly adjust their positions and orientations to add variability. It is worth noting that our experimental setup more accurately mirrors practical real-world scenarios compared to previous studies like Eisner et al. [22]. Our setup features diverse scenes and eliminates the necessity for clean object segmentation (Figure 10). This resemblance underscores the robustness and stability of our system in real-world scenarios.

The criteria for successful task completion varied. For "pickup" and "push" tasks, moving the object in the correct direction by more than 15cm is deemed successful. The "Put-down" action is deemed successful if the object is ultimately placed on the desktop and the orientation of the object is appropriate (for example, the mouth of a mug facing vertically



Scene1: Box **Scene2**: Laptop, Toy Car **Scene3**: Clothes

Scene4: Refrigerator **Scene5**: Drawer, Safe **Scene6**: Mug

Fig. 9: This figure illustrates the distribution of 8 objects across 18 tasks, encompassing various categories such as rigid, articulated, and soft bodies, arranged into 6 distinct scenes. Manipulated objects are highlighted within yellow bounding boxes, with each corresponding task denoted by a green box.

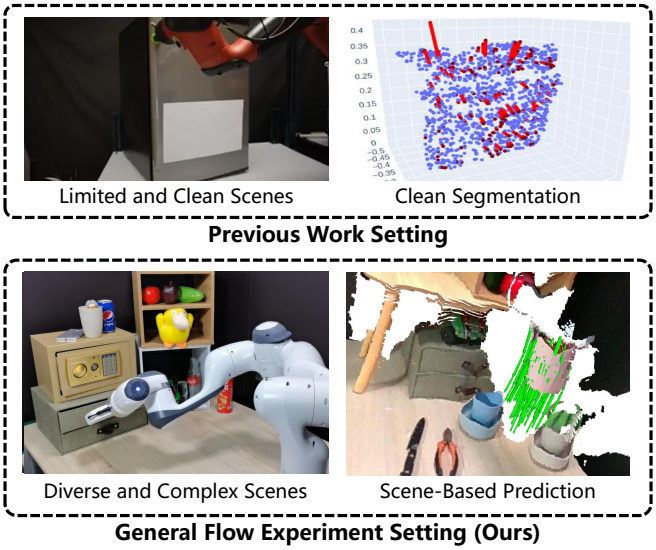


Fig. 10: Our environment setting is much closer to practical real-world situations compared to previous work [22].

upwards). For the "open" task of the revolute articulation structure, an opening of 80 degrees is considered a success. For its "close" task, bringing the object to within less than 5 degrees of the fully closed state is regarded as successful. 5 cm (to fully open or close state) is used as a criterion for prismatic structure. The "fold" is considered successful if one end of the garment reaches the other end.

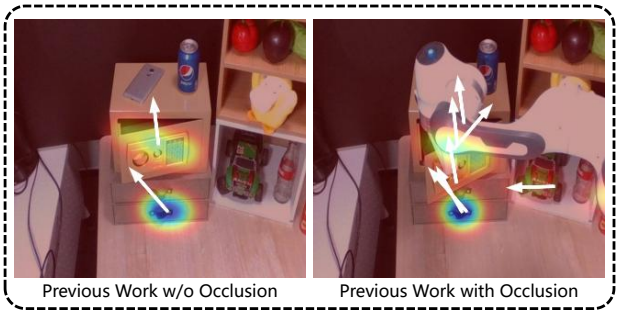


Fig. 11: Affordance prediction from our baseline model [4] in the real world. The text prompts for the grounding module are set to "safe" and "drawer".

H. Real-World Baseline

To the best of our knowledge, Bahl et al. [4] is *the only open-source work* that shares a similar setting with ours, which involves learning a low-level affordance model directly from real-world human videos. We deploy this model in our experimental environment, using 'safe' and 'drawer' as text prompts for the grounding module [54]. Figure 11 shows the visualization of the affordance prediction. Although it provides semantically meaningful predictions to some extent, it is limited by: (1) inadequate generalization for accurate motion direction prediction; (2) providing only 2D guidance without depth information; and (3) significant disturbances caused by embodiment occlusions. Given that the predicted post-grasp trajectories do not offer sufficient 3D guidance for closed-loop execution, we refrain from further robot execution trials.

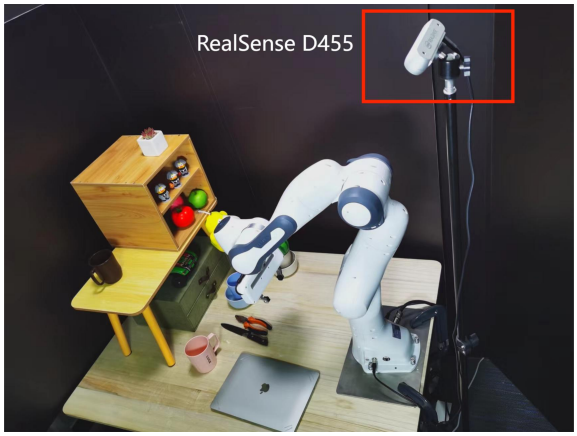


Fig. 12: Real-world deployment setting

I. Development of the Robotic System

Figure 12 shows a snapshot of our real-world deployment setup. We use a RealSense D455 RGBD camera to capture point cloud streams at a resolution of 1280 × 720, which is lower than the 1920 × 1080 resolution used in HOI4D [55]. Consequently, we opt for 0.01cm voxel downsampling during deployment, as opposed to the 0.02cm used in model training. The calibration parameters for our camera-robot system are (qw=0.911, qx=-0.015, qy=0.410, qz=-0.032) for orientation and (x=-0.265, y=0.260, z=1.095) for position. This configuration mimics a human ego-view manipulation perspective, beneficial not only for minimizing inference domain-gap but also aligning with practical applications in mobile robots. We utilize the RealSense camera's ROS driver for data acquisition.

The robot's base, static during manipulation, serves as a prompt for the FastSAM [96] model for robot segmentation. For enhanced accuracy, more prompt points or customized models [32] can be employed. Post-segmentation, we reconstruct 3D scene point clouds and select query points within 10cm of the gripper. These, along with the scene point clouds and the text instruction, are fed into our prediction model (ScaleFlow-B in our experiments) to obtain the anticipated general flow. We then apply the SVD algorithm[6] to obtain a robust transformation aligned with the predicted flow. The robot arm is driven by the Deoxys library [98] to follow the derived SE(3) transformation in a closed-loop manner, achieving a 0.4s inference latency (0.05s without FastSAM).

In practical applications, we note that the 6DoF controller for operational space in Deoxys lacks the necessary control precision for minute distances. Consequently, for trajectories shorter than 5 cm, we adopt a strategy of consolidating all steps into one and scaling this unified step to 5 cm in length. This method significantly enhances control accuracy over shorter distances, boosting the system's overall effectiveness and efficiency. Approximately 25% of predictions activate this workaround, which is considered acceptable given that the majority of tasks do not demand high levels of dexterity. Future improvements could include more precise controllers and calibration. The loop rate for our ROS system is set at 20 Hz. For safety, we manually confirm each planning step, although we find this almost unnecessary, as all experiments proceed with continuous pressing and confirmation without any delays.

TABLE V: The inference latency of each part in our pipeline. The results is the average value of 10 measurements.

Part	Time (ms)
Data Acquisition	3.2
FastSAM Segmentation	347.6
PointCloud Generation	30.8
Query Points Sampling	0.3
Flow Prediction (ScaleFlow-B)	22.1
Heuristic Policy Generation	1.7
Total (with Segment)	405.7 (2.5Hz)
Total (without Segment)	58.1 (17.5Hz)

J. Heuristic Policy Derivation

This section offers an in-depth explanation of our heuristic policy derivation. We begin by obtaining the gripper pose from the Deoxys API and projecting it into the camera's coordinates using calibration parameters. We select points within a 10cm distance from the gripper. Utilizing these points, we predict general flow and proceed to derive a 6DoF end-effector motion plan in camera space. For point clouds k_t and k_{t+1}, each containing N points and representing adjacent timestamps, our objective is to identify a 6DoF transformation with rotation $\hat{R}$ and translation $\hat{T}$ that fulfills the following criteria:

$$
\begin{aligned}
w_i &= \left(\frac{\frac{1}{d_i+\beta}}{\sum_{j=1}^{N} \frac{1}{d_j+\beta}} \right) \\
\hat{R}, \hat{T} &= \arg\min_{R,T} w_i \left\| k_{t+1}^i - (R \cdot k_t^i + T) \right\|^2
\end{aligned}
\tag{4}
$$

where w_i denotes the regression weight inversely proportional to the distance d_i between the i-th query points and the gripper position, with β set to 1. We solve Equation 4 using the SVD algorithm [6] for robust results. The acquired transformation $\mathcal{T} = (\hat{R}, \hat{T})$ is then projected back into the robot's coordinates and adjusted to the gripper's coordinates for controller execution.

K. Inference Latency

Table V details the average inference latency of each component in our pipeline, based on 10 measurements of the "open(grasp) Refrigerator" task. The significant bottleneck is the FastSAM segmentation, which contributes to 85.7% of the latency. This highlights the need for more efficient open-world segmentation models in future work. Without FastSAM segmentation, general flow prediction is not the sole system bottleneck; stream acquisition of point clouds also presents substantial room for improvement.

L. Failure Case Analysis

We analyze failure cases for tasks with success rates below 60% and suggest potential improvement methods:

- *"Push Toy Car" (50% success rate)*: The Toy Car's direction requires a sophisticated semantic understanding. To improve our model's capability in this area, additional data collection is essential. Due to the toy car's small size, integrating a wrist camera could also help mitigate significant occlusion problems and enhance performance.
- *"Open Drawer" (30% grasping, 40% pulling)*: The mixture of prismatic and revolute structures in the HOI4D [55] datasets leads to a slight tendency towards including rotational components in predictions. Its negative impact is amplified in our fabric cabinet with high friction. The leather handle on the drawer also poses a challenge, often slipping from the gripper. Future enhancements could include using larger datasets with more robust language semantics [34] and redesigning the gripper.
- *"Open Laptop" (50%)*: The laptop's thin lid often results in poor or incorrect RGBD point cloud generation. Utilizing point clouds fused from multiple camera views could ameliorate this issue.

Failure case videos are available on project website. In conclusion, most of these problems are addressable in future deployments. We systematically summarize these limitations and potential improvements in the next section.

M. Future Directions

We discuss the limitations of our current system and outline potential directions for future improvements:

- **Scaling Up with More Data:** Our model still lacks sufficient geometric guidance for complex tasks. Additionally, current data lacks fine-grained language control, for example, summarizing both revolute and prismatic structures as "Storage Furniture." More extensive data training could resolve these problems. Fortunately, the release of large-scale, semantically rich human manipulation datasets like [34] coincides with our work provide a rich resource for future research. Moreover, RGB datasets [16, 33] with depth estimation techniques [10, 50] could also be valuable training resources.
- **Enhanced Downstream Policy Derivation:** Our current system relies on a basic heuristic policy based on general flow prediction. This approach has limitations, such as failing when no points surround the gripper. Future developments could leverage both few-shot imitation learning [85] and reinforcement learning [4] for better policy derivation.
- **Advancements in Deployment Techniques:** There is considerable scope for improvement in real-world deployment. Incorporating wrist cameras and multi-view cameras could help alleviate occlusion issues and enhance point cloud quality. More accurate camera calibration and end-effector control could also be helpful for performance enhancement.

N. Codebases

We extend our gratitude to the following codebases for their support in the development of this work:

- The model training framework and the 3D backbone are based on the codebase from Qian et al. [72].
- For HOI4D point-cloud data processing, we adopt methods from Liu et al. [55].
- For Hand-Object-Interaction (HOI) detection, we utilize the 100DOH tools from Shan et al. [75].
- FastSAM (Zhao et al. [96]) is used for all segmentation in this work.
- The co-tracker from Karaev et al. [48] is employed to track points in pixel space.
- Implementations of ResNet, Vision Transformer, R3M, and the VAT-MART baseline are adopted from Bao et al. [5], Nair et al. [65], and Wu et al. [88].
- The ros-perception and data stream acquisition are based on Shridhar et al. [77].
- We inherit the SVD transformation solver directly from Zhong et al. [97].
- The impedance controller for the end-effector is adopted from the Doexys library (Zhu et al. [98]).