VidBot: Learning Generalizable 3D Actions from In-the-Wild 2D Human Videos for Zero-Shot Robotic Manipulation

Hanzhi Chen[1] Boyang Sun[2] Anran Zhang[1] Marc Pollefeys[2,3] Stefan Leutenegger[1,2]
[1] Technical University of Munich [2] ETH Zürich [3] Microsoft

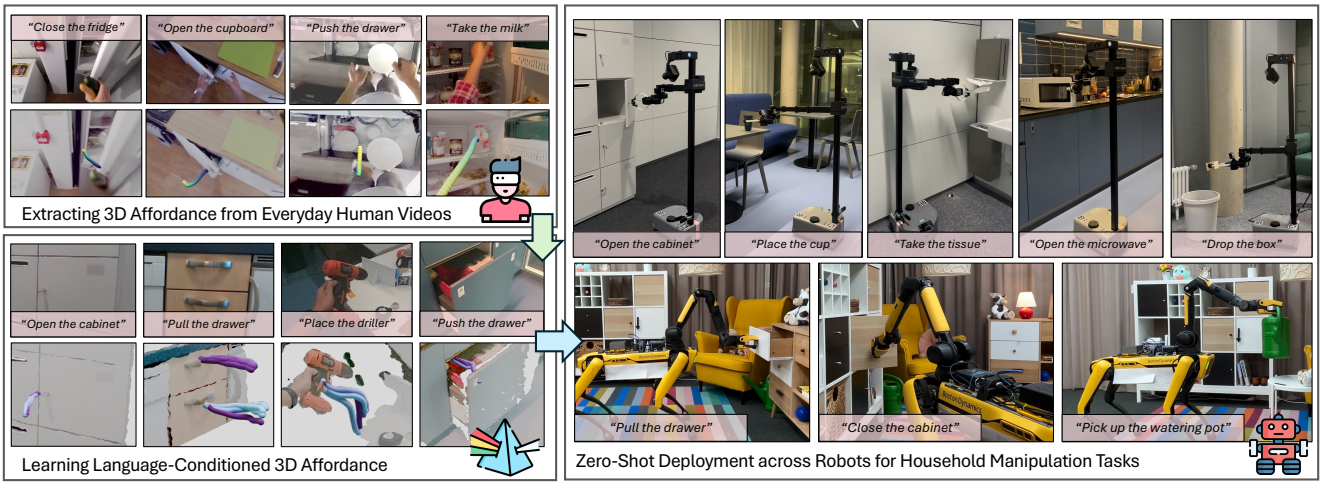

Figure 1. We present *VidBot*, a framework to learn interactions from in-the-wild RGB-only human videos. Our affordance model can be deployed across robots for daily manipulation tasks. Project cite: https://hanzhic.github.io/vidbot-project/

Abstract

Future robots are envisioned as versatile systems capable of performing a variety of household tasks. The big question remains, how can we bridge the embodiment gap while minimizing physical robot learning, which fundamentally does not scale well. We argue that learning from in-the-wild human videos offers a promising solution for robotic manipulation tasks, as vast amounts of relevant data already exist on the internet. In this work, we present VidBot, a framework enabling zero-shot robotic manipulation using learned 3D affordance from in-the-wild monocular RGB-only human videos. VidBot leverages a pipeline to extract explicit representations from them, namely 3D hand trajectories from videos, combining a depth foundation model with structure-from-motion techniques to reconstruct temporally consistent, metric-scale 3D affordance representations agnostic to embodiments. We introduce a coarse-to-fine affordance learning model that first identifies coarse actions from the pixel space and then generates fine-grained interaction trajectories with a diffusion model, conditioned on coarse actions and guided by test-time constraints for context-aware interaction planning, enabling substantial generalization to novel scenes and embodiments. Extensive experiments demonstrate the efficacy of VidBot, which significantly outperforms counterparts across 13 manipulation tasks in zero-shot settings and can be seamlessly deployed across robot systems in real-world environments. VidBot paves the way for leveraging everyday human videos to make robot learning more scalable.

1. Introduction

Advancements in AI are accelerating the development of personalized devices, such as smart glasses that offer virtual guidance to users [13, 22, 28, 62]. In the near future, robots will also become personalized systems, akin to smartphones or smart glasses, designed to provide physical assistance to humans. However, the diverse and novel forms of robotic embodiment pose a significant challenge for deploying AI to perform open-ended tasks in open-ended environments.

State-of-the-art approaches attempting to learn robot manipulation skills still rely heavily on human experts' teleoperated demonstrations, which are used to train robot policies under the Imitation Learning (IL) paradigm [35, 37, 64, 95]. However, this process remains costly, time-consuming, and labor-intensive. While recent efforts have gathered large-scale robotic demonstrations for everyday manipulation tasks–such as Open X-Embodiment [59] and DROID [39]–

scaling data collection remains challenging due to the combinatorial explosion of embodiments, tasks, and environments. We argue that human videos offer a promising scalable solution: there are already massive amounts of web videos capturing humans performing diverse tasks across various environments. Several previous approaches have explored human-to-robot skill transfer [2, 3, 65, 73, 78, 79, 88, 97]. Nevertheless, they face certain limitations, such as requiring static cameras or scenes, depth sensors, MoCap systems, etc. These constraints often result in in-lab settings lacking diversity in scenes, illuminations, or viewpoints. One line of research has explored leveraging internet human videos with rich scene contexts to boost robot learning tasks, focusing on learning visual representations for visuomotor policies [52, 57, 67, 85]. However, one major limitation is the reliance on humans to collect task-specific teleoperated data in every new environment with every new embodiment to fine-tune the pre-trained model. More recently, works like [4] have progressed by explicitly extracting agent-agnostic interaction trajectories. Nevertheless, these extracted motions are simplified as 2D vectors in pixel space, limiting their direct deployment to robots. We argue that, beyond visual representations or pixel-level action cues limited to the 2D image plane, 3D affordance–specifically, the contact points and interaction trajectories with spatial awareness–is crucial for unifying different embodiments to interpret action from perception. However, extracting general 3D affordance data from everyday human videos remains a significant challenge, impeding robots from learning manipulation skills by watching humans.

In this work, we aim to enable zero-shot robot learning from a large amount of unlabeled everyday human videos, tackling two key questions: (1) How can 3D actionable knowledge be extracted from raw RGB-only human videos? (2) How can this extracted knowledge be reliably transferred to novel environments and new robot embodiments in a zero-shot manner? To answer the first question, we adopt a principled approach by exploiting Structure-from-Motion (SfM) for robot learning, developing a gradient-based optimization pipeline that extracts 3D hand trajectories from in-the-wild videos. Our pipeline combines learned monocular depth with geometric constraints, ensuring temporally consistent, metric-scale reconstructions. This allows for the recovery of contact points and smooth hand trajectories in 3D, serving as agent-agnostic 3D affordance representations. We introduce a coarse-to-fine affordance learning framework for the second question to learn rich actions from the diverse extracted training data. At the coarse prediction stage, our affordance model identifies high-level actions from pixels, i.e., contact points and goal points, based on RGB-D observations and task instructions. In the fine prediction stage, we employ a diffusion model to generate fine-grained interaction trajectories conditioned on coarse-stage outputs and task observations. Rather than relying solely on learned action priors from humans, we incorporate several differentiable cost functions as test-time sampling guidance [32]. These cost functions guide the diffusion denoising process by perturbing outputs to satisfy test-time constraints during deployment. Differentiable objectives, such as multi-goal reaching and collision avoidance, capture the distribution of coarse outputs and leverage geometric cues from the scene context. The guidance terms improve the plausibility of interactions by accounting for the variance in scene contexts and robot morphology while providing intuitive heuristics for selecting the optimal plans.

We conducted extensive experiments in both simulation and real-world settings to evaluate the effectiveness of our framework. Utilizing in-the-wild 2D human videos only, our 3D affordance model outperformed several baselines trained using simulator exploration or pre-trained with teleoperated demonstrations. Furthermore, we demonstrated our model's versatility in various downstream robot learning tasks, such as visual goal-reaching and exploration, showcasing rapid convergence to superior performance.

2. Related Work

2.1. Visual Affordance Learning

Affordance centers around determining *where* and *how* an agent should interact with a given scene. One line of work regress affordance using manually annotated datasets [16, 19, 20, 55]. However, collecting affordance labels is highly costly. Hence, a more recent line of work addressed this challenge by deploying agents in simulated environments to explore effective interaction [12, 24, 54, 58, 83]. Despite offering a data collection alternative without human intervention, these methods often suffer from the cost of obtaining diverse virtual assets. In contrast, human videos have gained attention as a more general source of affordance priors. Several approaches [4, 27, 48, 56] predict per-pixel affordance scores by leveraging hand-object contact labels from human videos. However, these pipelines usually only identify contact regions or model interaction actions within the image plane, lacking spatial awareness. More recent works [7, 92] attempted to address this limitation by utilizing flows as spatially-aware affordance representations. However, these approaches require either goal images or initial contact regions given at test time. In contrast, our affordance model eliminates these requirements and directly infers contact points and interaction trajectories in 3D, as learned from in-the-wild RGB-only human videos.

2.2. Robot Learning from Humans

Previous works have explored utilizing human videos to aid robot learning tasks. One approach involves learning visual representations from human videos and us-

ing pre-trained visual encoders to train policy networks [6, 52, 57, 67, 82, 85, 89]. Another line of research focuses on learning reward functions from human videos [3, 10, 11, 43, 47, 75, 78, 86, 88]. Additionally, some works use motion attributes extracted from videos, such as estimating 3D hand poses or tracking wrist trajectories [5, 61, 65, 73, 74, 78, 79, 91]. However, these methods are typically restricted to in-lab setups and/or require further teleoperated demonstrations by human experts. [4] is the closest work to ours, which used everyday human videos to extract embodiment-agnostic actions. However, its inferred 2D pixel-level motions are oversimplified and ambiguous, limiting direct deployment to robots. In contrast, our framework leverages the same human video data as [4] for supervision but can predict affordance in 3D space, enabling zero-shot skill transfer to robots.

2.3. Diffusion Models in Robotics

Diffusion models are a powerful learning paradigm approximating complex data distributions through an iterative denoising process. Recently, they have achieved success across various generative modeling applications [15, 23, 31, 32, 40, 68–70, 81, 96]. In robotics, diffusion models have shown to be strong policy learning frameworks [1, 14, 36, 38, 45, 46, 51, 84]. Diffusion Policy [14] introduced a general framework for generating multimodal robot trajectories via a conditional denoising diffusion process. Diffuser [36] enhanced guided trajectory sampling by incorporating reward functions. Follow-up works [46, 51, 84] have proposed more factorized policy learning frameworks that allow diffusion models to generate smooth actions between key steps. However, these approaches focus on regressing highly limited in-domain teleoperation data with no modality or embodiment gaps during testing. In contrast, our approach learns policy from massive training data extracted from human videos. We introduce a coarse-to-fine affordance learning framework integrated with cost guidance to enhance generalization and test-time flexibility in novel scenes with new embodiments.

3. Method

3.1. Problem Definition

We aim to learn a factorized affordance model $\mathbf{a} = \pi(\{\tilde{\mathbf{I}}, \tilde{\mathbf{D}}\}, l)$ from everyday human videos, where $\{\tilde{\mathbf{I}}, \tilde{\mathbf{D}}\}$ is an RGB-D frame (image $\tilde{\mathbf{I}}$, depth $\tilde{\mathbf{D}}$), and l is language instruction. Note the depth frame can be obtained either from a depth sensor or a metric-depth foundation model [8, 90]. As the affordance representation is expected to be embodiment-agnostic, we formulate the final output affordance representation $\mathbf{a}$ as contact points $\mathbf{c}$ and interaction trajectories $\boldsymbol{\tau}$ following previous works [4, 48], while extending this formulation into 3D space. Specifically, $\mathbf{a} = \{\mathbf{c}, \boldsymbol{\tau}\}$, where $\mathbf{c} \in \mathbb{R}^{N_c \times 3}, \boldsymbol{\tau} \in \mathbb{R}^{H \times 3}$. Here, N_c is the number of contact points, and H is the trajectory horizon. Note $\mathbf{a}$ is represented in the observation camera's frame.

3.2. 3D Affordance Acquisition from Human Videos

We first design a pipeline to extract 3D hand trajectories from daily human videos recorded by a **moving** monocular camera, where each frame's pose and are **unknown**. Here, we introduce the key components in our pipeline.

Data Preparation. Given a video with color images $\{\hat{\mathbf{I}}_0, ..., \hat{\mathbf{I}}_T\}$ and language description l, an SfM system [71] is first employed to estimate camera intrinsics $\mathbf{K}$, per-frame scale-unaware poses $\{\mathbf{T}_{\mathrm{WC}_0}, ..., \mathbf{T}_{\mathrm{WC}_T}\}$ and sparse landmarks $\{{}_{\mathrm{W}}\mathbf{l}_0, ..., {}_{\mathrm{W}}\mathbf{l}_{N_l}\}$ expressed in the world frame. We harness a metric-depth foundation model [8, 33, 90] to predict each frame's dense depth $\{\hat{\mathbf{D}}_0, ..., \hat{\mathbf{D}}_T\}$. We further utilize a hand-object detection model [72] and segmentation models [41, 94] to acquire the masks of each frame's hand and in-contact object, i.e., $\{\mathbf{M}^{\mathrm{h}}_0, ..., \mathbf{M}^{\mathrm{h}}_T\}$, $\{\mathbf{M}^{\mathrm{o}}_0, ..., \mathbf{M}^{\mathrm{o}}_T\}$. With the hand masks provided, we further collect frames before $\mathbf{I}_0$ and their hand masks to obtain hand-less frames $\{\tilde{\mathbf{I}}_0, ..., \tilde{\mathbf{I}}_T\}$ with a video inpainting model [44].

Consistent Pose Optimization. Our first objective is to correct the camera poses to the metric-space scale. To achieve this goal, we optimize a global scale s_{g} for all frames by projecting the sparse landmarks to each image plane using camera intrinsics and its pose:

$$\min_{s_{\mathrm{g}}} \sum_{i,j} \tilde{\mathbf{M}}_i[\mathbf{u}_{ij}] \left\| \hat{\mathbf{D}}_i[\mathbf{u}_{ij}] - s_{\mathrm{g}}\, \mathrm{d}(\mathbf{T}^{-1}_{\mathrm{WC}_i}\, {}_{\mathrm{W}}\mathbf{l}_j) \right\|^2_2, \quad (1)$$

where $\tilde{\mathbf{M}}_i = \neg(\mathbf{M}^{\mathrm{h}}_i \bigcup \mathbf{M}^{\mathrm{o}}_i)$, denoting static regions, $\mathbf{u}_{ij}$ is the pixel coordinate of landmark j in camera i and $\mathrm{d}(\cdot)$ is the depth of the point. We then refine all frames' poses $\mathbf{T}_{\mathrm{WC}_i} \in \mathcal{T}$ and scales $s_i \in \mathcal{S}$ to compensate SfM reconstruction errors due to dynamic hand-object motions and simultaneously make the predicted depth more consistent across views by optimizing the following term:

$$\min_{\mathcal{T}\setminus\{\mathbf{T}_{\mathrm{WC}_k}\}, \mathcal{S}\setminus\{s_k\}} \sum_{i \neq k} \sum_{\mathbf{u}_i} \tilde{\mathbf{M}}_i[\mathbf{u}_i] \tilde{\mathbf{M}}_k[\mathbf{u}_k] \mathbf{E}[\mathbf{u}_i], \quad (2)$$

where k is the reference frame index yielding the highest co-visibility with the others, ${}_{\mathrm{C}_i}\mathbf{X}^n_i$ is the back-projected points using the camera intrinsics $\mathbf{K}$ and the depth $\hat{\mathbf{D}}_i$. $\mathbf{E}[\mathbf{u}_i] = \left\| s^{-1}_i \mathbf{T}_{\mathrm{C}_k\mathrm{C}_i\mathrm{C}_i} \mathbf{X}^n_i[\mathbf{u}_i] - s^{-1}_k\, {}_{\mathrm{C}_k}\mathbf{X}^n_k[\mathbf{u}_k] \right\|^2_2$, where $\mathbf{u}_k$ is the projective pixel correspondence to $\mathbf{u}_i$ computed with $s_i, s_k, \mathbf{K}, \hat{\mathbf{D}}_i$, and $\mathbf{T}_{\mathrm{C}_k\mathrm{C}_i} = \mathbf{T}^{-1}_{\mathrm{WC}_k} \mathbf{T}_{\mathrm{WC}_i}$. s_k is fixed to s_{g}.

Affordance Extraction. We obtain each frame's hand center point and transform it to the first frame with the refined poses and scales to compute the interaction trajectory $\hat{\boldsymbol{\tau}}$. We downsample hand points uniformly in the first frame to acquire contact points $\hat{\mathbf{c}}$, and them from the last frame to

Figure 2. Example trajectories extracted from raw human videos.

acquire goal points $\hat{\mathbf{g}}$ to supervise the intermediate prediction of our affordance model. Language description l, inpainted color $\tilde{\mathbf{I}}_0$ and its depth $\tilde{\mathbf{D}}_0$ from [90], together with the inpainted object image $\tilde{\mathbf{I}}_0^o$ cropped using $\mathbf{M}_0^o$, are used as model inputs. We leverage the EpicKitchens-100 Videos dataset [18] and its SfM results provided by EpicFields [76] to showcase the effectiveness of our pipeline. Fig. 2 exemplifies the extracted results.

3.3. Coarse-to-Fine Affordance Learning

The overview of our affordance model is shown in Fig. 3. We design the model with two key considerations: (1) It should capture the action distribution conditioned on observation and instruction from massive in-the-wild human affordance data. (2) It should leverage contextual information during test time to mitigate the embodiment gap and potentially noisy prediction due to imperfect training data, thereby enhancing the quality of the generated affordance.

To address the first aspect, we factorize the affordance model π into a coarse model π_c and a fine model π_f. In the coarse stage, π_c performs high-level scene understanding to infer a set of goal points $\mathbf{g}$ and contact points $\mathbf{c}$ conditioned on the RGB-D frame $\{\tilde{\mathbf{I}}, \tilde{\mathbf{D}}\}$ and instruction l, i.e., $\{\mathbf{g}, \mathbf{c}\} = \pi_\text{c}(\{\tilde{\mathbf{I}}, \tilde{\mathbf{D}}\}, l)$, $\mathbf{a}_\text{c} = \{\mathbf{g}, \mathbf{c}\}$. Given the coarse-stage outputs together with the task inputs, π_f plans fine-grained interaction trajectories at a low level with $\boldsymbol{\tau} = \pi_\text{f}(\{\tilde{\mathbf{I}}, \tilde{\mathbf{D}}\}, l, \boldsymbol{a}_\text{c})$. To achieve the second objective, we integrate multiple analytical cost functions for π_f, incorporating scene context and agent embodiment during the test time. These constraints guide the trajectory generation process, leading to more plausible and context-aware interaction trajectories.

The contact points $\mathbf{c}$ and the interaction trajectory $\boldsymbol{\tau}$ will be the final affordance outputs $\mathbf{a} = \{\mathbf{c}, \boldsymbol{\tau}\}$.

3.3.1. Coarse Affordance Prediction

At this stage, the coarse affordance model is designed to extract macro actionable information from high-dimensional image space. To achieve this, we represent the coarse action points in pixel space by learning probabilities of the coarse affordance, along with their corresponding depth, where applicable. As illustrated in Fig. 3, we first obtain a cropped RGB-D image of the objects of interest using an off-the-shelf open-set object detector [49, 87]. Our coarse model, π_c, consists of two networks: π_c^goal and π_c^cont, which predict the goal and contact points, respectively.

For goal points prediction, the context color image and a depth image filled with the median depth of the object of interest is fed to π_c^goal to obtain the goal heatmap activation with the exact resolution, together with the depth of the goal points, as the endpoint of an interaction trajectory tends to be distributed in free space. Given the *global context feature* $\mathbf{z}^\text{goal}$ encoded by the visual encoder of π_c^goal, we extract the object-centric embedding $\mathbf{z}_\text{o}^\text{goal}$ using RoI Pooling [26]. Further, we acquire bounding box positional feature $\mathbf{z}_b^\text{goal}$ from a multi-layer perceptron (MLP) and the feature embedding $\mathbf{z}_l$ of the language instruction using a frozen CLIP model [66]. Given the conditional feature $\mathbf{z}_\text{c}^\text{goal} = \{\mathbf{z}_\text{o}^\text{goal}, \mathbf{z}_\text{b}^\text{goal}, \mathbf{z}_\text{l}\}$, we leverage a Perceiver [34] module with several self-attention and cross-attention blocks, enabling the global context feature $\mathbf{z}^\text{goal}$ to attend to the conditional feature $\mathbf{z}_\text{c}^\text{goal}$. The fused *global context feature* will be first passed to the transformer encoder and MLP layers to predict the goal depth and forwarded to a visual decoder to predict the per-pixel goal probabilities.

The contact predictor follows a similar hourglass network architecture as the π_c^goal. However, we omit the prediction of contact points' depth as they tend to lie on the objects' surface with valid depth. Only the language feature $\mathbf{z}_\text{l}$ is fused into the *object contact feature* $\mathbf{z}^\text{cont}$ extracted by the visual encoder of π_c^cont.

Finally, using the camera intrinsics, sampled pixel coordinates from the predicted heatmaps, and their queried depth, we lift them to 3D to obtain the coarse affordance output, i.e., the goal points $\mathbf{g}$, and the contact points $\mathbf{c}$.

3.3.2. Fine Affordance Prediction

The fine affordance model is conditioned to infer a fine-grained interaction trajectory guided by the contact and goal points. This stage is modeled as a conditional diffusion denoising process inspired by [36].

Diffusion models preliminaries. Our fine affordance model follows the diffusion probabilistic model formulation [31]. Such a formulation is modeled with a *forward process* and a *reverse process*.

Given a sample $\boldsymbol{\tau}^0$ drawn from its underlying distribution $q(\boldsymbol{\tau})$, the *forward process* iteratively injects Gaussian noise in K steps as a Markovian process. Such a process can be expressed as:

$$\begin{aligned} q(\boldsymbol{\tau}^k|\boldsymbol{\tau}^{k-1}) &= \mathcal{N}(\boldsymbol{\tau}^k; \sqrt{1-\beta_k}\boldsymbol{\tau}^{k-1}, \beta_k\mathbf{I}), \\ q(\boldsymbol{\tau}^{1:K}|\boldsymbol{\tau}^0) &= \prod_{k=1}^{K} q(\boldsymbol{\tau}^k|\boldsymbol{\tau}^{k-1}), \end{aligned} \tag{3}$$

where β_k is acquired from a pre-defined scheduler.

In the *reverse process*, a denoising neural network ϕ learns distribution $p_\phi(\boldsymbol{\tau}^{k-1}|\boldsymbol{\tau}^k)$ to gradually remove the

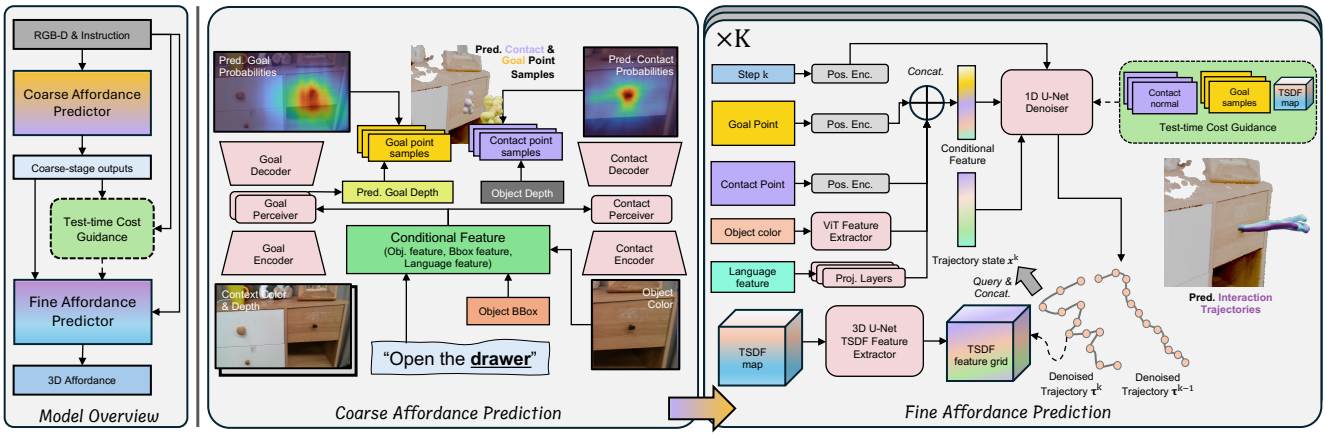

Figure 3. Overview of our affordance learning model. The affordance model is factorized into a coarse stage and a fine stage. We parse high-level contact and goal configurations from task inputs in the coarse stage. Supp. Mat. provides more detailed illustration of conditional feature extraction process. In the fine stage, we utilize the coarse stage outputs to guide the fine-grained interaction trajectory generation process through conditioning and cost functions. The color represents the final cost value, with darker shades indicating lower costs.

noise so as to recover $\boldsymbol{\tau}^0$:

$$p_\phi(\boldsymbol{\tau}^{k-1}|\boldsymbol{\tau}^k) = \mathcal{N}(\boldsymbol{\tau}^{k-1}; \boldsymbol{\mu}_\phi(\boldsymbol{\tau}^k, k), \boldsymbol{\Sigma}_k),$$
$$p_\phi(\boldsymbol{\tau}^{1:K}) = p(\boldsymbol{\tau}_K) \prod_{k=1}^{K} p_\phi(\boldsymbol{\tau}^{k-1}|\boldsymbol{\tau}^k), \tag{4}$$

where $\boldsymbol{\mu}_\phi(\boldsymbol{\tau}^k, k)$ is from the denoising neural network, and $\boldsymbol{\Sigma}_k$ is from a fixed scheduler.

The desired diffusion state $\boldsymbol{\tau}$ is the interaction trajectory in our setting. The denoising network learns a conditional distribution as $p_\phi(\boldsymbol{\tau}^{k-1}|\boldsymbol{\tau}^k, \boldsymbol{o})$ conditioned on task observations $\boldsymbol{o}$ to be explained, i.e., $\boldsymbol{\mu}^k = \boldsymbol{\mu}_\phi(\boldsymbol{\tau}^k, k, \boldsymbol{o})$.

Trajectory generation. The fine affordance model π_f is parameterized by a 1D U-Net similar to [69]. Specifically, we first acquire the goal point $\bar{\mathbf{g}}$ and contact point $\bar{\mathbf{c}}$ from $\mathbf{g}$ and $\mathbf{c}$ with the highest predicted probabilities. Moreover, we obtain the object's feature embedding $\mathbf{z}_\text{o}^\text{fine}$ extracted by a vision transformer [9]. Together with the language instruction embedding $\mathbf{z}_l$ extracted in the coarse stage, we establish the conditional embedding as $\boldsymbol{o} = \{\text{PE}(\mathbf{g}), \text{PE}(\mathbf{c}), \text{Proj}(\mathbf{z}_l), \mathbf{z}_\text{o}^\text{fine}\}$, PE is the positional encoding operator and Proj denotes blocks with a transformer encoder [77] and an MLP. To integrate spatial awareness into the inferred trajectory, we encode C_m-dimensional latent features from the voxelized TSDF map $\mathbf{U}$ acquired from RGB-D frame using a 3D U-Net [63] and compute the spatial feature $\mathbf{f}^k \in \mathbb{R}^{H \times C_m}$ for each waypoint from denoised trajectory $\boldsymbol{\tau}^k$ using trilinear interpolation. $\mathbf{f}^k$ and $\boldsymbol{\tau}^k$ are concatenated together as denoising inputs $\mathbf{x}^k = \{\boldsymbol{\tau}^k, \mathbf{f}^k\}$ fed to π_f. Instead of using noise-prediction, π_f directly infers the unnoised trajectory $\bar{\boldsymbol{\tau}}^0$ in each step k and uses it to compute $\boldsymbol{\mu}^k$ (*c.f.* Supp. Mat. for details), i.e., $\bar{\boldsymbol{\tau}}^0 = \pi_\text{f}(\mathbf{x}^k, \text{PE}(k), \boldsymbol{o})$.

3.3.3. Cost-Guided Trajectory Generation

The inferred trajectory could become erroneous if the conditioned goal point $\bar{\mathbf{g}}$ has an offset. This is expected since π_f essentially acts as a gap filler between the contact point $\bar{\mathbf{c}}$ and the goal point $\bar{\mathbf{g}}$. The optimal goal point for conditioning may not always be selected based on the predicted scores, and multiple goal points from the goal set $\mathbf{g}$ can yield more diverse and robust predictions. However, querying the affordance model multiple times by sampling different goal configurations is computationally inefficient. Therefore, we cast the multi-goal conditioning as a cost function to guide trajectory generation during test time. Specifically, we define the multi-goal cost function as:

$$\mathcal{J}_\text{goal} = \min_{\mathbf{g}_n \in \mathbf{g}} \left\| \mathbf{g}_n - \bar{\boldsymbol{\tau}}_H^0 \right\|_2^2, \tag{5}$$

where $\bar{\boldsymbol{\tau}}_H^0$ denotes the trajectory endpoint and $\mathbf{g}_n$ is the n-th goal point. We additionally formulate scene collision avoidance guidance and contact normal guidance as $\mathcal{J}_\text{collide}$ and $\mathcal{J}_\text{normal}$, where h denotes the trajectory horizon index:

$$\mathcal{J}_\text{collide} = \frac{1}{H' N_\text{p}} \sum_{h \neq 1, i} -\min\left(\mathbf{U}[\mathbf{p}_i + \bar{\boldsymbol{\tau}}_h^0 - \bar{\boldsymbol{\tau}}_1^0], 0\right),$$
$$\mathcal{J}_\text{normal} = \frac{1}{H'} \sum_{h \neq 1} \min_{s_\text{n} \in \{-1, 1\}} \left\| (\bar{\boldsymbol{\tau}}_h^0 - \bar{\boldsymbol{\tau}}_1^0) / \|\bar{\boldsymbol{\tau}}_h^0 - \bar{\boldsymbol{\tau}}_1^0\|_2 - s_\text{n} \mathbf{n} \right\|_2^2. \tag{6}$$

We sample N_p points from both the agent's hand model (e.g., robot gripper) and the object's surface (if it is portable) prior to interaction ($h = 1$), where each one is denoted as $\mathbf{p}_i$, and query their values differentiably from the pre-computed TSDF map $\mathbf{U}$. The normal vector $\mathbf{n}$ is computed from the contact points. The trajectory starting point $\bar{\boldsymbol{\tau}}_1^0$ is not optimized during guidance, hence $H' = H - 1$. The final cost function $\mathcal{J}$ for test-time guidance is formulated as:

$$\mathcal{J} = \lambda_\text{g} \mathcal{J}_\text{goal} + \lambda_\text{c} \mathcal{J}_\text{collide} + \lambda_\text{n} \mathcal{J}_\text{normal}, \tag{7}$$

where λ_g, λ_c, λ_n control the strength of each guidance term.

We adopt *reconstruction guidance* from [32, 69]. At each denoising step, gradients of $\mathcal{J}$ w.r.t. $\boldsymbol{\tau}^k$ will adjust the

unnoised predictions $\bar{\boldsymbol{\tau}}^0$ as $\boldsymbol{\tau}^0$:

$$\boldsymbol{\tau}^0 = \bar{\boldsymbol{\tau}}^0 - \boldsymbol{\Sigma}_k \nabla_{\boldsymbol{\tau}^k} \mathcal{J}. \tag{8}$$

Now we use $\boldsymbol{\tau}^0$ instead of $\bar{\boldsymbol{\tau}}^0$ to compute $\boldsymbol{\mu}^k$.

Introducing test-time guidance into the trajectory generation process offers several advantages: 1) Trajectories can better capture the goal distribution without extensive forward passes through the fine affordance model. 2) The morphology of novel embodiments and the geometry of previously unseen objects can be accounted for, providing collision-free hand trajectories readily integrated into downstream whole-body planning. 3) The final cost value $\mathcal{J}$ for each trajectory is an informative criterion for the agent to select the optimal interaction plan.

3.3.4. Affordance Models Training

To train the coarse affordance model, i.e., $\pi_{\text{c}}^{\text{goal}}$ and $\pi_{\text{c}}^{\text{cont}}$, we project the extracted goal points $\hat{\mathbf{g}}$ and contact points $\hat{\mathbf{c}}$ to the image plane and obtain ground-truth probabilities by fitting a Gaussian mixture model to them, which results in $\hat{\mathbf{H}}_{\text{g}}$ and $\hat{\mathbf{H}}_{\text{c}}$. The goal depth $\hat{D}_{\text{g}}$ is additionally regressed by $\pi_{\text{c}}^{\text{goal}}$, which is the median depth of the goal points. We include an auxiliary vector field regression loss $\mathcal{L}_{\text{v}}$ for coarse affordance model training, *c.f.* Supp. Mat. for details. L_{g} and L_{c} are used to supervise $\pi_{\text{c}}^{\text{goal}}$ and $\pi_{\text{c}}^{\text{cont}}$:

$$\mathcal{L}_{\text{g}} = \text{BCE}(\hat{\mathbf{H}}_{\text{g}}, \mathbf{H}_{\text{g}}) + \lambda_{\text{d}} \left\| \hat{D}_{\text{g}} - D_{\text{g}} \right\|_2^2 + \lambda_{\text{v}} \mathcal{L}_{\text{v}}, \tag{9}$$

where $\mathbf{H}_{\text{g}}$, D_{g} are predicted by the goal predictor $\pi_{\text{c}}^{\text{goal}}$, λ_{d} and λ_{v} are weighting factors.

$$\mathcal{L}_{\text{c}} = \text{BCE}(\hat{\mathbf{H}}_{\text{c}}, \mathbf{H}_{\text{c}}) + \lambda_{\text{v}} \mathcal{L}_{\text{v}}, \tag{10}$$

where $\mathbf{H}_{\text{c}}$ is outputted by the contact predictor $\pi_{\text{c}}^{\text{cont}}$, λ_{v} is a weighting factor same as the one used in $\mathcal{L}_{\text{g}}$.

The fine affordance model π_{f} is trained by supervising its output $\bar{\boldsymbol{\tau}}^0$ using the extracted trajectory $\hat{\boldsymbol{\tau}} \sim q(\boldsymbol{\tau})$:

$$\mathcal{L}_{\text{f}} = \mathbb{E}_{\boldsymbol{\epsilon}, k} \left[\left\| \hat{\boldsymbol{\tau}} - \bar{\boldsymbol{\tau}}^0 \right\|_2^2 \right], \tag{11}$$

where $k \sim \mathcal{U}\{1, ..., K\}$ is the diffusion step index, and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is used to corrupt $\hat{\boldsymbol{\tau}}$ to obtain $\boldsymbol{\tau}^k$ inputted to π_{f}.

4. Experiments

We aim to showcase the following aspects of our affordance model: (1) It significantly outperforms several strong baselines in zero-shot robot manipulation tasks. (2) Both coarse affordance prediction and test-time cost-guidance are crucial to guarantee its optimal performance. (3) It can enhance several downstream robot learning applications. (4) It can be seamlessly deployed across real robot systems.

4.1. Experimental Setups

Simulator Environments. We use IsaacGym [53] as our simulation platform for benchmarking, with environments developed based on [43]. We select 13 everyday household tasks from three widely-used benchmarks: FrankaKitchen [30], PartManip [24], and ManiSkill [29]. These tasks encompass action primitives such as opening, pushing, sliding, etc., and various objects, including cabinets, drawers, and kettles. Each task is evaluated from three different viewpoints. Each model generates five trajectories for each viewpoint, totaling 15 trials per task per model. Our evaluation protocol quantifies performance using the success rate (%) commonly adopted in previous works, where a successful interaction is defined as causing the task object's degree of freedom (DoF) to exceed a pre-specified threshold without colliding with other objects in the scene.

Baseline Models. We compare our model against several representative baselines publicly available. Specifically, GAPartNet [25] and Where2Act [54] are trained using virtual articulated assets collected (and interacted with) in simulators. Octo [59] is pre-trained on a large-scale teleoperated dataset [60] and further fine-tuned with our collected dataset. VRB [4], GFlow [92] and our model are trained using human videos, while GFlow [92] can access the ground-truth depth, camera parameters, and object poses from [50]. Hence, VRB [4] and ours operate in much more in-the-wild settings. We follow the strategy from [42] to lift pixel-level trajectories from VRB [4] to 3D using object normal clusters as cues. We observed that baselines such as VRB [4] and GFlow [92] could not accurately infer contact regions. To ensure a fair comparison, we used our model instead to infer and standardize contact configurations. Hence, the benchmark focuses on predicting accurate interaction trajectories, which are more challenging than contact regions. We detail the strategy to obtain robot actions from our predicted affordance in Supp. Mat.

4.2. Results and Discussions

As shown in Table. 1, our model achieves the best overall performance with an 88.2% success rate across all tasks, surpassing the runner-up method by nearly 20%. Among the baselines, VRB [4], GAPartNet [25], and Where2Act [54] exhibit inferior performance. We attribute this to their interaction policies being abstracted as directional vectors. While this approach works reasonably well for simpler tasks–such as pulling or pushing drawers along a straight line–it struggles with tasks requiring curved interactions, like opening a cabinet. The oversimplified motions tend to lead to gripper slip and subsequent task failures. Octo [59] improves upon these models by approximately 10%, achieving the second-best success rate of 69.2%. This improvement is expected, given that Octo is pre-trained on a large-scale teleoperated dataset [60] and demonstrates good

	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	Avg.
GAPartNet [25]	73.3	13.3	66.7	33.3	53.3	53.3	40.0	66.7	60.0	-	-	-	-	51.1
Where2Act [54]	86.7	53.3	60.0	0.0	46.7	66.7	60.0	**100.0**	53.3	-	-	-	-	58.5
Octo* [59]	93.3	33.3	93.3	**66.7**	60.0	46.7	80.0	80.0	53.3	**100.0**	66.7	53.3	73.3	69.2
VRB† [4]	**100.0**	20.0	**100.0**	46.7	66.7	53.3	40.0	26.7	46.7	**100.0**	60.0	46.7	60.0	59.0
GFlow [92]	**100.0**	33.3	**100.0**	6.7	60.0	53.3	6.7	26.7	73.3	93.3	**100.0**	60.0	80.0	61.0
Ours	**100.0**	**93.3**	**100.0**	**66.7**	**80.0**	**86.7**	**100.0**	66.7	**86.7**	**100.0**	**100.0**	**66.7**	**100.0**	**88.2**

Table 1. Quantitative results on 13 evaluated tasks evaluated on success rate (%) in simulators. **T01**: Close hinge cabinet, **T02**: Close slide cabinet, **T03**: Close microwave, **T04**: Open hinge cabinet, **T05**: Open microwave, **T06**: Pull drawer, **T07**: Push drawer, **T08**: Close dishwasher, **T09**: Open dishwasher, **T10**: Pick up kettle, **T11**: Pick up can from clutter, **T12**: Pick up box from clutter, **T13**: Lift lid. *: Fine-tuned on our extracted affordance data. †: Use strategy from [42] to lift affordance to 3D.

	AT01	AT02	AT03	AT04	AT05	AT06	Avg.
Ours [Full Model]	**93.3**	**66.7**	**80.0**	86.7	**86.7**	**100.0**	**85.6**
w/o coarse goal. pred. **[V1]**	66.7	33.3	73.3	53.3	60.0	60.0	57.8
w/o multi-goal guide. **[V2]**	80.0	40.0	60.0	93.3	66.7	**100.0**	73.3
w/o contact-normal guide. **[V3]**	86.7	33.3	**80.0**	**100.0**	60.0	**100.0**	76.7
w/o collision-avoid. guide. **[V4]**	**93.3**	60.0	**80.0**	80.0	80.0	73.3	77.8
w/o cost-informed heuristics **[V5]**	80.0	46.7	66.7	80.0	73.3	**100.0**	74.5

Table 2. Ablation results on 6 selected tasks evaluated on success rate (%). **AT01**: Close slide cabinet, **AT02**: Open hinge cabinet, **AT03**: Open microwave, **AT04**: Pull drawer, **AT05**: Open dishwasher. **AT06**: Pick up can from clutter.

generalization ability after being fine-tuned using our extracted dataset. However, the challenge of directly inferring complex actions from scene context may account for its lower performance than ours, as we quantitatively validate in Sec. 4.3. GFlow [92] suffers from erroneous trajectory scale predictions, leading to deteriorated performance.

The above results demonstrate that in-the-wild human videos can be a powerful data source for transferring manipulation skills to robots, provided that (pseudo) 3D interaction labels can be captured. However, the variance in scene contexts necessitates an effective affordance model capable of simultaneous high-level context understanding and low-level action planning that adapts to test-time constraints.

Notably, VRB [4] shares the same video data source [17] as ours for affordance supervision. However, by fully exploiting 3D priors and employing our proposed affordance learning strategy, we significantly improved—boosting the task success rate by *ca.* 30%. Fig. 5 demonstrates this enhancement using self-captured real-world data.

4.3. Ablation Studies

As in Table. 2, we conducted detailed ablation experiments on a subset of manipulation tasks to assess the necessity of various important design choices in our affordance model. Specifically, **V1** is a variant without coarse goal prediction, hence also without goal conditioning and multi-goal guidance $\mathcal{J}_{\text{goal}}$. **V2**-**V4** are variants that generate trajectories without multi-goal guidance $\mathcal{J}_{\text{goal}}$, contact-normal guidance $\mathcal{J}_{\text{contact}}$ or collision-avoidance guidance $\mathcal{J}_{\text{collide}}$, respectively. **V5** randomly selects the generated trajectories instead of using the final guidance cost value as heuristics.

Impact of Coarse Affordance Prediction. In **V1**, we observe a significant performance drop from 85.6% to 57.8%. This confirms that directly generating fine-grained actions from information-dense task observations is challenging. Inferring coarse goal configurations as *affordance cues* from high-dimensional observation space simplifies the generation of accurate interaction trajectories.

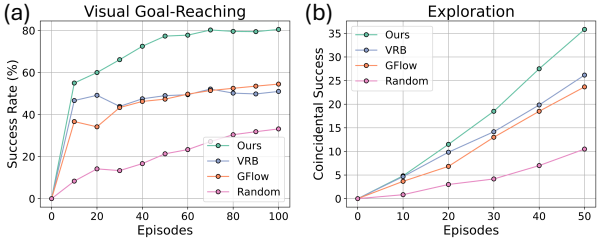

Figure 4. (a) Average success rate for the visual goal-reaching task. (b) Average coincidental success for the exploration task.

Impact of Cost Guidance. Multi-goal guidance (**V2**) is the most crucial factor boosting overall performance by 12.3%. Single-goal conditioning can mislead the generation process, whereas multi-goal conditioning through guidance helps correct accumulated errors from the coarse stage. Contact normal guidance (**V3**) is an intuitive hint for action generation, further improving performance. Collision avoidance guidance (**V4**) demonstrates its effectiveness, especially for the *pick-up* task of portable objects (**AT06**) with a 26.7% increase in success rate. These guidance terms, derived from test-time observations, enable more controllable trajectory generation under explicit constraints, enhancing the model's generalization toward unseen scenarios with new environments and embodiments.

Impact of Cost-informed Heuristics. The performance drop of 11.1% in **V5** underscores the importance of using the final guidance cost value as an intuitive criterion to select the optimal interaction plan.

4.4. Robot Learning Applications

Following the robot learning paradigms introduced in [4], we conducted several downstream application studies on the ablation tasks (**AT01**-**AT06**) to showcase the versatility of our affordance model as a strong prior.

Visual Goal-Reaching. The agent can additionally access an image of the object's desired configuration to enhance policy search supervision. Since our affordance model supports test-time guidance, we probabilistically replace the predicted goal points used in $\mathcal{J}_{\text{goal}}$ with those sampled from a buffer of successful trajectories in later episodes

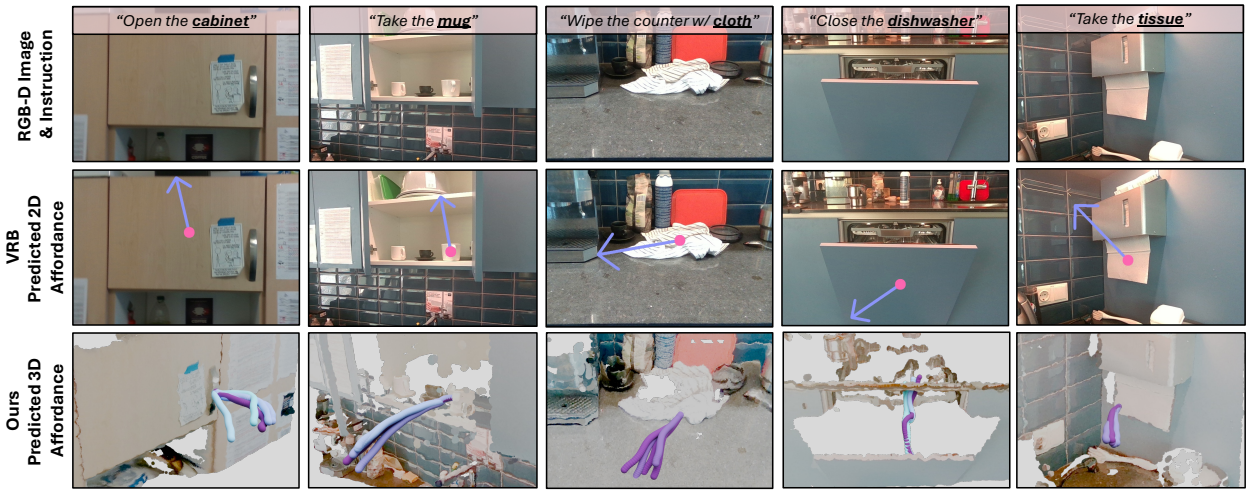

Figure 5. Predicted affordance by VRB [4] and ours given instruction and RGB-D image. Though using the same RGB-only human videos for training, our framework predicts much more accurate contact points and interaction trajectories in 3D space directly, outperforming VRB [4] with ambiguous prediction in the pixel space. We visualize the top five affordance samples inferred by our model, where colors represent the final cost values; darker shades indicate lower costs and, therefore, a higher rank for the agent to execute.

after collecting sufficient valid interactions. As shown in Fig. 4 (a), our method significantly outperforms other baselines trained with human videos, demonstrating faster convergence speeds and better overall performance.

Exploration. The agent seeks to maximize the environment changes when interacting with the scenes. We employ the coincidental success metric as in [4], i.e., the number of trajectories that bring the environment to the desired configurations without access to it. Similar to the goal-reaching tasks, we also use the previous successful trajectories to guide the sampling process of our affordance model. As illustrated in Fig. 4 (b), our method consistently demonstrates significant improvements over other baselines.

4.5. Real Robot Experiments

We validate the efficacy of our framework on two real-world mobile robot platforms: the Hello Robot Stretch 3 and the Boston Dynamics Spot (*c.f.* Fig. 6). Both robots are equipped with onboard RGB-D cameras for perception and receive language instructions for manipulation tasks. We test several household tasks within the robots' physical capabilities, such as pushing a drawer, opening a cabinet, and taking a tissue, across three different human-suited environments. Overall, the robots achieved a success rate of 80.0% over 55 trials, demonstrating the framework's embodiment-agnostic nature and zero-shot transferability. Additional quantitative results are available in Supp. Mat.

5. Conclusion

In this work, we introduce *VidBot*, a scalable and effective framework that enables robots to learn manipulation skills directly from in-the-wild RGB-only human videos. *VidBot* demonstrates substantial generalization capabilities, outperforming existing methods by 20% in success rate across 13 manipulation tasks in simulators in a zero-shot setting.

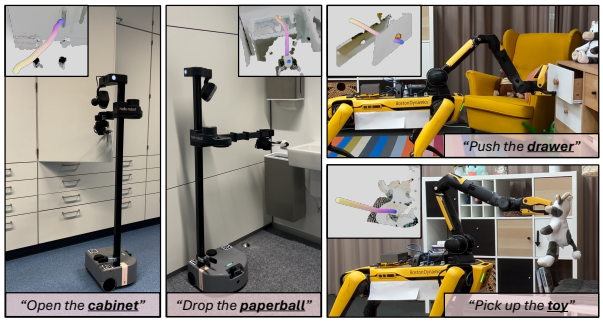

Figure 6. Real-world robotic manipulation tasks with inferred affordance displayed in the top panels.

Moreover, its embodiment-agnostic design allows for deployment across robot platforms, enabling successful executions of diverse household tasks in several real-world environments. The superior performance in downstream robot learning applications further underscores its versatility. One limitation of our framework is that the data quality is constrained by the accuracy of the depth foundation model and the SfM pipeline despite using the final optimization loss to filter low-quality labels. Given the method-agnostic nature of our data extraction pipeline, exploring recent learning-based SfM frameworks [21, 80, 93] could further enhance labels' quality. In future work, we plan to extract multi-modal affordance data using wearable devices, enabling robots to learn highly precise tasks like unscrewing caps, which currently remain challenging within our framework.

Acknowledgment

This work was funded by TUM Georg Nemetschek Institute (GNI) via project SPAICR as well as gift funding from Google LLC. We thank Simon Schaefer, Simon Boche, Yannick Burkhardt, Leonard Freissmuth, Daoyi Gao, Yao Zhong for proofreading and fruitful discussions.

References

[1] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? *arXiv preprint arXiv:2211.15657*, 2022. 3

[2] Arpit Bahety, Priyanka Mandikal, Ben Abbatematteo, and Roberto Martín-Martín. Screwmimic: Bimanual imitation from human videos with screw space projection. *arXiv preprint arXiv:2405.03666*, 2024. 2

[3] Shikhar Bahl, Abhinav Gupta, and Deepak Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022. 2, 3

[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. 2, 3, 6, 7, 8

[5] Homanga Bharadhwaj, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Zero-shot robot manipulation from passive human videos. *arXiv preprint arXiv:2302.02011*, 2023. 3

[6] Homanga Bharadhwaj, Abhinav Gupta, Vikash Kumar, and Shubham Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6904–6911. IEEE, 2024. 3

[7] Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. *arXiv preprint arXiv:2405.01527*, 2024. 2

[8] 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. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 5

[10] Matthew Chang, Aditya Prakash, and Saurabh Gupta. Look ma, no hands! agent-environment factorization of egocentric videos. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[11] Annie S Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. *arXiv preprint arXiv:2103.16817*, 2021. 3

[12] Hanzhi Chen, Binbin Xu, and Stefan Leutenegger. Funcgrasp: Learning object-centric neural grasp functions from single annotated example object. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1900–1906. IEEE, 2024. 2

[13] Jiaqi Chen, Boyang Sun, Marc Pollefeys, and Hermann Blum. A 3d mixed reality interface for human-robot teaming. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11327–11333. IEEE, 2024. 1

[14] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 3

[15] Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-sdf: Conditional generative modeling of signed distance functions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2262–2272, 2023. 3

[16] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 2

[17] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 7

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 4

[19] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[20] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018. 2

[21] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. 8

[22] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1

[23] Daoyi Gao, Dávid Rozenberszki, Stefan Leutenegger, and Angela Dai. Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 3

[24] Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, and He Wang. Partmanip: Learning cross-category generalizable part manipulation policy from point cloud observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2978–2988, 2023. 2, 6

[25] 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. 6, 7

[26] R Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015. 4

[27] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3293–3303, 2022. 2

[28] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 1

[29] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 6

[30] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019. 6

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 4

[32] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2, 3, 5

[33] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[34] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 4

[35] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022. 1

[36] Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, 2022. 3, 4

[37] Edward Johns. Coarse-to-fine imitation learning: Robot manipulation from a single demonstration. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 4613–4619. IEEE, 2021. 1

[38] Bingyi Kang, Xiao Ma, Chao Du, Tianyu Pang, and Shuicheng Yan. Efficient diffusion policies for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[39] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 1

[40] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

[41] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[42] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024. 6, 7

[43] Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. *arXiv preprint arXiv:2404.17521*, 2024. 3, 6

[44] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17562–17571, 2022. 3

[45] Zhixuan Liang, Yao Mu, Mingyu Ding, Fei Ni, Masayoshi Tomizuka, and Ping Luo. Adaptdiffuser: Diffusion models as adaptive self-evolving planners. *arXiv preprint arXiv:2302.01877*, 2023. 3

[46] Zhixuan Liang, Yao Mu, Hengbo Ma, Masayoshi Tomizuka, Mingyu Ding, and Ping Luo. Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16467–16476, 2024. 3

[47] Ziwei Liao, Binbin Xu, and Steven L Waslander. Toward general object-level mapping from sparse views with 3d diffusion priors. *arXiv preprint arXiv:2410.05514*, 2024. 3

[48] 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. 2, 3

[49] 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. 4

[50] 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 (CVPR)*, pages 21013–21022, 2022. 6

[51] Xiao Ma, Sumit Patidar, Iain Haughton, and Stephen James. Hierarchical diffusion policy for kinematics-aware multi-task robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18081–18090, 2024. 3

[52] 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. 2, 3

[53] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 6

[54] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 2, 6, 7

[55] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015. 2

[56] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 2

[57] 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. 2, 3

[58] Chuanruo Ning, Ruihai Wu, Haoran Lu, Kaichun Mo, and Hao Dong. Where2explore: Few-shot affordance learning for unseen novel categories of articulated objects. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[59] 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, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024. 1, 6, 7

[60] 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. 6

[61] Georgios Papagiannis, Norman Di Palo, Pietro Vitiello, and Edward Johns. R+ x: Retrieval and execution from everyday human videos. *arXiv preprint arXiv:2407.12957*, 2024. 3

[62] Sebeom Park, Shokhrukh Bokijonov, and Yosoon Choi. Review of microsoft hololens applications over the past five years. *Applied sciences*, 11(16):7259, 2021. 1

[63] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 5

[64] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988. 1

[65] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 2, 3

[66] 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. 4

[67] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023. 2, 3

[68] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[69] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13756–13766, 2023. 5

[70] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[71] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[72] 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. 3

[73] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR, 2023. 2, 3

[74] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. *arXiv preprint arXiv:2202.10448*, 2022. 3

[75] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019. 3

[76] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Larina, Diane Larlus, Dima Damen, and Andrea

Vedaldi. EPIC Fields: Marrying 3D Geometry and Video Understanding. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023. 4

[77] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 5

[78] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 2, 3

[79] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*, 2024. 2, 3

[80] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 8

[81] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say interact as you can: Language-guided human motion generation with scene affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 3

[82] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning, 2023. 3

[83] 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. 2

[84] Zhou Xian, Nikolaos Gkanatsios, Theophile Gervet, Tsung-Wei Ke, and Katerina Fragkiadaki. Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation. In *7th Annual Conference on Robot Learning*, 2023. 3

[85] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022. 2, 3

[86] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7827–7834. IEEE, 2021. 3

[87] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121, 2024. 4

[88] Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023. 2, 3

[89] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024. 3

[90] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 3, 4

[91] Jianglong Ye, Jiashun Wang, Binghao Huang, Yuzhe Qin, and Xiaolong Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *IEEE Robotics and Automation Letters*, 8(5):2882–2889, 2023. 3

[92] Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *Conference on Robot Learning*, 2024. 2, 6, 7

[93] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 8

[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. 3

[95] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5628–5635. IEEE, 2018. 1

[96] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[97] Yifeng Zhu, Arisrei Lim, Peter Stone, and Yuke Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024. 2