Hand-Object Interaction Pretraining from Videos

Anonymous Author(s)
Affiliation
Address
email

Abstract: We present an approach to learn *general robot manipulation priors* from 3D hand-object interaction trajectories. We build a framework to use in-the-wild videos to generate sensorimotor robot trajectories. We do so by lifting both the human hand and the manipulated object in a shared 3D space and retargeting human motions to robot actions. Generative modeling on this data gives us a task-agnostic base policy. This policy captures a general yet flexible manipulation prior. We empirically demonstrate that finetuning this policy, with both reinforcement learning (RL) and behavior cloning (BC), enables sample-efficient adaptation to downstream tasks and simultaneously improves robustness and generalizability compared to alternate approaches. Qualitative experiments are available at: hopretraining.site.

Keywords: Learning from videos, dexterous manipulation

1 Introduction

Reusable sensorimotor representations have the potential to give robots access to the versatility of their sensorimotor apparatus, thereby enabling them to achieve a wide variety of goals. Similar to advancements in other AI domains [1, 2], such representations are likely to be trained with unsupervised objectives on large datasets. In this work, we study the feasibility of training such representations using human videos in the context of dexterous manipulation.

Using videos as a data engine comes with several advantages: (1) they are abundant; (2) they cover a wide range of skills that we want robots to master; and (3) they reflect natural or socially acceptable behaviors that we want robots to emulate. However, training sensorimotor representations on videos is a challenging endeavor. First, videos only partially capture the nature of an agent's interaction with their surroundings. For instance, by looking at a person holding an object, it is almost impossible to estimate the force their fingers are exerting. In addition, the larger the embodiment gap between a human and a robot, the more their actions will differ to achieve the same objectives.

The difficulty of learning from videos led previous work to mostly focus on specific aspects of the problem. One line of research focused on training visual representations with off-the-shelf self-supervised vision algorithms on large vision datasets [3, 4, 5, 6, 7]. While simple and effective, such pretrained representations lack a motor component, making them less effective on downstream tasks [8]. Another line of work aims to extract both sensory and motor information from videos by estimating human motions in 3D [9, 10, 11, 12, 13]. However, these approaches require alignment between the training videos and the robot's downstream tasks, which compromises the generality of the learned representations. And most of these works overlook the trajectory of the object the human interacts with, resulting in representations that capture only the distribution of human motions.

In this paper, we present an approach to capture a general manipulation prior from in-the-wild videos. Such a prior is implicitly embedded in the weights of a causal transformer, pretrained with a conditional distribution matching objective on sensorimotor robot trajectories. These trajectories are generated by mapping 3D hand-object interactions to the robot's embodiment via a physically

Submitted to the 8th Conference on Robot Learning (CoRL 2024). Do not distribute.

Figure 1: Real world rollouts of the policy finetuned from HOP using less than 50 demonstrations. HOP enables sample-efficient downstream adaptaion by learning a general manipulation prior from human videos.

grounded simulator. The resulting prior can be quickly adapted to any manipulation task either with reinforcement learning or behavioral cloning. After adaptation, the prior takes the form of an *end-to-end* policy mapping the robot's multi-modal sensory stream to low-level joint commands. This policy can be directly executed on a physical robot (see Fig. 1).

We empirically study the advantages brought forward by pretraining with hand-object interactions in both simulation and real-world experiments. The findings of this study indicate that our manipulation prior considerably speeds up skill acquisition compared to previous methods, even if such skills are not represented in the training videos. Additionally, it improves generalization and robustness to disturbances in the downstream policy.

2 Overview

The objective of **H**and-**O**bject interaction **P**retraing (HOP) is to capture general hand-object interaction priors from videos, such as the maneuvers required to approach objects and the appropriate hand poses to hold them. In contrast to previous work, we do not assume a strict alignment of the human's intent in the video and the downstream robot tasks. Our key intuition is that the basic skills required for manipulation lie on a manifold whose axes are well covered by unstructured human-object interactions.

We extract sensorimotor information from videos by lifting the human hand and the manipulated object in a shared 3D space. We then bring such 3D representations to a physics simulator, where we map human motion to robot actions. There are several advantages to using a simulator as an intermediary between videos and robot sensorimotor trajectories: (i) we can add physics, inevitably lost in videos, back to the interactions; (ii) it enables the synthesis of large training datasets without putting the physical platform in danger; and (iii) we can add diversity to the data by randomizing the simulation environment, e.g., varying the friction between the robot's joints, the scene's layout, and the object's location relative to the robot.

We generate a dataset of robot-object interactions $\mathcal{D} = \{\tau_1, \tau_2, \dots, \tau_N\}$ where $\tau = \{(\boldsymbol{o}[o], \boldsymbol{a}[1]), (\boldsymbol{o}[1], \boldsymbol{a}[1]), \dots, (\boldsymbol{o}[T], \boldsymbol{a}[T])\}$ are the observation-action pairs of a single sensorimotor trajectory. An observation $\boldsymbol{o}[k] \in \mathbb{O}$ at time $k \in [0, \dots, T]$ consists of a depth image $\mathcal{I}[k]$ and robot's joint angles $\phi[k]$, *i.e.*, its proprioception. The action $\boldsymbol{a}[k] \in \mathbb{A}$ consists of continuous joint angles, which are converted to joint torques with a low-level PD controller. We use $\mathcal{D}$ to train a

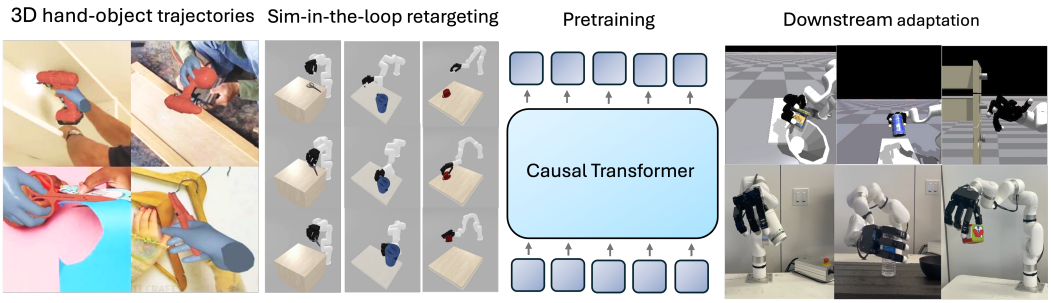

Figure 2: 3-D hand-object trajectories from in-the-wild human manipulation videos are re-targeted to a robot embodiment within a physics simulator, resulting in physically grounded robot data. General manipulation priors are learnt from this using generative modelling of trajectories. Such representation enables sample-efficient adaptation for new downstream tasks.

base policy $\boldsymbol{\pi}_b$ on the unsupervised objective of next-action prediction from a history of sensory observations, *i.e.*, $\hat{\boldsymbol{a}} = \boldsymbol{\pi}_b(\boldsymbol{o}[t : t - L])$, where L is a fixed context length. We finetune $\boldsymbol{\pi}_b$ to generate task-specific policies $\boldsymbol{\pi}_t$ either optimizing a reward with reinforcement learning or a behavioral cloning objective on few task-specific demonstrations. The next section presents each aspect of our method in detail.

3 Method

3.1 Lifting Hand-Object Interaction Videos to 3D

Recovering the underlying 3D structure of hand-object interactions from in-the-wild monocular videos is inherently ambiguous. To alleviate such ambiguity, previous work leveraged the insight that the human hand can be used as an anchor for the 3D location and scale of the manipulated object [14, 15, 16, 17, 18, 19]. Our setup to estimate hand-object interaction trajectories from videos builds upon recent advances in 3D vision. Our approach closely follows MCC-HO [14] with a few modifications to adapt it to our use case.

Given a single RGB image and an estimate of the 3D hand geometry from HaMeR [20], MCC-HO jointly infers hand-object geometry as point clouds. To fine-tune the quality of the prediction, MCC-HO finetunes the object's pose by fitting it to a CAD model. However, this finetuning assumes knowledge of the object the human is interacting with. To increase generality, we wave this assumption and skip the CAD-based post-processing. This simplification comes at the cost of reduced reconstruction quality and temporal smoothness. While we find the first problem not critical for pre-training, we increase temporal smoothness by anchoring object reconstructions to time-smoothed hand detections [20]. In addition, we make the simplifying assumption that the camera from which the video is collected is static. More details of our 3D estimation pipeline are provided in the appendix. The result of this pipeline is a sequence of 3D hand-object poses.

3.2 Mapping 3D Human-Object Interactions to Robot-Object Interactions

We formulate a non-linear optimization problem to generate a sensorimotor trajectory τ from a sequence of 3D hand-object poses. At each step k, we find the action $\boldsymbol{a}[k]$ by optimizing the following cost function:

$$\min_{\boldsymbol{a}[k]} \frac{1}{2}\|\boldsymbol{x}_h[k] - f(\boldsymbol{a}[k])\|^2 + \lambda\|\boldsymbol{a}[k] - \boldsymbol{\phi}[k-1]\|^2 \quad \text{s.t.} \quad \boldsymbol{a}[k] \in \mathbb{A}, \tag{1}$$

where f is the robot's forward kinematics, and $\boldsymbol{x}_h[k]$ are the 3D coordinates of a set of keypoints on the human hand. While there are approaches that include the object dynamics in the optimization [21, 22, 23, 24], they are challenging to apply to in-the-wild videos due to noise in object pose estimates. Therefore, for simplicity, we disregard the dynamics of the manipulated object and place

it on every step at the location observed in the video. While this can lead to physical implausibility in the object motion and possibly lack of force-closure grasps, the data quality does not deteriorate much in practice, as we can still learn useful behaviors.

The optimization is performed independently on each timestep k, and the resulting actions $\boldsymbol{a}[k]$ are executed in a high-fidelity simulator to generate $\boldsymbol{o}[k]$. We refer to this method of mapping human motion to robot sensorimotor trajectories τ as *simulator-in-the-loop retargeting*. The primary advantage of this approach is that the optimization in (1) can be conducted using a simplified forward kinematics f, reducing the computational burden. Despite this simplification, the actions are executed in a high-fidelity simulator, ensuring realistic behaviors and high-quality observations.

We randomize the simulated scene to increase data diversity. Specifically, we add obstacles like tables and walls to the scene and vary their positions relative to the robot. We use such a heuristic to generate different trajectories from a single video. Note that this approach to retargeting differs from previous methods, disregarding other objects in the scene and optimizing actions via physics-based constraints, e.g., minimum jerk [12, 25, 9] or minimum velocity [13].

The quality of the resulting robot trajectories decreases as the difference between the environment where the video was collected, the simulated scene, and f grows. However, given the non-convex optimization landscape of (1), we can obtain good trajectories by running the optimization multiple times with various initial positions and scene layouts. High-quality data is then obtained by pruning the trajectories on metrics like collision with obstacles and the tracking error between the hand's and the robot's keypoints.

3.3 Robot Trajectory Pretraining

The resulting trajectory dataset $\mathcal{T}$ contains knowledge that could be valuable to any manipulation tasks. For instance, $\mathcal{T}$ has information about object affordance, *i.e.*, where and how to grasp; some intuitive (although rudimentary) physics, *e.g.*, an object should be reached upon before being lifted; or wrist-hand coordination, *i.e.*, the behavior of orienting and shaping the hand simultaneously while moving the wrist to maximize efficiency [26].

We aim to incorporate this knowledge as useful behavioral priors into a policy π_b that can be finetuned to downstream tasks. Similar to previous work in language [2], vision [27], and robotics [8, 28, 29], we instantiate π_b as a transformer [30] and train it on a generative modeling objective. Specifically, we train π_b to capture the conditional distribution $\Pi(\boldsymbol{a}[t-L:t]|\boldsymbol{o}[t-L:t])$ by optimizing the following loss:

$$\mathcal{L}(\tau;\theta) = \mathbb{E}_{t\sim[1\ldots T]}\left[\|\boldsymbol{a}[t-L:t] - \pi_b(\boldsymbol{o}[t-L:t])\|_1\right]. \tag{2}$$

However, unlike previous work, our pretraining dataset $\mathcal{T}$ contains neither real-world demonstrations nor complete task executions. This is because our data is generated from unstructured 3D hand-object interactions, and we disregard the dynamics of the manipulated object during retargeting (Sec. 3.2). Yet, we find that the pre-training paradigm in (2) leads to the emergence of useful representations in π_b.

Downstream Finetuning. The pretrained policy π_b exhibits primitive manipulation skills, *e.g.*, reaching an object while occasionally lifting it. We finetune these skills to a task by optimizing a reward with reinforcement learning or a behavior cloning loss on limited demonstrations. We finetune the whole model for the task. Empirically, we find that finetuned policies use the information in π_b to train faster, are more robust to disturbances, and generalize better than policies trained from scratch and a set of baselines. In addition, we find that the finetuning process re-utilizes the information in π_b even for tasks not explicitly represented in the training videos.

4 Experimental Setup

Robot. We use a low-cost 7-DoF xArm robot with a 16-DoF Allegro hand [31] vertically mounted at its end effector. The proprioception observation $\boldsymbol{\phi}_k$ includes joint position from both robots.

While we don't make any specific assumption about the robot embodiment, we use a multi-fingered hand instead of a parallel joint gripper since demonstration quality increases as the embodiment gap between the robot and the human decreases. We empirically found that since the robot base is fixed, a 7-DoF arm can track much better human motions than a 6-DoF arm, which often encounters singularities during such trajectories. Visual sensing comes from a single stereo camera (Zed-2) mounted on the robot's right side.

Simulation Setup. Our simulation environment is developed with the IsaacGym [32] simulator. The robot morphology and action space are identical to the real setup. However, since rendering depth images is prohibitively expensive, we give the agent access to the ground-truth object point-could instead of a depth image (see Section 2). Specific details about the task setup and reward design can be found in the Appendix.

Video Datasets. Our pretraining dataset of 3D hand-object trajectories consists of sequences from two datasets: DexYCB [33] and 100 Days of Hands [34]. We use 250 videos from the DeXYCB dataset (right-hand only) annotated with ground truth hand-object trajectories as a source of high-quality data. We additionally use approximately 200 videos from the 100 Days of Hands dataset. Sixty percent of these videos were previously annotated with hand-object interaction trajectories [11], which we directly use. We annotate the remaining videos with our 3D estimation pipeline (Sec. 3.1). Overall, our combined dataset contains approximately 450 videos. We retarget these videos to obtain a pretraining dataset $\mathcal{T}$ of approximately $70,000$ trajectories.

Retargeting. We use low-storage BFGS [35] from the NLOpt library [36] for optimization. We perform simulation-in-the-loop retargeting in a simple simulated scene with a ground floor on which the robot and a static table are placed 65cm apart. Objects start their trajectories above the table with a random pose. We run the optimization 700 times for each video, randomizing the table location and the robot's initial joint state. We add a trajectory to $\mathcal{T}$ only if, at any time, their retargeting error (See Eq. (1)) is below 3cm and the arm does not collide with the table or the floor. Our code is built upon the implementation of Qin et al. [37].

Transformer. Similar to previous work [38], we represent the policy π_b with a GPT-2-style causal transformer. The policy takes proprioception and observation input from the past 16 timesteps and predicts the next action. Details about the architecture can be found in the Appendix.

Pretraining. We train the transformer with the objective in Eq. (2) on $\mathcal{T}$. While we could make the prediction autoregressive and add decoding heads and proxy losses for future proprioception and images (as in [13]), we empirically found these changes to be not very helpful in practice to our tasks. Therefore, we predict only future actions for simplicity. We use as optimizer AdamW [39] with initial learning rate of 10^{-4} and weight decay of 10^{-2}.

Finetuning. *In simulation*, we finetune the transformer with PPO [40] using the default hyperparameters from [41]. However, we add a few modifications inspired by [42] for effective fine-tuning: (1) we use a small initial exploration noise of 0.1; (2) the value and policy networks share the observation tokenizer, but the tokenizer's weights are not updated by the value function's gradients; (3) we warm up the value function's parameters for the first 200 gradient steps, keeping the actor parameters fixed. *In the real world*, we finetune the entire π_b on limited demonstrations with the same objective and hyperparameters used for pretraining.

Inference. At test time, the model operates in *closed-loop*: it receives the past and current observations as input and predicts the next action to execute. The prediction loop runs at 20Hz. The predicted action is sent to the xArm and Allegro low-level controllers, which operate at 120Hz and 300Hz, respectively.

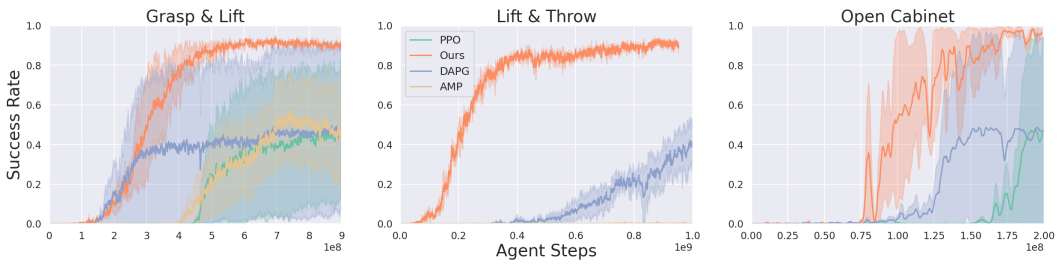

Figure 3: **Comparison of HOP-initialized actor with baselines.** HOP improves sample-efficiency of online RL across multiple tasks, particularly when the downstream task and the behaviors in the data are less aligned, as in *Lift & Throw*. Runs are averaged across two randomly chosen seeds.

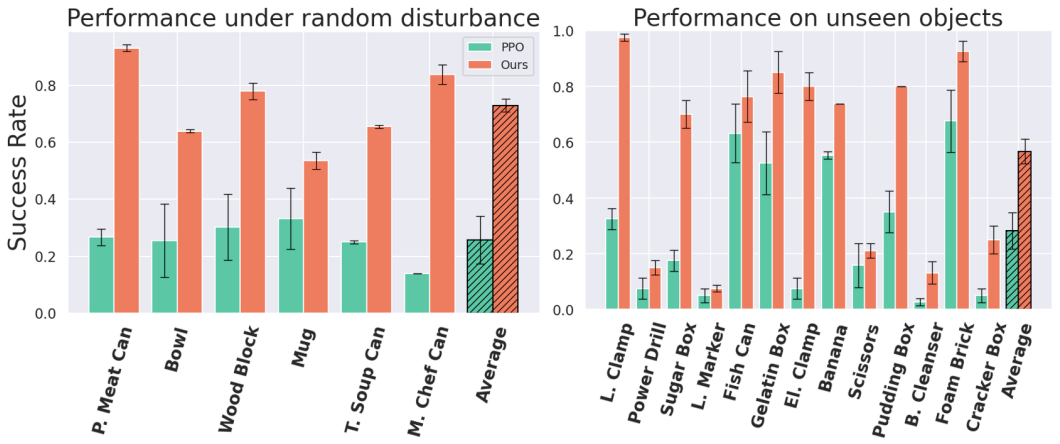

Figure 4: **Evaluating RL finetuning under out-of-distribution scenarios** (Left) To test grasp robustness in the task *Grasp & Lift*, we apply to the grasped objects, forces in random direction equal to their weights. When initialized with HOP, the resulting policy is more than $3\times$ more robust compared to training PPO from scratch. (Right) We evaluate grasp success on multiple objects from the YCB dataset that were not part of the training set. When initialized with HOP, the resulting policy is more than $2\times$ more robust compared to training PPO from scratch.

5 Experimental Results

We design an experimental procedure to analyze the advantages brought forward by HOP in terms of finetuning efficiency, generalization, and robustness to perturbations. Specifically, we ask the following questions: (i) *Is our pretraining-finetuning method more effective than popular approaches on demonstration-guided policy learning?* (ii) *Is the manipulation prior π_b still useful when finetuning on skills not present in the training hand-object interaction videos?* (iii) *Is the manipulation prior still useful when the downstream task involves more than a single object?* We answer these questions via controlled experiments in simulation and the physical world.

5.1 Simulation Experiments

Baselines. We compare our approach with three baselines: (1) training from scratch (*PPO*); (2) demonstration-guided reinforcement learning with a proxy imitation objective [43] (*DAPG*); and (3) using adversarial objectives to keep the policy close to the demonstrations [44] (*AMP*). Similarly to previous work [45], we found that training from scratch is unsuccessful using joint-position control as action space, which consistently leads the PPO baseline to fail. Therefore, we use the moving-average action space proposed by Petrenko et al. [41] to improve its performance.

Tasks and Metrics. We evaluate approaches on three tasks. The first requires to pick an object and place it at a specific location (*Grasp and Lift*). The second is to grasp an object and throw it in a

basket (*Grasp and Throw*). In the final task, the robot is required to open a cabinet. Our pre-training video dataset mostly comprises grasp-and-lift interactions, with no video demonstrating the other two tasks. We evaluate performance using success over 256 environments with different objects and report the mean and standard deviation over two seeds per approach. More details can be found in the appendix.

HOP enables sample-efficient RL and effective exploration As shown in Figure 3, HOP provides a 2-5X improvement in sample efficiency over training from scratch. Initializing with HOP allows for more informed exploration, which decreases variance in the policy gradients. Our method outperforms AMP and DAPG on all three tasks. The contrast is particularly marked when the pre-training corpus is not aligned to the task. This shows that the sensorimotor representations learned by our data modeling objective can be reused beyond the skills presented at pre-training time.

HOP learns robust and general behaviors Policies fine-tuned from HOP can potentially bias exploration toward human-like behavior, leading to more robustness against forces. This is shown in Fig. 4. Agents trained with our approach perform better when subject to forces than the ones trained from scratch. In addition, we show in Fig. 4 that our approach generalizes 3x better than the policy trained from scratch. Note that we train the scratch policy with two billion samples in these comparisons.

5.2 Real World Results

Our real-world experiments are presented in Fig. 5. We evaluate our approach on three tasks of increasing complexity. In the first task, *Grasp and Drop*, the robot needs to unstack a cube and put it in a bowl. This is a very easy task and even with a few demonstrations our approach can get an 80% success rate. The second is the *Grasp and Pour* task, where the robot needs to pick one of two bottles and point them towards a bowl. This requires more dexterity but is still applied to only two similarly-looking bottles. Therefore, only limited demonstrations were enough to achieve good performance. It is interesting to note that in the latter two tasks, the robot had to interact with more than one object. However, the pretraining prior only contained a single object. Despite this, finetuning could still benefit from the base policy π_b.

In the final task, *Grasp and Lift*, the robot must pick up one of 4 different-looking objects, all requiring different spatial affordances. One single model is trained to pick up all objects. In this task, we evaluate the ability of the approach to adapt with a few demonstrations on very different object shapes. This obviously required more demonstrations, and the final success rates were lower. We encourage the reader to check our supplementary website for videos of policy executions. More details about the scene setup and evaluation criteria can be found in the appendix.

We trained a policy from scratch on these demonstrations but obtained consistently zero performance in multiple experiments. In addition, since our setup is very different from classic imitation learning ones (with a single depth camera), we could not evaluate pre-trained models [4, 5].

6 Related Work

Learning Policies from Human Videos. In-the-wild videos hold the promise of solving the data problem in robotics. One of the pioneering efforts in this direction is by Yang et al. [46], where video data was used to generate action plans. Several works followed up on this idea, relying on pre-defined action primitives [47, 48, 49, 50, 51, 52]. However, waving the requirement for pre-defined primitives is challenging since in-the-wild videos lack motor information. One way to recover motor information from videos is training with a trajectory-matching objective [53, 54, 55, 56], possibly using intermediate representations like object segmentation or optical flow [57, 58, 59]. However, this approach requires collecting task- and environment-specific videos where humans and robots operate in the same workspace. Therefore, the trajectory-matching formulation largely constrains the number of videos that can be used for training. To overcome these constraints, researchers have

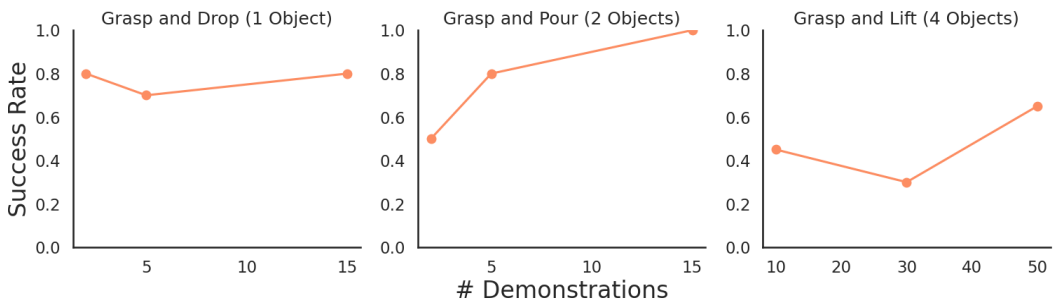

Figure 5: **Few-shot BC finetuning** Initialising the policy with HOP enables successful finetuning to new tasks with less than 50 demonstrations. For each task form left to right, we evaluate both the policy trained from scratch and the finetuned policy for 10, 10 and 20 trials respectively. The policy trained from scratch leads to no success across all tasks.

focused on either learning exclusively visual representations from videos or extracting 3D human poses and mapping them to robot actions. In the following, we cover these works in detail.

Visual Representation Learning for Robotics. Inspired by successes in computer vision [1] and natural language processing [60], the robot learning community has recently focused on pretraining representations on large video datasets like Ego4D [61] and fine-tuning these representations on downstream tasks [4, 7, 3, 5]. However, being the training objective based exclusively on image reconstruction, the representations focus primarily on low-level vision features, *e.g.*, shapes or edges. This gives them limited benefits compared to representations trained on standard vision datasets [6]. Overall, these works focus on visual generalization, *e.g.*, picking up two objects with the same shape but different colors. However, they have not yet demonstrated action generalization, where motor skills are adapted to accomplish novel objectives.

Actions from Videos via 3D. One common approach to extracting action information from videos is using 3D as an intermediate representation. If the embodiment gap is small, human motions can be mapped to robot actions via inverse kinematics. This is particularly effective when the videos are task-specific, *i.e.*, when the robot aims to mimic the human motion [10, 12, 62, 63, 64, 65, 11]. Instead of learning specific skills, other works focus on learning a re-usable sensorimotor prior from videos. However, this prior only captures human actions [66, 67, 9], disregarding the trajectory of the manipulated object. Conversely, our work aims to use 3D hand-object interactions from in-the-wild videos to learn a re-usable prior for object manipulation.

Dexterous Manipulation. Dexterous manipulation has been studied for decades [68, 69, 70, 71, 72]. In recently years, learning-based approaches make significant progress [73, 74]. They can be generally categorized to learning in simulation and then transferring to the real world (Sim-to-Real) [75, 76, 77, 78, 79, 80], and learning in the real-world [37, 81, 82, 83, 84]. Qin et al. [12] uses hand-object trajectories but only use it to collect demonstrations. Xu et al. [85] also does functional grasp generation, but the results are limited in simulation. Most of the aforementioned work learn policies from scratch and does not use any internet data as prior. Our approach studies learning hand-object interaction prior from human videos and is effective both in simulation and real-world.

7 Conclusion and Limitations

This work presents an approach to learning general yet flexible manipulation priors for robot policies from human videos. While our approach demonstrates a way to pre-train on a single object interaction, this can, in practice, be limiting. Indeed, human behavior in a video can potentially be conditioned on information encompassing multiple objects in the current and previous scenes. This leads to a loss of signal that could be extracted from the raw video. We predict that advances in 3-D reconstruction will enable us to use a more complex scene reconstruction and pretraining.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[3] Y. J. Ma, S. Sodhani, D. Jayaraman, O. Bastani, V. Kumar, and A. Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

[4] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022.

[5] A. Majumdar, K. Yadav, S. Arnaud, J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, T. Wu, J. Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36, 2024.

[6] S. Dasari, M. K. Srirama, U. Jain, and A. Gupta. An unbiased look at datasets for visuo-motor pre-training. In *Conference on Robot Learning*, pages 1183–1198. PMLR, 2023.

[7] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[8] I. Radosavovic, B. Shi, L. Fu, K. Goldberg, T. Darrell, and J. Malik. Robot learning with sensorimotor pre-training. In *Conference on Robot Learning*, pages 683–693. PMLR, 2023.

[9] K. Shaw, S. Bahl, and D. Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning*, pages 654–665. PMLR, 2023.

[10] X. B. Peng, A. Kanazawa, J. Malik, P. Abbeel, and S. Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics (TOG)*, 37(6):1–14, 2018.

[11] A. Patel, A. Wang, I. Radosavovic, and J. Malik. Learning to imitate object interactions from internet videos. *arXiv preprint arXiv:2211.13225*, 2022.

[12] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022.

[13] I. Radosavovic, B. Zhang, B. Shi, J. Rajasegaran, S. Kamat, T. Darrell, K. Sreenath, and J. Malik. Humanoid locomotion as next token prediction. *arXiv preprint arXiv:2402.19469*, 2024.

[14] J. Wu, G. Pavlakos, G. Gkioxari, and J. Malik. Reconstructing hand-held objects in 3d. *arXiv:2404.06507*, 2024.

[15] H. Choi, N. Chavan-Dafle, J. Yuan, V. Isler, and H. Park. Handnerf: Learning to reconstruct hand-object interaction scene from a single rgb image. *arXiv preprint arXiv:2309.07891*, 2023.

[16] Y. Ye, A. Gupta, and S. Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3895–3905, 2022.

[17] B. Tekin, F. Bogo, and M. Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019.

[18] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019.

[19] Z. Cao, I. Radosavovic, A. Kanazawa, and J. Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021.

[20] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[21] X. Zhu, J. Ke, Z. Xu, Z. Sun, B. Bai, J. Lv, Q. Liu, Y. Zeng, Q. Ye, C. Lu, et al. Diff-lfd: Contact-aware model-based learning from visual demonstration for robotic manipulation via differentiable physics-based simulation and rendering. In *Conference on Robot Learning*, pages 499–512. PMLR, 2023.

[22] A. S. Lakshmipathy, N. Feng, Y. X. Lee, M. Mahler, and N. Pollard. Contact edit: Artist tools for intuitive modeling of hand-object interactions. *ACM Transactions on Graphics (TOG)*, 42 (4):1–20, 2023.

[23] A. S. Lakshmipathy, J. K. Hodgins, and N. S. Pollard. Kinematic motion retargeting for contact-rich anthropomorphic manipulations. *arXiv preprint arXiv:2402.04820*, 2024.

[24] Y. Kim, H. Park, S. Bang, and S.-H. Lee. Retargeting human-object interaction to virtual avatars. *IEEE transactions on visualization and computer graphics*, 22(11):2405–2412, 2016.

[25] A. Sivakumar, K. Shaw, and D. Pathak. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube. *arXiv preprint arXiv:2202.10448*, 2022.

[26] L. A. Jones and S. J. Lederman. *Human hand function*. Oxford university press, 2006.

[27] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020.

[28] A. Kumar, A. Singh, F. Ebert, M. Nakamoto, Y. Yang, C. Finn, and S. Levine. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials. *arXiv preprint arXiv:2210.05178*, 2022.

[29] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023.

[31] WonikRobotics. Allegrohand. https://www.wonikrobotics.com/, 2013.

[32] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021.

[33] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. V. Wyk, U. Iqbal, S. Birchfield, J. Kautz, and D. Fox. Dexycb: A benchmark for capturing hand grasping of objects. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021. URL https://api.semanticscholar.org/CorpusID:233210016.

[34] D. Shan, J. Geng, M. Shu, and D. 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.

[35] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[36] S. G. Johnson. The NLopt nonlinear-optimization package. https://github.com/stevengj/nlopt, 2007.

[37] F. O. H. to Multiple Hands: Imitation Learning for Dexterous Manipulation from Single-Camera Teleoperation. Qin, yuzhe and su, hao and wang, xiaolong, 2022.

[38] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[39] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[40] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[41] A. Petrenko, A. Allshire, G. State, A. Handa, and V. Makoviychuk. Dexpbt: Scaling up dexterous manipulation for hand-arm systems with population based training. In *RSS*, 2023.

[42] R. Ramrakhya, D. Batra, E. Wijmans, and A. Das. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17896–17906, 2023.

[43] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations, 2018.

[44] X. B. Peng, Z. Ma, P. Abbeel, S. Levine, and A. Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20, 2021.

[45] Y. Jiang, C. Wang, R. Zhang, J. Wu, and L. Fei-Fei. Transic: Sim-to-real policy transfer by learning from online correction. *arXiv preprint arXiv: Arxiv-2405.10315*, 2024.

[46] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[47] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis. Translating videos to commands for robotic manipulation with deep recurrent neural networks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3782–3788. IEEE, 2018.

[48] J. Lee and M. S. Ryoo. Learning robot activities from first-person human videos using convolutional future regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–2, 2017.

[49] V. Arapi, C. Della Santina, D. Bacciu, M. Bianchi, and A. Bicchi. Deepdynamichand: a deep neural architecture for labeling hand manipulation strategies in video sources exploiting temporal information. *Frontiers in neurorobotics*, 12:86, 2018.

[50] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019.

[51] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. 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.

[52] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation, 2024.

[53] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. *arXiv preprint arXiv:2207.09450*, 2022.

[54] C. Finn, I. Goodfellow, and S. Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

[55] T. Yu, C. Finn, A. Xie, S. Dasari, T. Zhang, P. Abbeel, and S. Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. *arXiv preprint arXiv:1802.01557*, 2018.

[56] J. Jin, L. Petrich, M. Dehghan, Z. Zhang, and M. Jagersand. Robot eye-hand coordination learning by watching human demonstrations: a task function approximation approach. In *2019 international conference on robotics and automation (icra)*, pages 6624–6630. IEEE, 2019.

[57] H. Bharadhwaj, A. Gupta, V. Kumar, and S. Tulsiani. Towards generalizable zero-shot manipulation via translating human interaction plans. *arXiv preprint arXiv:2312.00775*, 2023.

[58] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning. *ArXiv*, abs/2401.00025, 2023. URL https://api.semanticscholar.org/CorpusID:266693687.

[59] S. Kumar, J. Zamora, N. Hansen, R. Jangir, and X. Wang. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*, pages 55–66. PMLR, 2023.

[60] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[61] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. 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.

[62] J. Ye, J. Wang, B. Huang, Y. Qin, and X. Wang. Learning continuous grasping function with a dexterous hand from human demonstrations. *IEEE Robotics and Automation Letters*, 8(5): 2882–2889, 2023.

[63] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022.

[64] Z. Q. Chen, K. Van Wyk, Y.-W. Chao, W. Yang, A. Mousavian, A. Gupta, and D. Fox. Learning robust real-world dexterous grasping policies via implicit shape augmentation. *arXiv preprint arXiv:2210.13638*, 2022.

[65] Z. Q. Chen, K. Van Wyk, Y.-W. Chao, W. Yang, A. Mousavian, A. Gupta, and D. Fox. Dextransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations. *arXiv preprint arXiv:2209.14284*, 2022.

[66] H. Bharadhwaj, A. Gupta, S. Tulsiani, and V. Kumar. Zero-shot robot manipulation from passive human videos. *arXiv:2302.02011*, 2023.

[67] Y. Ze, Y. Liu, R. Shi, J. Qin, Z. Yuan, J. Wang, and H. Xu. H-index: Visual reinforcement learning with hand-informed representations for dexterous manipulation. *NeurIPS*, 2024.

[68] R. Fearing. Implementing a force strategy for object re-orientation. In *ICRA*, 1986.

[69] L. Han and J. C. Trinkle. Dextrous manipulation by rolling and finger gaiting. In *ICRA*, 1998.

[70] A. M. Okamura, N. Smaby, and M. R. Cutkosky. An overview of dexterous manipulation. In *ICRA*, 2000.

[71] M. T. Ciocarlie and P. K. Allen. Hand posture subspaces for dexterous robotic grasping. *IJRR*, 2009.

[72] A. S. Morgan, K. Hang, B. Wen, K. Bekris, and A. M. Dollar. Complex in-hand manipulation via compliance-enabled finger gaiting and multi-modal planning. *RA-L*, 2022.

[73] OpenAI, M. Andrychowicz, B. Baker, M. Chociej, R. Józefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, J. Schneider, S. Sidor, J. Tobin, P. Welinder, L. Weng, and W. Zaremba. Learning dexterous in-hand manipulation. *IJRR*, 2019.

[74] C. Yu and P. Wang. Dexterous manipulation for multi-fingered robotic hands with reinforcement learning: a review. *Frontiers in Neurorobotics*, 2022.

[75] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang. Solving rubik's cube with a robot hand. *arXiv:1910.07113*, 2019.

[76] A. Handa, A. Allshire, V. Makoviychuk, A. Petrenko, R. Singh, J. Liu, D. Makoviichuk, K. Van Wyk, A. Zhurkevich, B. Sundaralingam, Y. Narang, J.-F. Lafleche, D. Fox, and G. State. Dextreme: Transfer of agile in-hand manipulation from simulation to reality. In *ICRA*, 2023.

[77] T. Chen, M. Tippur, S. Wu, V. Kumar, E. Adelson, and P. Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 2023.

[78] H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik. In-hand object rotation via rapid motor adaptation. In *CoRL*, 2022.

[79] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik. General in-hand object rotation with vision and touch. In *CoRL*, 2023.

[80] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak. Dexterous functional grasping. In *CoRL*, 2023.

[81] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu. Dexcap: Scalable and portable mocap data collection system for dexterous manipulation. In *RSS*, 2024.

[82] I. Guzey, B. Evans, S. Chintala, and L. Pinto. Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play. In *CoRL*, 2023.

[83] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto. See to touch: Learning tactile dexterity through visual incentives. In *ICRA*, 2024.

[84] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto. Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation. In *ICRA*, 2023.

[85] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, et al. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *CVPR*, 2023.