Towards Generalizable Zero-Shot Manipulation via Translating Human Interaction Plans

Homanga Bharadhwaj[1], Abhinav Gupta[*,2], Vikash Kumar[*,2], Shubham Tulsiani[*,2]

Abstract— We pursue the goal of developing robots that can interact zero-shot with generic unseen objects via a diverse repertoire of manipulation skills and show how passive human videos can serve as a rich source of data for learning such generalist robots. Unlike typical robot learning approaches which directly learn how a robot should act from interaction data, we adopt a factorized approach that can leverage large-scale human videos to learn how a human would accomplish a desired task (a human 'plan'), followed by 'translating' this plan to the robot's embodiment. Specifically, we learn a human 'plan predictor' that, given a current image of a scene and a goal image, predicts the future hand and object configurations. We combine this with a 'translation' module that learns a plan-conditioned robot manipulation policy, and allows following humans plans for generic manipulation tasks in a zero-shot manner with no deployment-time training. Importantly, while the plan predictor can leverage large-scale human videos for learning, the translation module only requires a small amount of in-domain data, and can generalize to tasks not seen during training. We show that our learned system can perform over 16 manipulation skills that generalize to 40 objects, encompassing 100 real-world tasks for table-top manipulation and diverse in-the-wild manipulation. https://homangab.github.io/hopman/

I. INTRODUCTION

A central goal in the rapidly growing area of robot learning is to develop generalist robots capable of performing a plethora of everyday manipulation tasks in diverse unseen real-world scenarios. In addition, to be practically useful, they should be able to accomplish these tasks out of the box when deployed in unseen scenarios. Towards this goal, our work pursues learning diverse core skills like manipulating articulated objects, picking, placing, scooping, pouring, twisting, stacking, and swiping, among others that humans can effortlessly perform during everyday interactions. Moreover, we want these skills to be generalizable to unseen scenes with new objects, and be executeable in a "zero-shot manner" i.e. without deployment-time training.

An unsophisticated way to attempt this goal is to collect a gigantic robot interaction dataset for imitation learning. Albeit simple, this is not scalable for diverse real-world generalization because it would require collecting data not just for different tasks but for interaction across different objects with different skills, and is bottle-necked by physical access constraints. Indeed, recent approaches that attempt at developing diverse manipulation capabilities require years of on-robot data collection [1], and are still largely limited to picking, placing, and pushing skills. Our solution is to

*equal contribution
[1] HB is with Carnegie Mellon University and FAIR, AI at Meta
[2] AG, VK, and ST are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA
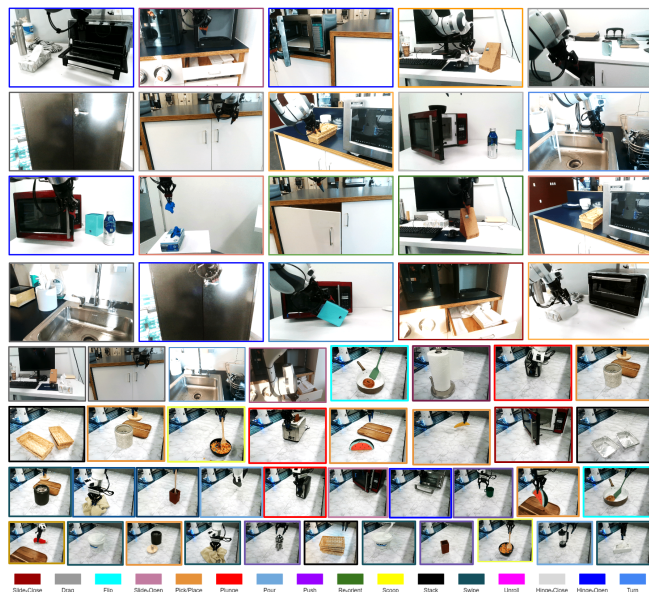
Fig. 1. A subset of different manipulation behaviors generated by our framework HOPMan . By learning task-agnostic *human-plan* prediction and *robot-action* translation models, our system can interact with generic objects and execute diverse skills e.g. unrolling, scooping, pouring, re-orientation, articulated object manipulation, etc. Videos are in the supplementary website https://homangab.github.io/hopman/

factorize the task of learning a generalizable policy into 1) learning an interaction plan that captures changes that the object and the manipulator can undergo, 2) translate the plan into actions that can be executed on a robot. Our key insight is that the first module can leverage non-robot data, and in particular large passive datasets of human videos on the web. Given this *human-interaction-plan*, acting in the real world reduces significantly in complexity as we only need to instantiate the human plan in a robot's context as *robot-actions*. This translation model can be trained with limited paired human-robot data and generalizes to objects and scenarios that are unseen in the robot data since the human-interaction plan generalizes by virtue of diverse training.

Some prior robot learning approaches have also investigated leveraging out-of-domain (human) data, primarily for learning visual representations [2, 3, 4] and robotic affordances [5, 6, 7, 8]. However, these approaches require *a lot of* further robot demonstrations for policy learning and typically also require a lot of deployment-time training. Other approaches learn task-specific action priors [9, 10] for a few categories of manipulation tasks, with separate policies for each task. Compared to these, our approach of factorizing the overall policy can enable zero-shot manipulation over a range of
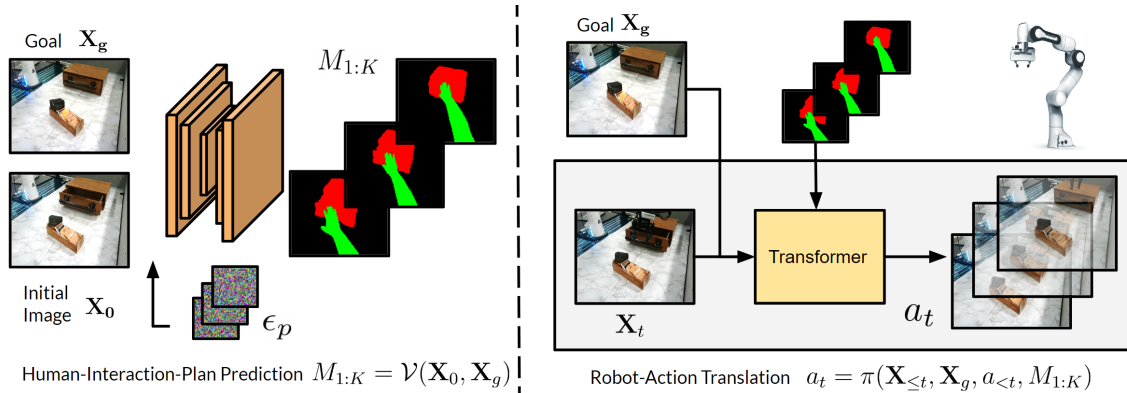
Fig. 2. `HOPMan` consists of a *human-interaction-plan prediction model* (left), and a *robot-action translation model* (right). Given an initial image of a scene $\mathbf{X_0}$ and a goal image $\mathbf{X_g}$, a diffusion model hallucinates plausible future hand and object masks $M_{1:K}$. These predictions along with current RGB observations of the scene $\mathbf{X}_t$ go as input to a translation model (instantiated as a closed-loop policy $\pi(\cdot)$) that outputs robot actions a_t for executing the motions on a robot. Additional details on the approach are in section III.

diverse tasks, with a single policy that can be appropriately goal-conditioned and doesn't require any deployment-time training.

We consider semantic masks of hands and objects as a structured space for defining the *human-plan*, since it abstracts out task-irrelevant details of the environment. Given an image of a scene and a goal image, we train the prediction model to predict the *human-plan* as plausible future hand and object masks. We train this model across clips in diverse passive videos on the web and show that it generalizes to new scenes in our real-robot experiments. In order to transform the predictions to a physical embodiment's *robot-actions* , we train a translation module on a small amount of paired data ($\sim$600 trajectories). We abbreviate our framework as `HOPMan` (**H**and **O**bject **P**lan for robotic **M**anipulation).

Through experiments on a set of 100 tasks, involving 16 skills and 40 objects, we show `HOPMan` can help distill information about manipulation from passive human videos on the web to physical scenes in a robot's workspace, as evaluated through generalization across five different axes. In summary, we make the following contributions:

- Present an approach for learning goal-conditioned prediction of hand-object interaction plans using everyday interaction videos.
- Develop a framework that casts robot manipulation as translation of (predicted) hand-object plans, thus allowing the use of easily available human videos for learning diverse manipulation.
- Demonstrate the overall framework across 100 manipulation tasks involving 40 objects with 16 skills, while evaluating generalization in a structured manner for table-top manipulation and in-the-wild manipulation in unseen scenes.

II. RELATED WORKS

Understanding human interactions from videos. Several recent approaches in computer vision have focused on understanding hand-object interactions in diverse everyday settings [11, 12, 13, 14, 15, 16, 17]. Specifically, prior work has investigated learning hand pose estimation [18, 19, 20, 21, 22, 23, 24, 25, 26, 27], object pose estimation [28,

29, 30, 31, 32], interaction hotspot prediction [5, 6, 33], prediction of plausible hand grasps [34, 35], and activity understanding [36, 37]. Our human-interaction-plan prediction module is inspired by these developments, where we focus on learning motions of hands and objects from passive human videos that are directly relevant for manipulation, and abstract out task-irrelevant visual details through semantic masks.

Learning Visual Representations for Manipulation. A growing body of recent works learn mappings from visual observations to robot actions for performing tasks [38, 39, 40]. One common way of using data beyond robot interactions for efficient learning is to pre-train the visual representations which serve as backbones for the policy models [2, 3, 4, 41, 42] with passive human videos [14, 43] and image data [44]. However, these methods still crucially rely on a lot of in-domain robot data or deployment-time training, and are restricted to learning task-specific policies.

Learning Affordances. Towards learning structure more directly related to manipulation, some works try to predict visual affordances in the form of where to interact in an image, and local information of how to interact [6, 8, 33, 34]. While these could serve as good initializations for a robotic policy, they are not sufficient on their own for accomplishing tasks, and so are typically used in conjunction with online learning, requiring several hours of deployment-time training and robot data [7, 8]. Our work differs from this in terms of predicting an approximate motion of how a human hand and the object is likely to move for the entire trajectory (not just at/near contacts unlike affordances) and is *zero-shot* in terms of not requiring any deployment-time training.

Manipulation without deployment-time training. With a goal similar to ours of using human videos to learn models that can be directly deployed, some approaches leverage curated data of human videos [9, 10] for learning task-specific policies (instead of a single model across generic tasks). Others that train a single policy across tasks require large in-domain perfectly aligned human-robot data [45, 46, 47] and are not capable of leveraging passive web data. Perhaps most closely aligned with ours, Bharadhwaj *et al.* [48] learn (human) action trajectories from passive web videos and

leverage a heuristic to convert these to robot trajectories. However, their actions are restricted to simple coarse open-loop motions that do not involve grasping, and hence can't exhibit diverse skills. Compared to these, our framework utilizes *diverse large-scale* passive human video data on the web, combined with a *small amount* of in-domain robot data, with a single model capable of tackling different manipulation tasks zero-shot.

III. HAND-OBJECT PLAN FOR ROBOTIC MANIPULATION

We aim to develop a robot manipulation system that can accomplish diverse skills zero-shot with a plethora of different unseen objects in the real world. Our key insight is to leverage a factorized policy model (see Fig. 2) that consists of two stages: a) a goal-conditioned human plan prediction model that predicts future masks for plausible hand and object motions, and b) a translation model that learns to transform the corresponding predicted plans into actions that can be executed with a robot for real-world manipulation. We show how we can train the human-plan prediction model on diverse passive human videos from existing large scale datasets, and use it for predicting plausible plans in a robot's environment. In contrast, the translation model can be trained with a small amount of paired human-robot data. This factorization allows us to generalize to scenarios that are unseen in the robot data, because the human-interaction-plan model with its diverse training generalizes well, and the translation model is tasked with a simpler job of converting these plans to the robot's embodiment.

A. The Human-Plan Prediction Model

Instead of predicting the future in the image space, we focus on predicting only the motion of the human hand and the object being interacted with, in terms of respective semantic masks. We enable this prediction through a diffusion model trained on diverse human videos on the web. For each video in the training data, we extract hand-object masks for each frame. Let $\mathbf{M_{1:K}}$ denote the respective mask frames from time steps 1 to K. For simplicity we consider each mask frame to be an image, where all the hand pixels are green, all the object pixels are red, and the rest of the pixels are black. Let $\mathbf{X_0}$ denote the first frame (RGB) of the video, $\mathbf{X_g}$ denote the last frame (RGB) of the video, which will act as a goal frame, and $\mathcal{V}(\mathbf{X_0}, \mathbf{X_g})$ denote the prediction model. In the forward diffusion process, all the mask frames $\mathbf{M_{1:K}}$
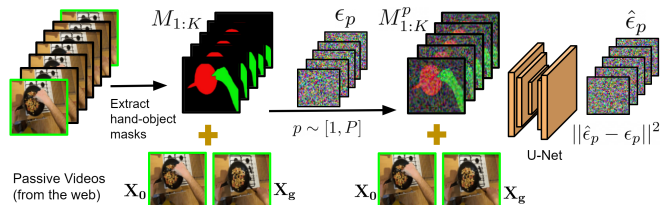

Fig. 3. Detailed illustration of a training pass through the future prediction model. This is a diffusion model, with a U-net that predicts per-frame noise at each step p of the diffusion process. Additional details on the model and training are in Section III-A.

are corrupted by incrementally adding noise, and converging

to a unit Gaussian distribution $N(\mathbf{0}, \mathbf{I})$. New samples can be generated by reversing the forward diffusion process, by going from Gaussian noise back to the space of mask frames. To solve the reverse diffusion process, we need to train a noise predictor $\epsilon_\theta(\cdot|t)$ which is a time-conditioned U-net [49, 50] trained to predict the noise at each step of the diffusion process. The input to the network at step t of the diffusion process is a channel-wise concatenation of the conditioning frames and noisy mask frames $[\mathbf{X_0}, \mathbf{X_g}, \mathbf{M_{1:K}^t}]$, the output is the predicted noise of same dimensionality as the input. Fig. 3 illustrates this visually, and equation 1 shows the training objective $\mathcal{L}(\theta)$.

$$\mathbb{E}_{t, [\mathbf{X_0}, \mathbf{X_g}, \mathbf{M_{1:K}}] \sim p_{\text{train}}, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}$$
$$\left[||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t} \mathbf{M_{1:K}} + \sqrt{1 - \overline{\alpha}_t} \epsilon | \mathbf{X_0}, \mathbf{X_g}, t)||^2 \right]$$

Here $\overline{\alpha}_t$ is a hyper-parameter that depends on the noise schedule of the diffusion process. During inference, given $\mathbf{X_0}, \mathbf{X_g}$ we obtain $M_{1:K} = \mathcal{V}(\mathbf{X_0}, \mathbf{X_g})$ through reverse diffusion.
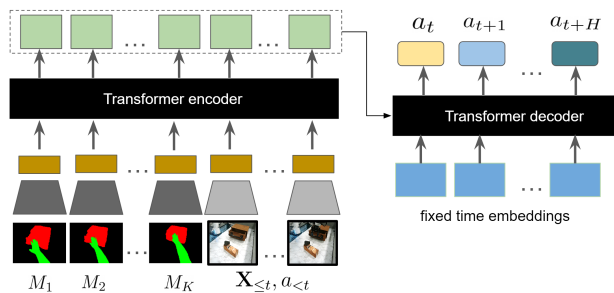
B. The Robot-Action Translation Model

Fig. 4. Architecture of the translation model that transforms predicted future hand-object masks to a robot trajectory, described in section III-B

We use the human-plan predictor discussed in Section III-A to hallucinate plausible future hand and object masks for inter-action in a robot's physical scene. However, this human-plan doesn't directly inform what actions the robot should execute to be able to perform the desired interaction. To enable robot manipulation in the context of the predicted plans, we learn a translation model. The translation model is a transformer that is conditioned on the outputs of the future prediction model $M_{1:K}$ and for each observation $\mathbf{X}_t$, and predicts actions for H steps in the future. The model behaves as a closed-loop policy $\pi(\mathbf{X}_{\leq t}, \mathbf{X}_g, a_{<t}, M_{1:K})$ that is queried at each time-step t during deployment. Predicting multiple time-steps H in the future and averaging actions during deployment, helps in executing smooth robot motions, with less compounding errors [51]. We describe the architecture of the translation model in Fig. 4 and additional details in Appendix C.

For training the translation model, we need some paired human-robot data, where we have pairs of trajectories that involve a robot manipulating an object, and a human manip-ulating a similar object. To obtain such paired trajectories, we develop two approaches:

Collecting paired demonstrations: A human operator tele-operates a robot in scene, and after reset, or in a parallel identical setup, a human manipulates a similar object with
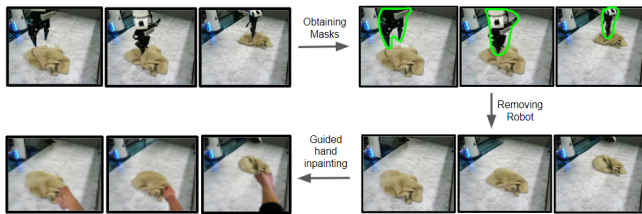
Fig. 5. Illustration of the different steps in generating hallucinated human hand trajectories from robot trajectories. This is an alternate data source for the translation model in addition to collecting paired human-robot data.
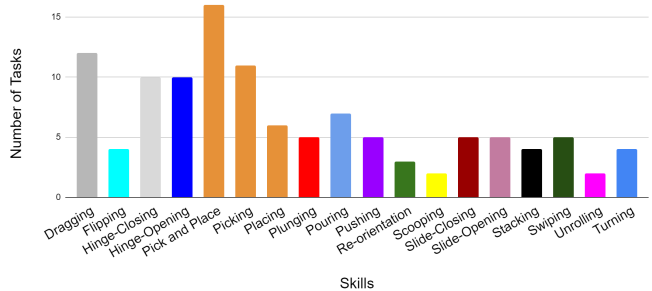

Fig. 6. Distribution of skills across tasks in our experiments. The diversity of skills is more representative of real-world distributions, compared to pushing/pick and place that is predominant in robot learning papers.

an approximately similar motion as the robot arm. Collecting this paired data is not very expensive, and we spent around 3 days to collect 600 trajectories.

Hallucinating paired data: To augment the paired demonstrations, we also propose to leverage (more easily collectable) robot-only data. To obtain hallucinated pairs, we can convert videos of a robot trajectory into a videos of a human trajectory through recent advances in hand in-painting techniques [49, 52]. Specifically, we obtain robot masks per frame through simulation, and perform inpainting to remove the robot from the scene. We then perform guided in-painting of a plausible human hand [52] around the location of the robot end-effector in the scene. Fig. 5 visually illustrates this process of hallucinated data generation. In the experiments, we show how hallucinated paired data generated through this approach can be used to boost the performance of the translation model. Additional details on the hallucinated data generation are in the Appendix D.

IV. EXPERIMENTS

Through experiments with diverse real-world objects in unseen scenarios, we demonstrate generalization of our framework for several robot manipulation tasks. Videos are in the website https://homangab.github.io/hopman/

A. Experiment Settings

We consider two different types of manipulation settings for experiments - table top scenarios with a fixed robot and camera, and in-the-wild manipulation with the same robot and camera on a mobile base.

Table-Top Manipulation. We consider several everyday objects with different plausible manipulations for our experiments. We demonstrate results on a total of 16 skills: pouring, plunging, pushing, picking/placing, slide-opening, slide-closing, hinge-opening, hinge-closing, swiping, dragging, flipping, scooping, in-place re-orientation, unrolling, and stacking, and 40 object types, with 2-3 instantiations per object type, comprising around 100 tasks. Detailed list of objects and tasks are in the Appendix section B.

In-the-Wild Manipulation. We drag a Franka Panda arm on a mobile base across natural kitchen and office scenes. The camera is also attached to the base, and moves along with it. For these experiments we fine-tune the translation model used for the table-top experiments, on ∼200 additional paired trajectories collected with the mobile robot. For evaluation, we consider the same generalization levels described above. This setting is much more challenging because in addition

to object and skill variations, we also have scene variations, including completely new scenes never seen in the paired data. Details of variations are in the supplementary website.

B. Training data

The training data for our framework consists of a large set of passive web videos, a small amount of paired human-robot in-domain data, and some unpaired robot-only data.

Passive Human Data. For the future prediction model, we use existing passive human videos [11, 14, 53] and obtain ground-truth semantic masks for the right hand and the object being interacted with the right hand in each frame [53, 54]. We sample short video clips, each lasting a few seconds and do not curate the videos in any way with tasks or language labels.

Paired Data. For the translation model, we use a small amount of paired collected by us (∼400 trajectories in-lab and ∼200 trajectories in-the-wild) and a larger robot-only data (∼1000 trajectories) combined with hallucinated hand masks through the approach described in section III-B. All the robot data are collected through an adaptation of the tele-operation stack proposed in [55].

C. Defining Tasks and Evaluating Generalization

Prior works in robot learning adopt widely different and oftentimes inconsistent definitions of generalization criteria. Some prior works [1, 9, 56, 57] consider seen vs. unseen objects, where the unseen objects often involve different instantiations of the seen objects, with shape, color, and texture variations, with skills (e.g. pushing, picking etc.) that are always seen in the training data. Others [58, 59] only consider generalization in terms of position and configuration variations of seen objects. In light of this, in this paper, we develop a structured criteria for evaluating generalization in terms of object categories, object instantiations, object configurations, and skills. We adopt the following definitions

- **Task definition:** Each task is a tuple consisting of (object category, object instance, skill). Here, *object category* denotes the type of the object e.g. 'drawer', 'mug', 'toaster' etc. While, *object instance* defines a particular object within a category, with a specific instantiation of color, shape, size, and texture. Finally, *skill* defines the particular behavior e.g. 'open', 'flip', 'push' etc. that can be done with an object.
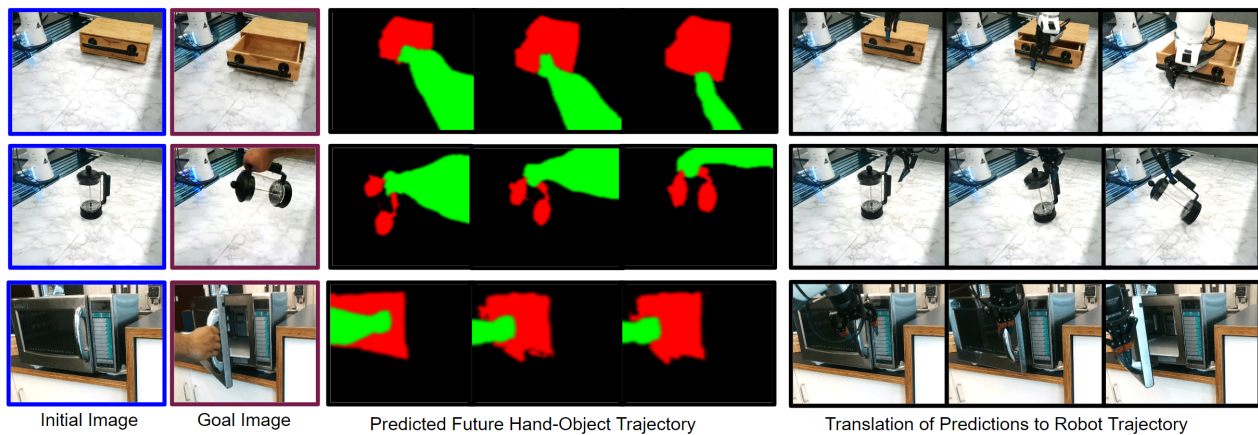
Fig. 7. **Qualitative results for the entire framework.** We show qualitative results for the predicted hand-object trajectory given an initial image of a scene and a goal image, followed by translation of the predictions to a robot trajectory for execution in the real world.
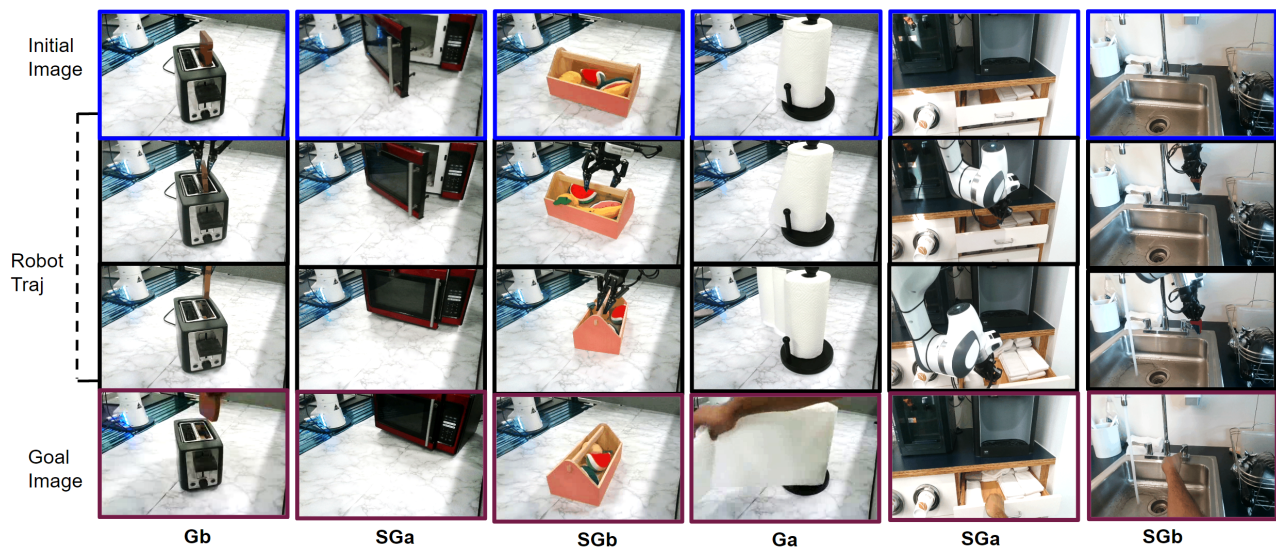

Fig. 8. **Examples of robot evaluations.** We show qualitative results for robot evaluations, with an intermediate image and the image corresponding to the final state reached by the robot, for a given initial scene and a goal image. Subscripts show the type of generalization for each evaluation, as described in sec IV-C. More robot videos of evaluations are in the linked website.

- **Mild generalization (MG):** This involves generalizing among unseen configurations (i.e. position and orientation variations) for seen object instance and seen skills, along with mild variations in the scene like lighting changes.

- **Standard generalization (G):** We have the following types of generalization in this category
  - **instance generalization (Ga):** In addition to variations in MG, in Ga we evaluate unseen object instance for seen skills. For example, only a red mug is seen with the push skill in training, and we generalize to pushing motions for green, and purple mugs of different shapes and textures.
  - **unseen combinations (Gb):** This includes scenarios with unseen (object category, skill) pairs but each seen independently in training. So atleast one instance of an object category is seen during training, and the skillis also seen during training but not in relation to this object. For example, 'open' is seen, and 'close door' is seen but 'open door' is not seen in training.

- **Strong Generalization (SG):** We categorize the following types of generalization that involve either a completely unseen object category or an unseen skill into this category. These are very challenging tests of generalization.
  - **object category completely unseen (SGa):** This includes scenarios where a particular object category e.g. microwave is never seen in training
  - **skill completely unseen (SGb):** This includes scenarios where a particular skill e.g. re-orientation is never seen in any context during training.

Note that our formalization of generalization is centered around objects being interacted with and the skills that are possible for interaction, and we do not consider scene variations of the background in the definitions, unlike some prior work [1, 57, 60, 61, 62]. However, for experiments, we consider diverse scenes, both for table-top manipulation and manipulation of objects in-the-wild in unseen kitchens and offices.
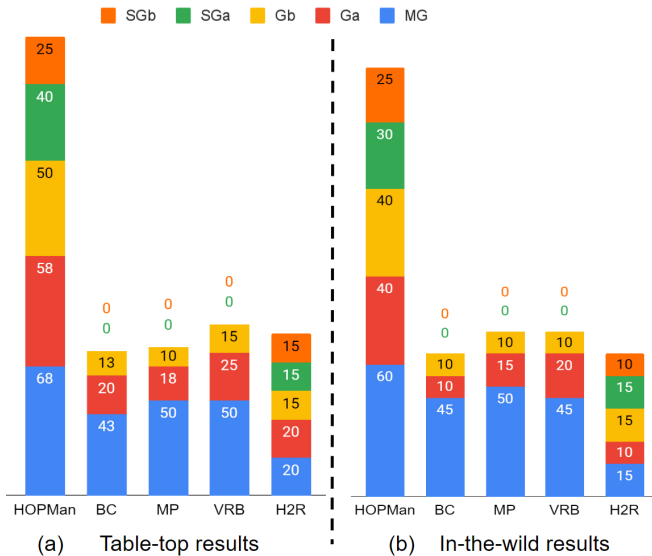
Fig. 9. **Summary of results.** The numbers represent success rates for goal-conditioned evaluations, in terms of % of trials that correspond to manipulating objects in the scene to bring them to the desired goal configurations. We perform evaluations separately for the table-top manipulation and in-the-wild manipulation experiments.
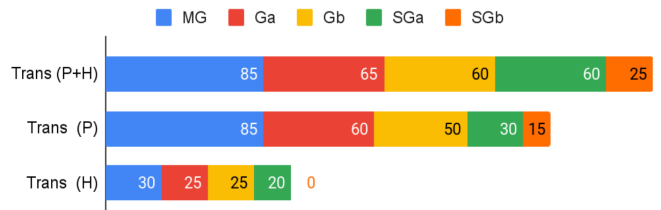

Fig. 10. **Translation model ablations.** Ablation results for the translation model alone with specified masked hand-object trajectories instead of future predictions. Here, P denotes paired data, and H denotes hallucinated data, described in section III-B. and the numbers represent success rates.

D. Baselines and Ablations:

We consider a goal-conditioned behavior cloning baseline (BC) trained on all the robot data (∼1600 trajectories). The architecture of the policy is a transformer similar to our translation model without the conditioning on human-interaction-plans. The next baseline (MP) uses paired human-robot data, and is an adaptation of [45]. We compare with VRB [8] by using the affordance model from the paper to do affordance conditioned imitation learning. We also consider a baseline that is trained entirely with passive human videos, for coarse manipulation (H2R) [48]. In addition to these, we consider variations of our translation model trained on only in-lab paired human-robot data (∼400 trajectories), only hallucinated data (∼1000 trajectories), and combined paired and hallucinated data (∼1400 trajectories).

E. Evaluating Goal-conditioned Manipulation

In this section, we evaluate `HOPMan` for robot manipulation. Given an image of a scene in the robot workspace and a goal image, we use the human-interaction-plan predictor to output a sequence of plausible hand-object masks, which are input to the translation model that performs closed-loop control for executing a sequence of actions on the robot. We evaluate across diverse unseen objects exhibiting several plausible skills, and unseen scenes in-the-wild, and tabulate success rates by aggregating over objects for each skill. We define success in terms of whether the object is brought to the desired configuration in the goal image.

Fig. 7 shows qualitative results for `HOPMan` where we see that the generated human-interaction-plans are plausible and correspond to manipulating the object to obtain the specified goal configuration. In Fig. 8 we show more robot evaluations in terms of an intermediate frame in the trajectory and the final frame reached at the end of robot evaluation, for different initial and goal images.

In Fig. 9 we summarize quantitative evaluations across the different generalization axes. For standard generalization `G` and strong generalization `SG`, we see that `HOPMan` achieves significantly high success rate. This demonstrates the effectiveness of learning plausible manipulation trajectories of hands and objects from internet videos combined with small paired data, for generalization to diverse settings, in comparison to relying on only in-domain data (BC, MP baselines), on predicting visual affordances combined with robot data (VRB) or on only passive data (H2R).

F. Ablations of the Translation Model

In this section, we evaluate the translation model in isolation independent from the prediction model. Specifically we evaluate how good is the translation model in translating the motion of a ground-truth hand-object trajectory into robot trajectories. Here, we introduce different objects in the scene and manually execute a motion with a human hand to reach the goal, and then pass the video through the hand-object segmentation model. We ablate over three variations of the the translation model, trained with paired data and hallucinated data, trained with only paired data, and trained with only hallucinated data, in table-top settings. From Fig. 10, we observe that training the model with combined paired and hallucinated data (P+H) leads to better performance than training with just paired data (P) indicating that the translation model is able to effectively utilize imperfect hallucinated trajectories for improving generalization.

V. DISCUSSION AND LIMITATIONS

In this work, we developed a framework for learning generalizable robot manipulation by combining internet-scale human videos of everyday interactions with limited in-domain robot demonstrations. Leveraging these, our framework can accomplish diverse tasks by predicting plausible hand-object plans and translating these to the robot's embodiment. Broadly, our work is indicative of how rich out-of-domain datasets like human videos can alleviate the data paucity that greatly bottlenecks robot learning by helping learn hand-object interaction plans, and enable wide generalization of manipulation skills to unseen scenarios. While our framework does allow strong generalization to unseen tasks, these are still limited in their complexity and it would be an interesting future direction to extend our approach for tacking long-horizon tasks that requiring composing multiple skills. Moreover, our framework may struggle with dexterous manipulation tasks as recovering precise hand and finger articulations from web videos remains a challenge in computer vision.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[2] 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.

[3] A. Majumdar, K. Yadav, S. Arnaud, Y. J. Ma, C. Chen, S. Silwal, A. Jain, V.-P. Berges, P. Abbeel, J. Malik, *et al.*, "Where are we in the search for an artificial visual cortex for embodied intelligence?" *arXiv preprint arXiv:2303.18240*, 2023.

[4] 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.

[5] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8688–8697.

[6] M. Goyal, S. Modi, R. Goyal, and S. Gupta, "Human hands as probes for interactive object understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3293–3303.

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

[8] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak, "Affordances from human videos as a versatile representation for robotics," *arXiv preprint arXiv:2304.08488*, 2023.

[9] 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," *arXiv preprint arXiv:2108.05877*, 2021.

[10] K. Shaw, S. Bahl, and D. Pathak, "Videodex: Learning dexterity from internet videos," in *6th Annual Conference on Robot Learning*.

[11] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, *et al.*, "The" something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5842–5850.

[12] P. Das, C. Xu, R. F. Doell, and J. J. Corso, "A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 2634–2641.

[13] F. De la Torre, J. Hodgins, A. Bargteil, X. Martin, J. Macey, A. Collado, and P. Beltran, "Guide to the carnegie mellon university multimodal activity (cmu-mmac) database," 2009.

[14] 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*, 2022, pp. 18 995–19 012.

[15] D. Shan, J. Geng, M. Shu, and D. Fouhey, "Understanding human hands in contact at internet scale," in *CVPR*, 2020.

[16] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.

[17] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 619–635.

[18] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *CVPR*, 2017.

[19] U. Iqbal, P. Molchanov, T. Breuel Juergen Gall, and J. Kautz, "Hand pose estimation via latent 2.5 d heatmap regression," in *ECCV*, 2018.

[20] A. Spurr, J. Song, S. Park, and O. Hilliges, "Cross-modal deep variational hand pose estimation," in *CVPR*, 2018.

[21] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. Yuan, "3d hand shape and pose estimation from a single rgb image," in *CVPR*, 2019.

[22] S. Baek, K. I. Kim, and T.-K. Kim, "Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering," in *CVPR*, 2019.

[23] A. Boukhayma, R. d. Bem, and P. H. Torr, "3d hand shape and pose from images in the wild," in *CVPR*, 2019.

[24] 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 *CVPR*, 2019.

[25] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou, "Weakly-supervised mesh-convolutional hand reconstruction in the wild," in *CVPR*, 2020.

[26] S. Liu, H. Jiang, J. Xu, S. Liu, and X. Wang, "Semi-supervised 3d hand-object poses estimation with interactions in time," in *CVPR*, 2021.

[27] Y. Rong, T. Shiratori, and H. Joo, "Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration," *arXiv preprint arXiv:2008.08324*, 2020.

[28] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," *ICCV*, 2017.

[29] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," *ICCV*, 2017.

[30] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv*, 2018.

[31] Y. Hu, J. Hugonot, P. Fua, and M. Salzmann, "Segmentation-driven 6d object pose estimation," *CVPR*, 2019.

[32] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," *CVPR*, 2020.

[33] S. Liu, S. Tripathi, S. Majumdar, and X. Wang, "Joint hand motion and interaction hotspots prediction from egocentric videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3282–3292.

[34] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6813–6823.

[35] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," *arXiv*, 2019.

[36] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2015, pp. 961–970.

[37] S. Tan, T. Nagarajan, and K. Grauman, "Egodistill: Egocentric head motion distillation for efficient video understanding," *arXiv preprint arXiv:2301.02217*, 2023.

[38] C. Finn, T. Yu, T. Zhang, P. Abbeel, and S. Levine, "One-shot visual imitation learning via meta-learning," in *Conference on robot learning*. PMLR, 2017, pp. 357–368.

[39] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," in *Conference on Robot Learning (CoRL)*, 2020.

[40] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, *et al.*, "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Conference on Robot Learning*. PMLR, 2018, pp. 879–893.

[41] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta, "The unsurprising effectiveness of pre-trained vision models for control," *arXiv preprint arXiv:2203.03580*, 2022.

[42] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv preprint arXiv:2203.06173*, 2022.

[43] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[44] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[45] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar, "Mimicplay: Long-horizon imitation learning by watching human play," *arXiv preprint arXiv:2302.12422*, 2023.

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

[47] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg, "Learning by watching: Physical imitation of manipulation skills from human videos," *arXiv*, 2021.

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

[49] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition, 2022, pp. 10 684–10 695.

[50] V. Voleti, A. Jolicoeur-Martineau, and C. Pal, "Masked conditional video diffusion for prediction, generation, and interpolation," *arXiv preprint arXiv:2205.09853*, 2022.

[51] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[52] Y. Ye, X. Li, A. Gupta, S. De Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu, "Affordance diffusion: Synthesizing hand-object interactions," in *CVPR*, 2023, pp. 22 479–22 489.

[53] A. Darkhalil, D. Shan, B. Zhu, J. Ma, A. Kar, R. Higgins, S. Fidler, D. Fouhey, and D. Damen, "Epic-kitchens visor benchmark: Video segmentations and object relations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 745–13 758, 2022.

[54] L. Zhang, S. Zhou, S. Stent, and J. Shi, "Fine-grained egocentric hand-object segmentation: Dataset, model, and applications," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX*. Springer, 2022, pp. 127–145.

[55] V. Kumar and E. Todorov, "Mujoco haptix: A virtual reality system for hand manipulation," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 657–663.

[56] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, *et al.*, "Open-world object manipulation using pre-trained vision-language models," *arXiv preprint arXiv:2303.00905*, 2023.

[57] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran, and V. Kumar, "Cacti: A framework for scalable multi-task multi-scene visual imitation learning," *arXiv preprint arXiv:2212.05711*, 2022.

[58] S. Haldar, V. Mathur, D. Yarats, and L. Pinto, "Watch and match: Super-charging imitation with regularized optimal transport," in *Conference on Robot Learning*. PMLR, 2023, pp. 32–43.

[59] Z. J. Cui, Y. Wang, N. Muhammad, L. Pinto, *et al.*, "From play to policy: Conditional behavior generation from uncurated robot data," *arXiv preprint arXiv:2210.10047*, 2022.

[60] Z. Chen, S. Kiami, A. Gupta, and V. Kumar, "Genaug: Retargeting behaviors to unseen situations via generative augmentation," *arXiv preprint arXiv:2302.06671*, 2023.

[61] T. Yu, T. Xiao, A. Stone, J. Tompson, A. Brohan, S. Wang, J. Singh, C. Tan, J. Peralta, B. Ichter, *et al.*, "Scaling robot learning with semantically imagined experience," *arXiv preprint arXiv:2302.11550*, 2023.

[62] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," *arXiv preprint arXiv:2309.01918*, 2023.

[63] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[64] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IROS*, 2012.

A. Robot Evaluation Videos

Robot videos are in the website https://homangab.github.io/hopman/

B. List of Tasks

TASK		
Object Category	Num of Instances	Skill
Toaster	2	Plunging
Toaster,Toast	4	Picking
Drawer	3	Slide-Opening
Drawer	3	Slide-Closing
Toaster Oven	3	Hinge-Opening
Toaster Oven	3	Hinge-Closing
Towel	3	Swiping
Bowl	3	Pushing
French Press	2	Pouring
Bagel	3	Flipping
Tool Container	2	Pick and Place
Paper Towel	2	Unrolling
Tissue Box	2	Picking
Tea Bags	2	Picking
Spice Container	2	Pick and Place
Tea Cup	3	Pick and Place
Mug	2	Pick and Place
Ketchup Bottle	1	Pick and Place
Tea Cup	3	Dragging
Mug	2	Dragging
French Press	2	Pushing
Tool Container	2	Dragging
French Press	3	Plunging
Toaster,Toast	4	Placing
Ketchup Bottle, Wooden Board	2	Pick and Place
Spatula, Spatula Holder	2	Picking
Spatula, Spatula Holder	2	Placing
Tea Cup	3	Pouring
Glass	2	Pouring
Watermelon	2	Pick and Place
Banana	2	Pick and Place
Strainer	1	Dragging
Microwave	1	Hinge-Opening
Microwave	1	Hinge-Closing
Spatula Holder	2	Dragging
Spatula Holder	2	Pick and Place
Bun	1	Flipping
Watermelon, Wooden Board	2	Pick and Place
Box of wipes	1	Picking
Box	2	Dragging
Basket	1	Dragging
Door (vertical hinge)	1	Hinge-Opening
Door (vertical hinge)	1	Hinge-Closing
Tool Container	2	Re-orientation
Cereal, Bowl, Spoon	2	Scooping
Bowls	2	Stacking
Boxes	2	Stacking
TOTAL TASKS =	100	

Fig. 11. Summary of the different tasks for table-top manipulation experiments in terms of object types, number of instantiations per object type (variations in shape, size, color ,texture) and verbs denoting the type of possible skill with each object type

C. Additional details on the models

1) Human-Plan Prediction model:: Instead of predicting the future in the image space, we focus on predicting only the motion of the human hand and the object being interacted with, in terms of respective semantic masks. We enable this prediction through a diffusion model trained on diverse human videos on the web. For each video $\mathcal{V}$ in the training data, we extract hand-object masks for each frame . Let $M_{1:K}$ denote the respective mask frames from time steps 1 to K. We set the value of $K = 7$ for our experiments, which amounts to choosing 7 uniformly space frames in a 2 second window of a video clip. For simplicity we consider each mask frame to be an image, where all the hand pixels are green, all the object pixels

are red, and the rest of the pixels are black. the Let X_0 denote the first frame (RGB) of the video, and X_g denote the last frame (RGB) of the video, which will act as an optional goal frame. The diffusion model operates at a resolution of 64x64 for the predicted masked frames.

In the forward diffusion process, all the mask frames $M_{1:K}$ are corrupted by incrementally adding noise, and converging to a unit Gaussian distribution $N(\mathbf{0}, \mathbf{I})$. New samples can be generated by reversing the forward diffusion process, by going from Gaussian noise back to the space of mask frames. To solve the reverse diffusion process, we need to train a noise predictor $\epsilon_\theta(\cdot|t)$ which is a time-conditioned U-net trained to predict the noise at each step of the diffusion process. The input to the network at step t of the diffusion process is a channel-wise concatenation of the conditioning frames and noisy mask frames $[\mathbf{X_0}, \mathbf{X_g}, \mathbf{M_{1:K}^t}]$, and the output is the predicted noise of same dimensionality as the input. The training objective is as follows:

$$\mathbb{E}_{t,[\mathbf{X_0},\mathbf{X_g},\mathbf{M_{1:K}}]\sim p_{\text{train}},\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})} \left[||\epsilon - \epsilon_\theta(\sqrt{\overline{\alpha}_t}\mathbf{M_{1:K}} + \sqrt{1-\overline{\alpha}_t}\epsilon|\mathbf{X_0}, \mathbf{X_g}, t)||^2 \right]$$

Here α is a constant hyper-parameter that depends on the noise schedule of the diffusion process. The architecture of the U-Net for the Diffusion model is based on prior works [50, 63], and it uses a combination of 2D convolutions, multi-head self-attention layers, and adaptive group-norm. The noise levels ($p \in [0, 1]$) use positional encodings that are adapted to the correct dimensionality for each residual block through fully connected layers. The individual residual blocks in the U-Net consist of GroupNorm, conv layers, fully connected layers, and dropout, and follow the architecture in [50].

For training the prediction model we obtain 2 second video clips from EpicKitchens [16] and Ego4D [14]. To obtain ground-truth hand-object masks, we use Visor annotations [53] for EpicKitchens and an off-the-shelf predictor [54] for obtaining the masks from Ego4D videos. In total, we curate around 150,000 video clips for training. The prediction model takes about 70 hours to train for 250,000 iterations on 8 2080Ti GPUs with a batch size of 64, and learning rate 1e-5.

2) The Translation model: The translation model is a transformer that is conditioned on the outputs of the future prediction model $M_{1:K}$ and for each observation O_t, predicts actions for H steps in the future. The model behaves as a closed-loop policy that is queried at each time-step t during deployment. The horizon lengths for each trajectory is 40, and we predict for $H = 10$ horizon at each time-step. The observations are of resolution 224x224, and we process them with ResNet18 backbones to obtain features. We upsample the predicted masks from 64x64 to to 224x224 dimension images and process them also with ResNet18 CNNs. At each time-step we feed in a history of 3 steps, i.e. the past two observations and actions, and the current observation. The actions are of dimension 8 (7 for joint positions, and the 8th dimension for end-effector open/close). We directly predict target joint positions instead of delta positions, as shown to be helpful by recent work [51]. The transformer encoder has 4 self-attention blocks, and the decoder has 7 cross-attention blocks, and the hidden dimensions are of size 512. We use a learning rate of 1e-5, batch size of 32, and dropout 0.1.

D. Hallucinated Data Generation Details

To augment the paired demonstrations, we develop an approach of using (more easily collectable) robot-only data. Given a robot trajectory video, we want to obtain a corresponding video where the robot in the scene is replaced by a human hand. To obtain such hallucinated pairs, we convert videos of robot trajectories into videos of human trajectories through recent advances in hand in-painting techniques [49, 52]. Given a robot trajectory, we first obtain robot masks per frame by bringing the robot to the specific joint position per-timestep in MuJoCo simulation [64]. Based on the per-frame robot masks, we inpaint each image in the trajectory to remove the robot from the scene, using an off-the-shelf inpainting model [49]. After we have removed the robot from each frame of the trajectory, we want to place a human hand where the robot end-effector used to be in each image. To do this, we perform guided in-painting of a plausible human hand around the location of the robot end-effector in the scene, using the approach in [52]. So finally, a video of a robot trajectory is thus converted to a video of an approximate human trajectory, such that the robot arm is replaced with a human hand at approximately the location where the end-effector used to be.

E. Baselines and Ablations

We consider a goal-conditioned behavior cloning baseline that is not conditioned on the predicted masks, and is directly trained on all the robot data collected in-lab ($\sim$1400 trajectories). For the in-the-wild experiments, we additionally fine-tune the model with the 200 paired trajectories collected for these experiments. The architecture of the policy is a transformer similar to our translation model without the conditioning on hand-object masks, and keeping everything else the same.

We consider another baseline that uses paired in-lab human-robot data, to be an adaptation of MimicPlay [45] . We train the latent planner model of MimicPlay (MP) with the human-data in the paired data of 400 trajectories we have collected for the experiments. For the in-the-wild experiments, we additionally fine-tune the model with the 200 paired trajectories collected for these experiments. Note that in the original paper [45], there are a limited number of tasks (14) and human hand data is collected for 10 minutes per scene. In comparison, our paired data of 400 trajectories is much smaller and encompass around 40 tasks, since we focus mostly on learning from out-of-domain passive human videos from the web. We

cannot use this large passive data for MimicPlay baseline as their framework relies on having the human videos in the exact same setup as the robot teleop data.

We compare with two baselines that use passive human videos in different ways. The first comparison is with VRB [8] by using the affordance model from the paper to do affordance conditioned imitation learning. The second comparison is a baseline that is trained entirely with passive human videos, for coarse manipulation (H2R) [48].

In addition to these, for the table-top experiments we consider variations of our translation model trained on only paired human-robot data (∼400 trajectories), only hallucinated data (∼1000 trajectories), and combined paired and hallucinated data (∼1400 trajectories). These ablations are on the same translation model architecture, and use manually specified hand trajectories transformed to hand-object masks through [54]. We manually provide masks instead of the predictions from the human plan prediction model, in order to evaluate the translation model in isolation independent from the prediction model.

F. Table-Top Robot Experiment Setup Details

For the robot experiments, we use several everyday objects like doors, microwaves, bowls, spatulas, boxes, french presses etc. (Fig. 11 has the overall list of objects), a fixed Intel Realsense camera in the scene, and a Franka Emika Panda arm operated through joint position control. We do not impose any artificial constraints on the robot's motions beyond what is possible without reaching joint limits. The action space of the translation model is 8 dimensional (7 for joint controls, and the 8th dimension for open/close of the gripper) We attach a Robotiq gripper to the arm with two festo finger grippers (for flexible grasps), so the overall end-effector is a two-finger gripper. As is the convention with image goals in real-robot experiments, we evaluate success by manually inspecting proximity of the final object configuration after robot execution, with that in the corresponding goal image.

G. In-The-Wild Robot Experiment Setup Details

We use the same Franka Emika Panda arm with flexible two finger grippers as the previous table-top experiments. The only difference is that the robot is now mounted on a mobile base with four wheels that can be moved around. The same Intel Realsense camera is mounted next to the robot on the mobile base. We drag the robot across different kitchen and office scenes and perform experiments with the same setup described previously. Importantly, we do not modify the scenes and directly test on existing office and kitchen scenes. Please refer to the evaluation videos on the website for the diversity of manipulation skills and behaviors we are able to demonstrate with our framework. https://homangab.github.io/hopman/