Robotics: Science and Systems 2021
Held Virtually, July 12–16, 2021

Learning Generalizable Robotic Reward Functions from "In-The-Wild" Human Videos

Annie S. Chen, Suraj Nair, Chelsea Finn
Stanford University

***Abstract*—We are motivated by the goal of generalist robots that can complete a wide range of tasks across many environments. Critical to this is the robot's ability to acquire some metric of task success or reward, which is necessary for reinforcement learning, planning, or knowing when to ask for help. For a general-purpose robot operating in the real world, this reward function must also be able to generalize broadly across environments, tasks, and objects, while depending only on on-board sensor observations (e.g. RGB images). While deep learning on large and diverse datasets has shown promise as a path towards such generalization in computer vision and natural language, collecting high quality datasets of robotic interaction at scale remains an open challenge. In contrast, "in-the-wild" videos of humans (e.g. YouTube) contain an extensive collection of people doing interesting tasks across a diverse range of settings. In this work, we propose a simple approach, Domain-agnostic Video Discriminator (DVD), that learns multitask reward functions by training a discriminator to classify whether two videos are performing the same task, and can generalize by virtue of learning from a *small amount of robot data* with a *broad dataset of human videos*. We find that by leveraging diverse human datasets, this reward function (a) can generalize zero shot to unseen environments, (b) generalize zero shot to unseen tasks, and (c) can be combined with visual model predictive control to solve robotic manipulation tasks on a real WidowX200 robot in an unseen environment from a single human demo.**

I. Introduction

Despite recent progress in robotic learning on tasks ranging from grasping [24] to in-hand manipulation [31], the long-standing goal of the "generalist robot" that can complete many tasks across environments and objects has remained out of reach. While there are numerous challenges to overcome in achieving this goal, one critical aspect of learning general purpose robotic policies is the ability to learn *general purpose reward functions*. Such reward functions are necessary for the robot to determine its own proficiency at the specified task from its on-board sensor observations (e.g. RGB camera images). Moreover, unless these reward functions can generalize across varying environments and tasks, an agent cannot hope to use them to learn generalizable multi-task policies.

While prior works in computer vision and NLP [11, 12, 4] have shown notable generalization via large and diverse datasets, translating these successes to robotic learning has remained challenging, partially due to the dearth of broad, high-quality robotic interaction data. Motivated by this, a number of recent works have taken important steps towards the collection of large and diverse datasets of robotic interaction [28, 22, 10, 53] and have shown some promise in enabling generalization [10]. At the same time, collecting such interaction data on real robots at a large scale remains challenging for a number of reasons, such as needing to balance data quality with scalability, and maintaining safety without relying heavily on human supervision and resets. Alternatively, YouTube and similar sources contain enormous amounts of "in-the-wild" visual data of humans interacting in diverse environments. Robots that can learn reward functions from such data have the potential to be able to generalize broadly due to the breadth of experience in this widely available data source.

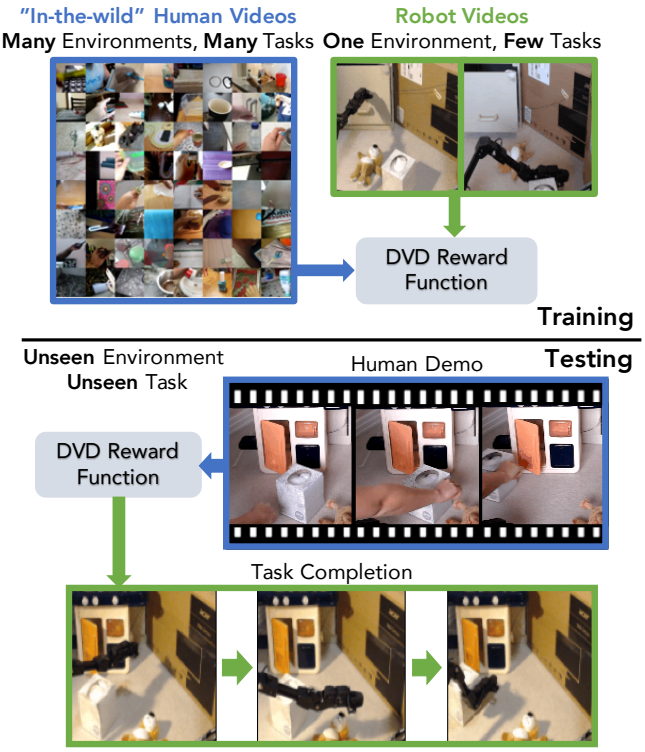

Figure 1: **Reward Learning and Planning from In-The-Wild Human Videos.** During training (**top**), the agent learns a reward function from a small set of robot videos in one environment, and a large set of in-the-wild human videos spanning many tasks and environments. At test time (**bottom**), the learned reward function is conditioned upon a task specification (a human video of the desired task), and produces a reward function which the robot can use to plan actions or learn a policy. By virtue of training on diverse human data, this reward function generalizes to unseen environments and tasks.

Of course, using such "in-the-wild" data of humans for robotic learning comes with a myriad of challenges. First, such data often will have tremendous domain shift from the robot's observation space, in both the morphology of the agent and the

visual appearance of the scene (e.g. see Figure 1). Furthermore, the human's action space in these "in-the-wild" videos is often quite different from the robot's action space, and as a result there may not always be a clear mapping between human and robot behavior. Lastly, in practice these videos will often be low quality, noisy, and may have an extremely diverse set of viewpoints or backgrounds. Critically however, this data is *plentiful* and already exists, and is easily accessible through websites like YouTube or in pre-collected academic datasets like the Something-Something data set [21], allowing them to be incorporated into the robot learning process with little additional supervision cost or collection overhead.

Given the above challenges, how might one actually learn reward functions from these videos? The key idea behind our approach is to train a classifier to predict whether two videos are completing the same task or not. By leveraging the activity labels that come with many human video datasets, along with a modest amount of robot demos, this model can capture the functional similarity between videos from drastically different visual domains. This approach, which we call a Domain-agnostic Video Discriminator (DVD), is simple and therefore can be readily scaled to large and diverse datasets, including heterogeneous datasets with both people and robots and without any dependence on a close one-to-one mapping between the robot and human data. Once trained, DVD conditions on a human video as a demonstration, and the robot's behavior as the other video, and outputs a score which is an effective measure of task success or reward.

The core contribution of this work is a simple technique for learning multi-task reward functions from a mix of robot and in-the-wild human videos, which measures the functional similarity between the robot's behavior and that of a human demonstrator. We find that this method is able to handle the diversity of human videos found in the Something-Something-V2 [21] dataset, and can be used in conjunction with visual model predictive control (VMPC) to solve tasks. Most notably, we find that by training on diverse human videos (even from unrelated tasks), our learned reward function is able to more effectively generalize to unseen environments and unseen tasks than when only using robot data, yielding a 15-20% absolute improvement in downstream task success. Lastly, we evaluate our method on a real WidowX200 robot, and find that it enables generalization to an unseen task in an unseen environment given only a single human demonstration video.

II. Related Work

A. Reward Learning

The problem of learning reward functions from demonstrations of tasks, also known as inverse reinforcement learning or inverse optimal control [1], has a rich literature of prior work [35, 58, 50, 17, 18]. A number of recent works have generalized this setting beyond full demonstrations to the case where humans provide only desired outcomes or goals [19, 45]. Furthermore, both techniques have been shown to be effective for learning manipulation tasks on real robots in challenging high dimensional settings [17, 45, 57]. Unlike these works, which study single task problems in a single environment, the focus of this work is in learning generalizable *multi-task* reward functions for visual robotic manipulation that can produce rewards for different tasks by conditioning on a single video of a human completing the task.

B. Robotic Learning from Human Videos

A number of works have studied learning robotic behavior from human videos. One approach is to explicitly perform some form of object or hand tracking in human videos, which can then be translated into a sequence of robot actions or motion primitives for task execution [26, 52, 30, 25, 36]. Unlike these works, which hand-design the mapping from a human sequence to robot behaviors, we aim to learn the functional similarity between human and robot videos through data.

More recently, a range of techniques have been proposed for end-to-end learning from human videos. One such approach is to learn to translate human demos or goals to the robot perspective directly through pixel based translation with paired [27, 44] or unpaired [47] data. Other works attempt to infer actions, rewards, or state-values of human videos and use them for learning predictive models [40] or RL [14, 39]. Learning keypoint [51, 8] or object/task centric representations from videos [42, 38, 34] is another promising strategy to learning rewards and representations between domains. Simulation has also been leveraged as supervision to learn such representations [32] or to produce human data with domain randomization [3]. Finally, meta-learning [54] and subtask discovery [41, 20] have also been explored as techniques for acquiring robot rewards or demos from human videos. In contrast to the majority of these works, which usually study a small set of human videos in a similar domain as the robot, we explicitly focus on leveraging "in-the-wild" human videos, specifically large and diverse sets of crowd-sourced videos from the real world from an existing dataset, which contains many different individuals, viewpoints, backgrounds, objects, and tasks.

Our approach is certainly not the first to study using such in-the-wild human videos. Works that have used object trackers [52], simulation [32], and sub-task discovery [20] have also been applied on in-the-wild video datasets like YouCook [9], Something-Something [21], and ActivityNet [15]. Learning from such videos has also shown promise for navigation problems [6]. Most related to this work is Concept2Robot [43], which learns robotic reward functions using videos from the Something-Something dataset [21] by using a pretrained video classifier. Unlike Concept2Robot, our method learns a reward function that is conditioned on a human video demo, and thus can be used to generalize to new tasks. Furthermore, in Section IV-D, we empirically find that our proposed approach provides a reward that generalizes to unseen environments with much greater success than the Concept2Robot classifier.

C. Robotic Learning from Large Datasets

Much like our work, a number of prior works have studied how learning from broad datasets can enhance generalization in robot learning [16, 33, 56, 13, 22, 24, 10, 5]. These works

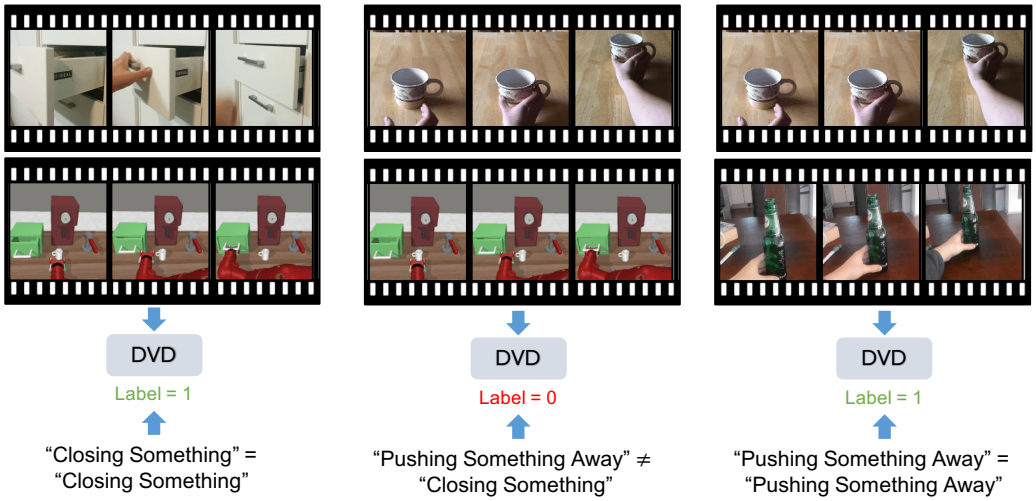

Figure 2: **Training DVD**. DVD is trained to predict if two videos are completing the same task or not. By leveraging task labels from in-the-wild human video datasets and a small number of robot demos, DVD is trained compare a video of a human to that of a robot (**left, middle**) and to compare pairs of human videos which may have significant visual differences, but may still be doing the same task (**right**). By training on these visually diverse examples, DVD is forced to learn the *functional* similarity between the videos.

have largely studied the problem of collecting large and diverse robotic datasets in scalable ways [28, 22, 10, 53, 7] as well as techniques for learning general purpose policies from this style of data in an offline [13, 5] or online [33, 29, 24] fashion. While our motivation of achieving generalization by learning from diverse data heavily overlaps with the above works, our approach fundamentally differs in that it aims to sidestep the challenges associated with collecting diverse robotic data by instead leveraging existing human data sources.

III. Learning Generalizable Reward Functions with Domain-Agnostic Video Discriminators

In this section, we describe our problem setting and introduce Domain-agnostic Video Discriminators (DVD), a simple approach for learning reward functions that leverage in-the-wild human videos to generalize to unseen environments and tasks.

A. Problem Statement

In our problem setting, we consider a robot that aims to complete K tasks $\{\mathcal{T}_i\}_{i=1}^K$, each of which has some underlying task reward function $\mathcal{R}_i$. As a result, for any given task i, our robotic agent operates in a fixed horizon Markov decision process (MDP) $\mathcal{M}_i^r$, consisting of the tuple $(\mathcal{S}, \mathcal{A}^r, p^r, \mathcal{R}_i, T)$ where $\mathcal{S}$ is the state space (in our case RGB images), $\mathcal{A}^r$ is the robot's action space, $p^r(s_{t+1}|s_t, a_t^r)$ is the robot environment's stochastic dynamics, $\mathcal{R}_i$ indicates the reward for task $\mathcal{T}_i$, and T is the episode horizon. Additionally, for each task $\mathcal{T}_i$, we consider a human operating in an MDP $\mathcal{M}_i^h$, consisting of the tuple $(\mathcal{S}, \mathcal{A}^h, p^h, \mathcal{R}_i, T)$ where $\mathcal{A}^h$ is the human's action space and $p^h(s_{t+1}|s_t, a_t^h)$ is the human environment's stochastic dynamics. Note that the human and robot MDPs for task i share a state space $\mathcal{S}$, reward function $\mathcal{R}_i$, and horizon T, but may have different action spaces and transition dynamics.

We assume that the task reward functions $\mathcal{R}_i$ are unobserved, and need to be inferred through a video of the task. Note that for many tasks these rewards will not be Markovian–for example for the task of "move two objects apart", the reward depends not only on the current state, but on how close together the objects were initially. We instead assume that the reward at time t is only dependent on the last $H < T$ timesteps, specifically states $s_{t-H:t}$. Our goal then is to learn a parametric model which estimates the underlying reward function for each task, conditioned on a task-specifying video. That is, given (1) a sequence of H states $s_{1:H}$ and (2) a video demonstration $d_i = s^*_{1:t_{d_i}}$ of variable length for each task $\mathcal{T}_i$, we aim to learn a reward function $\mathcal{R}_\theta(s_{1:H}, d_i)$ that approximates $\mathcal{R}_i(s_{1:H})$ for each i. Such a non-Markovian reward can then be optimized using a number of strategies, ranging from open-loop planners to policies with memory or frame stacking.

For training the reward function $\mathcal{R}_\theta$, we assume access to a dataset $\mathcal{D}^h = \{\mathcal{D}^h_{\mathcal{T}_i}\}_{i=1}^N$ of videos of humans doing $N < K$ tasks $\{\mathcal{T}_i\}_{i=1}^N$. There are no visual constraints on the viewpoints, backgrounds or quality of this dataset, and the dataset does not need to be balanced by task. We are also given a limited dataset $\mathcal{D}^r = \{\mathcal{D}^r_{\mathcal{T}_i}\}_{i=1}^M$ of videos of robot doing M tasks $\{\mathcal{T}_i\}_{i=1}^M$ where $\{\mathcal{T}_i\}_{i=1}^M \subset \{\mathcal{T}_i\}_{i=1}^N$, and so $M \leq N$. Both datasets are partitioned by task. Since human data is widely available, we have many more human video demonstrations than robot video demonstrations per task and often many more tasks that have human videos but not robot videos, in which case $M << N$. Importantly, the reward is inferred only through visual observations and does not assume any access to actions or low dimensional states from either the human or robot data, and we do not make any assumptions on the visual similarity between the human and robot data. As a result, there can be a large domain shift between the two datasets.

During evaluation, the robot is tasked with inferring the reward $\mathcal{R}_\theta$ based on a new demo d_i specifying a task $\mathcal{T}_i$. The goal is for this reward to be effective for solving a task $\mathcal{T}_i$. Furthermore, we aim to learn $\mathcal{R}_\theta$ in a way such that it can generalize to unseen tasks $\mathcal{T}_{new} \notin \{\mathcal{T}_i\}_{i=1}^N$ given a task demonstration d_{new}.

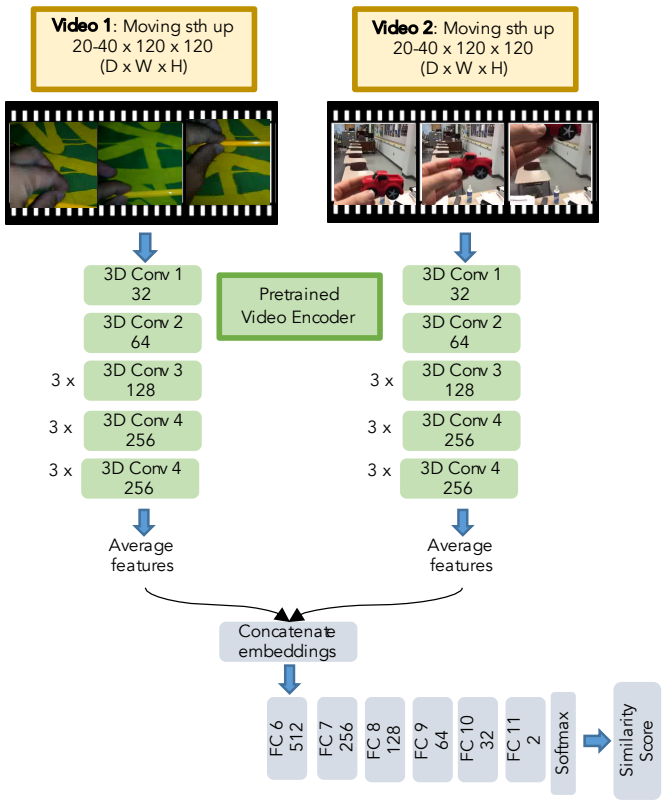

Figure 3: **DVD Architecture.** We use the same video encoder architecture as [43]. For each 3D convolution layer, the number of filters is denoted, and all kernels are $3 \times 3 \times 3$ except for the first, which is $3 \times 5 \times 5$. All conv layers have stride 1 in the temporal dimension, and conv layers 1, 3, 6, 9 and 11 have stride 2 in the spatial dimensions, the others having stride 1. All conv layers are followed by a BatchNorm3D layer and all layers except the last FC are followed by a ReLU activation.

B. Domain-Agnostic Video Discriminators

How exactly do we go about learning $\mathcal{R}_\theta$? Our key idea is to learn $\mathcal{R}_\theta$ that captures functional similarity by training a classifier which takes as input two videos d_i from $\mathcal{T}_i$ and d_j from $\mathcal{T}_j$ and predicts if $i = j$. Both videos can come from either $\mathcal{D}^h$ or $\mathcal{D}^r$, and labels can be acquired since we know which demos d_i correspond to which tasks $\mathcal{T}_i$ (See Figure 2).

To train $\mathcal{R}_\theta$, we sample batches of videos (d_i, d_i', d_j) from $\mathcal{D}^h \cup \mathcal{D}^r$, where d_i and d_i' are both labelled as completing the same task $\mathcal{T}_i$, and d_j is completing a different task $\mathcal{T}_j$. The output of $\mathcal{R}_\theta$ represents a "similarity score" that indicates how similar task-wise the two input videos are. More formally, $\mathcal{R}_\theta$ is trained to minimize the following objective, which is the average cross-entropy loss over video pairs in the distribution of the training data:

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathcal{D}^h \cup \mathcal{D}^r}[\log(\mathcal{R}_\theta(d_i, d_i')) + \log(1 - \mathcal{R}_\theta(d_i, d_j))]. \quad (1)$$

Since in-the-wild human videos are so diverse and visually different from the robot environment, a large challenge lies in bridging the domain gap between the range of human video environments and the robot environment. In optimizing Equation 1, $\mathcal{R}_\theta$ must learn to identify functional behavior in the robot videos and associate it with actions in human videos.

Algorithm 1 DOMAIN-AGNOSTIC VIDEO DISCRIMINATOR (DVD)

// Training DVD
Require: $\mathcal{D}^h$ human demonstration data for N tasks $\{\mathcal{T}_n\}$
Require: $\mathcal{D}^r$ robot demonstration data for M tasks $\{\mathcal{T}_m\} \subseteq \{\mathcal{T}_n\}$
Require: Pre-trained video encoder f_{enc}
Randomly initialize θ
while training **do**
  Sample anchor video $d_i \in \mathcal{D}^h \cup \mathcal{D}^r$
  Sample positive video $d_i' \in \{\mathcal{D}^h_{\mathcal{T}_i}\} \cup \{\mathcal{D}^r_{\mathcal{T}_i}\} \setminus d_i$
  Sample negative video $d_j \in \{\mathcal{D}^h_{\mathcal{T}_j}\} \cup \{\mathcal{D}^r_{\mathcal{T}_j}\} \forall j \neq i$
  Update $\mathcal{R}_\theta$ with d_i, d_i', d_j according to Eq. 1
// Planning Conditioned on Video Demo
Require: Trained reward function $\mathcal{R}_\theta$ & video prediction model p_ϕ
Require: Human video demo d_i for task $\mathcal{T}_i$
for trials $1, ..., n$ **do**
  Sample $\{a^{1:G}_{1:H}\}$ & get predictions $\{\tilde{s}^g_{1:H}\} \sim \{p_\phi(s_0, a^g_{1:H})\}$
  Step $a^*_{1:H}$ which maximizes $\mathcal{R}_\theta(\tilde{s}^g_{1:H}, d_i)$

C. DVD Implementation

We implement our reward function $\mathcal{R}_\theta$ as

$$\mathcal{R}_\theta(d_i, d_j) = f_{sim}\left(f_{enc}(d_i), f_{enc}(d_j); \theta\right) \quad (2)$$

where $h = f_{enc}$ is a pretrained video encoder and $f_{sim}(h_i, h_j; \theta)$ is a fully connected neural network parametrized by θ trained to predict if video encodings h_i and h_j are completing the same task. Specifically, we encode each video using a neural network video encoder f_{enc} into a latent space, and then train f_{sim} as a binary classifier trained according to Equation 1. See Figure 3 for the detailed architecture. f_{enc} is pretrained on the entire Sth Sth V2 dataset and fixed during training (as in [43]), while f_{sim} is randomly initialized. While the training dataset contains many more human videos than robot videos, we sample the batches so that they are roughly balanced between robot and human videos; specifically, each of (d_i, d_i', d_j) are selected to be a robot demonstration with 0.5 probability.

D. Using DVD for Task Execution

Once we've trained the reward function R_θ, how do we use it to select actions that will successfully complete a task? While in principle, this reward function can be combined with either model-free or model based reinforcement learning approaches, we choose to use visual model predictive control (VMPC) [49, 16, 13, 23], which uses a learned visual dynamics model to plan a sequence of actions. We condition R_θ on a human demonstration video d_i of the desired task $\mathcal{T}_i$ and then use the predicted similarity as a reward for optimizing actions with a learned visual dynamics model (See Figure 4).

Concretely, we first train an action-conditioned video prediction model $p_\phi(s_{t+1:t+H}|s_t, a_{t:t+H})$ using the SV2P model [2]. We then uses the cross-entropy method (CEM) [37] with this dynamics model p_ϕ to choose actions that maximize similarity with the given demonstration. More specifically, for each iteration of CEM, for an input image s_t, we sample G action trajectories of length H and roll out G corresponding predicted trajectories $\{s_{t+1:t+H}\}^g$ using p_ϕ. We then feed each

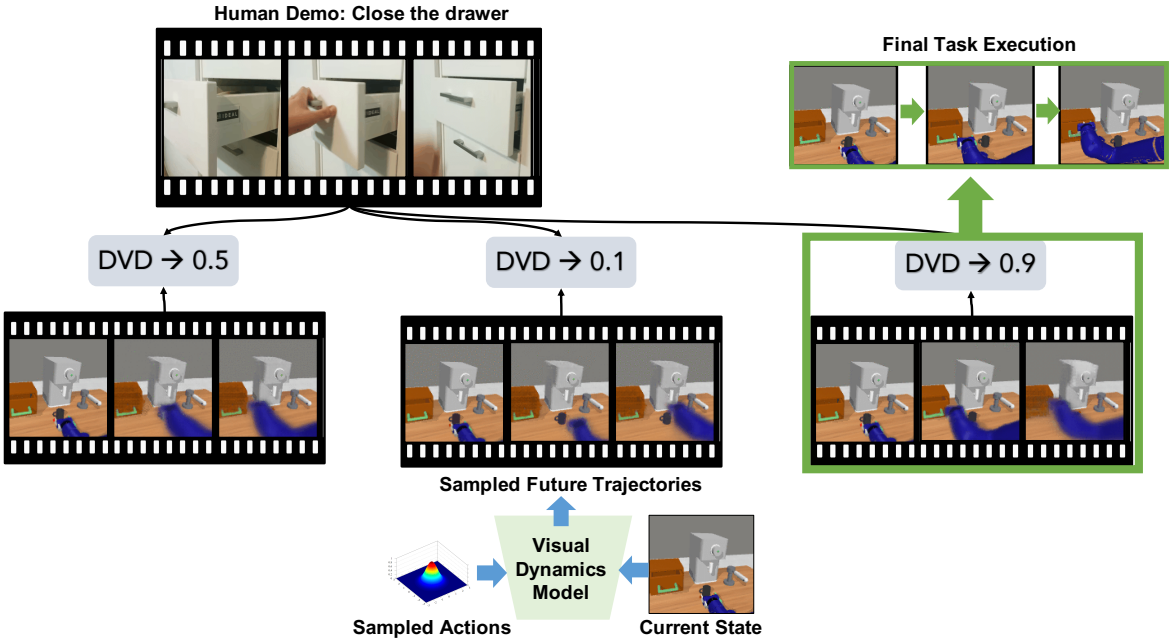

Figure 4: **Planning with DVD**. To use DVD to select actions, we perform visual model predictive control (VMPC) with a learned visual dynamics model. Specifically, we sample many action sequences from an action distribution and feed each through our visual dynamics model to get many "imagined" future trajectories. For each trajectory, we feed the predicted visual sequence into DVD along with the human provided demonstration video, which specifies the task. DVD scores each trajectory by its functional similarity to the human demo video, and steps the highest scored action sequence in the environment to complete the task.

predicted trajectory and demonstration d_i into $\mathcal{R}_\theta$, resulting in G similarity scores corresponding to the task-similarity between d_i and each predicted image trajectory. The action trajectory corresponding to the image sequence with the highest predicted probability is then executed to complete the task. The full algorithm with all stages is described in Algorithm 1.

IV. Experiments

In our experiments, we aim to study how effectively our method DVD can leverage diverse human data, and to what extent doing so enables generalization to unseen environments and tasks. Concretely, we study the following questions:

1) By leveraging human videos is DVD able to more effectively **generalize to new environments?**
2) By leveraging human videos is DVD able to more effectively **generalize to new tasks?**
3) Does DVD enable robots to generalize from a single human demonstration **more effectively than prior work**?
4) Can DVD infer rewards from a human video on a **real robot**?

In the following sections, we first describe our experimental setup and then investigate the above questions. For videos please see https://sites.google.com/view/dvd-human-videos.

A. Simulated Experimental Set-Up

a) Environments: For our first 3 experimental questions, we utilize a MuJoCo [48] simulated tabletop environment adapted from Meta-World [55] that consists of a Sawyer robot arm interacting with a drawer, a faucet, and a coffee cup/coffee machine. We use 4 variants of this environment to study

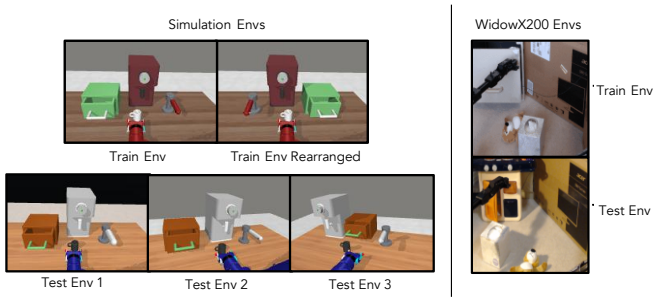

Figure 5: **Environment Domains.** We consider various simulated tabletop environments that have a drawer, a faucet, a coffee cup, and a coffee machine, as well as a real robot environment with a tissue box, stuffed animal, and either a file cabinet or a toy kitchen set. In the simulation experiments, half of the robot demonstrations that are used for training come from the train env and the other half from the rearranged train env.

environment generalization, each of which is progressively more difficult, shown in Figure 5. These include an original variant (Train Env), from which we have task demos, as well as a variant with changed colors (Test Env 1), changed colors and viewpoint (Test Env 2), and changed colors, viewpoint, and object arrangement (Test Env 3).

b) Tasks: We evaluate our method on three target tasks in simulation, specifically, (1) closing an open drawer, (2) turning the faucet right, and (3) pushing the cup away from the camera to the coffee machine. Each task is specified by an unseen in-the-wild human video completing the task (See Figure 6).

c) Training Data: For human demonstration data, we use the Something-Something-V2 dataset [21], which contains 220,837 total videos and 174 total classes, each with humans

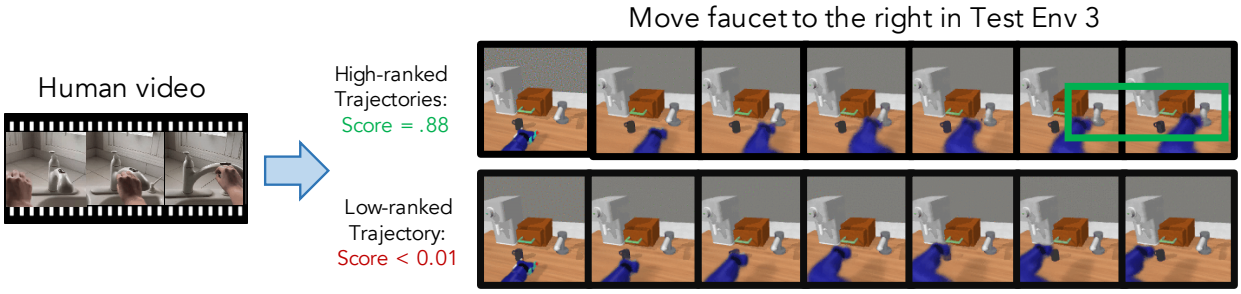

Figure 6: **Example Rankings During Planning.** Examples of predicted trajectories that are ranked high and low for the task of moving the faucet to the right in the test env 3 with the similarity scores that were outputted by DVD. DVD associates high functional similarity with trajectories that complete the same task as specified in the human video and low scores to trajectories that do not, despite the large visual domain shift between the given videos and the simulation environments.

performing a different basic action with a wide variety of different objects in various environments. Depending on the experiment, we choose videos from up to 15 different human tasks for training DVD, where each task has from 853-3170 videos (See Appendix for details). For our simulated robot demonstration data, we assume 120 video demonstrations of 3 tasks *in the training environment only* (See Figure 5). We ablate the number of robot demos needed in Section IV-F.

B. Experiment 1: Environment Generalization

In our first experiment, we aim to study how varying the amount of human data used for training impacts the reward function's ability to generalize across environments. To do so, we train DVD on robot videos of the 3 target tasks from the training environment, as well as varying amounts of human data, and measure task performance across *unseen environments*. One of our core hypotheses is that the use of diverse human data can improve the reward function's ability to generalize to new environments. To test this hypothesis, we compare training DVD on only the robot videos (**Robot Only**), to training DVD on a mix of the robot videos and human videos from K tasks (**Robot +** K **Human Tasks**). Note that the first 3 human tasks included are for the same 3 target tasks in the robot videos, and thus $K > 3$ implies using human videos for *completely unrelated* tasks to the target tasks. To evaluate the learned reward functions, we report the success rate from running visual MPC with respect to the inferred reward, where we train the visual dynamics model on data that is autonomously collected in the test environment. (See Appendix A for details). All methods infer the reward from a single human video. However, to provide an even stronger comparison, we also evaluate the Robot Only DVD model with a robot demo at test time **Robot Only (Robot Demo)**, since this model has only been trained with robot data.

In Figure 7, we report the success rate using each reward function, computed over 3 randomized sets of 100 trials. Our *first* key observation is that training with human videos significantly improves environment generalization performance over using only robot videos (20% on average), even when the robot only comparison gets the privileged information of a robot demonstration. *Additionally*, we observe that DVD is

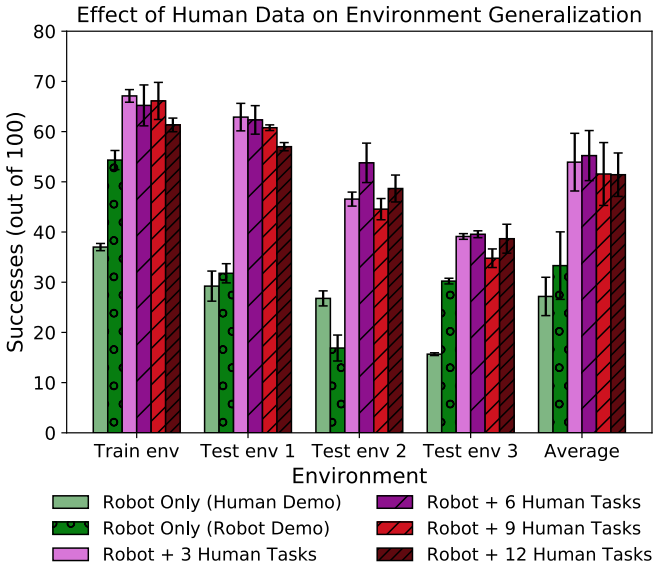

Figure 7: **Effect of Human Data on Environment Generalization.** We compare DVD's performance on seen and unseen environments when trained on only robot videos compared to varying number of human videos. We see that training with human videos provides significantly improved performance over only training on robot videos, and that DVD is generally robust to the number of different human video tasks used. Each bar shows the average success rate over all 3 target tasks, computed over 3 seeds of 100 trials, with error bars denoting standard error.

generally robust to the number of human tasks included, even if these tasks are *unrelated* to the target tasks. Even when using 9 completely unrelated tasks, performance greatly exceeds not using any human videos. Qualitatively, in Figure 6, we observe that DVD gives high similarity scores to trajectories that are completing the task specified by the human video demo and low scores to trajectories that have less relevant behavior.

C. Experiment 2: Task Generalization

In our second experiment, we study how including human data for training affects the reward function's ability to generalize to new tasks. In this case, we do not train on any (human or robot) data from the target tasks, and instead train DVD on robot videos of the 3 *different* tasks from the training environment, namely (1) opening the drawer, (2) moving

something from right to left, (3) not moving any objects, as well as varying amounts of human data. To again test how human videos affect generalization, we compare the same methods as in the previous experiment. Since we are testing task generalization, all evaluation is in the training environment.

In the bottom section of Table I, we report the success rate using DVD with varying amounts of human data, computed over 3 randomized sets of 100 trials. Similar to the conclusions of the environment generalization experiment, *first* we find that training with human videos significantly improves task generalization performance over using only robot videos (by roughly 10% on average), even with the robot only comparison conditioned on a robot demonstration. Given a human video demonstration, Robot Only does well at closing the drawer, but is completely unable to move the faucet to the right, suggesting that it is by default moving to the same area of the environment and is unable to actually distinguish tasks. This is unsurprising considering the reward function is not trained on any human videos. *Second*, we observe that on average, including human videos for 6 unrelated human tasks can significantly improve performance, leading to more than a 20% gap over just training with robot videos, suggesting that training with human videos from more unrelated tasks is particularly helpful for task generalization.

D. Experiment 3: Prior Work Comparison

In this experiment, we study how effective DVD is compared to other techniques for learning from in-the-wild human videos. The most related work is **Concept2Robot** [43], which uses a pretrained 174-way video classifier on only the Sth Sth V2 dataset (no robot videos) as a reward. Since this method is not naturally conducive to one-shot imitation from a video demonstration, during planning we follow the method used in the original paper and take the classification score for the target task from the predicted robot video as the reward (instead of conditioning on a human video). Unlike the open-loop trajectory generator used in the original paper, we use the same visual MPC approach for selecting actions for a fair comparison of the learned reward function; we expect the relative performance of the reward functions to be agnostic to this choice. In addition, we also compare to a demo-conditioned **behavioral cloning** method, similar to what has been used in prior work [54, 3, 46]. We train this approach using behavior cloning on the 120 robot demonstrations and their actions for 3 tasks conditioned on a video demo of the task from either a robot or a human. See Appendix A for more details on this comparison. We also include a comparison to a **random** policy.

In Figure 8 we compare DVD with 6 human videos to these prior methods on the environment generalization experiment presented in Section IV-B. Across all environments, DVD performs significantly better than all three comparisons on the target tasks, and 20% better on average than the best-performing other method. In Table I, we make the same comparison, now on the experiment of task generalization presented in Section IV-C. Since Concept2Robot is not demo-conditioned and is already trained on all 174 possible human video tasks in the Sth Sth

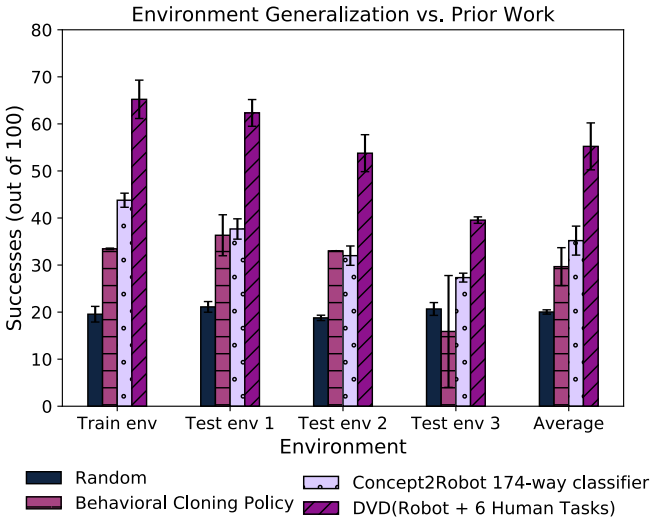

Figure 8: **Environment Generalization Prior Work Comparison.** Compared to Concept2Robot, the most relevant work leveraging "in-the-wild" human videos, as well as a demo-conditioned behavioral cloning policy and a random policy, DVD performs significantly better across all environments, and over 20% better on average. Each bar shows the average success rate over all 3 target tasks, computed over 3 seeds of 100 trials, with error bars denoting standard error.

V2 dataset, there is no natural method for testing generalization to an unseen task specified by a human video. We see that DVD outperforms both other baselines by over 30%.

First, DVD's significant improvement over Concept2Robot suggests that in learning reward functions which address the human-domain robot gap, using *some* robot data, even in small quantities, is important for good performance. *Second*, methods like demo-conditioned behavior cloning likely require many more robot demonstrations to learn good policies, as prior work in demo-conditioned behavior cloning often use on the order of thousands of demonstrations [46]. DVD on the other hand, uses the demos only to learn a reward function and offloads the behavior learning to visual MPC. *Lastly*, when examining the performance of demo-conditioned behavioral cloning on each individual task, we see the policy learns to ignore the conditioning demo and mimics one trajectory for one of the target tasks, doing well for only that task but completely failing at other tasks, suggesting that the policy struggles to infer the task from the visually diverse human videos.

E. Experiment 4: Real Robot Efficacy

To answer our last experimental question, we study how DVD with human data enables better environment and task generalization on a real WidowX200 robot. We consider a similar setup as described in Sections IV-B and IV-C, where DVD is now trained on 80 robot demos from each of 2 training tasks in a training environment and human videos. Then during testing, DVD is used as reward for visual MPC in an unseen environment, performing both a seen and unseen task.

Specifically, in our real robot setup, the training environment consists of a file cabinet, and in the testing environment, it is replaced with a toy kitchen set (See Figure 5). The training

Method	Close drawer	Move faucet to right	Push cup away from the camera	Average
Random	20.00 (3.00)	9.00 (1.73)	32.33 (8.08)	20.44 (2.78)
Behavioral Cloning Policy	0.00 (0.00)	45.33 (38.84)	1.00 (0.00)	15.44 (12.95)
Concept2Robot 174-way classifier	n/a	n/a	n/a	n/a
DVD, Robot Only (Human Demo)	**67.33 (4.51)**	1.00 (1.00)	29.67 (0.58)	32.67 (1.53)
DVD, Robot Only (Robot Demo)	29.33 (14.99)	23.67 (1.53)	28.33 (0.58)	27.11 (5.23)
DVD, Robot + 3 Human Tasks	**66.33 (6.03)**	19.33 (0.58)	40.00 (6.93)	41.89 (3.10)
DVD, Robot + 6 Human Tasks	59.00 (5.29)	17.00 (7.94)	56.33 (11.06)	44.11 (1.39)
DVD, Robot + 9 Human Tasks	57.67 (0.58)	**52.67 (1.15)**	55.00 (5.57)	**55.11 (2.04)**
DVD, Robot + 12 Human Tasks	31.67 (9.02)	49.00 (6.24)	**57.33 (2.08)**	46.00 (2.60)

Table I: Task generalization results in the original environment. DVD trained with human videos performs significantly better on average than with only robot videos, a baseline behavioral cloning policy, and random. We report the average success rate for all 3 target tasks, computed over 3 seeds of 100 trials, as well as the standard deviation in parentheses.

Method (Out of 20 Trials)	Test Env	Test Env + Unseen Task
Random	5	5
Concept2Robot 174-way classifier	4	n/a
DVD, Robot Only (Human Demo)	5	6
DVD, Robot Only (Robot Demo)	5	8
DVD, Robot + 2 Human Tasks	7	7
DVD, Robot + 6 Human Tasks	**13**	**14**
DVD, Robot + 9 Human Tasks	9	11
DVD, Robot + 12 Human Tasks	10	9

Table II: **Env and task generalization results on a real robot.** We report successes out of 20 trials on a WidowX200 in an unseen environment on two different tasks, one on closing a toy kitchen door and another on moving a tissue box to the left. On both, DVD performs significantly better when trained with human videos than with only robot demonstrations.

tasks are "Closing Something" and "Pushing something left to right" and the test tasks are "Closing Something" (seen) and "Pushing something right to left" (unseen).

We compare DVD with varying amounts of human data to only robot data and baselines in Table II, where we report the success rate out of 20 trials when used with visual MPC conditioned on a human demo of the task. DVD trained with human videos has about twice the success rate when leveraging the diverse human dataset than when relying only on robot videos. In particular, DVD trained with 6 tasks worth of human videos succeeds over 65-70% of the time whereas robot only succeeds at most 40%. We also observe that in general using human videos from unrelated tasks improves over only using human videos for the training tasks. Finally, we see qualitatively in Figure 14 in Appendix C that DVD captures the functional task being specified, in this case closing the door.

F. Ablation on Amount of Robot Data for Training

In our previous simulation experiments, we use 120 robot demonstrations per task. While this is a manageable number of robot demonstrations, it would be better to rely on fewer demonstrations. Hence, we ablate on the number of robot demonstrations used during training and evaluate environment generalization. In Figure 9, we see that the success rate of DVD decreases by only 6% when using **as few as 20 robot demonstrations** per task and by 13% when using no robot data, suggesting that DVD is robust to small amounts of robot data, but *some* robot data is important for good performance.

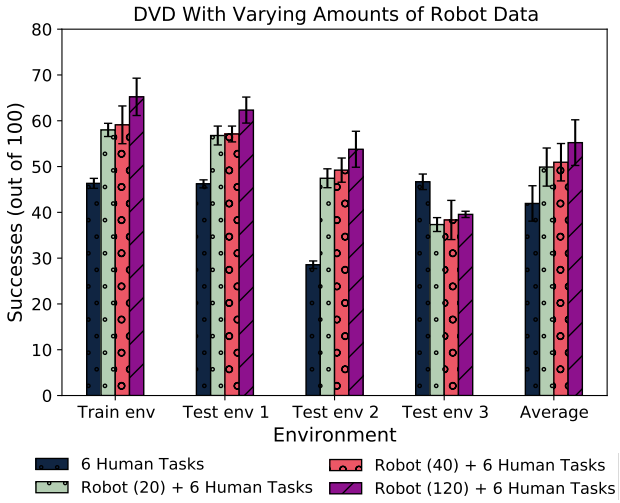

Figure 9: **Ablation on Amount of Robot Data Used for Training.** While using 120 robot demonstrations per task slightly benefits performance over using only 0, 20, or 40, DVD still performs comparably with fewer robot demos.

V. Limitations and Future Work

We presented an approach, domain-agnostic video discriminator (DVD), that leverages the diversity of "in-the-wild" human videos to learn generalizable robotic reward functions. Our experiments find that training with a large, diverse dataset of human videos can significantly improve the reward function's ability to generalize to unseen tasks and environments, and can be combined with visual MPC to solve tasks.

There are multiple limitations and directions for future work. First, our method focuses only on learning reward functions that generalize and does not learn a generalizable policy or visual dynamics model directly. This is a necessary next step to achieve agents that broadly generalize and is an exciting direction for future work. Second, while limited in quantity, our work assumes access to some robot demonstrations and task labels for these demos and for all of the human videos. Techniques that can sidestep the need for this supervision would further enhance the scalability of DVD. Lastly, so far we have only tested DVD on coarse tasks that don't require fine-grained manipulation. Designing more powerful visual models and testing DVD with them on harder, more precise tasks is another exciting direction for future work.

ACKNOWLEDGMENTS

The authors would like to thank Ashvin Nair as well as members of the IRIS lab for valuable discussions. This work was supported in part by Schmidt Futures, by ONR grant N00014-20-1-2675, and by an NSF GRFP. Chelsea Finn is a CIFAR Fellow in the Learning in Machines & Brains program.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, page 1, 2004.

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.

[3] Alessandro Bonardi, Stephen James, and Andrew J Davison. Learning one-shot imitation from humans without humans. *IEEE Robotics and Automation Letters*, 2020.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv:2005.14165*, 2020.

[5] Serkan Cabi, Sergio Gómez Colmenarejo, Alexander Novikov, Ksenia Konyushkova, Scott Reed, Rae Jeong, Konrad Zolna, Yusuf Aytar, David Budden, Mel Vecerik, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning. *arXiv:1909.12200*, 2019.

[6] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. In *NeurIPS*, 2020.

[7] Annie S. Chen, HyunJi Nam, Suraj Nair, and Chelsea Finn. Batch exploration with examples for scalable robotic reinforcement learning. *IEEE Robotics and Automation Letters*, 2021.

[8] Neha Das, Sarah Bechtle, Todor Davchev, Dinesh Jayaraman, Akshara Rai, and Franziska Meier. Model-based inverse reinforcement learning from visual demonstrations, 2021.

[9] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason 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*, pages 2634–2641, 2013.

[10] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, 2019.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv:1812.00568*, 2018.

[14] Ashley D Edwards and Charles L Isbell. Perceptual values from observation. *arXiv preprint arXiv:1905.07861*, 2019.

[15] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[16] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[17] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International conference on machine learning*, pages 49–58. PMLR, 2016.

[18] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations*, 2018.

[19] Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Advances in Neural Information Processing Systems*, 2018.

[20] W. Goo and S. Niekum. One-shot learning of multi-step tasks from observation via activity localization in auxiliary video. In *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz 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*, pages 5842–5850, 2017.

[22] Abhinav Gupta, Adithyavairavan Murali, Dhiraj Prakashchand Gandhi, and Lerrel Pinto. Robot learning in homes: Improving generalization and reducing dataset bias. In *Advances in Neural Information Processing Systems*, 2018.

[23] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben

Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pages 2555–2565. PMLR, 2019.

[24] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018.

[25] Jangwon Lee and Michael 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.

[26] Kyuhwa Lee, Yanyu Su, Tae-Kyun Kim, and Yiannis Demiris. A syntactic approach to robot imitation learning using probabilistic activity grammars. *Robotics and Autonomous Systems*, 61(12):1323–1334, 2013.

[27] YuXuan Liu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1118–1125. IEEE, 2018.

[28] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, 2018.

[29] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In *Advances in Neural Information Processing Systems*, 2018.

[30] Anh Nguyen, Dimitrios Kanoulas, Luca Muratore, Darwin G Caldwell, and Nikos 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.

[31] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation, 2019.

[32] Vladimír Petrík, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Learning object manipulation skills via approximate state estimation from real videos, 2020.

[33] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *IEEE international conference on robotics and automation (ICRA)*, 2016.

[34] Sören Pirk, Mohi Khansari, Yunfei Bai, Corey Lynch, and Pierre Sermanet. Online object representations with contrastive learning, 2019.

[35] Nathan D. Ratliff, J. Andrew Bagnell, and Martin A. Zinkevich. Maximum margin planning. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 729–736, 2006.

[36] Jonas Rothfuss, Fabio Ferreira, Eren Erdal Aksoy, You Zhou, and Tamim Asfour. Deep episodic memory: Encoding, recalling, and predicting episodic experiences for robot action execution. *IEEE Robotics and Automation Letters*, 3(4):4007–4014, 2018.

[37] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media, 2013.

[38] Rosario Scalise, Jesse Thomason, Yonatan Bisk, and Siddhartha Srinivasa. Improving robot success detection using static object data. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019.

[39] Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In *CoRL*, 2020.

[40] Karl Schmeckpeper, Annie Xie, Oleh Rybkin, Stephen Tian, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Learning predictive models from observation and interaction. In *ECCV*, 2020.

[41] Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. *Proceedings of Robotics: Science and Systems (RSS)*, 2017.

[42] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*, 2018.

[43] Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[44] P. Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. In *NeurIPS*, 2019.

[45] Avi Singh, Larry Yang, Chelsea Finn, and Sergey Levine. End-to-end robotic reinforcement learning without reward engineering. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, June 2019.

[46] Avi Singh, Eric Jang, Alexander Irpan, Daniel Kappler, Murtaza Dalal, Sergey Levinev, Mohi Khansari, and Chelsea Finn. Scalable multi-task imitation learning with autonomous improvement. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2167–2173. IEEE, 2020.

[47] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. AVID: Learning Multi-Stage Tasks via Pixel-Level Translation of Human Videos. In

Proceedings of Robotics: Science and Systems, Corvalis, Oregon, USA, July 2020.

[48] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[49] Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: a locally linear latent dynamics model for control from raw images. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2746–2754, 2015.

[50] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning, 2016.

[51] Haoyu Xiong, Quanzhou Li, Yun-Chun Chen, Homanga Bharadhwaj, Samarth Sinha, and Animesh Garg. Learning by watching: Physical imitation of manipulation skills from human videos, 2021.

[52] Yezhou Yang, Yi Li, Cornelia Fermüller, and Yiannis Aloimonos. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *AAAI*, pages 3686–3693, 2015.

[53] Sarah Young, Dhiraj Gandhi, Shubham Tulsiani, Abhinav Gupta, Pieter Abbeel, and Lerrel Pinto. Visual imitation made easy. In *CoRL*, 2020.

[54] Tianhe Yu, Chelsea Finn, Sudeep Dasari, Annie Xie, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-shot imitation from observing humans via domain-adaptive meta-learning. In *Proceedings of Robotics: Science and Systems*, Pittsburgh, Pennsylvania, June 2018.

[55] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, 2020.

[56] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[57] Henry Zhu, Justin Yu, Abhishek Gupta, Dhruv Shah, Kristian Hartikainen, Avi Singh, Vikash Kumar, and Sergey Levine. The ingredients of real world robotic reinforcement learning. In *International Conference on Learning Representations*, 2020.

[58] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *Proc. AAAI*, pages 1433–1438, 2008.

APPENDIX

A. Training Details

a) Dataset Details: Depending on the experiment, we choose videos from the following 15 different human tasks in the Something-Something-V2 dataset for training DVD, where each task has from 853-3170 training videos: 1) Closing sth, 2) Moving sth away from camera, 3) Moving sth towards camera, 4) Opening sth, 5) Pushing sth left to right, 6) Pushing sth right to left, 7) Poking sth so lightly it doesn't move, 8) Moving sth down, 9) Moving sth up, 10) Pulling sth from left to right, 11) Pulling sth from right to left, 12) Pushing sth with sth, 13) Moving sth closer to sth, 14) Plugging sth into sth, and 15) Pushing sth so that it slightly moves. We used these tasks because they are appropriate for a single-arm setting and cover a diverse range of various actions. The first seven of those tasks were chosen as they are relevant to tasks possible in the simulation environments, but the other tasks were not chosen for any particular reason, i.e. could have been replaced by a different set of 8 other appropriate tasks. For an experiment, if human videos for a task are used, we use *all* of the human videos available in the Sth Sth V2 training set for that task to train DVD.

For the simulation experiments, DVD is also trained on 120 robot video demonstrations of 3 tasks, half of which are collected in the original training environment and the other half in the rearranged training environment. These videos are collected via model predictive control with random shooting, a ground truth video-prediction model, and a ground-truth shaped reward particular to each task. For the real robot experiments on the WidowX200, in addition to varying amounts of human videos, DVD is trained on 80 robot video demonstrations of 2 tasks, which are collected in the original training environment. These demonstrations are collected via a hard-coded script with uniform noise between -0.02 and 0.02 added to each action.

To evaluate DVD's training progress, we use a validation set consisting of all of the human videos available in the Sth Sth V2 validation set for the chosen tasks as well as 48 robot video demonstrations for the same 3 tasks with robot demos in the training set, with half of these coming from the original training environment and the other half from the rearranged. For the WidowX200 experiments, we add 8 robot video demonstrations for each of the 2 tasks into the validation set.

b) Hyperparameters: For DVD, the similarity discriminator is trained with a learning rate of 0.01 using stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.00001. We use a batch size of 24, where each element of the batch consists of a triplet with two videos having the same task label and the third having a different label. Each version of DVD in the experiments is trained for 120 epochs, where one epoch consists of 200 optimizer steps. For each epoch, the video clips fed into DVD for training are sequences of consecutive frames with random length between 20 and 40 frames taken from the original video. If the original video has fewer than the randomly selected amount of frames, the last frame is repeated to achieve the desired number of frames for the training clip. Additionally, during training, each input video is first randomly rotated between -15 and 15 degrees, scaled so that the height has size 120, and then randomly cropped to have size $120 \times 120 \times 3$. At planning time, the video demonstration is spliced so that it is between 30 and 40 frames, rescaled to have a height of 120 pixels, and then center-cropped to have size $120 \times 120 \times 3$. The demo-conditioned behavioral cloning baseline uses the same hyperparameters and training details except that it uses weight decay 0.0001.

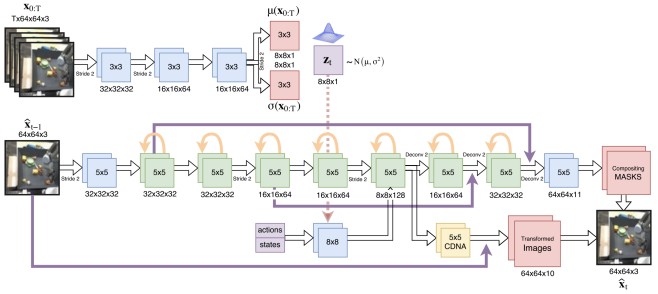

Figure 10: **SV2P Architecture.** We use video prediction models trained via SV2P with the reward from DVD in order to complete tasks specified by a given human video. Figure taken from the original paper [2].

c) SV2P Training: To evaluate DVD's performance on potentially unseen tasks with a given human video, we employ visual model predictive control with SV2P visual prediction models trained on datasets autonomously collected in each environment. SV2P learns an action-conditioned video prediction model by sampling a latent variable and subsequently generating an image prediction with that sample. We use the same architecture, which is shown in Figure 10, and default hyperparameters as the original paper [2].

For each of the four simulation environments in which we evaluate DVD, we collect 10,000 random episodes, each with 60 total frames, of the agent interacting in that environment and train SV2P for 200,000 epochs on all of the data. The models are trained to predict the next fifteen frames given an input of five frames. To evaluate DVD in the robot test environment, we train SV2P for 160,000 epochs on 58,500 frames worth of autonomously collected robot interaction in the original train environment, and then we finetune the model for another 60,000 epochs on 21,000 frames worth of autonomously collected data in the test environment that has the toy kitchen door. Collecting this data on the WidowX200 took a total of roughly 75 hours but was entirely autonomous.

d) Additional DVD Details: Here we expand on some of the DVD implementation details touched upon in Section III-C, particularly the way that batches are sampled during training. Each batch consists of triplets (d_i, d_i', d_j), where d_j is labeled as a different task as d_i and d_i', which are labeled as the same task. In each triplet, d_i is randomly sampled with 0.5 probability of being a robot demonstration. Then, if d_i is from a task with only human data, d_i' will be chosen from the remaining human data for that task; otherwise it is chosen to be a robot video from that task with 0.5 probability. Finally, d_j is randomly sampled repeatedly (usually just once) with 0.5 probability of

being a robot demonstration until a video with a different task label from d_i and d_i' is sampled.

e) Comparisons: For the Concept2Robot comparison, we use the same 174-way classifier that the paper used and do not alter it. For the demo-conditioned behavioral cloning comparison, we use a method similar to [3], [46], and [54]. We train a model that takes in as input the concatenated encodings of a conditioning video and the image state from one of the robot demonstrations in the training set and outputs an action that aims to lead the agent from the given image state to completing the same task as shown in the conditioning demo. During training, the model is trained on batches of (conditioning video, robot demonstration) pairs, where the conditioning video is randomly taken from the combined human and robot dataset and a robot demonstration with the same task label is randomly chosen. Because there are many more human videos than robot demonstrations, the conditioning video is chosen to be a robot demonstration with 50% probability, which is analogous to the balancing of batches used in DVD. Note that this method cannot naturally use human videos from tasks for which there are no robot demonstrations.

The behavior cloning model uses the same pretrained video encoder as DVD to encode the conditioning demo as well as a pretrained ResNet18 for the image state. The resulting features are concatenated and passed into an MLP that takes an input of size [1512] and has fully connected layers $[512, 256, 128, 64, 32, a]$, where each layer except the last is followed by a ReLU activation and a corresponds to the number of action dimensions. The model is trained to minimize mean squared error between the output action and the true action.

B. Experimental Details

a) Domains: For the simulation domains, we use a Mujoco simulation built off the Meta-World environments [55]. In simulation, the state space is the space of RGB image observations with size [180, 120, 3]. We use a continuous action space over the linear and angular velocity of the robot's gripper and a discrete action space over the gripper open/close action, for a total of five dimensions. For the robot domain, we consider a real WidowX200 robot interacting with a file cabinet, a tissue box, a stuffed animal, and a toy kitchen set. The state space is the space of RGB image observations with size [120, 120, 3], and the action space consists of the continuous linear velocity of the robot's gripper in the x and z directions as well as the gripper's y-position, for a total of three dimensions.

b) Simulation Experiments: For all environment and task generalization experiments, in each trial we plan 3 trajectories of length 20. For each trajectory, we sample 100 action sequences uniformly randomly and randomly choose one of the top 5 predicted trajectories with the highest functional similarity score given by DVD to execute in the environment. For Concept2Robot, we take one of the top 5 predicted trajectories with the highest classification score for the specified task, and for the behavioral cloning policy, we simply take the predicted action at each state. We evaluate on the following three target tasks: 1) Closing the drawer, which is defined as the last frame in the 60-frame trajectory having the drawer pushed in to be less than 0.05, where it starts open at 0.07, 2) Turning the faucet to the right more than 0.01 distance, where it starts at 0, and 3) Moves cup to be less than 0.07 distance to the coffee machine, where the cup starts out at least 0.1 away. We run 100 trials for 3 different seeds for each task for every method in all experiments.

c) Real Robot Experiments: On the WidowX200, for all experiments, in each trial we plan 1 trajectory of length 10. For this trajectory, we run 2 iterations of the cross-entropy method (CEM), sampling 100 action sequences and refitting to the top 20 repeatedly. We then choose one of the top 5 predicted trajectories with the highest functional similarity score given by DVD to execute in the environment. We evaluate on the following two target tasks: 1) Closing the toy kitchen door, where a success is recorded for any trial where the robot arm completely closes the door, and 2) Pushing the tissue box to the left, where the robot arm must clearly push the tissue box left of its original starting position. We run 20 trials for each task for each method.

C. Additional Experimental Results

In our simulation environment generalization experiments, we evaluate on the three tasks of 1) Closing the drawer, 2) Turning the faucet to the right, and 3) Pushing the cup away from the camera. In Section IV, we reported the average performance across all the three tasks. In Figure 11, we present the individual task results for DVD trained with varying amounts of human data. The conclusions of these experiments are the same as those in Section IV, in that leveraging diverse human videos in DVD allows for more effective generalization across new environments rather than relying only on robot videos.

In Figure 12, we present results on the individual tasks across all four environments for DVD trained with 6 tasks worth of human videos compared with our three comparisons: Concept2Robot [43], a demo-conditioned behavior cloning policy, and a random policy. On average over all of the environments, DVD performs over 40% better on the drawer task and 30% better on the faucet task than the next best performing method. It also performs reasonably on the cup task; Concept2Robot just performs particularly well on that task since it often chooses to push the cup away no matter which task is specified. The behavioral cloning policy has somewhat erratic behavior, mimicking the trajectory for one of the target tasks in each environment and doing well on that task but not on the other tasks. Hence, we see that both Concept2Robot and the behavioral cloning policy are not able to provide effective *multi-task* reward signals for each environment.

Additionally, in Figure 13, we show the accuracy curves on the training and validation sets while training DVD. We see unsurprisingly that the model trained only on three tasks of robot demonstrations (Robot Only) has the highest validation accuracy at 99%. However, while adding human videos significantly increases the difficulty of the optimization,

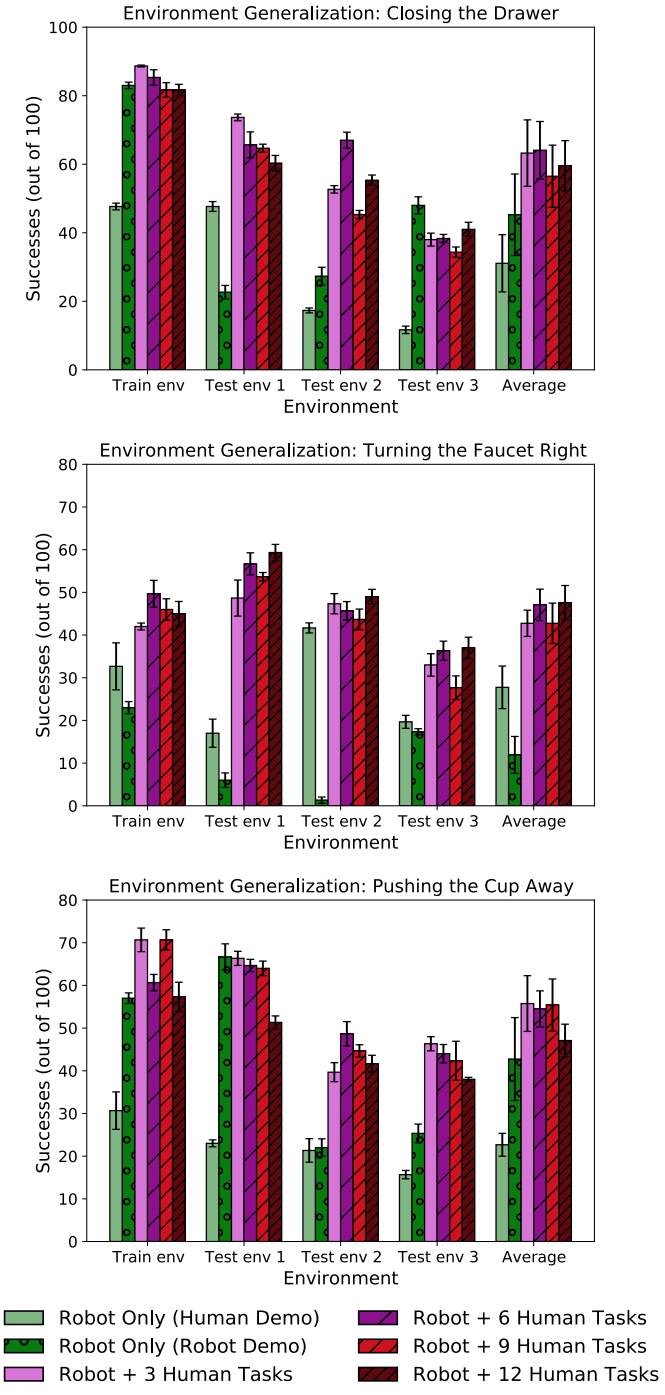

Figure 11: **Effect of Human Data on Environment Generalization.** We compare DVD's performance on seen and unseen environments when trained on only robot videos compared to varying number of human videos. Across all three tasks, we see that training with human videos provides significantly improved performance over only training on robot videos, and that DVD is generally robust to the number of different human video tasks used. Each bar shows the average success rate over all 3 target tasks, computed over 3 seeds of 100 trials, with error bars denoting standard error.

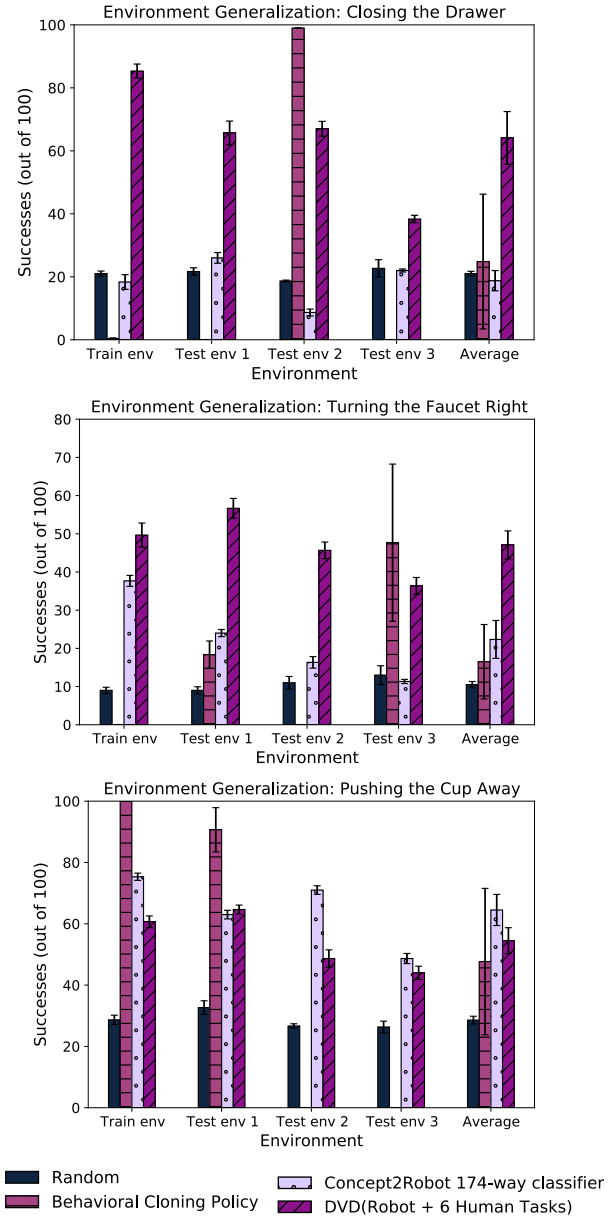

Figure 12: **Environment Generalization Prior Work Comparison.** We compare DVD's performance to Concept2Robot, the most relevant work, a demo-conditioned behavioral cloning policy, and a random policy. On average across environments, DVD performs around or over 30% better than the next-best performing method on two of the three tasks. Each bar shows the average success rate over all 3 target tasks, computed over 3 seeds of 100 trials, with error bars denoting standard error.

the models remain generally robust, with DVD trained on robot data and 12 tasks worth of human videos still obtaining 89% validation accuracy. We find in our experiments in Section IV that this trade-off in discriminator accuracy from adding human videos to the training set results in much greater ability to generalize to unseen environments and tasks.

Finally, in Figure 14, we include examples on the real Widowx200 of predicted trajectories and their similarity scores with a human video demonstration given by DVD. We see that

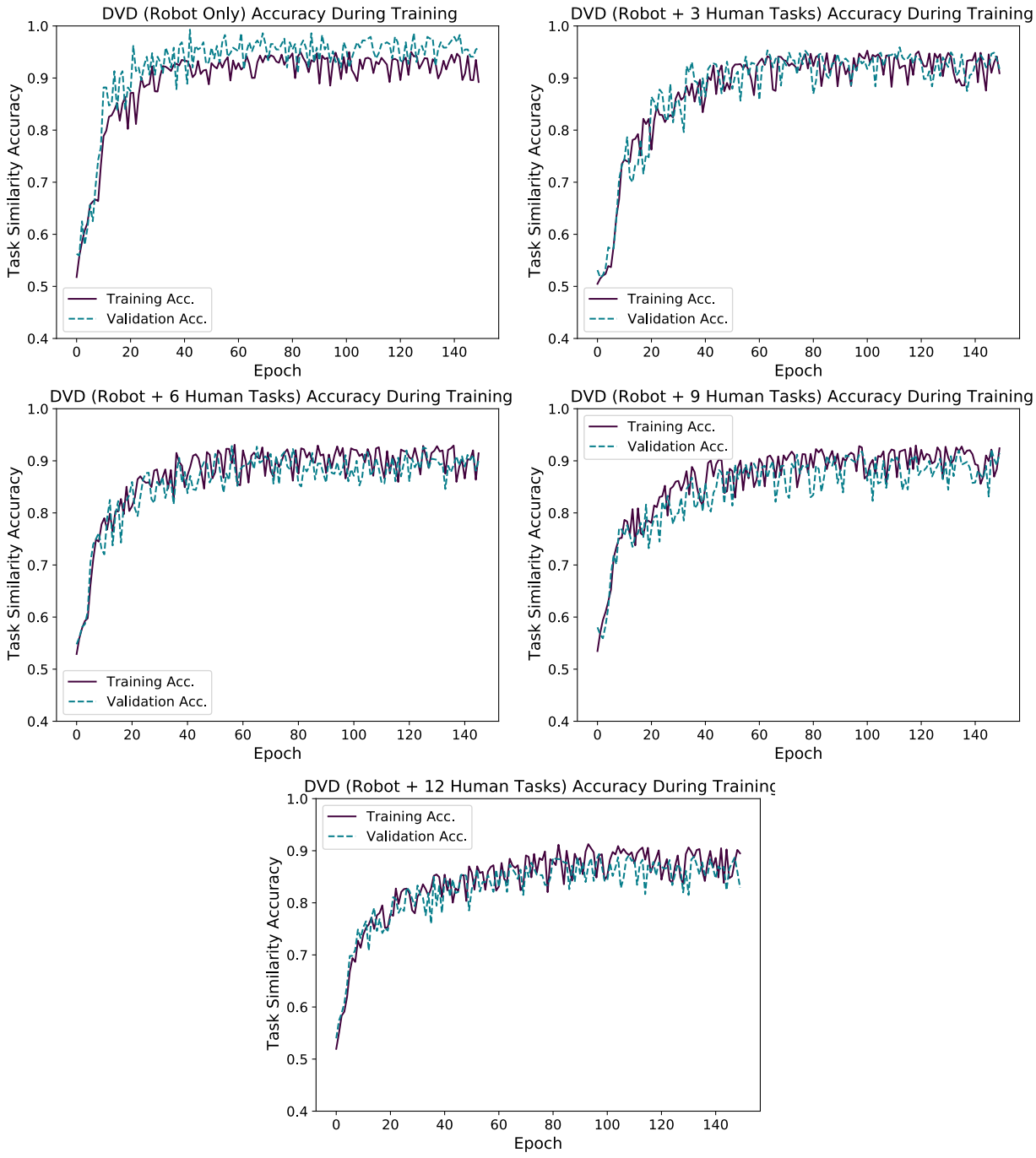

Figure 13: **Accuracy Curves During DVD Training.** We plot both training and validation accuracies over the course of training DVD for 150 epochs with varying amounts of human data. The accuracies gradually decrease as more human videos are added, but we find that this trade-off is worthwhile for greater generalization capabilities.

Figure 14: **Rankings on the real robot.** Examples of predicted trajectories on the WidowX200 that are ranked high and low for the task of closing an unseen toy kitchen door. DVD gives the predicted trajectory where the door is closed a high similarity score and the predicted trajectory where the door stays open a low similarity score.

DVD highly ranks trajectories that are completing the same task as demonstrated in the given human video.