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.

1 INTRODUCTION

Despite recent progress in robotic learning on tasks ranging from grasping (Kalashnikov et al., 2018) to in-hand manipulation (OpenAI et al., 2019), 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 themselves 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 (Deng et al., 2009; Devlin et al., 2019; Brown et al., 2020) 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 (Mandlekar et al., 2018; Gupta et al., 2018; Dasari et al., 2019; Young et al., 2020) and have shown some promise in enabling generalization (Dasari et al., 2019). 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 strong dependence 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 could 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.

Of course, using such "in-the-wild" data of humans to enable better 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 (Goyal et al., 2017), allowing them to be incorporated into the robot learning process with little additional supervision cost or collection overhead.

Given the extremely diverse and noisy nature of "in-the-wild" human videos (See Figure 1), how might one actually learn reward functions from them? 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 which come with many human video datasets as supervision, along with a small amount of robot demos, we can train this model to 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 one-to-one mapping between the robot and human data. Once trained, DVD can condition upon a human video as a demonstration, and the robots behavior as the other video, and outputs a score which is an effective measure of task success and reward.

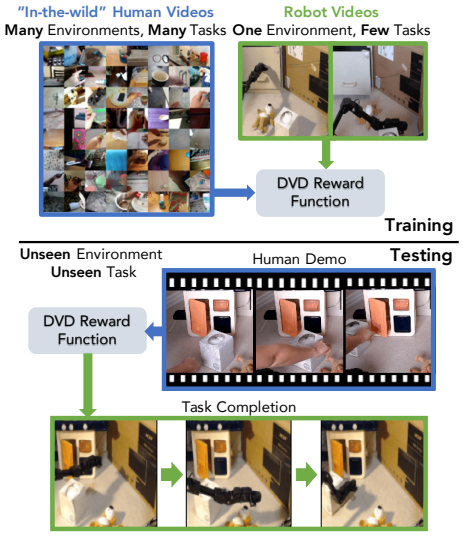

Figure 1: **Reward Learning and Planning from In-The-Wild Human Videos.** During training (**top**), the agent learns a reward function from a small set of robot videos in one environment, and a large set of in-the-wild human videos spanning many tasks and environments. At test time (**bottom**), the 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. Furthermore, by virtue of training on diverse human data, this reward function is able to generalize to unseen environments and tasks.

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, that can then be used to provide a robot with a reward that measures the functional similarity between its behavior and that of a human demonstrator. We find that this reward function is able to handle the diversity of human videos found in the Something-Something-V2 (Goyal et al., 2017) 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% improvement in downstream task success. Lastly, we evaluate our method on a real WidowX200 robot, and find that it enables zero shot generalization to an unseen task in an unseen environment given only a single human demonstration.

2 Related Work

2.1 Robotic Learning from Human Videos

Our approach is certainly not the first to study using such in-the-wild human videos. Works which have used object trackers (Yang et al., 2015), simulation (Petrík et al., 2020), and sub-task discovery (Goo and Niekum, 2019) have also been applied on in-the-wild video datasets like YouCook (Das et al., 2013), Something-Something (Goyal et al., 2017), and ActivityNet (Fabian Caba Heilbron and Niebles, 2015). Learning from such in-the-wild videos has also shown promise as an approach for navigation (Chang et al., 2020). Most related to our work is Concept2Robot (Shao et al., 2020) which also studies learning robotic reward functions from in-the-wild human videos from the Something-

Something dataset (Goyal et al., 2017), specifically by using a pretrained video classifier as a reward function for robotic RL. Unlike Concept2Robot, our method learns a single reward function that is conditioned on a human video demo, and thus can be used to generalize to new tasks. Furthermore, in Section 4.4 we empirically compare our approach to Concept2Robot and find that our proposed approach provides a more effective reward for generalizing to unseen environments. See Appendix A for a detailed discussion of related work.

3 Learning Generalizable Reward Functions with Domain-agnostic Video Discriminators

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

3.1 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 Markov decision process (MDP) $\mathcal{M}_i^r$, consisting of the tuple $(\mathcal{S}, \mathcal{A}^r, p^r, \mathcal{R}_i)$ where $\mathcal{S}$ is the state space, $\mathcal{A}^r$ is the robot's action space, $p^r(s_{t+1}|s_t, a_t)$ is the robot environment's stochastic dynamics, and $\mathcal{R}_i$ indicates the reward for task $\mathcal{T}_i$. Additionally, for each task $\mathcal{T}_i$, we consider a human operating in a human MDP $\mathcal{M}_i^h$, consisting of the tuple $(\mathcal{S}, \mathcal{A}^h, p^h, \mathcal{R}_i)$ where $\mathcal{S}$ is the state space, $\mathcal{A}^h$ is the human's action space, $p^h(s_{t+1}|s_t, a_t)$ is the human environment's stochastic dynamics, and $\mathcal{R}_i$ indicates the reward for task $\mathcal{T}_i$. Note that the human and robot MDPs for task i share a state space $\mathcal{S}$ and a reward function $\mathcal{R}_i$, 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 demonstrations of each task. Our goal then is to learn a parametrized model which estimates the underlying reward function for each task, conditioned on a task-specifying video. Concretely, we aim to learn a reward function that for all i approximates $\mathcal{R}_i(s_{t:t+H})$ with $\mathcal{R}_\theta(s_{t:t+H}, d_i)$ where d_i is a video demonstration solving task $\mathcal{T}_i$, and the learned reward function is parametrized by θ.

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. As much more human data is readily 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 $M << N$. Each video d in either dataset simply consists of a sequence of image observations $(s_1, ..., s_{t_d})$, which have varying lengths, and we do not require any low-dimensional state information. Importantly, the reward is inferred only through observations and does not assume any access to actions 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}.

3.2 Learning the Reward Function

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

Concretely, we define our reward function

$$R_\theta(s_{t:t+H}, d_i) = f_{sim}((f_{enc}(d_i), f_{enc}(s_{t:t+H})); \theta) \tag{1}$$

where $h = f_{enc}(d)$ 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. To train f_{sim}, 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 $\mathcal{T}_j$. We encode these videos using a neural network video encoder f_{enc} into a latent space, and then train f_{sim} as a binary classifier with the encodings for the pair (d_i, d_i^*) concatenated and labeled as positive and the encodings for the pair (d_i, d_j) concatenated and labeled as negative (See Figure 7 in Appendix B.2 for details). In this way, the output of f_{sim} represents a "similarity score" that indicates how similar task-wise the two input videos are. More formally, f_{sim} is trained to minimize the following objective, which is the average cross-entropy loss over video pairs in the distribution of the training data P:

$$\mathcal{J}(d_i, d_i^*, d_j, f_{sim}) = \mathbb{E}_P[\log(f_{sim}(f_{enc}(d_i), f_{enc}(d_i^*))) + \log(1 - f_{sim}(f_{enc}(d_i), f_{enc}(d_j)))]. \tag{2}$$

The video encoder for f_{enc} is the same as the one used in Shao et al. (2020). It is pretrained on the entire Sth Sth V2 dataset and not modified during training. The discriminator f_{sim} is randomly initialized.

Figure 2: **Training DVD**. DVD is trained to predict if two videos are completeing the same task or not. By leveraging task labels from in-the-wild human video datasets and a small number of robot demos, DVD learns to look at a video of a human doing a task and a robot doing a task, and predict when they are doing the same task (**left, middle**). Additionally, DVD is trained on 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.

Since in-the-wild human videos are so diverse and are so 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. The similarity discriminator must learn to distinguish various tasks in the robot environment and associate them with actions in human videos, as a human video demonstration will be given at planning time in the second stage. Because the training dataset contains many more human videos than robot videos, in order to leverage the limited amount of robot data, the batches sampled are roughly balanced between robot and human videos: each of (d_i, d_i^*, d_j) are selected to be a robot demonstration with 0.5 probability, resulting in roughly 25% robot-robot or human-human (d_i, d_i^*) or (d_i, d_j) pairs and 50% human-robot pairs.

3.3 One-shot Planning with Visual MPC

Once we've trained the reward function R_θ how do we use it to select actions which will successfully complete a task? We take a visual model predictive control (VMPC) approach, where we first condition R_θ on a human demonstration video of the desired task, then use it as a planning cost to select actions using a learned visual dynamics model, as shown in Figure 3. Concretely, we leverage our trained similarity discriminator f_{sim} to specify a reward function $R(s_{t:t+H}, d_i)$ for task $\mathcal{T}_i$. Given any human video demonstration d_i of the task, we would like our robotic agent to take actions that lead to states $s_{t:t+H}$ that are completing the same task.

First, we train an action-conditioned visual dynamics model $p_\phi(s_{t+1:t+H}|s_t, a_{t:t+H})$ using a state of the art video prediction framework SV2P Babaeizadeh et al. (2018), where the states are images. DVD then uses the cross-entropy method (CEM) Rubinstein and Kroese (2013) using this dynamics model p_ϕ to choose actions that maximize similarity with the given demonstration. More specifically, for each iteration of CEM, at an input image s_t, we first 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 use the video encoder f_{enc} to encode the demonstration d_i and each predicted trajectory $s_{t+1:t+H}$, concatenate the encodings, and feed into f_{sim}, 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 similarity score is then chosen and executed in order to complete the task. The full algorithm with both stages is laid out in Algorithm 1.

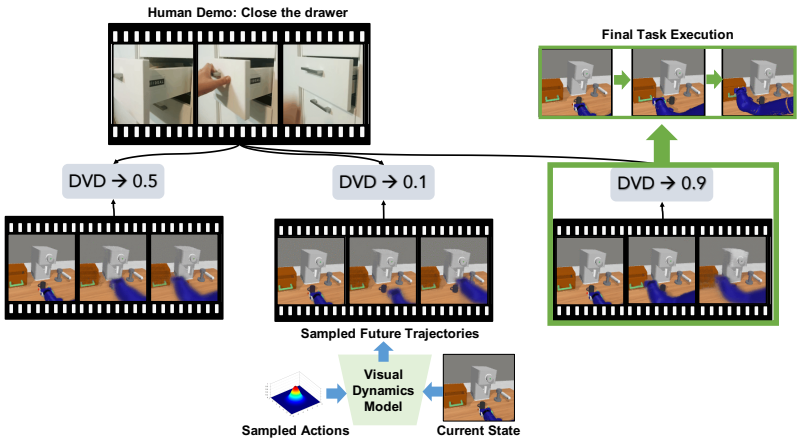

Figure 3: **Planning with DVD**. To use DVD to select actions, we perform visual model predictive control (VMPC) with a learned visual dynamics model. Specifically, we sample many action sequences from an action distribution and feed each through our visual dynamics model to get many "imagined" future 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, ans teps the highest scored action sequence in the environment to complete the task.

4 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 experimental questions:

1. By leveraging human videos is DVD able to more effectively **generalize across environments?**
2. By leveraging human videos is DVD able to more effectively **generalize across tasks?**
3. Does DVD enable robots to generalize from a single human demonstration **more effectively than prior work**?
4. Does DVD enable robotic imitation from a few human videos on a **real robot?**

In the following sections, we first describe our experimental domains and comparisons and then investigate each of the above questions.

4.1 Experimental Set-Up

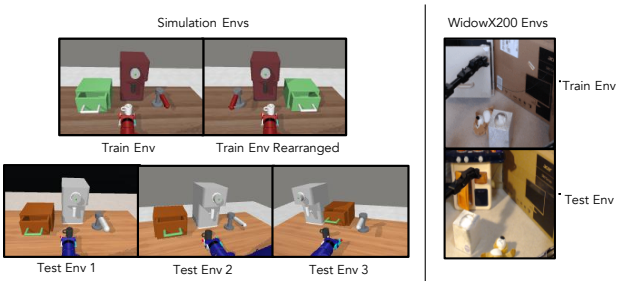

Figure 4: **Environment Domains.** We consider various simulated tabletop environments that have a drawer, a faucet, and a coffee cup/coffee machine, as well as a real robot environment with a tissue box, stuffed animal, and either a file cabinet or a toy kitchen set. In the simulation experiments, half of the robot demonstrations that are used for training come from the train env and the other half from the rearranged train env.

Environments For our first 3 experimental questions, we utilize a MuJoCo Todorov et al. (2012) simulated tabletop environment which consists of a Sawyer robot arm interacting with a drawer, a faucet, and a coffee cup/coffee machine. We study 4 variants of this environment to study environment generalization, each of which is progressively more difficult, shown in Figure 4. 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). Additionally, we study experimental question 4 on a real robot setup using a WidowX200 robot, in which the training environment includes a file cabinet, a tissue box, and a stuffed animal, and the test environment involves a toy kitchen set.

Tasks We evaluate our method on three target tasks in simulation. These tasks are (a) closing an open drawer, (b) turning the faucet right, and (c) pushing the cup away from the camera to a coffee machine. Each task is specified by an unseen human video performing the task in a different domain (See Figure 8 in Appendix D). On the real robot, we evaluate on the following two tasks: (1) Closing a toy kitchen door, and (2) Pushing a tissue box to the left.

Training Data We assume access to human demonstration data for some tasks, as well as small amounts of robot demonstrations for some subset of these tasks. For human demonstration data, we use the Sth Sth V2 dataset Goyal et al. (2017), which contains 220,837 total videos and 174 total classes, each with humans 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 4). Additionally, we ablate the number of robot demonstrations needed in Section D.2.

4.2 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 robot'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. All success rates are determined by running visual model predictive control[1] using the learned reward as described in Section 3.3 conditioned on a human demo of the task, except for (**Robot Only (Robot Demo)**) which receives the privileged information of a robot demo as a fairer comparison, since this reward function has never seen any human videos.

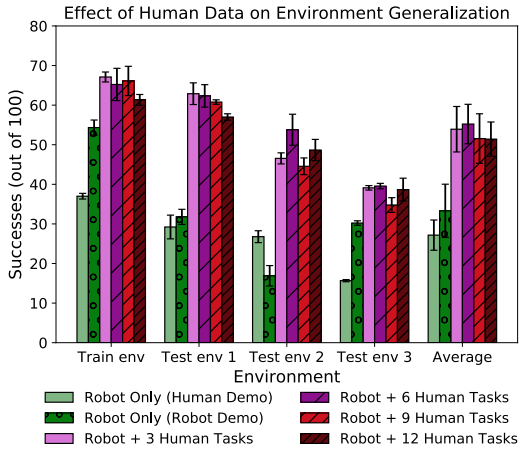

Figure 5: **Effect of Human Data on Environment Generalization.** We compare DVD's performance on seen and unseen environments when trained on only robot videos compared to varying number of human videos. We see that training with human videos provides significantly improved performance over only training on robot videos, and that DVD is generally robust to the number of different human video tasks used. Success rates computed over 3 seeds of 100 trials.

In Figure 5 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. Interestingly, this performance improvement exists even in the training environment, suggesting that not only does the diversity of the human videos improve unseen environment generalization, but it also improves robustness on the seen environments. *Second*, we observe that on average, including human videos for 3 unrelated human tasks can improve performance. *Lastly*, we see that adding several (6 or 9) unrelated tasks worth of human videos slightly hurts performance, as it can make learning the reward function more challenging. However, it still performs significantly better than not using human videos, suggesting that in general reward function performance is robust to the particular human videos or tasks used. Qualitatively, in Figure 8 in Appendix D, we observe that DVD gives high similarity scores to trajectories that are

[1]Note that while the unseen environment is totally unseen to the reward function, we do assume access to a trained visual dynamics model in the unseen environment for planning.

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	NA	NA	NA	NA
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 1: 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.

completing the task specified by the human video demo and low scores to trajectories that have less relevant behavior.

4.3 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) does not move any objects, as well as varying amounts of human data. Similar to our previous experiment, we compare training on only the robot videos (**Robot Only**), to training on a mix of the robot videos and human videos from K tasks (**Robot +** K **Human Tasks**), conditioned on a human demo of the unseen task. As before, **Robot Only (Robot Demo)** receives the privileged information of a robot demo for the target task as a fairer comparison, since the reward function is not trained on any human videos.

In Table 1, 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 (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 no matter the task specified by the conditioning demonstration and is unable to actually distinguish tasks. This is not surprising 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.

4.4 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.

Comparisons. The most related work is **Concept2Robot** Shao et al. (2020), 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). The only change to the method used in the original paper is instead of open-loop trajectory generator, we use the same visual MPC approach to selecting actions as DVD for a fair comparison of the learned reward function. In addition, we also compare to a demo-conditioned **behavioral cloning** method, similar to the first part of the method used in MILI Singh et al. (2020). We train this baseline using behavior cloning on the 120 robot demonstrations and their actions for 3 tasks conditioned on the combined on a video demo of the task from either a robot or a human. We also include a comparison to a **random** policy.

In Figure 6 we compare DVD with 6 human videos to these baselines on the environment generalization experiment presented in Section 4.2. 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. Nevertheless, Concept2Robot and the behavioral cloning policy still do perform significantly better than random. When examining the performance of the behavioral cloning baseline on each individual task, we see that in each environment, the policy learned ignores the conditioning demo and mimics one trajectory for one of the target tasks, doing well for that task but not for either of the other two. This is likely the case because the human videos are so diverse that behavior cloning is unable to extract the necessary task information from the demonstration.

In Table 1, we make the same comparison, now on the experiment of task generalization presented in Section 4.3. Since Concept2Robot is not demo-conditioned and is already trained on all 174 possible human video tasks in the Sth Sth 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%, as the behavioral cloning method ignores the conditioning demonstration.

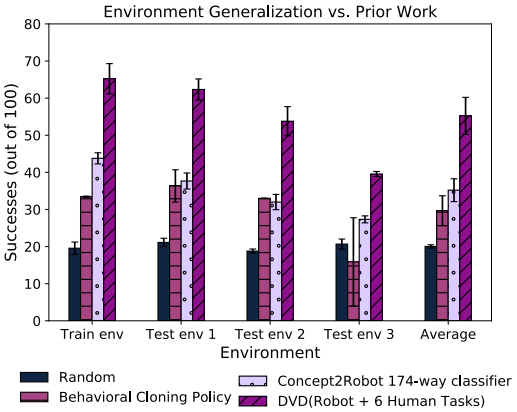

Figure 6: **Environment Generalization Prior Work Comparison.** Compared to Concept2Robot, the most relevant work leveraging "in-the-wild" human videos, as well as a baseline behavioral cloning policy and a random policy, DVD performs significantly better across all environments, and over 20% better on average.

4.5 Experiment 4: Real Robot Efficacy

To answer our last main experimental question, in Table 2, we study how DVD with human data enables better environment and task generalization on a real WidowX200 robot. We train DVD with varying amounts of human videos as well as 80 robot demonstrations from an original train env for the two tasks of 1) Closing a file cabinet and 2) Moving a tissue box to the right. We then evaluate DVD's environment generalization capabilities in a new environment with the task of closing a toy kitchen door. We also evaluate in a combined environment and task generalization setting on an unseen task of moving the tissue box to the left in the new environment, with DVD having not been trained on any videos (human or robot) of moving objects to the left. We also compare to a random policy and Concept2Robot for the environment generalization setting. For all settings, we report the success rate out of 20 trials.

In both settings, DVD trained with human videos succeeds much more often when leveraging the diverse human dataset than when relying only on robot videos. In particular, DVD with 6 tasks worth of human videos as well as robot demonstrations for 2 of those tasks succeeds over 65-70% of the time whereas robot only succeeds at most 40% of the time. While adding additional human videos makes learning DVD more challenging, the performance still outweighs not adding any human videos. We also see in Figure 9 in Appendix D that DVD captures the functional aspect of the specified task.

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

Table 2: **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.

5 Limitations

We have presented an approach, domain-agnostic video discriminator (DVD), that leverages the diversity in "in-the-wild" human videos to learn generalizable robotic reward functions. While our experiments demonstrate that training with a large, diverse dataset of human videos can significantly improve one-shot imitation performance to unseen tasks and in unseen environments, there are many limitations and directions for future work. First, our method focuses on learning reward functions, so it does not learn a generalizable policy directly. Second, we assume access to some robot demonstrations in a training environment, task labels for all training data, and a video prediction model in test environment for planning. Lastly, the learned tasks are still coarse, and we have not extended the method to more fine-grained tasks.

REFERENCES

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

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

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

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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

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.

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

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.

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

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*, 2018b.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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

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.

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.

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.

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.

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

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

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

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.

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.

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

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.

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.

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.

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.

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

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.

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.

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

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.

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.

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.

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.

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

A Extended Related Work

A.1 Reward Learning

The problem of learning reward functions from demonstrations of tasks, also known as inverse reinforcement learning or inverse optimal control Abbeel and Ng (2004), has a rich literature of prior work Ratliff et al. (2006); Ziebart et al. (2008); Wulfmeier et al. (2016); Finn et al. (2016); Fu et al. (2018a). A number of recent works have generalized this setting beyond full demonstrations to the case where humans provide only desired outcomes or goals Fu et al. (2018b); Singh et al. (2019). Furthermore, both techniques have been shown to be effective for learning manipulation tasks on real robots in challenging high dimensional settings Finn et al. (2016); Singh et al. (2019); Zhu et al. (2020). Unlike the majority of these works, which study single task reinforcement learning problems in a single fixed MDP, the focus of this work is in learning generalizable *multitask* 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.

A.2 Robotic Learning from Human Videos

In addition to the works on robot learning from human videos that were mentioned in Section 2, there are other works that have studied learning robotic behavior from human videos. One approach to this problem is to explicitly perform some form of object or hand tracking in human videos, which can then be translated into a sequence of robotic actions or motion primitives for task execution Lee et al. (2013); Yang et al. (2015); Nguyen et al. (2018); Lee and Ryoo (2017); Rothfuss et al. (2018). Unlike these works which hand design the mapping from a human sequence to robot behaviors, we aim to implicitly 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 Liu et al. (2018); Sharma et al. (2019) or unpaired Smith et al. (2020) data. Other works attempt to infer actions, rewards, or state-values of human videos and use them for learning predictive models Schmeckpeper et al. (2020b) or RL Schmeckpeper et al. (2020a); Edwards and Isbell (2019). Learning keypoint Xiong et al. (2021); Das et al. (2021) or object/task centric representations from videos Sermanet et al. (2018); Scalise et al. (2019); Pirk et al. (2019) is another promising strategy to learning rewards and representations between domains. Simulation has also been leveraged as supervision to learn such representations Petrík et al. (2020) or to produce human data using domain randomization Bonardi et al. (2020). Finally, meta-learing Yu

et al. (2018) and subtask discovery Sermanet et al. (2017); Goo and Niekum (2019) 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 the setting of "in-the-wild" human videos, specifically large and diverse sets of crowd-sourced videos from the real world, and contain many different individuals, viewpoints, backgrounds, objects, and tasks.

A.3 Robotic Learning from Large Datasets

Much like our work, a number of prior works have studied the problem of general purpose robotic agents, and learning from large and diverse data as a means to accomplishing this goal Finn and Levine (2017); Pinto and Gupta (2016); Zeng et al. (2018); Ebert et al. (2018); Gupta et al. (2018); Kalashnikov et al. (2018); Dasari et al. (2019); Cabi et al. (2019). These works have largely studied the problem of collecting large and diverse robotic datasets in scalable ways Mandlekar et al. (2018); Gupta et al. (2018); Dasari et al. (2019); Young et al. (2020); Chen et al. (2021) as well as techniques for learning general purpose policies from this style of data in an offline Ebert et al. (2018); Cabi et al. (2019) or online Pinto and Gupta (2016); Nair et al. (2018); Kalashnikov et al. (2018) fashion. While our motivation of achieving generalization through 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 easier to collect human data sources for accomplishing the same goal.

B Additional Method Details

B.1 Domain-agnostic Video Discriminator (DVD) Pseudocode

We provide pseudocode of our method in Algorithm 1.

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 similarity discriminator f_{sim}
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 f_{sim} with d_i, d_i^*, d_j according to Eq. 2
// Planning Conditioned on Video Demo
Require: Pre-trained video encoder f_{enc} & video prediction model p
Require: Trained similarity discriminator f_{sim}
Require: Human video demo d_i for task $\mathcal{T}_i$
for trials $1, ..., n$ **do**
    Sample $\{a^{1:G}_{1:H}\}$ & get predictions $\{\tilde{s}^g_{1:H}\} \sim \{p_\phi(s_0, a^g_{1:H})\}$
    Calculate $\mathcal{R}^g_\theta = f_{sim}(f_{enc}(\tilde{s}^g_{1:H}), f_{enc}(d_i); \theta)$
    Step $a^*_{1:H}$ which maximizes corresponding $\mathcal{R}^g$

B.2 Architecture Details

In Figure 7, we detail the model architecture used for DVD.

C Training Details

C.1 Dataset Details

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 1) Closing sth, 2) Moving sth away from camera, 3) Moving sth towards camera, 4) Opening sth, 5) Push left to right, 6) Push 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.

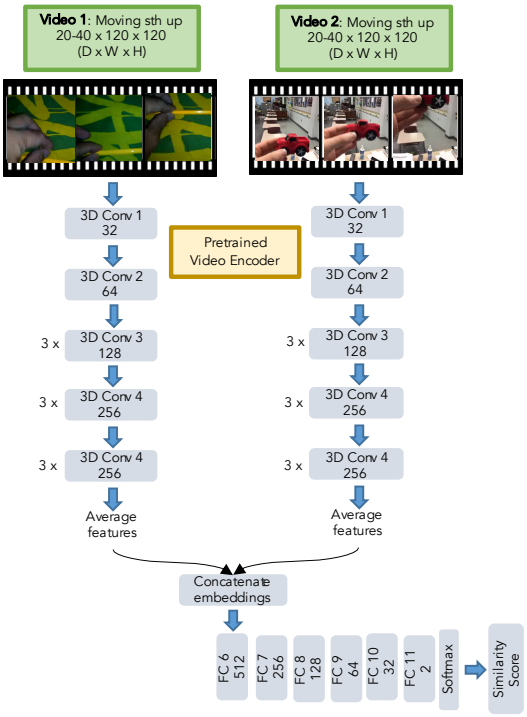

Figure 7: **DVD Architecture.** We use the same video encoder architecture as Shao et al. (2020). For each 3D convolution layer, the number of filters is denoted, and all kernels are $3 \times 3 \times 3$ except for the first, which is $3 \times 5 \times 5$. All conv layers have stride 1 in the temporal 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.

C.2 Experimental Details

Domains: For the simulation domains, we use a Mujoco simulation built off the Meta-World environments Yu et al. (2020). 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.

D Additional Experimental Results

D.1 Qualitative Results

In Figures 8 and 9, we provide examples of predicted trajectories and their similarity scores with a human video demonstration given by DVD. We see that DVD highly ranks trajectories that are completing the same task as demonstrated in the given human video.

D.2 Ablation on Amount of Robot Data for Training

In our main experiments, we use 120 robot demonstrations per task in the simulated environments and 80 per task in the real robot environment. While this is a manageable number of robot demonstrations, it would be better to rely on fewer demonstrations. Hence, to better understand the role of robot demonstrations for DVD, we ablate on the number of robot demonstrations used during training and evaluate in the environment generalization setting.

In Figure 10, we see that the performance of DVD decreases by only a small margin when using **as few as 20 robot demonstrations** per task. By leveraging the diversity in the human data, DVD

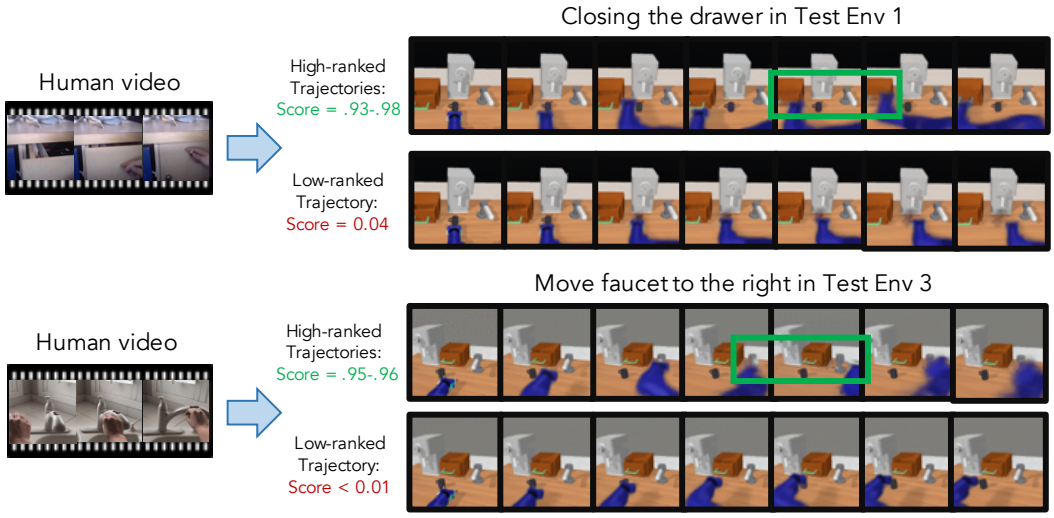

Figure 8: **Example Rankings During Planning.** Examples of predicted trajectories that are ranked high and low for the tasks of closing the drawer in test env 1 and moving the faucet to the right in the test env 3. DVD gives high similarity scores to trajectories that complete the same task specified by 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.

Figure 9: **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.

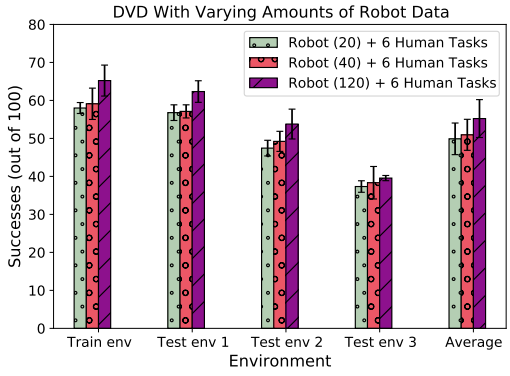

Figure 10: **Ablation on Amount of Robot Data Used for Training.** While using 120 robot demonstrations per task slightly benefits performance over using only 20 or 40, DVD still performs comparably with fewer robot demos.

is still able to complete tasks in challenging, unseen environments with little reliance on training environment robot data, which may allow more scalable training of tasks.