Action-Free Reasoning for Policy Generalization

Jaden Clark, Suvir Mirchandani, Dorsa Sadigh, Suneel Belkhale
Department of Computer Science
Stanford University United States
{jvclark, smirchan, dorsa, belkhale}@stanford.edu

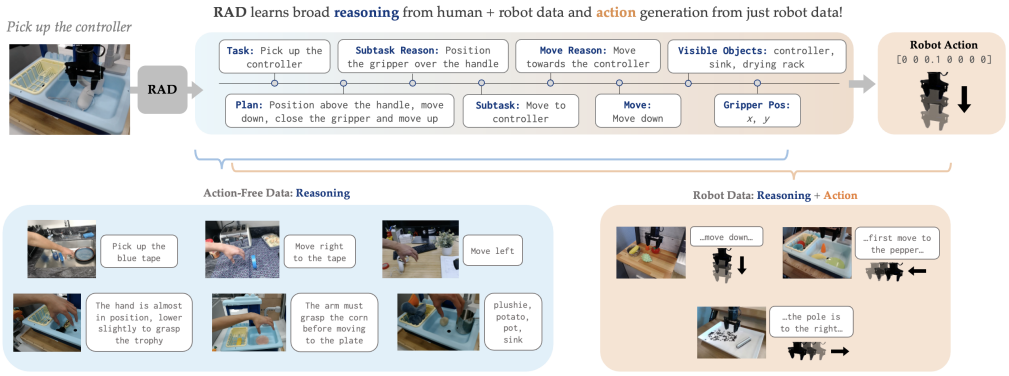


Figure 1: RAD learns from human and robot data through reasoning. RAD learns how to reason through high level task plans, subtasks, and movements from human data, and how to map reasonings to action from robot data. Thus, RAD generalizes to tasks unseen in human and robot data.

Abstract: End-to-end imitation learning offers a promising approach for training robot policies. However, generalizing to new settings—such as unseen scenes, tasks, and object instances—remains a challenge. Although large-scale robot demonstration datasets have shown potential for inducing generalization, they are resource-intensive to scale. In contrast, human video data is abundant and diverse, but human videos lack action labels, complicating their use in imitation learning. Existing methods attempt to extract grounded action representations (e.g., hand poses), but resulting policies struggle to bridge the embodiment gap between human and robot actions. We propose an alternative approach: leveraging language-based reasoning from human videos to train generalizable robot policies. Our method, Reasoning through Action-free Data (RAD), learns from both robot demonstration data (with reasoning and action labels) and action-free human video data (with only reasoning labels). The robot data teaches the model to map reasoning to low-level actions, while the action-free data enhances reasoning capabilities. Our experiments demonstrate that RAD enables effective transfer across the embodiment gap, allowing robots to perform tasks seen only in action-free data (+30% success). Furthermore, scaling up action-free reasoning data significantly improves policy performance and generalization to novel tasks (+25% success). Additionally, we are releasing a dataset of 3,377 human-hand demonstrations compatible with the Bridge V2 benchmark, including chain-of-thought reasoning annotations and hand-tracking data to help facilitate future work. See website with videos: https://rad-generalization.github.io.

Keywords: Imitation Learning, Embodied Reasoning, Human Videos

1 Introduction

Training visuomotor policies via imitation learning is a promising paradigm for robot control. However, current end-to-end learning methods struggle to generalize to new settings beyond their training

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

data, such as new scenes, new task instructions, and new object instances. For example, a robot that learns to pick up a video game controller in a lab setting should be able to generalize if it encounters the same controller on a couch in a home setting. Generalizing to novel scenarios is essential for deploying learning-based policies in the diverse and unpredictable scenarios in the real world.

One approach to achieving generalizable policies is to collect diverse large-scale robot demonstration data to train expressive multi-task policies [1, 2, 3, 4]. while there are promising signs of scaling up datasets being the solution, we have yet to reach the scale needed for comprehensive generalization, and collecting data at even larger scales is even more expensive.

On the other hand, many see tapping into human video datasets, consisting of humans directly performing tasks as opposed to collecting robot data, as the answer [5, 6, 7]. This data is cheap to collect and already present at scale in Internet datasets. However, human videos lack action labels, making supervised learning methods like imitation learning very difficult. Some works tackle this challenge by extracting *grounded action-like* representations from video as action labels, for example hand poses or object affordances [8, 9, 10, 11]. However, extracting grounded actions from human videos often makes assumptions about the scene and the embodiment gap (e.g., how the hand pose maps to the robot action or relying on paired human and robot data) which can limit their practicality at scale.

Instead of extracting grounded actions from videos and the restrictive assumptions that come with it, we ask: is there any other *behavioral information*—representations that directly influence robot actions—that we can extract from human videos? Our insight is that human videos contain vast amounts of *higher-level reasoning* that guide robot action prediction. For example, if the task is to pick up a cup, a human might reason about moving the hand towards the cup, then grasping the cup, and then lifting the cup. Prior works have shown the generalization benefits of this style of language reasoning, however they often learn reasoning from just robot demonstrations [12, 13]: our key idea is to instead extract such reasoning from *action-free* human videos—significantly scaling up data that informs robot actions.

We introduce our method, Reasoning through Action-free Data (RAD), a robot policy that leverages reasoning traces extracted from action-free data. RAD trains a large transformer model on a mixture of robot demonstration data with both reasoning and robot action labels, and action-free (human video) data labeled with *just* reasoning. The robot data teaches the model to autoregressively go from reasoning to low-level actions, while the action-free data augments the reasoning capabilities of the model. We label reasoning traces by leveraging pretrained vision-language models such as Gemini [14] with hindsight knowledge as done in prior work [12].

We experimentally validate that learning from action-free reasoning data transfers well across the embodiment gap—showing 20% better performance on tasks only seen in the action-free data over models not finetuned with RAD. Additionally, we find that action-free reasoning data improves the capacity of RAD to generalize to tasks that have never been seen in both robot *and* human data, with RAD outperforming baselines by 15%. Finally, we trained a RAD model on a portion of the Something-Something V2 dataset [15], and found that this model was able to to learn new skills beyond those exhibited in the vanilla RAD model (+15% success rate).

2 Related Work

In this section, we situate our work among prior work on the use of language as a representation of low-level actions in robot learning, vision-language-action models (VLAs) as a recipe for language-conditioned robot policies, and approaches that leverage human videos for robot learning.

Language as an Action Representation. Language is commonly used as a high-level representation in imitation learning, either for conditioning multi-task policies on specific instructions [16, 17, 18, 19, 3], or as a way to decompose high-level, long-horizon instructions into lower-level subtask instructions [20, 21, 22]. More recently, several works have studied the role of more fine-grained language such as "language motions" as intermediate representations to predict [13] or

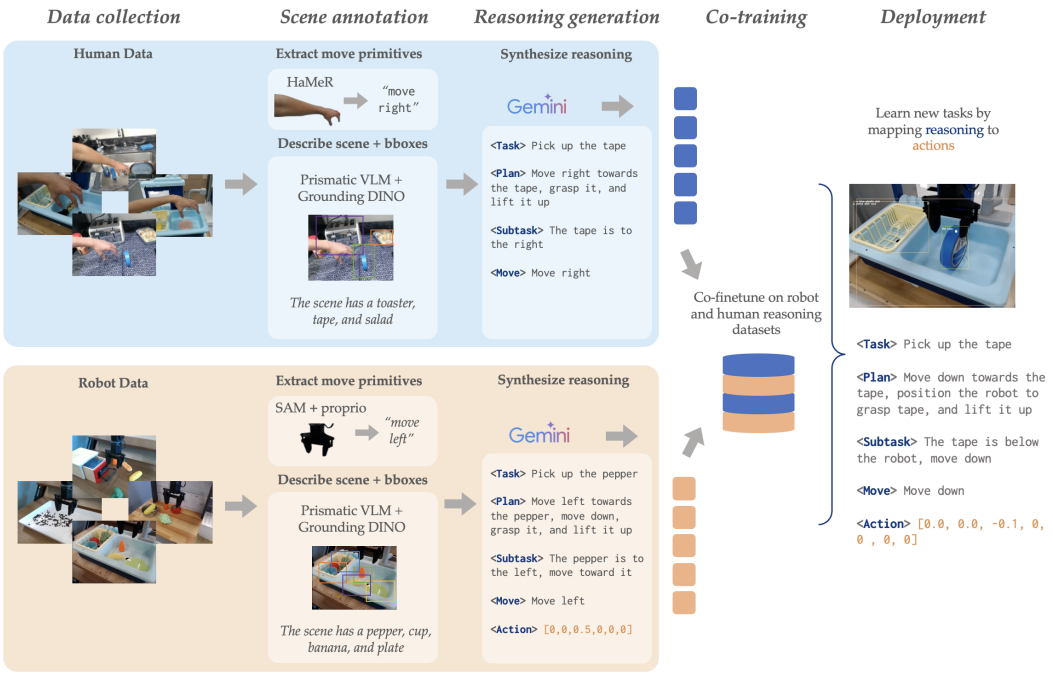


Figure 2: RAD generates reasonings on both human and robot data using a suite of pretrained models. For both human and robot data, scene descriptors and object bounding boxes are generated using Prismatic VLM and Grounding DINO. For robot data, SAM and proprioception can be used to generate movement primitives; for human data, RAD relies on HaMeR to track human hands for primitive generation. For both data types, the scene descriptions, bounding boxes, and movement primitives (and actions for robot data) are synthesized by Gemini into a language reasoning. Reasonings are tokenized and fed into a dataset containing both human and robot data for co-finetuning.

explicitly reason over language as well as other visually-grounded features such as bounding boxes as a way of guiding large pretrained policies [12]. In contrast to prior works which use language as a goal representation, we explore how reasoning in language can be used as an action representation for human video data in addition to robot data.

Vision Language Action Models. Recent works have explored the use of pre-trained Vision-Language Models (VLMs) as backbones for Vision-Language Action Models (VLAs) which directly predict low-level robot actions. For example, RT-2-X [2] fine-tunes the 55B-parameter PaLI-X VLM [23] on the Open-X Embodiment dataset [2], and OpenVLA [3] uses a 7B-parameter Llama 2 LLM backbone with a vision encoder based on DINOv2 [24] and SigLIP [25]. The promise of VLAs for manipulation is to build off of generalization of VLMs which have been trained on Internet-scale vision-language data. An additional way to achieve transfer of VLM capabilities to VLAs is to take advantage of their textual reasoning abilities. For example, Embodied Chain of Thought (ECoT) uses multiple steps of reasoning prior to predicting robot actions by training on synthetic reasoning data [12].

Learning from Human Video. A large number of prior works in imitation learning for robotics focus on learning from demonstrations collected via teleoperation by expert operators. This method of collecting data is costly, so a number of prior works have investigated ways to leverage existing data sources of human videos to improve robot policy learning — for example, by pre-training visual representations [26, 27, 28] or learning reward functions [29, 30, 31]. However, bridging the gap between human videos and robot actions can be challenging due to embodiment differences and diversity in videos. Several works learn priors from human video datasets and/or in-domain human videos [32, 33, 6, 10] or aligning paired/unpaired examples of human videos and robot demonstration videos [34, 35, 36, 37] or simulations [38]. These works are still fundamentally limited by the

quantity of robot demonstrations. Another line of work leverages intermediate representations for predicting robot actions downstream, but make assumptions about the human hand behavior, which is not necessarily the same as the robot [39, 8]. Our work goes beyond existing methods that rely on generating intermediate representations for action predictions by generating detailed reasoning steps about human video demonstrations.

3 Reasoning through Action-free Data

In this section, we will first describe our problem setting and lay out our assumptions, and then we will outline our method for learning from action-free data using language reasoning chains. As an overview, RAD involves two major steps. First, annotate action-free data with language reasoning (Section 3.3). Second, train a reasoning-based policy on a combination of robot demonstration data with both actions and reasoning chains and action-free data with only reasoning chains (Section 3.4).

3.1 Problem: Learning Reasoning in Action-free Data

In multi-task imitation learning, we are given a dataset $\mathcal{D} = \{(o_i, a_i, g_i)\}_{i=1}^N$ of observations $o \in \mathcal{O}$, actions $a \in \mathcal{A}$, and task specifications $g \in \mathcal{G}$ (e.g., natural language goals). The objective is to learn the expert policy $P(a \mid o, g)$.

In the *reasoning-based* multi-task imitation learning setting, we assume actions are mediated by a chain of C intermediate reasoning steps $(l^1, \ldots, l^C)$, where each l^j is a language description that depends on (o, g) and previous reasoning steps $(l^1, \ldots, l^{j-1})$. The final action a depends on the full reasoning chain and (o, g). Our goal is to learn the joint distribution: $P(a, l^1, \ldots, l^C \mid o, g)$. We parameterize this with model P_θ and maximize the log-likelihood over the dataset $\mathcal{D}$:

$$
\begin{aligned}
L(\theta) &= \sum_{i=1}^{N} \log P_\theta(a_i, l_i^1, \ldots, l_i^C \mid o_i, g_i) \\
&= \sum_{i=1}^{N} \left[\log P_\theta(a_i \mid l_i^1, \ldots, l_i^C, o_i, g_i) + \sum_{j=1}^{C} \log P_\theta(l_i^j \mid l_i^{<j}, o_i, g_i) \right]
\end{aligned}
$$

where $l_i^{<j}$ denotes $(l_i^1, \ldots, l_i^{j-1})$. We refer to the two terms as $L_{\texttt{action}}(\theta)$ and $L_{\texttt{reasoning}}(\theta)$.

Our key insight in RAD is that *action-free* datasets—such as human videos—can supervise learning the reasoning component ($L_{\texttt{reasoning}}(\theta)$). Specifically, we assume access to a dataset $\tilde{\mathcal{D}} = \{(\tilde{o}_i, \tilde{g}_i, \tilde{l}_i^1, \ldots, \tilde{l}_i^{C_i})\}_{i=1}^M$, where each sample provides a partial reasoning chain of length $C_i \geq 1$. Each sample might have different C_i, differing based on annotation confidence levels or quality.

To incorporate this action-free data, we optimize an auxiliary reasoning objective:

$$
\tilde{L}_{\texttt{reasoning}}(\theta) = \sum_{i=1}^{M} \sum_{j=1}^{C_i} \log P_\theta(\tilde{l}_i^j \mid \tilde{l}_i^{<j}, \tilde{o}_i, \tilde{g}_i)
$$

By training on both $\mathcal{D}$ and $\tilde{\mathcal{D}}$, we aim to improve the reasoning component of the policy, enabling generalization to new tasks from action-free data.

3.2 Reasoning Steps in RAD

While this setup can in principle work with different formulations of language reasoning steps, we instantiate our algorithm with the following reasoning steps from prior work [12]: `TaskPlan` (l^1), `SubtaskReasoning` (l^2), `Subtask` (l^3), `MoveReasoning` (l^4), `MovePrimitive` (l^5), `GripperPosition` (l^6), `VisibleObjects` (l^7), and finally the action itself (see Section 5.1 for detailed breakdowns of each reasoning step.

These reasoning steps trace through information at an increasing amount of physical and spatial groundedness—beginning with high-level scene reasoning over tasks and subtasks, transitioning to reasoning over language motions, followed by spatial information about the gripper and objects, and

concluding with the low-level robot action. We take advantage of this fact in designing a pipeline to label reasoning in action-free data, as we describe in the following section.

3.3 Labeling Reasoning in Action-free Data

In order to construct $\tilde{D}$—our dataset of observations, goals and action-free reasoning—we need to generate labels for the reasoning steps above from human videos. Our pipeline is similar to the automated procedure used by Embodied Chain-of-Thought (ECoT) [12] for generating reasoning over robot demonstrations, with some key modifications to handle human videos. To obtain reasoning labels for robot demonstrations, ECoT first generates `GripperPositions` and `VisibleObjects` tags using off-the-shelf object detectors to obtain bounding boxes. Then, it extracts `MovePrimitive` (e.g. "move to the left") directly from actions using an automated heuristic. Conditioned on these more grounded reasoning steps (l^5, l^6, l^7) and the image observation o, it queries Gemini [14] to label the prior reasoning steps, from `TaskPlan` through `MoveReasoning` $(l^1, \ldots, l^4)$.

In the action-free setting with human videos, we can still extract high-level reasoning with Gemini, as well as extract `VisibleObjects` with off-the-shelf object detectors. However, generating the more action-grounded reasoning steps is challenging: we can no longer extract `MovePrimitives` or `GripperPositions` automatically because we lack explicit action labels. In order to overcome this, we extract the `MovePrimitives` and `GripperPositions` using HaMeR [40], a hand keypoint and pose tracking method. Given these predictions, we can extract the `MovePrimitives` from changes in the hand pose information: first, we study each axis of the change in hand poses for each frame; then, we label the move primitive based on the dominant axis of motion. In this work we focus on tracking gripper and positional movement primitives, but also show tracking rotational movement is feasible with RAD in 5.3. We outline this labeling procedure in Fig. 2.

3.4 Training on Partial Reasoning Chains

To train on mixtures of demonstration and action-free data, we use the ECoT and OpenVLA [12, 3] architecture, which trains a 7B parameter VLM transformer – pretrained on Internet-scale vision-language tasks – to predict sequences of language reasoning and then action tokens. In RAD, we reuse this paradigm for the robot demonstration data, but for the new action-free data, our "labels" for training contain only reasoning as described in Section 3.3.

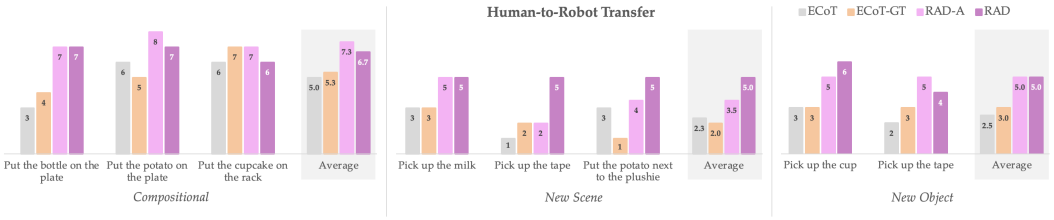


Figure 3: RAD outperforms baselines where human video data was trained on, but no new robot data was provided. RAD-A is RAD trained only on human video data for the given axis of generalization. ECoT-GT is finetuned on the same data as RAD, but only using human hand locations (and not the full reasoning data).

4 Experiments

In this section, we evaluate how RAD enables transfer from human videos to robot policies and generalization beyond settings in the human videos or robot demonstration data. Specifically, we seek to answer the following questions:

Q1 – Human-to-Robot Transfer: Can RAD enable learning new tasks seen only in the human video data and not the robot demonstration data?

Q2 – Reasoning Generalization: Does reasoning in RAD enable generalization to novel tasks beyond both the robot demonstration data and human video data it was trained on?

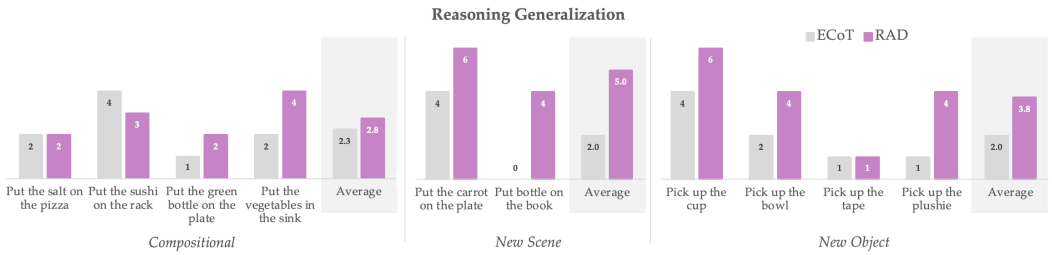


Figure 4: RAD compared to ECoT for tasks contained in neither human or robot data. RAD shows improved performance across all three axes of generalization.

Q3 – Cross-Environment Transfer: Can RAD learn new tasks from human video data in out-of-domain environments?

4.1 Evaluating Generalization

Next, we discuss the environments, tasks, and model baselines we use to evaluate the reasoning generalization capabilities of RAD.

Real-World Environments: We use a 6-DoF WidowX robot arm for our experiments. We perform all evaluations in Section 4.2 and Section 4.3 on the Toy Sink setup from [41], to ensure fair comparison with existing pre-trained models. All human video data for Section 4.2 and Section 4.3 was also collected in the Toy Sink setup (1616 demonstration videos), using both the standard Bridge V2 camera setup, as well as an additional camera for better hand tracking. Notably, the Bridge V2 setup is comprised of mostly miniature toy replicas of real world objects such as small kitchen supplies, blocks, and home supplies. Therefore, we also seek to assess how RAD responds to data from real-world human environments, and learns to interact with realistically sized objects. We thus collect data in two additional environments: a plain tabletop and a cluttered desk, as well as various real home and kitchen environments. This data was used to assess how RAD responds to data from unstructured environments in Section 4.4.

Generalization Tasks: We evaluate RAD across a variety of generalization tasks. These tasks comprise three main axes of generalization:

1. **Compositional Generalization**: In this axis, the objects, tasks, and scenes are all seen in pre-training data (Bridge V2 data), but not in those particular configurations. For example, pizza and salt both exist in Bridge V2, but salt is never placed on the pizza.
2. **New Object Generalization**: This axis introduces unseen objects for known behaviors (e.g., *pick cup* $\rightarrow$ *pick plushie*).
3. **New Scene Generalization**: This axis requires generalizing to novel backgrounds and distractor objects for seen tasks; for example, picking up a known object with a pot in the background.

Note that the Compositional Generalization axis tests the model's ability to *interpolate* the training data, while New Object and New Scene axes test the model's ability to *extrapolate* from the training data. Exact tasks for each axis can be found in Section 5.3.

Methods: To test the efficacy of reasoning in learning from human video data, we evaluate the following models in our generalization scenarios. **Embodied Chain-of-Thought (ECoT)** [12], is a state-of-the-art action reasoning model trained on Bridge V2, but without any human video data. **ECoT w/ Gripper Tracking (ECoT-GT):** is ECoT finetuned on the same human video data as RAD, but only generates the `GripperPosition` portion of the reasoning chain. This is analogous to how prior work learns from extracted pose information only in human videos, but does not extract higher level language reasoning [39, 10, 9]. **RAD (Ours)** is ECoT finetuned on the full chain of reasonings generated from human video data. **RAD-A (Ours)** is the same as RAD, but trained on only human videos from one *axis* of generalization at a time (the axes are described in Section 4.1).

4.2 Can RAD enable transfer from human-to-robot embodiments?

First, we assess if RAD can learn accurate reasonings and robot actions on new tasks that are present only in human video demonstrations. We train the axis-specific models (RAD-A) only on human video data for that axis (8-12 tasks with a total of 320-500 videos per axis). We evaluate these axis-specific models against zero-shot ECoT, as well as RAD (trained on human video data from all three axes) and ECoT-GT models trained on our full human video dataset.

In Fig. 3, we find that despite having no new robot demonstration data for these new tasks, RAD-A achieves consistently higher success rates than zero-shot ECoT and ECoT-GT across all areas of generalization (**Q1**).

Compositional: On compositionally new tasks, RAD-A outperforms ECoT by 23% and ECoT-GT by 20%. RAD outperforms ECoT and ECoT-GT by 17% and 13% respectively. Qualitatively, RAD models demonstrates significantly better reasoning capability, particularly in the second step of pick place tasks (such as placing the object of interest in the desired location).

New Object: On tasks with new objects, RAD and RAD-A both improves on ECoT and ECoT-GT by 25% and 20%, respectively. RAD models demonstrate substantially better ability to reason about grasp points on new objects, such as moving towards the sides of large cups instead of the middle.

New Scene: RAD models also substantially outperform baselines on novel scenes (containing distractors and other scene modifications). RAD-A outperforms ECoT by 12% and ECoT-GT by 15%. The full RAD model had stronger performance, outperforming ECoT by 27% and ECoT-GT by 30% - potentially due to improved ability to ignore distractors as a result of training on a larger more diverse dataset. Reasoning traces on RAD models also appeared to be more accurate, with ECoT often becoming distracted and generating non-sensical reasonings. These results indicate that augmenting chain-of-thought models with reasoning from human video data improves their ability to reason and infer robot actions on previously unseen task configurations.

4.3 Can RAD train more generalizable policies?

Ultimately, training on large datasets of human video data should enable VLAs to generalize not only to human demonstrated tasks, but also to completely unseen scenarios. To explore if RAD enables training more general models, we evaluate our model against ECoT on 10 novel tasks (unseen in both human and robot data) comprising all three generalization axes. Results are presented in Fig. 4.

Compositional: On novel compositional tasks, RAD outperforms ECoT by 5%. RAD reasoned better than ECoT over multi-step tasks, such as knowing where to place the salt after picking it up.

New Object: RAD substantially improves performance on tasks with unseen objects, such as bowls and large cups, despite not seeing such objects in human or robot training data. RAD achieves 30% higher success compared to ECoT.

New Scene: In novel scenes (environments with large distractors in the scene, such as cloth, pots, and a large plushie), RAD reached 18% higher success rate than ECoT. Qualitatively, ECoT struggled to reason about the new scene and would often generate poor reasonings and execute seemingly random actions, whereas RAD generated correct reasoning which informed action prediction.

This indicates that reasoning in RAD enables better generalization to a variety of unseen tasks, without training on any new human or robot data (**Q2**).

4.4 Can RAD leverage data from new environments?

The previous experiments demonstrate strong generalization from human data collected in the same environment – but to truly leverage large-scale video data, generalist robot policies must learn from demonstrations in diverse scenes. Thus, we first train RAD with small-scale human video data in unseen environments to see how well it can incorporate this data, and we study its scaling properties

Table 1: Cross-Environment and Internet-Data Results

(a) Cross-Environment Transfer

Task	ECoT	RAD	ECoT-GT
pick up the cup	3/10	6/10	4/10
put the sushi on the book	4.5/10	6.5/10	5/10
pick up the tiger	3/10	3/10	3/10
pick up the controller	2/10	3.5/10	2/10

(b) Internet Data Transfer

Task	ECoT	RAD	RAD-SS
Hold the bowl with the markers on it	1/10	2.5/10	3.5/10
Cover the car key with the cloth	2/10	1.5/10	3.5/10
Lift the computer charger	1/10	3/10	5/10

compared to data from the same environment. Then, we study the performance of RAD with larger scale in-the-wild human video data.

Human Videos from New Environment: We first collect human data for two unseen tasks in a brand new tabletop setup (unseen in Bridge V2 data). Then, we evaluate models trained on this new enviroment data in the original Bridge Toy Sink environment. In Table 1a, we see that models trained on this data outperform ECoT by 16% and ECoT-GT by 13%. Similarly to Section 4.2 and Section 4.3, RAD models reasoned better about grasp points (e.g. where to pick up the controller) despite the data being in a different environment (**Q3**).

In-distribution vs. Out-of-Distribution Human Data: Next, we assess how RAD performance scales with increased data for the same tasks collected in-distribution (in the miniature Toy Sink setup) versus out-of-distribution (various real world kitchen and office environments). To do so, we collected 100 additional demos for the *pick up the tape* task in the Toy Sink setup. We also collected 250 out-of-domain demos for *pick up the tape* in novel environments such as real kitchens, countertops, and desks. Then, we trained RAD on two different data mixtures: (1) The original RAD data mix (which already contains 40 *pick up the tape* demos) + in-distribution data and (2) the original RAD data mix + out-of-domain data.

Results for both mixtures are shown in Table 2. We find that RAD models trained on both in-domain (+30% success) and out-of-domain data (+25% success) show improved performance over the original model (**Q3**). Qualitatively, RAD models were better able to reason about when to bring the gripper to the level of the tape, with ECoT models often moving too low and knocking over the tape, which is taller than objects in Bridge V2.

Table 2: Data Scaling

Data	Model	Success Rate
Original Model (40 Demos)	ECoT	2/10
	ECoT-GT	3/10
	RAD	4/10
	RAD-A	5/10
Same Environment (+100 ID Demos)	RAD	7/10
	ECoT-GT	4/10
New Environments (+250 OOD Demos)	RAD	6.5/10
	ECoT-GT	5/10

Leveraging In-the-Wild Data: Finally, we labeled reasoning for 31,656 videos from the Something-Something V2 dataset as described in Section 5.1. We trained RAD on this data mixed with Bridge V2 and our original mixture, and we call this model RAD-SS. We evaluated this model on 3 out-of-distribution tasks containing real world objects like a car key and computer charger as well as new task descriptions such as "lift" or "cover". We found RAD-SS outperformed both the vanilla RAD model (+15%) and ECoT (+26.7%) on these tasks as seen in Table 1b. Qualitatively, RAD-SS showed improved reasoning on semantically new tasks like "cover" or "hold" in addition to robustness to distractors.

5 Discussion

In this work we present RAD, a new way to train generalist robot policies from human video data. RAD learns to predict *reasoning*, which can be labeled on both robot and human video data. We find that RAD enables VLAs to cross the embodiment gap, and to learn tasks represented in only human video data. Models trained with RAD are also able to generalize to completely unseen tasks (not present in either robot or human data). Finally, we find RAD responds positively to data from out-of-domain environments and even in-the-wild datasets, enabling models to learn new tasks from environments completely separate from the target domain. These results demonstrate that RAD is a promising step towards training generalist robot policies, laying the groundwork for models that can leverage both robot data and large-scale human video data.

Limitations: Our work demonstrates the promise of using human video data to improve generalization in robot policies; however, there are key challenges to address before scaling up the method to fully tap into larger and noisier datasets of human videos, such as those found on the Internet. As human hand pose estimation methods become more accurate, we anticipate that this limitation will be partially mitigated and allow us to better leverage more natural videos of human hands, as well as to expand the set of language motions in our labeling pipeline. Additionally, we scope our work to focus our study of generalization on pick-and-place tasks with rigid objects, characteristic of the tasks in prior work on reasoning-based imitation learning [12]. Expanding the set of tasks to include more fine-grained and dexterous manipulation provides a rich area for future work.

References

[1] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

[2] O. X.-E. Collaboration, A. O'Neill, A. Rehman, A. Gupta, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, A. Tung, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Gupta, A. Wang, A. Kolobov, A. Singh, A. Garg, A. Kembhavi, A. Xie, A. Brohan, A. Raffin, A. Sharma, A. Yavary, A. Jain, A. Balakrishna, A. Wahid, B. Burgess-Limerick, B. Kim, B. Schölkopf, B. Wulfe, B. Ichter, C. Lu, C. Xu, C. Le, C. Finn, C. Wang, C. Xu, C. Chi, C. Huang, C. Chan, C. Agia, C. Pan, C. Fu, C. Devin, D. Xu, D. Morton, D. Driess, D. Chen, D. Pathak, D. Shah, D. Büchler, D. Jayaraman, D. Kalashnikov, D. Sadigh, E. Johns, E. Foster, F. Liu, F. Ceola, F. Xia, F. Zhao, F. V. Frujeri, F. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Yang, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, H. Bharadhwaj, H. Walke, H. Fang, H. Ha, I. Mordatch, I. Radosavovic, I. Leal, J. Liang, J. Abou-Chakra, J. Kim, J. Drake, J. Peters, J. Schneider, J. Hsu, J. Vakil, J. Bohg, J. Bingham, J. Wu, J. Gao, J. Hu, J. Wu, J. Wu, J. Sun, J. Luo, J. Gu, J. Tan, J. Oh, J. Wu, J. Lu, J. Yang, J. Malik, J. Silvério, J. Hejna, J. Booher, J. Tompson, J. Yang, J. Salvador, J. J. Lim, J. Han, K. Wang, K. Rao, K. Pertsch, K. Hausman, K. Go, K. Gopalakrishnan, K. Goldberg, K. Byrne, K. Oslund, K. Kawaharazuka, K. Black, K. Lin, K. Zhang, K. Ehsani, K. Lekkala, K. Ellis, K. Rana, K. Srinivasan, K. Fang, K. P. Singh, K.-H. Zeng, K. Hatch, K. Hsu, L. Itti, L. Y. Chen, L. Pinto, L. Fei-Fei, L. Tan, L. J. Fan, L. Ott, L. Lee, L. Weihs, M. Chen, M. Lepert, M. Memmel, M. Tomizuka, M. Itkina, M. G. Castro, M. Spero, M. Du, M. Ahn, M. C. Yip, M. Zhang, M. Ding, M. Heo, M. K. Srirama, M. Sharma, M. J. Kim, N. Kanazawa, N. Hansen, N. Heess, N. J. Joshi, N. Suenderhauf, N. Liu, N. D. Palo, N. M. M. Shafiullah, O. Mees, O. Kroemer, O. Bastani, P. R. Sanketi, P. T. Miller, P. Yin, P. Wohlhart, P. Xu, P. D. Fagan, P. Mitrano, P. Sermanet, P. Abbeel, P. Sundaresan, Q. Chen, Q. Vuong, R. Rafailov, R. Tian, R. Doshi, R. Mart'in-Mart'in, R. Baijal, R. Scalise, R. Hendrix, R. Lin, R. Qian, R. Zhang, R. Mendonca, R. Shah, R. Hoque, R. Julian, S. Bustamante, S. Kirmani, S. Levine, S. Lin, S. Moore, S. Bahl, S. Dass, S. Sonawani, S. Tulsiani, S. Song, S. Xu, S. Haldar, S. Karamcheti, S. Adebola, S. Guist, S. Nasiriany, S. Schaal, S. Welker, S. Tian, S. Ramamoorthy, S. Dasari, S. Belkhale, S. Park, S. Nair, S. Mirchandani, T. Osa, T. Gupta, T. Harada, T. Matsushima, T. Xiao, T. Kollar, T. Yu, T. Ding, T. Davchev, T. Z. Zhao, T. Armstrong, T. Darrell, T. Chung, V. Jain, V. Kumar, V. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Pang, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, Y. Dou, Y. Cho, Y. Lee, Y. Cui, Y. Cao, Y.-H. Wu, Y. Tang, Y. Zhu, Y. Zhang, Y. Jiang, Y. Li, Y. Li, Y. Iwasawa, Y. Matsuo, Z. Ma, Z. Xu, Z. J. Cui, Z. Zhang, Z. Fu, and Z. Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. In *International Conference on Robotics and Automation (ICRA)*, 2024.

[3] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, et al. Openvla: An open-source vision-language-action model. *Conference on Robot Learning (CoRL)*, 2024.

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

[5] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

[6] C. Wang, L. Fan, J. Sun, R. Zhang, L. Fei-Fei, D. Xu, Y. Zhu, and A. Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning (CoRL)*, 2023.

[7] H. Bharadhwaj, D. Dwibedi, A. Gupta, S. Tulsiani, C. Doersch, T. Xiao, D. Shah, F. Xia, D. Sadigh, and S. Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

[8] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation. *arXiv preprint arXiv:2405.01527*, 2024.

[9] J. Ren, P. Sundaresan, D. Sadigh, S. Choudhury, and J. Bohg. Motion tracks: A unified representation for human-robot transfer in few-shot imitation learning. *arXiv preprint arXiv:2501.06994*, 2025.

[10] M. Lepert, R. Doshi, and J. Bohg. Shadow: Leveraging segmentation masks for cross-embodiment policy transfer. In *8th Annual Conference on Robot Learning*.

[11] M. Xu, Z. Xu, C. Chi, M. Veloso, and S. Song. Xskill: Cross embodiment skill discovery. In *Conference on Robot Learning*, pages 3536–3555. PMLR, 2023.

[12] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine. Robotic control via embodied chain-of-thought reasoning. In *Conference on Robot Learning (CoRL)*, 2024.

[13] S. Belkhale, T. Ding, T. Xiao, P. Sermanet, Q. Vuong, J. Tompson, Y. Chebotar, D. Dwibedi, and D. Sadigh. RT-H: Action hierarchies using language. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[14] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv*, 2023.

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

[16] S. Stepputtis, J. Campbell, M. Phielipp, S. Lee, C. Baral, and H. Ben Amor. Language-conditioned imitation learning for robot manipulation tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[17] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. BC-Z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning (CoRL)*, 2022.

[18] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, T. Jackson, S. Jesmonth, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, K.-H. Lee, S. Levine, Y. Lu, U. Malla, D. Manjunath, I. Mordatch, O. Nachum, C. Parada, J. Peralta, E. Perez, K. Pertsch, J. Quiambao, K. Rao, M. Ryoo, G. Salazar, P. Sanketi, K. Sayed, J. Singh, S. Sontakke, A. Stone, C. Tan, H. Tran, V. Vanhoucke, S. Vega, Q. Vuong, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. Rt-1: Robotics transformer for real-world control at scale. *arXiv*, 2022.

[19] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *arXiv*, 2023.

[20] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, et al. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning*, pages 287–318, 2023.

[21] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, 2023.

[22] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn. Yell at your robot: Improving on-the-fly from language corrections. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[23] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, B. Changpinyo, J. Wu, C. Riquelme, S. Goodman, X. Wang, Y. Tay, S. Shakeri, M. Dehghani, D. Salz, M. Lučić, M. Tschannen, A. Nagrani, H. F. Hu, M. Joshi, B. Pang, C. Montgomery, P. Pietrzyk, M. Ritter, A. Piergiovanni, M. Minderer, F. Pavetić, A. Waters, G. Li, I. Alabdulmohsin, L. Beyer, J. Amelot, K. Lee, A. Steiner, Y. Li, D. Keysers, A. Arnab, Y. Xu, K. Rong, A. Kolesnikov, M. Seyedhosseini, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut. Pali-x: On scaling up a multilingual vision and language model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[25] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[26] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

[27] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik. Masked visual pre-training for motor control. *arXiv*, 2022.

[28] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

[29] L. Shao, T. Migimatsu, Q. Zhang, K. Yang, and J. Bohg. Concept2Robot: Learning manipulation concepts from instructions and human demonstrations. *The International Journal of Robotics Research (IJRR)*, 2021.

[30] A. S. Chen, S. Nair, and C. Finn. Learning generalizable robotic reward functions from" in-the-wild" human videos. In *Proceedings of Robotics: Science and Systems (RSS)*, 2021.

[31] P. Mandikal and K. Grauman. DexVIP: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning (CoRL)*, 2022.

[32] K. Shaw, S. Bahl, and D. Pathak. Videodex: Learning dexterity from internet videos. In *Conference on Robot Learning (CoRL)*, 2023.

[33] S. Bahl, A. Gupta, and D. Pathak. Human-to-robot imitation in the wild. In *Proceedings of Robotics: Science and Systems (RSS)*, 2022.

[34] P. Sharma, D. Pathak, and A. Gupta. Third-person visual imitation learning via decoupled hierarchical controller. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[35] L. Smith, N. Dhawan, M. Zhang, P. Abbeel, and S. Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. In *Proceedings of Robotics: Science and Systems (RSS)*, 2019.

[36] H. Xiong, Q. Li, Y.-C. Chen, H. Bharadhwaj, S. Sinha, and A. Garg. Learning by watching: Physical imitation of manipulation skills from human videos. In *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[37] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, et al. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers. *arXiv preprint arXiv:2403.12943*, 2024.

[38] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, 2022.

[39] G. Papagiannis, N. Di Palo, P. Vitiello, and E. Johns. R+ x: Retrieval and execution from everyday human videos. *arXiv preprint arXiv:2407.12957*, 2024.

[40] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik. Reconstructing hands in 3D with transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[41] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

[42] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Appendix

We outline the dataset collection and reasoning generation procedure in Section 5.1. The models, training procedure, and baselines are described in detail in Section 5.2. Finally, Section 5.3 provides examples of results and description of reported success rates.

5.1 Dataset Details

Reasoning Steps: The exact reasoning steps we use in RAD follow prior work [12]:

- `TaskPlan` (l^1): describes a list of subtasks to achieve g.
- `SubtaskReasoning` (l^2): reasons about which subtask currently needs to be executed in the plan.
- `Subtask` (l^3): predicts the subtask that currently needs to be executed.
- `MoveReasoning` (l^4): reasons about the motion needed to achieve the subtask in the scene.
- `MovePrimitive` (l^5): predicts a movement primitive in language.
- `GripperPosition` (l^6): predicts the pixel position of the end-effector.
- `VisibleObjects` (l^7): predicts the bounding box coordinates of objects in the scene.
- `Action` (a): predicts the low-level robot action as an end-effector position delta.

Data Collection: Our main human video data collection was on the Bridge V2 Toy Sink setup. We aligned one camera based on the original Bridge V2 scene. We also set up a second camera from directly behind the WidowX gripper to better track hand movement as seen in Fig. 5. Example tasks are shown in Fig. 6. We used HaMeR to track the hand using the secondary camera perspective. We used the average location of the thumb tip and index finger tip points tracked by HaMeR as the gripper location. Based on the delta gripper position between frames, we characterized every frame as "stop", "move forward", "move backward", "move left", "move right", "move up", or "move down" movement primitives. We used the average distance between the thumb tip and index tip to determine "close gripper" and "open gripper" primitives. For reasoning generation on the human videos, we followed the the pipeline of [12], but used this HaMeR tracking in place of proprioception and SAM to generate movement primitives and gripper locations.

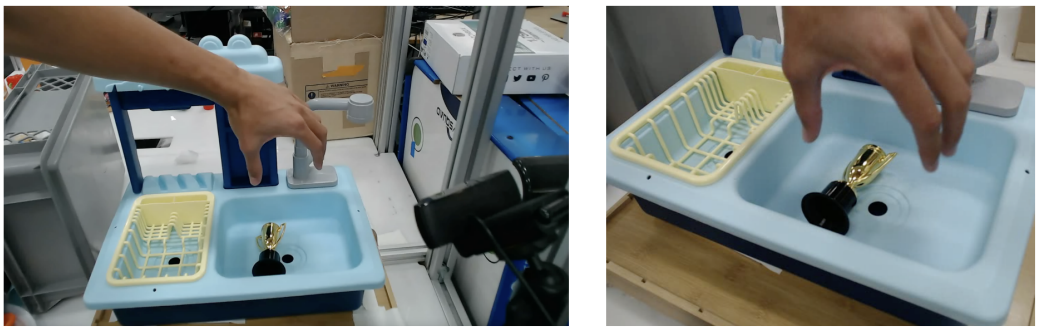
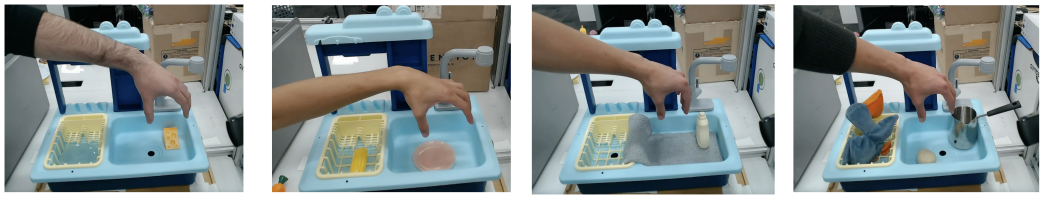


Figure 5: The main Bridge V2 perspective (right) versus the secondary perspective used for hand tracking (left).

Data Mixtures: For RAD-A models in Section 4.2 we collected 392 demonstrations for the compositional generalization dataset, 304 demonstrations for the new object dataset, and 280 demonstrations for the new scene dataset. The full RAD model as well as ECoT-GT model were both trained on all three of these datasets as well as 640 additional demos to make 1616 total demonstrations.



Put the cheese on the plate *Put the corn on the plate* *Put the bottle on the rack* *Put the potato on the plushie*

Figure 6: Example human video tasks collected.

Pick up the tiger *Put the sushi on the book* *Pick up the controller*

Figure 7: Task demonstrations collected in environments outside of Bridge V2 to assess how RAD responds to data from different types of scenes.

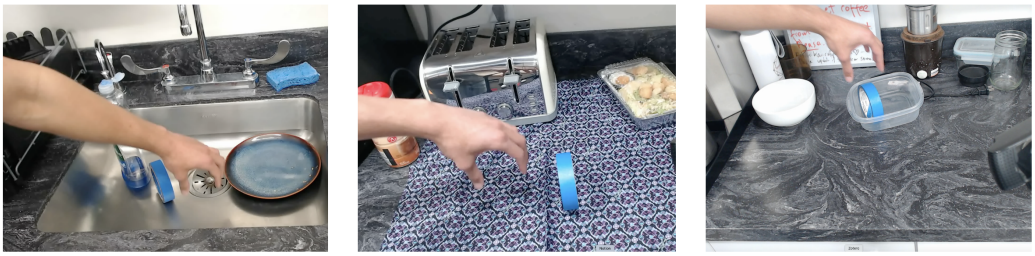


Figure 8: Real world environment data RAD is trained with for Section 4.4.

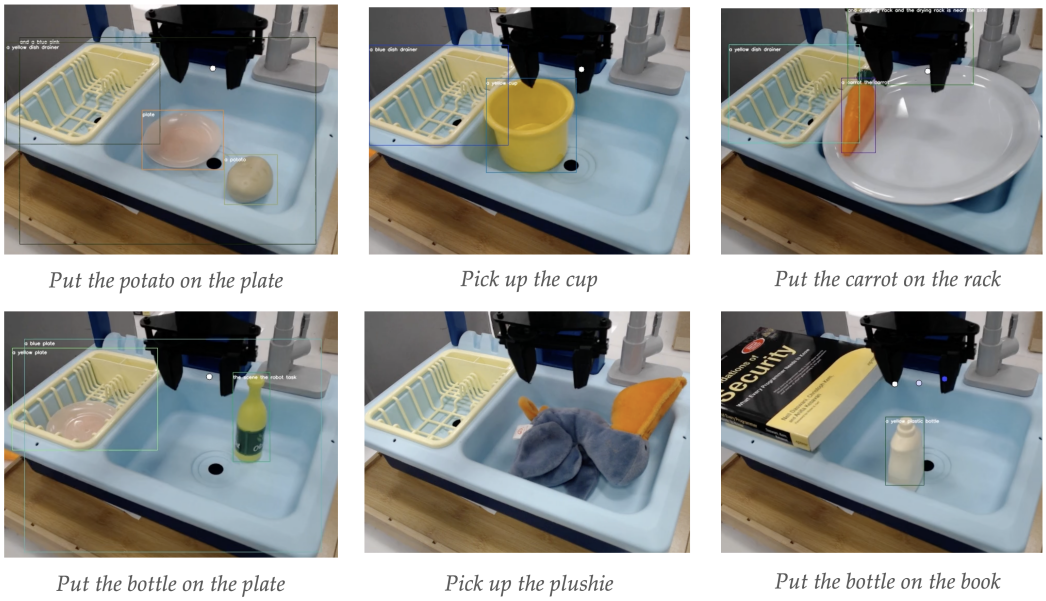
Data for Table 1a was collected from two new tabletop environments as shown in Fig. 7. Each task in Table 1a had 40 total demos collected. For Table 2 we collected 100 additional demos in the Toy Sink setup for the "in-distribution" evaluation. For the "OOD" data, we collected 50 demos from 5 different scenes as show in Fig. 8. In general we weighted data mixtures so that human video data and robot data were weighted approximately 50 percent each.

Something-Something V2 Data: We selected a subset of the Something-Something V2 data we deemed relevant based for robotics pick and place tasks on task types. For example, we selected tasks with tasks involving movements like "cover", "place", or "hold" but did not include tasks for movements a robot could not complete such as "move the camera". We used the same annotation pipeline for reasonings for both Something-Something V2 data as well as our other data mixes.

5.2 Training Details

RAD uses the Prismatic VLM [35] architecture from OpenVLA [3], which fuses pre-trained SigLIP [25] and/or DinoV2 [24] features for the visual encoder, and a LLaMA 2 7B [42] language backbone. We adopt the same model architecture (7B Prismatic VLM), action space (7-DoF discrete action space), observation space (image and natural language task), and prediction horizon (single step) as OpenVLA [3]. All models are fine-tuned to convergence with a learning rate of 2e-4, a LoRA batch size of 2, and anywhere from 2 to 8 GPUs (L40s or A40). Training of the ECoT-GT baseline is the

same as RAD except the loss term for the stop token is omitted and we also adjust the query prompt from "What action should the robot take to [*task*]?" to "Where is the robot hand in the image?".



Put the potato on the plate | *Pick up the cup* | *Put the carrot on the rack*

Put the bottle on the plate | *Pick up the plushie* | *Put the bottle on the book*

Figure 9: Example tasks for compositionally new tasks (left), new objects (middle), and new scenes (right).

5.3 Results

Rotation experiments: We conducted an additional experiment to assess if RAD could respond to data with significant hand rotations. We adjusted the reaosning generation pipeline to learn gripper rotation primitives, and then designed three tasks that required the gripper to rotate. We found that RAD showed a similar performance boost over ECoT to in-distribution results in 4.2 with a 38.3% boost in success rate as shown in 3.

Table 3: Rotation Experiments

Task	**ECoT**	**RAD**
pick up the corn	2/10	6/10
pick up the carrot and rotate counterclockwise	1.5/10	4/10
rotate to pick up the cereal box	2/10	6/10

Evaluation Details: Every task was evaluated 10 times. Objects were randomly placed throughout the scenes in a different spot for all 10 trials. For pick and place tasks, partial credit (0.5) was given for successfully picking up the object, but placing in the wrong location. For pick objects, no partial credit was given except for the "pick up the controller" task, which had an exceptionally high payload. Thus partial credit was given for grasping the object, even if the object slipped out of grasp upon being lifted.