RT-Affordance: Reasoning about Robotic Manipulation with Affordances

Anonymous Author(s)
Affiliation
Address
email

Abstract: We explore how policy input interfaces can facilitate generalization by providing intermediate guidance on how to perform manipulation tasks. Existing interfaces such as language, goal-image, and trajectory sketches have been shown to be helpful, but these representations either do not provide enough context or provide over-specified context that yields less robust policies. We propose conditioning policies on affordances, which capture the pose of the robot at key stages of the task. Affordances offer expressive yet lightweight abstractions, are easy for users to specify, and facilitate efficient learning by transferring knowledge from large internet datasets. Our method, RT-Affordance is a hierarchical model that first proposes an affordance plan given the task language, and then conditions the policy on this affordance plan to perform manipulation. Our affordance model can flexibly bridge diverse sources of supervision, including large web datasets, robot trajectories, and cheap-to-collect in-domain datasets, allowing us to learn new tasks with minimal effort. We show on a diverse set of novel tasks how RT-Affordance exceeds the performance of existing methods by over 50%, and we empirically demonstrate that affordances are robust to novel settings.

Keywords: Manipulation, VLMs, Affordances

1 Introduction

Vision-language-action (VLA) models [1, 2], pretrained with large-scale robot data on top of vision-language models (VLMs) [3] come with the promise of generalization to new objects, scenes, and tasks. However, VLAs are not yet reliable enough to be deployed outside of the narrow lab settings on which they are trained. While these shortcomings can be mitigated by expanding the scope and diversity of robot datasets, this is highly resource intensive and challenging to scale.

Alternatively, there are various ways of interfacing with the policy that can potentially facilitate generalization by operating in a lower-dimensional and generalizable input-space. Examples of these *policy interfaces* include language specifications [1, 4], goal images [5], goal sketches [6], and trajectory sketches [7]. These interfaces introduce mid-level abstractions that can provide intermediate guidance on how to perform the task, and can shield the policy from reasoning in a higher dimensional input space — leading to policies that can generalize over these intermediate representations. While one of the most common policy interfaces is conditioning on language, in practice most robot datasets are labeled with underspecified descriptions of the task. Alternatively, goal image-conditioned policies provide detailed spatial context about the final goal configuration of the scene. However, goal-images are high-dimensional, which presents learning challenges due to over-specification issues [6, 8]. This has lead to exploration of other intermediate representations — trajectory or goal sketches [7, 6], or keypoints [9, 10] — that attempt to provide spatial plans for the policy. While these spatial plans are informative about how to perform the task, or which point on an object to manipulate, they still lack sufficient information for the policy on *how* to manipulate.

In this work, we seek a policy interface that provides expressive yet lightweight abstractions for learning robust manipulation polices. We propose RT-Affordance, which is a policy conditioned on both language specifications and *visual affordances*. The visual affordances show the pose of the robot end effector at key stages of the task, vi-

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

sually projected onto the image input of the policy. By conditioning on affordances, the robot will have access to precise yet concise guidance on how to manipulate objects. To allow a seamless experience for the human user, we employ a hierarchical model that only requires task language from the user. The model first predicts the affordances given a task specification in language, and then leverages the affordances as an intermediate representation to steer the policy. The initial affordance prediction module can be trained on existing robot trajectories and web-scale datasets labeled with spatial information and affordances [11] as well as a small dataset of in-domain images with annotated ground truth affordances. Predicting affordances serves to bridge learning across these diverse sources of data (see Figure 1) and by harnessing all of these data sources, we can learn novel tasks efficiently.

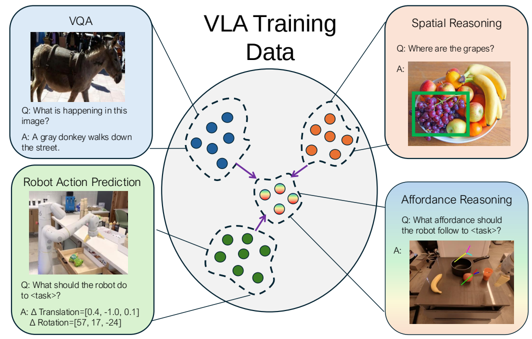

Figure 1: **Bridging the gap between robot data and internet data.** We propose using affordances as a means to bridge the gap between robot and web data.

We perform extensive experiments, where we show that RT-Affordance is effective across a broad range of real world tasks, achieving 69% overall success rate compared to 15% success rate for language-conditioned policies. We show how combining on web data and cheap-to-collect affordance images allows us to learn novel tasks *without collecting any additional robot demonstrations*. We also demonstrate that the resulting affordance prediction model is robust to distribution shifts.

2 RT-A: Affordance-Based Policy Learning

2.1 Affordance-conditioned policies

We are given a dataset of robot trajectories $\mathcal{D} = \{l, \{(o_i, e_i, g_i, a_i)\}_{i=0}^{T}\}$; each trajectory consists of a language instruction l and a sequence of images o_i, actions a_i, end-effector poses e_i and gripper states g_i. We learn an affordance-conditioned policy $\pi(a|l, o, q)$ that generates actions given the language instruction l, current image o, and additionally the *affordance plan* q. We define the affordance plan as the sequence of robot end effector poses corresponding to key timesteps in the trajectory, $q = (e_{t_1}, e_{t_2}, ..., e_{t_n})$. These timesteps capture critical stages in the task execution, for example when the robot is about to come in contact with objects or encounters bottleneck states. We can employ a variety of approaches to extract these timesteps. In practice we adopt a simple and scalable solution: we define key timesteps as when the gripper state changes from open to close ($g_{i-1} > \alpha$ and $g_i < \alpha$ for some constant α) or vice versa from close to open, or the final timestep of the trajectory. However, solely conditioning on affordance plans may not reveal full context about the task, and we thus opt to condition the policy on both affordance plans and language. This ensures that we retain the full expressiveness of language-conditioned policies, while benefiting from the additional context provided by affordance plans.

We train the affordance-conditioned policy via behavioral cloning and additionally co-train on web datasets, in a similar manner as in RT-2. We can represent these affordances either as tokenized text values passed as input to the policy, by overlaying them onto the image using a visual operator $\psi(o, q)$, following similar techniques in prior work [12, 7]. In our implementation we visually project the outline of the robot hand at the poses e_i onto the image. See Figure 2 for an illustration. We designate unique colors to each of the affordances overlaid onto the image to capture temporal ordering. Note that this projection step assumes knowledge of the robot camera intrinsics and extrinsics which is readily available for many robot platforms. If this information is not available, we can opt to condition the policy on the affordance plan directly as tokenized text values.

2.2 Learning to predict affordances

We can deploy the affordance-conditioned policy by asking the human user to provide affordance plans and language goals to the policy at inference time. We can also learn models to predict affordance plans automatically, sidestepping the need for humans to provide affordances at all at

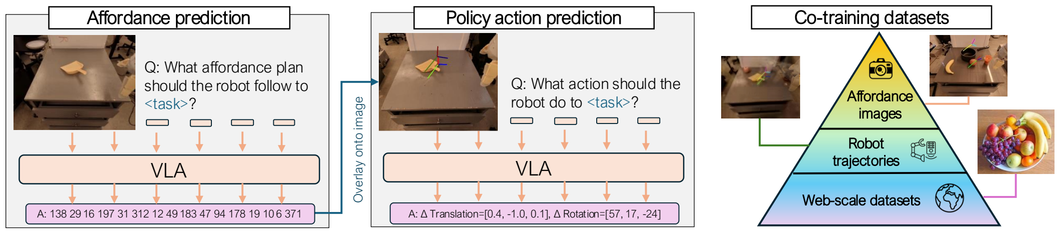

Figure 2: **Model overview.** Our hierarchical model first predicts the affordance plan given the task language and initial image of the task. We co-train the model on web datasets (largest data source), robot trajectories, and a modest number of cheap-to-collect images labeled with affordances.

test time. We learn an affordance prediction model $\phi(q|l, o)$ which predicts the affordance plan given the language task instruction l and initial image of the scene o. To train the model we extract (o, l, q) tuples from the same robot dataset $\mathcal{D}$ used to train π and we also co-train the model with web datasets. In some applications, training on these datasets may not yield adequate performance and we may seek additional training data to further improve the capabilities of the model. Instead of collecting additional demonstrations through expensive robot teleoperation, we can collect a set of images with corresponding task labels, ie. $\mathcal{D}_{\text{aug}} = \{(o_i, l_i)\}_{i=0}^{n}$. We can collect hundreds or thousands of these images at a fraction of the cost compared to costly teleoperation. After this data collection process we can annotate each image with the affordance plan through a posthoc labeling procedure efficiently without expensive hardware or teleoperation.

3 Experiments

3.1 Experiment implementation

We use the robot manipulator from RT-1 [13]. The arm is controlled via Cartesian end-effector control. Our robot demonstration datasets comprise three phases of data collection: (1) the RT-1 dataset [13] which focuses on basic manipulation skills, (2) the MOO dataset [14] which focuses on picking diverse objects, and (3) an additional set of trajectories targeting more dexterous tasks. We use the same web datasets from RT-2 for co-training. We adopt PaLM-E 2 [15, 16] as the underlying model and use the 1-billion parameter variant, unless otherwise noted. We train the affordance prediction model with the hindsight affordance labels from the robot trajectory datasets, in addition to a set of ~750 cheap-to-collect images manually annotated with affordance labels. These images include the tasks and objects from our grasping tasks and additional tasks beyond grasping which. We collect all of these images in approximately one hour and dedicate an additional two hours annotating them with affordance labels afterwards. We then train a separate affordance model VLA on these images, employing the same 1-billion model, trained to predict affordances from the language task prompt.

	RT-2	GC-RT-2	**RT-A (Oracle Aff)**	**RT-A (Ours)**
Pick dustpan	1/5	1/5	3/5	**4/5**
Pick kettle	1/5	1/5	**4/5**	**4/5**
Pick pot	0/5	1/5	**4/5**	1/5
Pick box	**4/5**	1/5	**4/5**	**4/5**
Pick headphones	1/5	2/5	**4/5**	**4/5**
Average	28%	24%	**76%**	68%

Table 1: **Experimental results on grasping.** State-of-the-art VLAs achieve success rates of under 30%. In contrast our affordance-conditioned policy paired with oracle human-provided affordances achieves 76% performance, and 68% when employing an affordance prediction model.

3.2 Learning to grasp novel objects efficiently

In our first experiment we investigate how affordances facilitate learning to grasp novel objects. We design a benchmark of picking diverse household objects, including dustpans, kettles, pots, boxes, and headphones. Note that our benchmark focuses on unseen object categories, meaning that they are not present in any of our robot trajectory datasets. We run comprehensive evaluations comparing our method to prior state-of-the-art approaches, with five rollouts per object category. See Table 1.

First we compare to **RT-2** [1], a state-of-the-art language-conditioned robot policy learning model notable for its impressive capabilities in understanding novel semantic concepts and objects. Despite these capabilities, we find that it struggles on our suite of evaluations, achieving an average success rate of just 28%. We observe that the policy is generally capable in identifying the correct object on the table and reaching the vicinity of the object but is unable to grasp the object at the appropriate location. Similar for picking the pot the robot tries to grasp around the base of the pot rather than handle. However, it is generally capable of picking boxes. We also tried to prompt the policy with specific language instructions indicating how to grasp the object (eg. "pick the dustpan by the handle") but the policy failed to follow these instructions effectively.

We also evaluate a goal-conditioned variant of RT-2 (**GC-RT-2**), which replaces language-conditioning for image goal-conditioning. We use the larger 24-billion variant PaLM 2 backbone to accommodate the additional goal-image passed into the policy. We run evaluations on the same objects, and for each episode we manually take a snapshot of the robot having grasped and lifted the object in the air at the final goal configuration. We observe an average success rate of just 24%. While the goal image conveys the precise pose at which to grasp the object, the policy is unable to precisely grasp the object at this pose.

Next we compare our hindsight affordance model RT-A. We condition the policy with the language instruction and visual affordances overlaid on top of the current image. We first evaluate the model with oracle affordances, ie. for each trial we manually provide the pregrasp and goal affordance poses of the robot. We call this self-baseline of our method **RT-A (Oracle Aff)**. We observe a significant improvement in policy performance, achieving 76% average success. The policy is faithful in executing the human provided affordance poses, and failures are only due to small imperfections from the robot policy in following the given affordance poses. Again, we highlight that none of these object categories are present in the robot trajectory datasets, making this a effective method for grasping a broad set of objects.

Finally we compare to the full hierarchical variant of our method in which we predict affordance plans before conditioning the policy on these plans (**RT-A**). We see an average performance of 68%, which is close to the performance of the policy conditioned on oracle affordances. Compared to the oracle affordance self-baseline we see similar performance across all object categories except picking the pot.

	RT-2	**RT-A (Oracle Aff)**	**RT-A (Ours)**
Place apple into pot	0/5	**4/5**	3/5
Place peach onto plate	1/5	**4/5**	**4/5**
Place bell pepper into basket	0/5	3/5	**4/5**
Place eggplant into box	0/5	2/5	**3/5**
Close the cubby	0/5	**4/5**	**4/5**
Turn sink faucet	0/5	**4/5**	3/5
Average	3%	**70%**	**70%**

Table 2: **Beyond grasping.** RT-A is applicable to a broad set of tasks and outperforms RT-2 by a wide margin.

3.3 Beyond object picking

We demonstrate that these findings are not exclusive to grasping tasks but can be extended to a range of manipulation tasks. We compare RT-A to the next best baseline from the previous experiments, the language-conditioned RT-2 model, on an additional set of manipulation tasks. We consider tasks involving placing objects into receptacles and articulated manipulation. Again, we highlight that these tasks are *unseen* in the robot trajectory datasets. See Table 2 for results. Surprisingly, the RT-2 baseline performs quite poorly in this setting achieving only 3% success rate. With RT-A we see a significant improvement of performance, with 70% success rate using our affordance prediction model. These results show that affordances are a flexible form of task specification that can describe a broad set of tasks. In cases where the user provides oracle affordances at evaluation, we can solve novel tasks without any additional data, and training our affordance prediction model to infer affordances automatically only incurs a small budget to collect and annotate images.

See Appendix sections B and C for additional experiments and section D for related work.

4 Conclusion

We have presented RT-Affordance, a hierarchical method that uses affordances as an intermediate representation for policies. We have shown empirically that affordance-conditioned policies can perform a wide range of novel tasks without requiring additional human demonstrations. In the future, we are interested in exploring the complementary strengths of different policy interfaces and combining their capabilities into a single model that can share knowledge across interfaces.

References

[1] 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 *7th Annual Conference on Robot Learning*, 2023.

[2] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[3] G. Team. Gemini: A family of highly capable multimodal models, 2024.

[4] 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 *Robotics: Science and Systems (RSS)*, 2024.

[5] K. Black, M. Nakamoto, P. Atreya, H. R. Walke, C. Finn, A. Kumar, and S. Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *The Twelfth International Conference on Learning Representations*.

[6] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman, D. Sadigh, J. Bohg, and S. Schaal. Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches, 2024. URL https://arxiv.org/abs/2403.02709.

[7] J. Gu, S. Kirmani, P. Wohlhart, Y. Lu, M. G. Arenas, K. Rao, W. Yu, C. Fu, K. Gopalakrishnan, Z. Xu, P. Sundaresan, P. Xu, H. Su, K. Hausman, C. Finn, Q. Vuong, and T. Xiao. Rt-trajectory: Robotic task generalization via hindsight trajectory sketches, 2023.

[8] R. Shah, R. Martín-Martín, and Y. Zhu. Mutex: Learning unified policies from multimodal task specifications. In *7th Annual Conference on Robot Learning*, 2023. URL https://openreview.net/forum?id=PwqiqaaEzJ.

[9] W. Yuan, J. Duan, V. Blukis, W. Pumacay, R. Krishna, A. Murali, A. Mousavian, and D. Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics, 2024. URL https://arxiv.org/abs/2406.10721.

[10] K. Fang, F. Liu, P. Abbeel, and S. Levine. Moka: Open-world robotic manipulation through mark-based visual prompting. *Robotics: Science and Systems (RSS)*, 2024.

[11] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, M. Martin, T. Nagarajan, I. Radosavovic, S. K. Ramakrishnan, F. Ryan, J. Sharma, M. Wray, M. Xu, E. Z. Xu, C. Zhao, S. Bansal, D. Batra, V. Cartillier, S. Crane, T. Do, M. Doulaty, A. Erapalli, C. Feichtenhofer, A. Fragomeni, Q. Fu, C. Fuegen, A. Gebreselasie, C. Gonzalez, J. Hillis, X. Huang, Y. Huang, W. Jia, W. Khoo, J. Kolar, S. Kottur, A. Kumar, F. Landini, C. Li, Y. Li, Z. Li, K. Mangalam, R. Modhugu, J. Munro, T. Murrell, T. Nishiyasu, W. Price, P. R. Puentes, M. Ramazanova, L. Sari, K. Somasundaram, A. Southerland, Y. Sugano, R. Tao, M. Vo, Y. Wang, X. Wu, T. Yagi, Y. Zhu, P. Arbelaez, D. Crandall, D. Damen, G. M. Farinella, B. Ghanem, V. K. Ithapu, C. V. Jawahar, H. Joo, K. Kitani, H. Li, R. Newcombe, A. Oliva, H. S. Park, J. M. Rehg, Y. Sato, J. Shi, M. Z. Shou, A. Torralba, L. Torresani, M. Yan, and J. Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022.

[12] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, et al. Pivot: Iterative visual prompting elicits actionable knowledge for vlms. In *Forty-first International Conference on Machine Learning*.

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

[14] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, C. Finn, and K. Hausman. Open-world object manipulation using pre-trained vision-language models. In *arXiv preprint*, 2023.

[15] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, E. Chu, J. H. Clark, L. E. Shafey, Y. Huang, K. Meier-Hellstern, G. Mishra, E. Moreira, M. Omernick, K. Robinson, S. Ruder, Y. Tay, K. Xiao, Y. Xu, Y. Zhang, G. H. Abrego, J. Ahn, J. Austin, P. Barham, J. Botha, J. Bradbury, S. Brahma, K. Brooks, M. Catasta, Y. Cheng, C. Cherry, C. A. Choquette-Choo, A. Chowdhery, C. Crepy, S. Dave, M. Dehghani, S. Dev, J. Devlin, M. Díaz, N. Du, E. Dyer, V. Feinberg, F. Feng, V. Fienber, M. Freitag, X. Garcia, S. Gehrmann, L. Gonzalez, G. Gur-Ari, S. Hand, H. Hashemi, L. Hou, J. Howland, A. Hu, J. Hui, J. Hurwitz, M. Isard, A. Ittycheriah, M. Jagielski, W. Jia, K. Kenealy, M. Krikun, S. Kudugunta, C. Lan, K. Lee, B. Lee, E. Li, M. Li, W. Li, Y. Li, J. Li, H. Lim, H. Lin, Z. Liu, F. Liu, M. Maggioni, A. Mahendru, J. Maynez, V. Misra, M. Moussalem, Z. Nado, J. Nham, E. Ni, A. Nystrom, A. Parrish, M. Pellat, M. Polacek, A. Polozov, R. Pope, S. Qiao, E. Reif, B. Richter, P. Riley, A. C. Ros, A. Roy, B. Saeta, R. Samuel, R. Shelby, A. Slone, D. Smilkov, D. R. So, D. Sohn, S. Tokumine, D. Valter, V. Vasudevan, K. Vodrahalli, X. Wang, P. Wang, Z. Wang, T. Wang, J. Wieting, Y. Wu, K. Xu, Y. Xu, L. Xue, P. Yin, J. Yu, Q. Zhang, S. Zheng, C. Zheng, W. Zhou, D. Zhou, S. Petrov, and Y. Wu. Palm 2 technical report, 2023. URL https://arxiv.org/abs/2305.10403.

[16] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, et al. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*, pages 8469–8488. PMLR, 2023.

[17] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman. Mt-opt: Continuous multi-task robotic reinforcement learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[18] K. Hausman, J. T. Springenberg, Z. Wang, N. Heess, and M. Riedmiller. Learning an embedding space for transferable robot skills. In *International Conference on Learning Representations*, 2018.

[19] S. James, M. Bloesch, and A. J. Davison. Task-embedded control networks for few-shot imitation learning. In *Conference on robot learning*, pages 783–795. PMLR, 2018.

[20] K. Pertsch, Y. Lee, and J. Lim. Accelerating reinforcement learning with learned skill priors. In *Conference on robot learning*, pages 188–204. PMLR, 2021.

[21] 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*, 2021.

[22] 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*, 33:13139–13150, 2020.

[23] J. Zhang, K. Pertsch, J. Zhang, and J. J. Lim. Sprint: Scalable policy pre-training via language instruction relabeling. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9168–9175. IEEE, 2024.

[24] S. Nair, E. Mitchell, K. Chen, S. Savarese, C. Finn, et al. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *Conference on Robot Learning*, pages 1303–1315. PMLR, 2022.

[25] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan. Vima: General robot manipulation with multimodal prompts. In *International Conference on Machine Learning*, 2023.

[26] P. Sundaresan, S. Belkhale, D. Sadigh, and J. Bohg. Kite: Keypoint-conditioned policies for semantic manipulation. *Robotics: Science and Systems (RSS)*, 2023.

[27] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning. *Robotics: Science and Systems (RSS)*, 2024.

[28] K. Bousmalis, G. Vezzani, D. Rao, C. M. Devin, A. X. Lee, M. B. Villalonga, T. Davchev, Y. Zhou, A. Gupta, A. Raju, et al. Robocat: A self-improving generalist agent for robotic manipulation. *Transactions on Machine Learning Research*, 2023.

[29] Y. Chebotar, K. Hausman, Y. Lu, T. Xiao, D. Kalashnikov, J. Varley, A. Irpan, B. Eysenbach, R. C. Julian, C. Finn, et al. Actionable models: Unsupervised offline reinforcement learning of robotic skills. In *International Conference on Machine Learning*, pages 1518–1528. PMLR, 2021.

[30] S. Nasiriany, V. Pong, S. Lin, and S. Levine. Planning with goal-conditioned policies. *Advances in neural information processing systems*, 32, 2019.

[31] Z. J. Cui, Y. Wang, N. M. M. Shafiullah, and L. Pinto. From play to policy: Conditional behavior generation from uncurated robot data. In *The Eleventh International Conference on Learning Representations*.

[32] Y. Cui, S. Niekum, A. Gupta, V. Kumar, and A. Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In R. Firoozi, N. Mehr, E. Yel, R. Antonova, J. Bohg, M. Schwager, and M. Kochenderfer, editors, *Proceedings of The 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pages 893–905. PMLR, 23–24 Jun 2022. URL https://proceedings.mlr.press/v168/cui22a.html.

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

[34] S. Dasari and A. Gupta. Transformers for one-shot visual imitation. In *Conference on Robot Learning*, pages 2071–2084. PMLR, 2021.

[35] V. Jain, M. Attarian, N. J. Joshi, A. Wahid, D. Driess, Q. Vuong, P. R. Sanketi, P. Sermanet, S. Welker, C. Chan, I. Gilitschenski, Y. Bisk, and D. Dwibedi. Vid2robot: End-to-end video-conditioned policy learning with cross-attention transformers, 2024.

[36] P. Ardon, E. Pairet, K. S. Lohan, S. Ramamoorthy, and R. P. A. Petrick. Affordances in robotic tasks – a survey, 2020. URL https://arxiv.org/abs/2004.07400.

[37] A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *International Conference on Computer Vision (ICCV)*, 2019.

[38] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. 2021.

[39] H.-S. Fang, C. Wang, M. Gou, and C. Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020.

[40] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. 2017.

[41] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics (T-RO)*, 2023.

[42] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, pages 540–562. PMLR, 2023.

[43] C. Tang, D. Huang, L. Meng, W. Liu, and H. Zhang. Task-oriented grasp prediction with visual-language inputs. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4881–4888. IEEE, 2023.

[44] J. Duan, W. Yuan, W. Pumacay, Y. R. Wang, K. Ehsani, D. Fox, and R. Krishna. Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*, 2024.

[45] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.

[46] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[47] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[48] 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, 2024.

[49] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[50] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics. 2023.

[51] M. K. Srirama, S. Dasari, S. Bahl, and A. Gupta. Hrp: Human affordances for robotic pre-training. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

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

Appendix

A Model Inference

Our model inference procedure is as follows. We are given the initial image of the scene o_{init} and a natural language task instruction l. We can either prompt a human or the affordance prediction model $\phi(q|l, o_{\text{init}})$ to provide the affordance plan q. Then we can prompt the policy $\pi(a|l, \psi(o, q))$ with the language instruction and affordance plan to execute the task. We can optionally replan updated affordance plans at fixed or adaptive intervals to handle novel scenarios that arise during the execution of the policy

B Robustness to out of distribution factors

Next, we perform an analysis of the affordance prediction model. In order for the affordance prediction model to be useful it needs to be robust to a wide range of out-of-distribution (OOD) settings. To better understand this, we perform a comprehensive evaluation on the grasping tasks from Table 1 comparing the following settings:

- **In-distribution**: evaluating the model under the same distribution it was trained on. ie. same object instances, same camera view, and same environment background.
- **OOD: novel objects**: evaluating the model with novel object instances on which it was not trained on.
- **OOD: novel camera view**: evaluating the model with images taken with significant camera shift.
- **OOD: novel background**: evaluating the model with novel object textures.

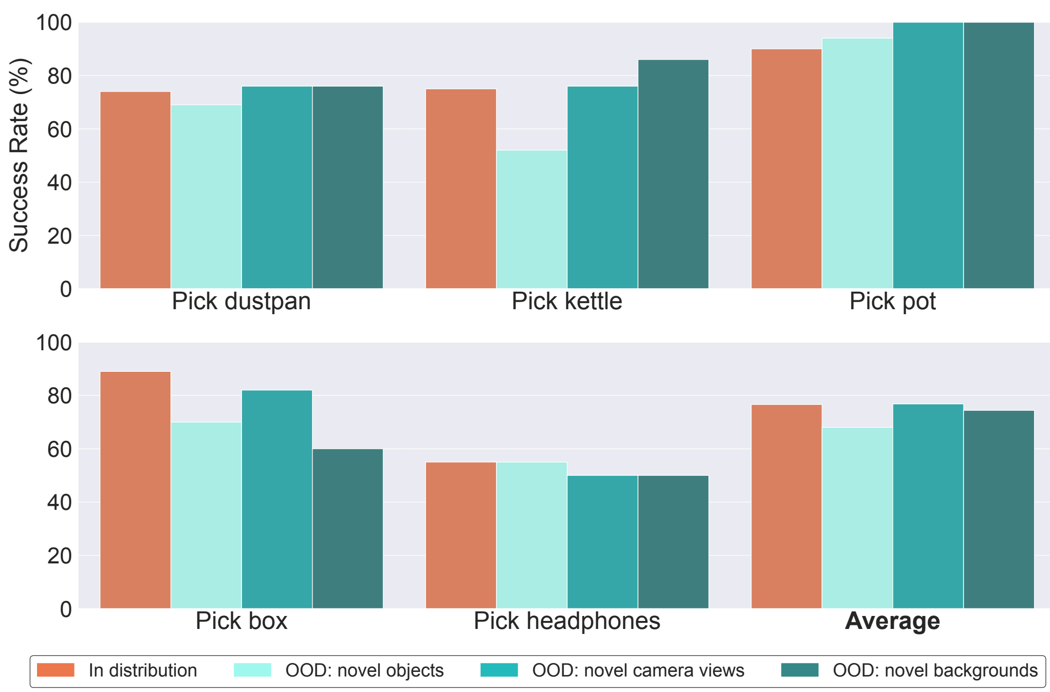

Figure 3: **Evaluation of the affordance prediction model on out of distribution scenarios.** We perform a comprehensive evaluation of the affordance prediction model on in-distribution and out-of-distribution (OOD) and observe a graceful degradation of performance in OOD settings.

We perform a comprehensive offline evaluation over hundreds of test images, where for each image we assess whether the model's predicted affordance would result in a successful grasp, assuming that

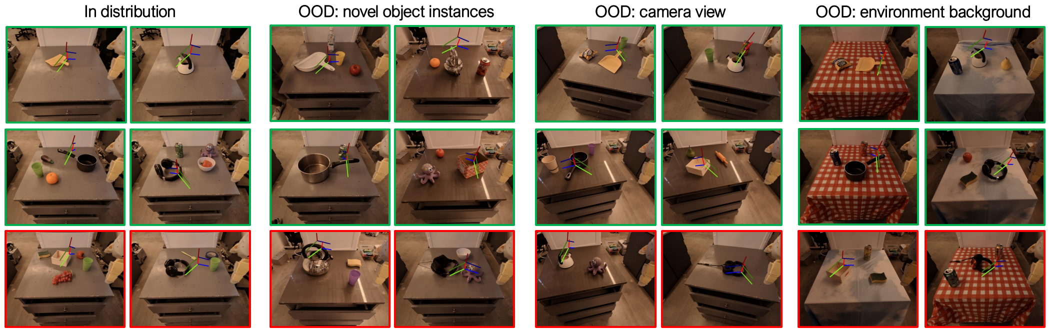

Figure 4: **Robustness to out of distribution factors** We show examples of successful and incorrect predictions of our affordance prediction model across in-distribution and out-of-distribution settings. Successful predictions are highlighted in green and incorrect predictions are highlighted in red.

the policy can follow the given affordances perfectly. We report the results in Figure 3. First, we see that the affordance prediction model is general capable in in-distribution settings, with 77% of trials classified as success. Across the OOD settings model performance degrades gracefully, falling no more than 10% compared to the in-distribution setting. Some factors affect model performance more than others. With novel camera views the performance is nearly identical at 77%, and with novel backgrounds performance only falls at 3% on average. However with novel object instances the performance drops the most, especially for grasping novel instances of kettles and boxes. We provide illustrative examples in Figure 4.

C Ablation study

We perform an ablation study on our affordance prediction model, where we study the impact of different data sources on the model. Our model is trained on the full data mixture including (1) robot trajectories, (2) web datasets, and (3) the 750 additional augmented affordance images we collected. We perform ablations where we (a) exclude the augmented data (**No aug data)** and (b) exclude web datasets (**No web data**). We compare these settings on the same in-distribution evaluation suite outlined in Section B, and we report results in Table 3. We see that removing these sources of data leads to a large drop in performance. We hypothesize that large web datasets play an important role for training robust models, and that our augmented data is needed to train performant models for specific downstream tasks.

	Ours	**No aug data**	**No web data**
Pick dustpan	**74%**	20%	3%
Pick kettle	**75%**	30%	10%
Pick pot	**90%**	10%	14%
Pick box	**89%**	33%	11%
Pick headphones	**55%**	28%	16%
Average	**77%**	24%	11%

Table 3: **Ablation study.** We perform an ablation study of our affordance prediction model the same in-distribution evaluations as Figure 3. We find that removing the augmented dataset of affordance images significantly diminishes performance, and removing web datasets for co-training diminishes performance even further.

D Related Work

Prior works have studied how multi-task robot manipulation policies can be conditioned on various types of representations and interfaces to perform different manipulation skills. Popular interfaces have included one-hot task vectors [17], latent skill or task embeddings [18, 19, 20], templated or natural language [21, 13, 22, 23, 24], object-centric representations [25, 14, 26, 10], trajectories [7, 27], goal images or sketches [28, 29, 6, 30, 5, 31, 32], and videos [33, 34, 35]. Our method leverages affordances represented visually or textually as an interface, which strikes a balance between flexibility, expressivity, and data efficiency.

Affordances for robot manipulation. Affordances [36] and grasp pose predictions have been heavily leveraged in robotics research for motion planning, grasping, and hierarchical control. Modern data-driven methods [37, 38] build upon prior works which leverage optimization-based approaches, and achieve performant grasp pose prediction capabilities given large-scale grasping datasets [39] and point-cloud [40] or geometry based inductive biases [41]. More recently, robot manipulation systems propose combining vision-language models (VLMs) with affordance or grasp prediction models and downstream control policies [42, 43, 44, 45]. In contrast, our method RT-Affordance does not rely on large-scale offline grasp pose specific datasets, 3D point clouds at training or test time, or simulation-based geometric planning.

Learning pre-trained representations from non-action data. Similar to trends seen in scaling up VLMs [46], there has also been exploration in robotics on adapting large-scale internet data for improving perception and reasoning capabilities [16] which are important for downstream robot policy learning, particularly with the usage of vision-language-action (VLA) models [1]. Non-robotics interaction datasets have been particularly of interest, due to the substantial cost of real-world robotics action data such as teleoperated expert demonstrations [47, 48]; representation learning methods which learn affordance prediction from internet data and human videos [49, 11] have been proposed [50, 51, 52]. Most similar to our method is RoboPoint [9], which proposes fine-tuning a VLM to predict points which represent spatial affordances by leveraging procedural 3D scene generation in simulation. Our method RT-Affordance also studies predicting spatial affordances, but proposes a more descriptive affordance representation beyond a single point, and also does not require large-scale simulated scene generation.