

Appendix

Table of Contents

A Experimental Setup	12
A.1 Experimental Domain Details	12
A.2 Implementation Details of Baseline Methods and Ablations	13
A.3 Retrieval Threshold Selection	14
B Additional Baselines and Ablation Results	14
B.1 Additional Baselines	14
B.2 Using Proprioception for Retrieval	14
C Retrieval Visualization	15
D Ablating Retrieval Strategies	16
D.1 KNN-based Retrieval Strategy	16
D.2 Retrieving with Pre-trained Representations	16

A Experimental Setup

A.1 Experimental Domain Details

Below we provide full details about the experimental domain we used in our experiments:

- **Square Assembly:** The goal is to pick up and place a square nut into the square peg. $\mathcal{D}_{\text{target}}$ consists of 10 demonstrations. $\mathcal{D}_{\text{prior}}$ consists of two different types of data: 200 episodes of *useful* demonstrations placing the nut into the goal peg, and 200 episodes of *adversarial* demonstrations that place the nut into the wrong peg. Note that the initial phase of the adversarial data also consists of useful motion for learning to pick up the nut. In our setup, the target task is in the same environment as the adversarial data while the useful data has a different background. We generated optimal demonstrations using scripted policies.
- **LIBERO-Can:** The goal is to pick up a can and place it into a basket. Corresponding task description is "pick up the alphabet soup and place it in the basket" and the environment is LIVING_ROOM_SCENE1 (from the LIBERO-90 suite). $\mathcal{D}_{\text{target}}$ consists of 10 human demonstrations from the LIBERO benchmark, and $\mathcal{D}_{\text{prior}}$ consists of 18 selected tasks from LIBERO-90 and each with 50 trajectories, resulting in a total of 900 trajectories. The list of task IDs in LIBERO-90 that was selected to form the prior dataset is:

```
"LIVING_ROOM_SCENE1_pick_up_the_ketchup_and_put_it_in_the_basket"  
"LIVING_ROOM_SCENE2_pick_up_the_butter_and_put_it_in_the_basket"  
"LIVING_ROOM_SCENE2_pick_up_the_orange_juice_and_put_it_in_the_basket"  
"LIVING_ROOM_SCENE2_pick_up_the_tomato_sauce_and_put_it_in_the_basket"  
"LIVING_ROOM_SCENE2_pick_up_the_milk_and_put_it_in_the_basket"  
"LIVING_ROOM_SCENE2_pick_up_the_alphabet_soup_and_put_it_in_the_basket"  
"LIVING_ROOM_SCENE3_pick_up_the_alphabet_soup_and_put_it_in_the_tray"  
"LIVING_ROOM_SCENE3_pick_up_the_butter_and_put_it_in_the_tray"  
"LIVING_ROOM_SCENE4_stack_the_left_bowl_on_the_right_bowl_and_place_them_in_the_tray"  
"LIVING_ROOM_SCENE6_put_the_chocolate_pudding_to_the_left_of_the_plate"  
"KITCHEN_SCENE1_open_the_bottom_drawer_of_the_cabinet"  
"KITCHEN_SCENE2_put_the_middle_black_bowl_on_the_plate"  
"KITCHEN_SCENE3_turn_on_the_stove_and_put_the_frying_pan_on_it"  
"KITCHEN_SCENE8_turn_off_the_stove"  
"KITCHEN_SCENE10_close_the_top_drawer_of_the_cabinet"  
"STUDY_SCENE1_pick_up_the_book_and_place_it_in_the_front_compartment_of_the_caddy"  
"STUDY_SCENE2_pick_up_the_book_and_place_it_in_the_back_compartment_of_the_caddy"  
"STUDY_SCENE3_pick_up_the_white_mug_and_place_it_to_the_right_of_the_caddy"
```

- **Bridge-Pot and Bridge-Microwave:** In *Bridge-Pot* the robot picks up a pot on the burner and places into the sink. In *Bridge-Microwave*, the robot grasps the handle of the microwave door and pulls it open. We leverage Bridge-V2 [7] as $\mathcal{D}_{\text{prior}}$ and collected $\mathcal{D}_{\text{target}}$ in our own setup with a ViperX arm [36], which is similar to some environments in Bridge-V2 but does not exist in the prior dataset. We use a ViperX arm [36] instead of the WidowX [37] as in Bridge-V2 dataset, which introduces additional domain difference for transferring useful pattern in prior data. We collected 10 demonstration for each of the target tasks using VR teleoperation and use a subset of Bridge-V2 as prior data. The full list of environments in the prior dataset is:

```

datacol1_toykitchen1
datacol1_toykitchen6
datacol2_folding_table
datacol2_robot_desk
datacol2_toykitchen1
datacol2_toykitchen5
datacol2_toykitchen7
datacol2_toysink2
deephought_robot_desk
deephought_toykitchen1
deephought_toykitchen2
minsky_folding_table_white_tray

```

- **Franka-Pen-in-Cup:** The goal is to pick up a marker and put it into a cup. $\mathcal{D}_{\text{target}}$ consists of 10 human demonstrations. We tested two instances of $\mathcal{D}_{\text{prior}}$: 1) *PnP*: 105 trajectories of pick-and-place tasks from the same robot that we collected ourselves through VR teleoperation, and 2) *Wild*: 400 randomly sampled trajectories from the DROID [8] dataset (filtered to roughly match the viewpoint in target task). Pen-in-Cup is a task that exists in DROID but the demonstrations were collected in very different environments.

A.2 Implementation Details of Baseline Methods and Ablations

We have two sets of implementations of FLOWRETRIEVAL and baselines for our experiments: 1) for experiments in simulation as well as the *Franka-Pen-in-Cup* task, we adapted the diffusion policy implementation of Chi et al. [33]; we use the U-Net variant of diffusion policy and load pretrained ImageNet weights for initializing the encoder network; 2) for Bridge-* tasks, we use the diffusion model implementation provided by Walke et al. [7]³; we also load pretrained ImageNet weights for the encoder of the policy.

For **BR**, we re-implemented the pretraining logic for the VAE and reused the hyperparameters as documented in Du et al. [2]. For **SR** results in the paper, we re-implemented the pretraining logic for the latent skill space with the provided hyperparameters in Nasiriany et al. [1]. We also ran our simulation experiments using the full implementation of SAILOR provided by Nasiriany et al. [1], but fail to obtain non-zero success rates. One potential reason is that the tasks we consider in our experimental setup is relatively short, while SAILOR was designed for tackling long-horizon tasks (4-6 steps per task), such that our experimental domains may require a completely different set of hyperparameters than those used in the original work.

ProprioRetrieval is a method where we use proprioceptive data for computing similarity at the retrieval stage. We also apply flow loss during policy learning, and therefore it can be considered as an ablation of the retrieval space of FLOWRETRIEVAL. Specifically, we leverage end-effector Cartesian position difference between s_t and s_{t+k} to represent motion. While it is desirable to match the rotational motion as well, we find using the full end-effector Cartesian pose to compute similarity requires tuning scaling factors for balancing positional state versus rotational state (orientation) and not only does that require manual evaluations, but also existing datasets often use different conventions for orientation and therefore introduces additional challenges for comparing proprioceptive states across datasets directly. Hence in our implementation of **ProprioRetrieval**, we use only delta of the position of the end effector to retrieve from prior data. The feature e_t for datapoint s_t is computed as:

$$e_t = (s_t, s_{t+k}[:3] - s_t[:3]) \quad (7)$$

The similarity score is then computed as the negative ℓ_2 distance between feature vectors.

³https://github.com/rail-berkeley/bridge_data_v2

Method	SquareAssembly	LIBERO-Can	Bridge-Micro	Bridge-Pot	Pen-in-Cup
BC	7%	64%	56%	20%	23%
BC-Co	3%	16%	56%	32%	12%
FlowBC	7%	71%	48%	8%	40%
ProprioRetrieval-	44%	73%	24%	40%	-
ProprioRetrieval	54%	81%	16%	12%	44%
FLOWRETRIEVAL	55%	90%	68%	64%	56%

Table 1: Success rates of baselines and ablations of FLOWRETRIEVAL.

A.3 Retrieval Threshold Selection

In FLOWRETRIEVAL, we retrieve the top $\delta\%$ from the prior data. However, the optimal threshold can be different for different target task and prior dataset. Intuitively, if the similarity metric truly ranks the prior datapoints by usefulness to target task, we want to retrieve just the right amount of data that supports the learning of target task. Du et al. [2] observed a bell-shaped curve between the relationship of retrieval threshold and policy performance. Due to constraints on computing resources, we only sparsely searched over a small range of threshold values for 3 tasks (2 in sim, 1 in real): Square Assembly, LIBERO-Can, and Bridge-Pot. In our experiments, we retrieve 35% from prior data in Square Assembly, 10% in LIBERO-Can, and 1% in Bridge tasks. We then reuse the same threshold from bridge-Pot for the Bridge-Microwave task. For Franka Pen-in-Cup, we reuse the threshold of found in the Square Assembly task.

B Additional Baselines and Ablation Results

B.1 Additional Baselines

We present the full evaluations of **BC-Co** and **FlowBC** in Table 1. **BC-Co** directly cotrains with the entire prior dataset. **FlowBC** applies auxiliary flow loss to **BC**. We see that while each could improve the performance from vanilla **BC** in certain domains, they can also hurt performance in others, depending on the composition of prior dataset.

B.2 Using Proprioception for Retrieval

Low level motion features from proprioception are simple to compute and can be effective for extracting similar actions from past experience. Therefore we ablate our retrieval method with using proprioceptive information, and denote this method as **ProprioRetrieval**. However, as proprioceptive features are entirely agnostic to the visual observation and therefore may retrieve irrelevant data such as those from very different viewpoints if the prior dataset has different viewpoints, e.g. OXE [13] and DROID [8].

ProprioRetrieval still applies the auxiliary flow loss during policy learning. We additionally evaluate a baseline version that does not use the flow loss as a comparison point, and use **ProprioRetrieval-** to denote this baseline. The success rates of both proprioception-based methods are reported in Table 1. We see that **ProprioRetrieval** can lead to high success rates when camera viewpoints are consistent (e.g. in Square Assembly and LIBERO-Can), but does not retrieve enough useful data from *Wild* to match the performance of FLOWRETRIEVAL in the Franka Pen-in-Cup task, and performs poorly in the Bridge tasks that have a large variety of prior tasks. Note that the downstream policy learning does not use the low-level action data to update the policy branch in the Franka Pen-in-Cup task when retrieving from the *Wild* dataset, due to the large variation in camera viewpoints. However, we still use the low level actions from the retrieved prior data of Bridge since the viewpoints are better aligned, but may introduce additional multimodality in policy learning.

C Retrieval Visualization

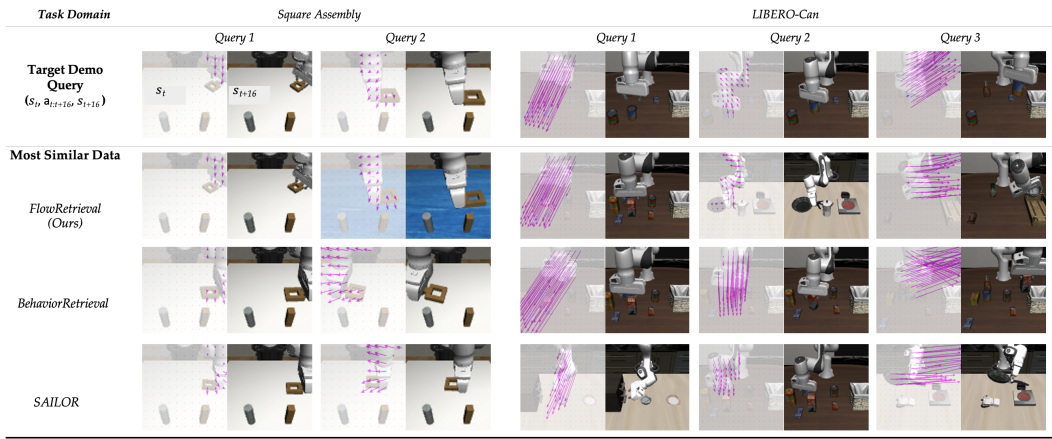


Figure 7: Visualization of paired retrieved data and query target datapoint by different methods in Square Assembly and LIBERO-Can.

Fig. 7 shows additional queries and retrieved data by different retrieval method in the two simulated tasks. BR and SR both encodes visual observations and low level actions in the latent space used for retrieval. We see that BR often overly focuses on visual scene similarity, and motion similarity is likely a second order feature (only effective if the visual scene is similar in the first place). Therefore in all cases, it cannot ignore the effect of the environment (background) and end up retrieving potentially adversarial data. SR can latch on either of the similarities, sometimes focusing on visual scene (Square Assembly), and sometimes retrieving based on action similarity (LIBERO-Can). In contrast, FLOWRETRIEVAL consistently focuses on visual motion similarity.

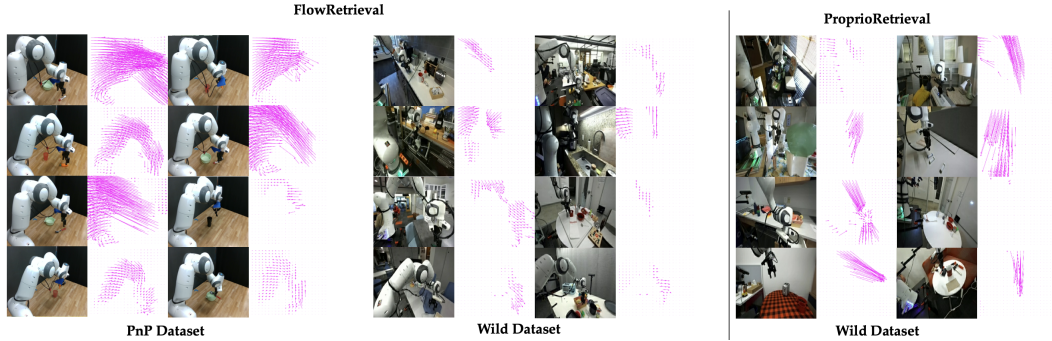


Figure 8: Visualization of example data points retrieved by FLOWRETRIEVAL and ProprioRetrieval in Franka Pen-in-Cup task.

Fig. 8 shows the example data points retrieved by FLOWRETRIEVAL and ProprioRetrieval in Franka Pen-in-Cup task. When retrieving from the *PnP* dataset, FLOWRETRIEVAL focuses on the pick-up and transfer stage of the target task, and does not retrieve the placing motions from the prior dataset, effectively filtering adversarial prior data. When retrieving from the *Wild* dataset, we see that FLOWRETRIEVAL retrieves viewpoints better aligned with that in the target task, while ProprioRetrieval retrieves very different viewpoints (sometimes the robot is not even in the view – see example in rightmost column, second from bottom). These data points are less informative for the downstream learning of the target task but may still provide additional regularization on learning a diverse set of visual features.

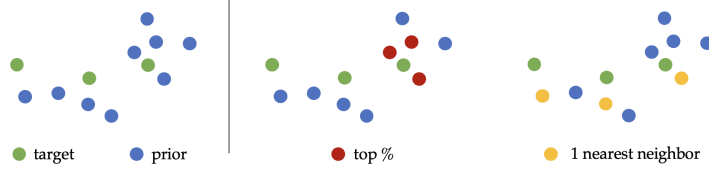


Figure 9: **Illustration of two retrieval strategies.** Depending on the distribution of target and prior data points, top-% may retrieve data only close to certain part of the target trajectory while KNN would retrieve uniformly.

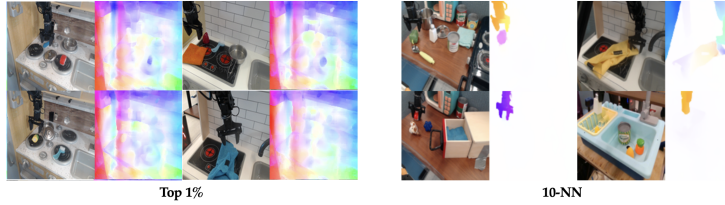


Figure 10: Visualization of the retrieved datapoints in Bridge tasks using two different strategies.

D Ablating Retrieval Strategies

D.1 KNN-based Retrieval Strategy

In our implementation of FLOWRETRIEVAL, we follow the practice of prior works and retrieve base on a threshold of the similarity score, taking the top $\delta\%$ closest datapoints from prior dataset to any point in target data. However, this does not augment the target task trajectory uniformly. As a result, when there are pauses in the dataset (which is true in most human-demonstrated datasets if not post-processed), we observe that the retrieved data contain a large portion of small motion data where the robot arm barely moves (Fig. 10 left).

One intuitive way to solve this issue is to retrieve top- k data points for each state in the target task data. See an illustration of the two different retrieval strategies in Fig. 9. We visualize samples of the data retrieved by 10-NN in Bridge tasks in Fig. 10 (right) and see that it is able to retrieve meaningfully similar data to target task. However, we find that, when retrieving similar amount of total data as top 1%, this approach surprisingly does not lead to as performant policies as the existing approach, achieving 40% in Square Assembly (-15%), 80% in LIBERO-Can (-10%), and 60% (-8%) in Bridge-Microwave. One potential reason is that when we force each datapoint to have at least some datapoints retrieved from prior data, we might end up retrieving dissimilar and potentially adversarial data if the k is not carefully selected for each datapoint.

D.2 Retrieving with Pre-trained Representations

We evaluated retrieval in the Square Assembly task with pretrained visual representations to see if off-the-shelf models could be leveraged to retrieve motion-similar data. Specifically we take the delta between the features of s_t and s_{t+k} and use that to represent motion as proposed in prior work [38]. Fig. 11 shows the detailed analysis of data retrieved by Voltron [21], R3M [39], CLIP [40], and Dino-v2 [41]. We see that in general these models focus on visual features more than the motion

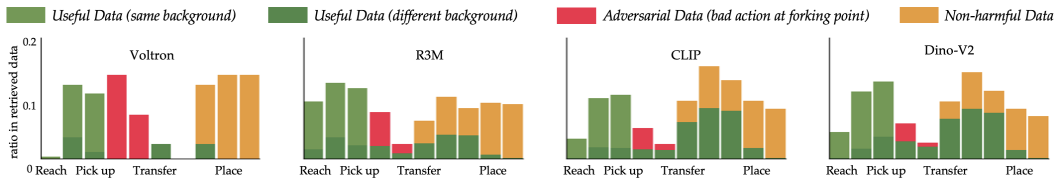


Figure 11: Retrieval Analysis for pretrained visual representations.

itself and retrieving with such embeddings cannot bypass adversarial data in this task, with motion language-aligned models (Voltron and R3M) suffering the most.

References

- [1] S. Nasiriany, T. Gao, A. Mandlekar, and Y. Zhu. Learning and retrieval from prior data for skill-based imitation learning. In *Conference on Robot Learning (CoRL)*, 2022.
- [2] M. Du, S. Nair, D. Sadigh, and C. Finn. Behavior retrieval: Few-shot imitation learning by querying unlabeled datasets. In *Robotics: Science and Systems (RSS)*, 2023.
- [3] 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. *arXiv preprint arXiv:2403.01823*, 2024.
- [4] 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*, 40(12-14): 1419–1434, 2021.
- [5] L. Zha, Y. Cui, L.-H. Lin, M. Kwon, M. G. Arenas, A. Zeng, F. Xia, and D. Sadigh. Distilling and retrieving generalizable knowledge for robot manipulation via language corrections. In *International Conference on Robotics and Automation*, 2024.
- [6] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Conference on Neural Information Processing Systems*, 2023.
- [7] 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.
- [8] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Bajjal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, D. A. Herrera, M. Heo, K. Hsu, J. Hu, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O’Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.
- [9] S. Dasari, M. K. Srirama, U. Jain, and A. Gupta. An unbiased look at datasets for visuo-motor pre-training. In *Conference on Robot Learning*, pages 1183–1198. PMLR, 2023.
- [10] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pages 416–426. PMLR, 2023.
- [11] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, D. Sadigh, C. Finn, and S. Levine. Octo: An open-source generalist robot policy. <https://octo-models.github.io>, 2023.
- [12] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *arXiv*, 2024.
- [13] O. X.-E. Collaboration, A. O’Neill, A. Rehman, 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. 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. Stulp, G. Zhou, G. S. Sukhatme, G. Salhotra, G. Yan, G. Feng, G. Schiavi, G. Berseth, G. Kahn, G. Wang, H. Su, H.-S. Fang, H. Shi, H. Bao, H. B. Amor, H. I. Christensen, H. Furuta, 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. 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. Boher, 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. 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. Vanhoucke, W. Zhan, W. Zhou, W. Burgard, X. Chen, X. Wang, X. Zhu, X. Geng, X. Liu, X. Liangwei, X. Li, Y. Lu, Y. J. Ma, Y. Kim, Y. Chebotar, Y. Zhou, Y. Zhu, Y. Wu, Y. Xu, Y. Wang, Y. Bisk, 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, and Z. Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. <https://arxiv.org/abs/2310.08864>, 2023.
- [14] S. Belkhale, Y. Cui, and D. Sadigh. Data quality in imitation learning. *Advances in Neural Information Processing Systems*, 36, 2024.
 - [15] G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv: Arxiv-2305.16291*, 2023.
 - [16] C. Lynch and P. Sermanet. Language conditioned imitation learning over unstructured data. *Proceedings of Robotics: Science and Systems (RSS)*, 2021.
 - [17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
 - [18] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.
 - [19] A. Rashid, S. Sharma, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg. Language embedded radiance fields for zero-shot task-oriented grasping. In *Conference on Robot Learning*, pages 178–200. PMLR, 2023.
 - [20] Y. J. Ma, W. Liang, V. Som, V. Kumar, A. Zhang, O. Bastani, and D. Jayaraman. Liv: Language-image representations and rewards for robotic control. *arXiv preprint arXiv:2306.00958*, 2023.
 - [21] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang. Language-driven representation learning for robotics. In *Robotics: Science and Systems (RSS)*, 2023.
 - [22] S. A. Sontakke, J. Zhang, S. Arnold, K. Pertsch, E. Biyik, D. Sadigh, C. Finn, and L. Itti. Roboclip: One demonstration is enough to learn robot policies. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
 - [23] S. Mirchandani, S. Karamcheti, and D. Sadigh. Ella: Exploration through learned language abstraction. *Advances in neural information processing systems*, 34:29529–29540, 2021.
 - [24] J. Hejna, P. Abbeel, and L. Pinto. Improving long-horizon imitation through instruction prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7857–7865, 2023.
 - [25] M. Srivastava, C. Colas, D. Sadigh, and J. Andreas. Policy learning with a language bottleneck. *arXiv preprint arXiv:2405.04118*, 2024.
 - [26] N. D. Palo and E. Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
 - [27] C. Wen, X. Lin, J. So, K. Chen, Q. Dou, Y. Gao, and P. Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.
 - [28] M. Vecerik, C. Doersch, Y. Yang, T. Davchev, Y. Aytaç, G. Zhou, R. Hadsell, L. Agapito, and J. Scholz. Robotap: Tracking arbitrary points for few-shot visual imitation. *arXiv*, 2023.
 - [29] P.-C. Ko, J. Mao, Y. Du, S.-H. Sun, and J. B. Tenenbaum. Learning to Act from Actionless Video through Dense Correspondences. *arXiv:2310.08576*, 2023.

- [30] H. Bharadhwaj, R. Mottaghi, A. Gupta, and S. Tulsiani. Track2act: Predicting point tracks from internet videos enables diverse zero-shot robot manipulation, 2024.
- [31] Y. Zhu, A. Lim, P. Stone, and Y. Zhu. Vision-based manipulation from single human video with open-world object graphs. *arXiv preprint arXiv:2405.20321*, 2024.
- [32] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, and D. Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8121–8130, 2022.
- [33] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [34] M. Reuss, M. Li, X. Jia, and R. Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. *arXiv preprint arXiv:2304.02532*, 2023.
- [35] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *arXiv preprint arXiv:2108.03298*, 2021.
- [36] T. Robotics. Viperx 300 robot arm, . URL <https://www.trossenrobotics.com/viperx-300-robot-arm.aspx>.
- [37] T. Robotics. Widowx 250 robot arm, . URL <https://www.trossenrobotics.com/widowx-250-robot-arm-6dof.aspx>.
- [38] Y. Cui, S. Niekum, A. Gupta, V. Kumar, and A. Rajeswaran. Can foundation models perform zero-shot task specification for robot manipulation? In *Learning for dynamics and control conference*, pages 893–905. PMLR, 2022.
- [39] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pages 892–909. PMLR, 2023.
- [40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [41] 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.