

APPENDIX OF DIFFUSION TRANSFORMER POLICY

Anonymous authors

Paper under double-blind review

A ADDITIONAL TRAINING AND MODEL DETAILS

The structure of our Model is shown in Figure 2. In this section, we will talk about some extra details of our model. The language instruction is encoded by a pretrained clip model and freeze the encoder in the training loop. We then resize the input images into 224*224 and feed it into a pretrained ViT model. The selected ViT is the base version of DinoV2. All the parameters in DinoV2 are trained. After the above process, we use a Q-Former to decrease the size of image features. The Q-Former is from scratch with a depth of 4. In each block, we insert the text token as a FiLM Condition to get the image features containing language information. The query length of the image features will reduce to 32 when out of the Q-Former. Then, we concatenate the processed text features and image features, together with the action noised by a DDPM scheduler with 100 time-step. The multimodal inputs then pass through a causal Transformer and predict the added noise, which will execute on the robot arm after a series of post-processing. The Transformer is a from scratch Llama2-type model with 12 self-attention blocks. The hidden size is set to 768. All the modules mentioned are trained except the text encoder of clip. In summary, we have 334M parameters in total and 221M trained. This is pioneering to get this performance with such a small-sized model.

B ADDITIONAL DETAILS ABOUT PRETRAINING DATA

We choose 15 large datasets from Open X-Embodiment Padalkar et al. (2023) as illustrated in Table 1. We mainly follow Team et al. (2024); Kim et al. (2024) to set the weights. Following Kim et al. (2024), we further resize the image to the size of 224.

fractal Brohan et al. (2022)	DobbE Shafiuallah et al. (2023)	Droid Khazatsky et al. (2024)
16.15	1.94	13.69
Robo Set Vikash Kumar	Viola Zhu et al. (2023b)	Kuka Kalashnikov et al. (2018)
2.99	1.30	17.47
BridgeV2 Walke et al. (2023)	NYU Franka Cui et al. (2022)	Furniture Bench Heo et al. (2023)
21.86	1.14	6.73
StanfordHydra Belkale et al. (2023)	DLR EDAN Quere et al. (2020)	BerkeleyFanuc Zhu et al. (2023a)
6.11	0.08	1.07
Jaco Dass et al. (2023)	LanguageTable Lynch et al. (2023)	toto Zhou et al. (2023)
0.67	6.01	2.78

Table 1: The training dataset mixture

C ANALYSIS IN CALVIN

As illustrated in the main paper, we use a common learning rate scheduler to decay the learning rate in the experiments in Calvin, rather than a fixed learning rate of 0.0001 in our pre-training stage. We demonstrate that this can slightly improve the performance in Table 2.

D VISUALIZATION ANALYSIS

D.1 ZERO-SHOT GENERALIZATION

Figure 1 presents the proposed method is able to grasp the object, while OpenVLA and Octo fails. We observe OpenVLA and Octo fail to rightly approach the right grasp position. This experiment

Table 2: The Ablation of learning rate scheduler on Calvin Benchmark.

strategy	No. Instructions in a Row (1000 chains)					
w lr decay	94.5%	82.5%	72.8%	61.3%	50.0%	3.61
w/o lr decay	91.8%	80.0%	68.0%	56.9%	45.9%	3.43



Figure 1: Visualization of zero-shot generalization of different models. The first row is Diffusion Transformer Policy(ours). The second row of demonstration is OpenVLA Kim et al. (2024), the third row is Octo Team et al. (2024). Our Model can complete the tasks successfully while both of the others fails.

demonstrates the proposed Diffusion Transformer structure achieves more robust policy learning compared to discretized actions or diffusion action head. The denoising transformer model has built a better mapping from the image observation to corresponding action chunks.

D.2 FAILURE CASES

We visualize some failure cases and make some further research to explore why model fails on the cases. Figure 2 shows the visualizations of some of these cases. We realize that, models can localize the object and move forward to it in nearly all of the cases, but it is in trouble with predicting a correct pose for the end-effector to grasp the object. Most of the cases fails due to the indiscernible differences on the image or the leak of depth information sometimes. We will explore how to increase the fineness of the prediction for the grasp pose and add the depth information to the model in the future.

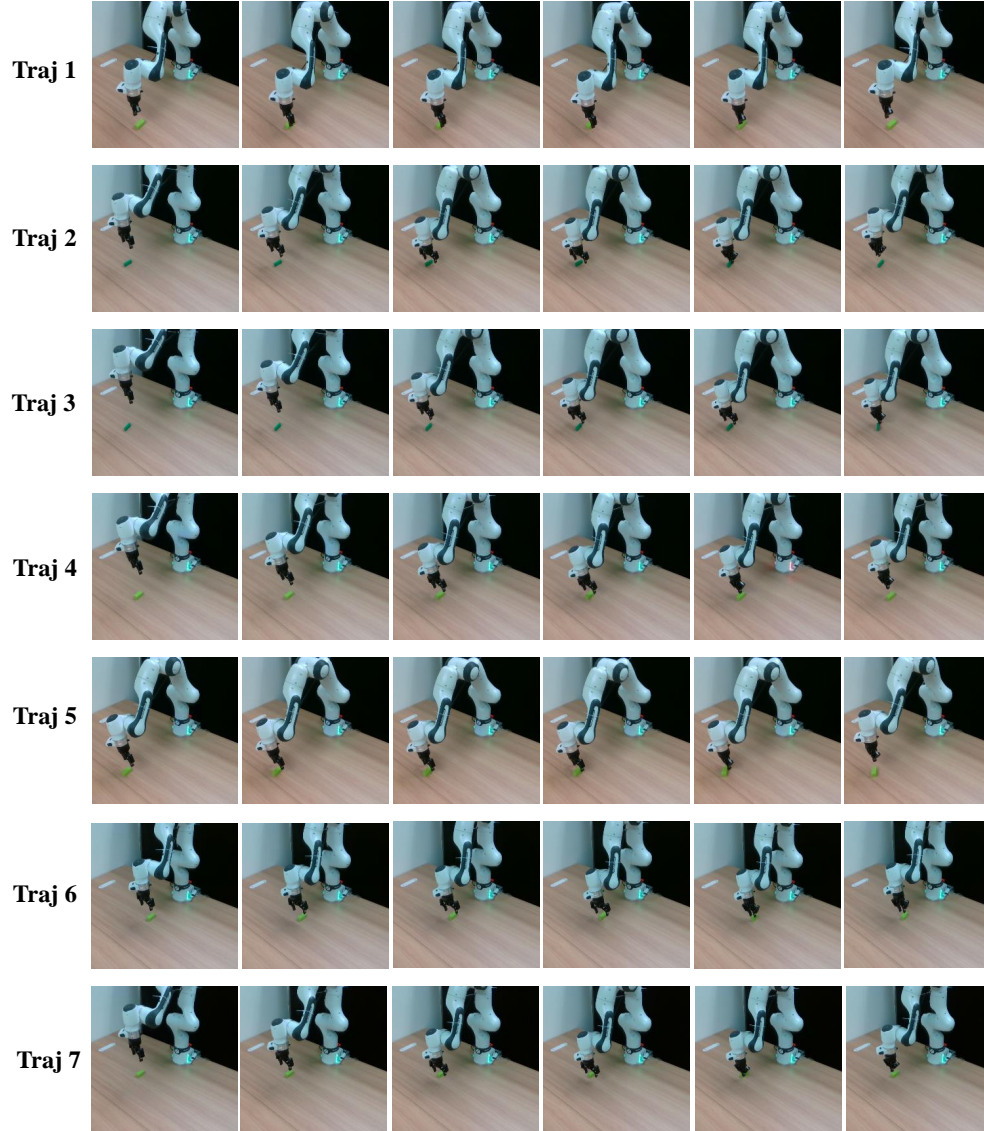


Figure 2: Failure examples on Real Robot. Models can not predict a totally correct pose for grasping. Many trajectories fail due to indiscernible differences or the lack of depth information.

REFERENCES

- Suneel Belkale, Yuchen Cui, and Dorsa Sadigh. Hydra: Hybrid robot actions for imitation learning. In *Conference on Robot Learning*, pp. 2113–2133. PMLR, 2023.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.
- Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncured robot data. *arXiv preprint arXiv:2210.10047*, 2022.
- Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023. URL https://github.com/clvr-ai/clvr_jaco_play_dataset.
- Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *arXiv preprint arXiv:2305.12821*, 2023.
- Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation, 2018. URL <https://arxiv.org/abs/1806.10293>.
- Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.
- Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023.
- Abhishek Padalkar, Acorn Pooley, Ajinkya Jain, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anikait Singh, Anthony Brohan, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023.
- Gabriel Quere, Annette Hagenruber, Maged Iskandar, Samuel Bustamante, Daniel Leidner, Freek Stulp, and Jörn Vogel. Shared control templates for assistive robotics. In *2020 IEEE international conference on robotics and automation (ICRA)*, pp. 1956–1962. IEEE, 2020.
- Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.
- Gaoyue Zhou Vincent Moens Vittorio Caggiano Jay Vakil Abhishek Gupta Aravind Rajeswaran Vikash Kumar, Rutav Shah. Robohive – a unified framework for robot learning.
- Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pp. 1723–1736. PMLR, 2023.
- Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, et al. Train offline, test online: A real robot learning benchmark. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9197–9203. IEEE, 2023.

Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot, 2023a.

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pp. 1199–1210. PMLR, 2023b.