Supplementary: Bimanual Dexterity for Complex Tasks

Anonymous Author(s)
Affiliation
Address
email

1 Detailed Cost Analysis

Please see Table 1 and Table 2 for a detailed Bill of Materials and breakdown of the cost to create BiDex. This is accurate pricing as of the paper submission. While we assert that BiDex is low cost, we acknowledge that it is still not affordable for everyone such as hobbyists. We believe that the price of motion capture gloves will continue to decrease over time as technology improves and demand increases in our field as well as other adjacent fields.

Object	Quantity	Total
Manus Meta Glove	1	$6000
Dynamixel XL330-M288	12	$300
U2D2 Control PCB	1	$20
5v 20A Power Supply	2	$25
14 AWG Cabling	1	$20
PLA Printer Plastic	N/A	$10
Total		$6375

Table 1: We present the bill of materials of BiDex for two arms and hands. The total cost is around $6000, mostly due to the Manus Meta gloves.

Object	Quantity	Total
xArm 6	2	$18000
Ubuntu Laptop	1	$2000
Mobile Base	1	$6000
Zed Camera	3	$1200
LEAP Hand or DLA Hand	2	$4000
Total		$31,2000

Table 2: We present the bill of materials of the mobile robot setup. The robot and BiDex costs around $35,000 which we believe is reasonable for a dexterous bimanual robot hand setup with 50+ degrees of freedom.

2 Assembly Instructions and Software

The assembly instructions will be released upon acceptance of the paper at our website. The hardware system and software will be a useful to recreate BiDex and create variants of it using high quality motion capture gloves. Our high quality teacher arm teleoperation is based off of [1] but the strength is increased to allow for the weight of the gloves and its mounting system.

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

3 About Manus Glove

We use the Manus Meta Quantum Metagloves [2] which is an $6000 tracking Mocap glove. Each finger is tracked by the glove and returns the fingertip positions as xyz-quaterion and also 4 different angles for each finger $\theta_{\mathsf{MCP}_{\mathsf{side}}}, \theta_{\mathsf{MCP}_{\mathsf{fwd}}}, \theta_{\mathsf{PIP}}, \theta_{\mathsf{DIP}}$ using hall effect sensors with very high accuracy and at 120hz. We use their Windows API (Linux is not available at time of release) and will release our version of that which sends the software to a Linux machine running ROS. These gloves are available for purchase at `https://www.manus-meta.com/`.

4 SteamVR Baseline

For the wrist tracking SteamVR baseline, we use the Manus Meta SteamVR trackers which connect to the gloves and seamlessly route the data through the aforementioned Windows API. They are wireless but require SteamVR Lighthouses setup around the perimeter of the workspace. In our test we mount the 4 SteamVR trackers on the ceiling to avoid as many occlusions as possible. We also mount the 4 trackers in a 16ft square around where the teleoperator would stand which is the recommended configuration. We will release this code for others to recreate in their comparison study.

5 Apple Vision Pro Baseline

The Apple Vision Pro baseline is based off of [3]. With this data, we control the hand using the same inverse kinematics as with the Manus Glove. For the arm, we scale, translate and rotate for the robot embodiment and then pass through inverse kinematics to control the arm. Our accompanying code for our two xArms will be available on our website upon accepance of the paper.

6 Behavior Cloning Policy Architecture and Hyperparameters

We illustrate our policy architecture in Figure 1. Our behavior cloning policy takes as input a RGB image and current hand joint angles (proprioception). We obtain tokens for the image observation via a ViT [4] and a token for joint proprioception via a linear layer. The weights of ViT is initialized from the Soup 1M model from [5]. The tokens then pass through a action chunking transformer [6], a encoder-decoder transformer, to output a sequence of actions. The action space is the absolute joint angles of two arms and two hands. A key decision that greatly improves policy generalization is to exclude current arm joints from the proprioception. Intuitively, this may force the model to extract object information from image observations, rather than overfitting to predict actions close to current arm states.

We list key hyperparameters for our behavior policy training Table 3. In general, we are able to obtain well-performing policies with 20-50 demonstrations and 1 hour of wall-clock time training on a RTX4090. With our easy-to-use teleoperation system, we are able to obtain diverse policies for complex bimanual dexterous tasks quickly.

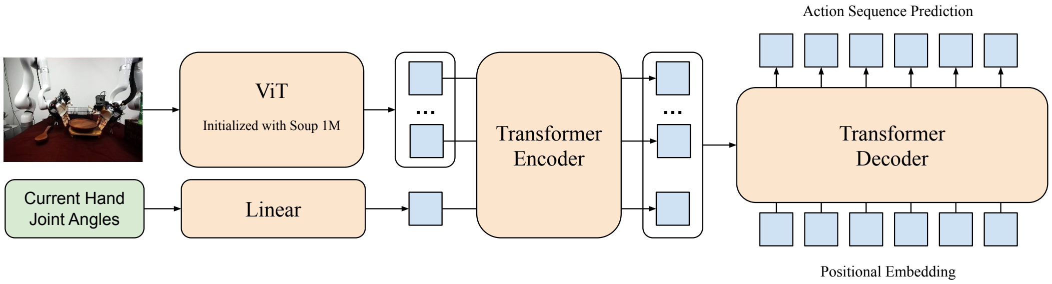

Figure 1: Behavior Cloning Policy Architecture

Hyperparameter	**Value**
Behavior Policy Training	
optimizer	AdamW
base learning rate	3e-4
weight decay	0.05
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
batch size	64
learning rate schedule	cosine decay
total steps	10000
warmup steps	500
augmentation	GaussianBlur, Normalize, RandomResizedCrop
GPU	RTX4090 (24 gb)
Wall-clock time	$\sim$ 1 hour
Visual Backbone ViT Architecture	
patch size	16
#layers	12
#MHSA heads	12
hidden dim	768
class token	yes
positional encoding	sin cos
Action Chunking Transformer Architecture	
# encoder layers	6
# decoder layers	6
#MHSA heads	8
hidden dim	512
feedforward dim	2048
dropout	0.1
positional encoding	sin cos
action chunk	100

Table 3: Hyperparameters for Behavior Cloning Policy Training

References

[1] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. *arXiv preprint arXiv:2309.13037*, 2023.

[2] Manus meta. https://www.manus-meta.com/.

[3] Y. Park and P. Agrawal. Using apple vision pro to train and control robots, 2024. URL https://github.com/Improbable-AI/VisionProTeleop.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] 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.

[6] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.