Technical Report of HCB-Hand Team for Dexterous HO Tracker Challenge on HANDS 2025 Challenge

Hang Xu[1], Yang Xiao[1†], Changlong Jiang[1], Haohong Kuang[1], Yangfan Deng[1], Qihang Zhou[1]
Ran Wang[2,3], Min Du[4], Zhiwen Fang[5,6], Zhiguo Cao[7], Joey Tianyi Zhou[8,9]
[1]National Key Laboratory of Multispectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2]School of Journalism and Information Communication, Huazhong University of Science and Technology, Wuhan 430074, China
[3]School of Future Technology, Huazhong University of Science and Technology, Wuhan 430074, China
[4]PICO, ByteDance, Beijing 100089, China
[5]Guangdong Provincial Key Laboratory of Medical Image Processing,
School of Biomedical Engineering, Southern Medical University, Guangzhou 510515, China
[6]Department of Rehabilitation Medicine, Zhujiang Hospital, Southern Medical University, Guangzhou 510280, China
[7]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[8]Centre for Frontier AI Research, Agency for Science, Technology and Research (A*STAR), Singapore
[9]Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore
{hang_xu, Yang_Xiao, cljiang, haohong_kuang,
yangfan_deng, qhzhou, rex_wang, zgcao}@hust.edu.cn,
bingwen.ai@bytedance.com, fzw310@smu.edu.cn, joey_zhou@a-star.edu.sg

Abstract

In this report, we introduce the method proposed for the Dexterous HO Tracker challenge in HANDS@ICCV2025. The challenge aims to transfer human-recorded hand-object interaction (HOI) motion capture sequences to dexterous robotic hands in simulation. To overcome the limitations and high learning difficulty of a single policy network, as used in the baseline ManipTrans, we employ a Mixture of Experts (MoE) architecture. Specifically, our model comprises 4 sparse experts and 1 shared expert. A router network is designed to process the current observation to activate 1 sparse expert, enabling different experts to specialize in specific observations and decomposing the complex problem into more specialized sub-problems. On official evaluation sequences, our method achieves a success rate of 76.2% on single-hand tasks and 52.8% on bimanual tasks, securing the first place in the challenge.

1. Introduction

The field of embodied AI has developed rapidly in recent years. Recognizing that human hands are a primary medium for interaction with the physical world, many current studies [1] now focus on dexterous robotic hand manipulation. This growing research trend has, in turn, created a significant demand for large-scale, high-precision datasets that capture human-like dexterous manipulations. While some studies rely on reinforcement learning (RL) [9] or teleoperation [3] to generate dexterous manipulation sequences, an alternative choice is to transfer human motion capture data to dexterous robotic hands via imitation learning [6, 7]. The latter methods show significant promise, driven by the increasing availability of hand motion capture datasets and advances in hand pose estimation technology [4, 5].

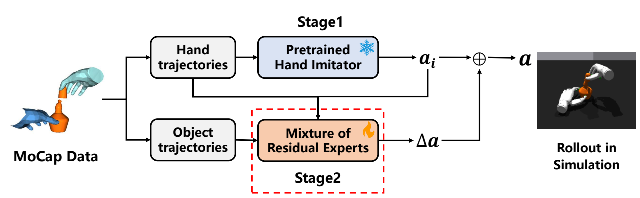

Figure 1. Our main framework follows the baseline method ManipTrans, which decomposes the transfer task into two stages. Compared with ManipTrans, we use the Mixture of Residual Experts instead of the single policy network in the second stage.

ManipTrans [6] is a powerful method for efficiently transferring human bimanual skills to simulated dexterous

[†]Yang Xiao is the corresponding author (Yang_Xiao@hust.edu.cn).

robotic hands, adopting a two-stage process. The first stage involves pre-training a robust generalist hand imitator using only hand trajectories. This imitator learns to accurately mimic human finger motions with resilience to noise and provides the initial actions. In the subsequent fine-tuning stage, the hand imitator is frozen and a residual policy network is trained. This network takes joint hand-object observations as input to learn interaction dynamics and outputs residual actions to refine the initial actions.

Although this two-stage approach partially alleviates the difficulty of learning complex hand-object interaction patterns, we believe that a single policy network still struggles to handle the full spectrum of complex observations. Therefore, we introduce the core of our solution: a Mixture of Residual Experts (MoRE). This architecture decomposes the complex policy learning task into more specialized subproblems, thereby alleviating the overall learning difficulty. Ultimately, our approach proves to be effective, securing first place in the Dexterous HO Tracker challenge.

2. Method

Our main pipline is shown in Fig. 1, which is mainly based on ManipTrans. We use the pre-trained hand imitator provided by ManipTrans and focus on learning the residual action in the second stage. Specifically, we design a Mixture of residual experts structure to replace the original single policy network, to alleviate its learning difficulty.

The structure of MoRE is shown in Fig. 2, which includes 4 sparse experts E_k, where $k \in [0, 3]$ denotes the expert id, 1 shared expert E_{shared}, and a router network that is designed to process the current observation to activate 1 sparse expert, enabling different experts to specialize in specific observations. The input of MoRE is the same as that of ManipTrans, denoted as obs, which is first fed into an obs encoder composed of MLP to extract the observation features f_{obs}:

$$f_{obs} = encoder(obs). \tag{1}$$

Then, the observation features are fed into the router network to generate the activation weights w_k for each sparse expert:

$$w_k = router(f_{obs}). \tag{2}$$

To ensure sparse expert activation and that only the most relevant Top-K experts contribute to the output, the weights of experts outside the Top-K are set to zero:

$$w'_k = \begin{cases} w_k, & \text{if } w_k \in TopK(w) \\ 0, & \text{otherwise} \end{cases}. \tag{3}$$

The module's output y is calculated as the weighted sum over all outputs of activated experts and the shared expert:

$$y = \sum_{k=0}^{3} w'_k \cdot E_k(x) + E_{shared}(x), \tag{4}$$

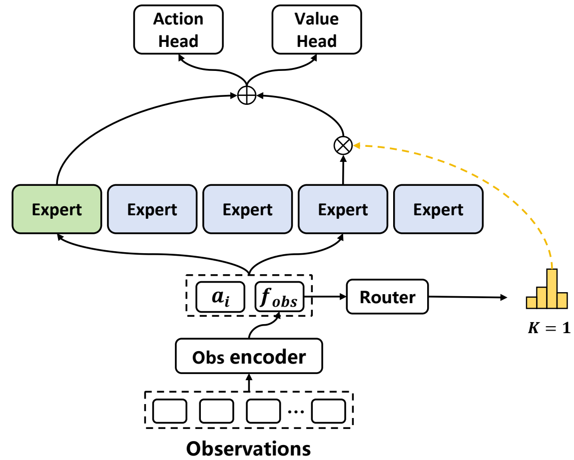

Figure 2. The structure of Mixture of Residual Experts. It contains 1 shared expert and 4 sparse experts, which are activated by a router network driven by the observations.

$$x = concat(f_{obs}, a_i), \tag{5}$$

where a_i denotes the initial action generated by the hand imitator in the first stage. Finally, we use an action head to predict the distribution mean of the residual action Δa and a value head to predict the value.

To maintain the load balance among experts and prevent expert degradation, we use a load balancing loss introduced by [2]. In addition, we also penalize the routing entropy of samples to enhance the determinism of the routing network:

$$L_{entropy} = -\lambda \sum_{n=1}^{N} w_n log(w_n), \tag{6}$$

where w_n denotes the activation weights of each sample.

3. Experiment

Our training strategy mainly follows ManipTrans. We train the MoRE using the Actor-Critic PPO algorithm [8], with a training horizon of 32 frames, a minibatch size of 2048 and 1000 epochs on the official evaluation samples, which contain 36 single-hand tasks and 39 bimanual tasks. All experiments are conducted in Isaac Gym, simulating 6144 environments on an NVIDIA RTX 3090 GPU. Optimization employs Adam with an initial learning rate of 5×10^{-4} and a decay scheduler.

	$E_r(\downarrow)$	$E_t(\downarrow)$	$E_j(\downarrow)$	$E_{ft}(\downarrow)$	$SR(\uparrow)$
Left	8.702	0.220	1.762	1.026	0.852
Right	9.807	0.323	2.429	1.650	0.731
Bimanual	9.470	0.728	2.486	1.626	0.528
Overall	9.498	0.490	2.360	1.550	0.762/0.528

Table 1. Performance on HANDS@ICCV2025 HO-Tracker challenge evaluation sequences.

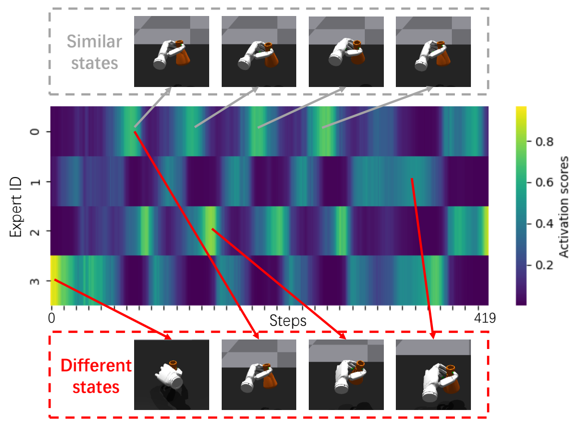

Figure 3. The expert activation patterns in our MoE architecture. It can be seen that: (1) The experts are used relatively evenly overall. (2) The activated experts correspond to higher scores, showing that the router makes confident selections. (3) Each expert focuses on similar types of observations, and different experts handle different states, highlighting their specialization.

We conduct our experiments on HANDS@ICCV2025 HO-Tracker challenge evaluation sequences. The results are shown in Table 1. It can be observed that our method achieves a success rate of 76.2% on single-hand tasks and 52.8% on bimanual tasks, which exceeds the official baseline and proves the effectiveness of our method.

We also further visualize and analyze the expert activation patterns on the sequence 39e5e@5, as shown in Fig. 3. This analysis yields three key observations. First, the utilization of each expert is relatively even across the rollout steps, demonstrating that our load balancing loss effectively prevents expert degradation. Secondly, the activated experts consistently correspond to higher weights, indicating that the router network selects experts with high confidence. This behavior, reinforced by our entropy penalty, aligns with the characteristics of PPO's online learning, where the model progressively learns a more deterministic and successful strategy. Finally, the visualization confirms the core hypothesis of our method: each expert focuses on similar types of observations, while different experts specialize in handling distinct states, highlighting their specialization.

4. Conclusion

In this report, we present our solution for the Dexterous HO Tracker challenge. We replace the baseline's single policy with a Mixture of Residual Experts to better handle complex HOI dynamics, decomposing the task and enabling experts to specialize in different observations. Our approach achieves 76.2% (single-hand) and 52.8% (bimanual) success rates. The Analysis confirmed MoRE's effectiveness, showing both expert specialization and balanced utilization.

A primary limitation is the significant performance gap between single-hand and bimanual tasks. This discrepancy is because our MoRE is a general framework designed to decompose observations and lacks specialized designs to address the unique complexities of bimanual coordination.

5. Acknowledgement

This work was jointly supported by the National Natural Science Foundation of China under Grant Nos. 62271221 and 62371219, the Taihu Lake Innovation Fund for Future Technology, Huazhong University of Science and Technology (HUST) under Grant 2023-B-8, and Guangdong Basic and Applied Basic Research Foundation under Grant Nos. 2024A1515140070 and 2023A1515011260.

References

[1] Yuanpei Chen, Chen Wang, Yaodong Yang, and C Karen Liu. Object-centric dexterous manipulation from human motion data. *arXiv preprint arXiv:2411.04005*, 2024. 1

[2] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024. 2

[3] Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024. 1

[4] Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, and Joey Tianyi Zhou. A2j-transformer: Anchor-to-joint transformer network for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8846–8855, 2023. 1

[5] Changlong Jiang, Yang Xiao, Jinghong Zheng, Haohong Kuang, Cunlin Wu, Mingyang Zhang, Zhiguo Cao, Min Du, Joey Tianyi Zhou, and Junsong Yuan. 3d hand pose estimation via articulated anchor-to-joint 3d local regressors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[6] Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6991–7003, 2025. 1

[7] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 1

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[9] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3891–3902, 2023. 1