

---

# Openpi Comet: Competition Solution For 2025 BEHAVIOR Challenge

Team Comet



[https://huggingface.co/sunshk/comet\\_submission](https://huggingface.co/sunshk/comet_submission)



<https://github.com/mli0603/openpi-comet>

## Abstract

The 2025 BEHAVIOR Challenge is designed to rigorously track progress toward solving long-horizon tasks by physical agents in simulated environments. BEHAVIOR-1K focuses on everyday household tasks that people most want robots to assist with and these tasks introduce long-horizon mobile manipulation challenges in realistic settings, bridging the gap between current research and real-world, human-centric applications. This report presents our solution to the 2025 BEHAVIOR Challenge in a very close 2nd place and substantially outperforms the rest of the submissions. Building on  $\pi_{0.5}$ , we focus on systematically building our solution by studying the effects of training techniques and data. Through careful ablations, we show the scaling power in pre-training and post-training phases for competitive performance. We summarize our practical lessons and design recommendations that we hope will provide actionable insights for the broader embodied AI community when adapting powerful foundation models to complex embodied scenarios.

## 1 Introduction

Vision-Language-Action (VLA) models (Brohan et al., 2022; 2023; Octo Model Team et al., 2024; Kim et al., 2024; Qu et al., 2025; Black et al., 2024) have recently emerged as a unifying paradigm for robotic policy learning, leveraging large-scale robot datasets to acquire robust and generalizable manipulation and navigation capabilities. By integrating perception, language understanding, and control within a single end-to-end framework, VLAs bypass the need for hand-engineered modules and have demonstrated strong performance across a variety of embodied AI benchmarks. Despite this progress, most existing VLA systems are primarily optimized for short-horizon tasks, and their ability to scale to complex, temporally extended activities remains limited.

Long-horizon manipulation (Zhao et al., 2025; Zawalski et al., 2024) introduces additional difficulties that fundamentally challenge current VLA designs. Such tasks require orchestrated sequences of interdependent behaviors, where compounding errors and shifting state distributions can degrade performance over time. A common approach is to decompose tasks into subtasks (Lin et al., 2022; Shi et al., 2023; Tie et al., 2025) and train separate local policies. However, this strategy does not resolve the skill chaining problem (Chen et al., 2024; Konidaris & Barto, 2009), which involves modeling and executing reliable transitions between subtasks while mitigating error accumulation. In addition, many solutions proposed for skill chaining rely on online adaptation or modular architectures, and these methods are often incompatible with the large-scale, offline, end-to-end training paradigm that underpins modern VLA models. Consequently, achieving reliable long-horizon performance while preserving scalability and generality remains an open challenge.

The BEHAVIOR Challenge, built upon the BEHAVIOR-1K (Li et al., 2024) benchmark, provides an rigorous benchmark for this problem. It features realistic household environments containing complex object interactions, and evaluates agents on 50 long-horizon tasks that reflect human-centered daily activities. Each task requires multi-step reasoning, precise manipulation, and coordinated navigation, making success highly dependent on robust long-horizon policy execution. With 10,000 expert demonstrations and a standardized evaluation protocol, the challenge places strong emphasis on generalization, control robustness, and error tolerance. These capabilities remain difficult for current VLA models to achieve consistently.

In this report, we examine how far a strong publicly available VLA backbone can be pushed on long-horizon tasks using careful data, training, and inference design within a simple end-to-end training pipeline. We treat the BEHAVIOR Challenge as a case study in adapting powerful but generic foundation policies to a complex embodied benchmark. Through systematic exploration of training configurations, pre-training choices, and inference strategies, we show that our solution completes 22 tasks out of the 50 household tasks, achieving a Q-score of 0.2514 in the competition [Table 1](#).

Table 1: Results of 2025 BEHAVIOR Challenge for standard track. Q-score for the test set (**bold**) is used for final ranking.

Rank	Team	Full Task Success Rate		Q-Score	
		Validation	Test	Validation	Test
1	Robot Learning Collective	0.1120	0.1240	0.2605	<b>0.2599</b>
2	Comet (ours)	<b>0.1440</b>	<b>0.1140</b>	<b>0.1830<sup>1</sup></b>	<b>0.2514</b>
3	SimpleAI Robot	0.1400	0.1080	0.1943	<b>0.1591</b>
4	The North Star	0.1280	0.0760	0.1702	<b>0.1204</b>

## 2 Architecture

As shown in [Figure 1\(a\)](#), we adopt the  $\pi_{0.5}$  as the base policy of our system.  $\pi_{0.5}$  follows the standard VLA design paradigm, combining a visual encoder for multi-view robot observations with a language encoder for task instructions, and fusing these modalities into a shared representation that conditions the action expert. The action expert is implemented as a transformer-style network that ingests features and denoises the low-level continuous control actions at each timestep. This end-to-end architecture allows perception, language understanding, and control to be trained jointly from large-scale robot datasets, and  $\pi_{0.5}$  further enhances generalization by being pretrained on heterogeneous data spanning multiple embodiments, environments, and tasks. The dataset includes the **1k hours human demonstrations officially provided by the BEHAVIOR Challenge**, as well as **our additional motion-planner trajectories and offline RL rollouts**, which together supply rich long-horizon behaviors and diverse manipulation strategies crucial for robust policy learning.

For post-training, we adopt an iterative RFT procedure as illustrated in [Figure 1\(b\)](#). Starting from the official human demonstrations, we introduce random pose perturbations and roll out the pretrained policy under these disturbed initial conditions. Successful episodes are retained as additional training data, progressively forming an offline data flywheel that continually improves the robustness and coverage of the policy.

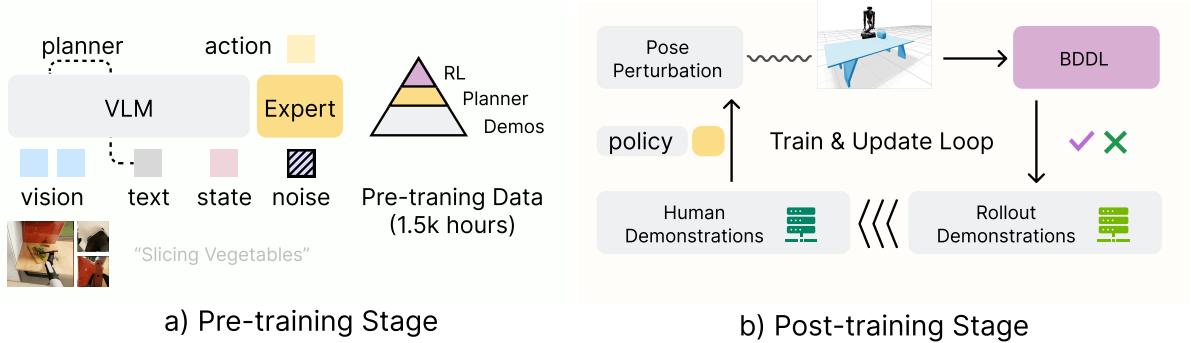


Figure 1: (a) Pre-training on large-scale heterogeneous data, including 1.1K hours of human demonstrations and  $\sim 0.4$ K hours of additional planner and offline RL trajectories. (b) RFT post-training: perturb initial poses, roll out the policy, and retain successful episodes to iteratively augment the dataset.

## 3 Dataset

The BEHAVIOR-1K benchmark provides 1,000 realistic household activities instantiated in 50 fully interactive 3D scenes with over 10k objects, designed to stress long-horizon mobile manipulation and high-level reasoning in human-centric environments. The NeurIPS 2025 BEHAVIOR Challenge selects 50 representative tasks from BEHAVIOR-1K and supplies 10,000 teleoperated expert demonstrations (200 per task, over 1,200 hours) with multi-modal observations and fine-grained skill annotations. Policies are evaluated in simulation by the task success rate defined via BDDL goal predicates, emphasizing completion of entire activities rather than short-horizon subroutines.

Beyond the official teleoperated dataset, we further incorporate  $\sim 3.6$ K trajectories composed of motion-planner demonstrations and offline RL rollouts. The planner data provides precise low-noise manipula-

<sup>1</sup>Due to limited time by the submission deadline, we could not finish evaluation across all tasks on time. Thus, our submitted validation Q-score is lower than the actual validation Q-score reported in [Table 2](#).

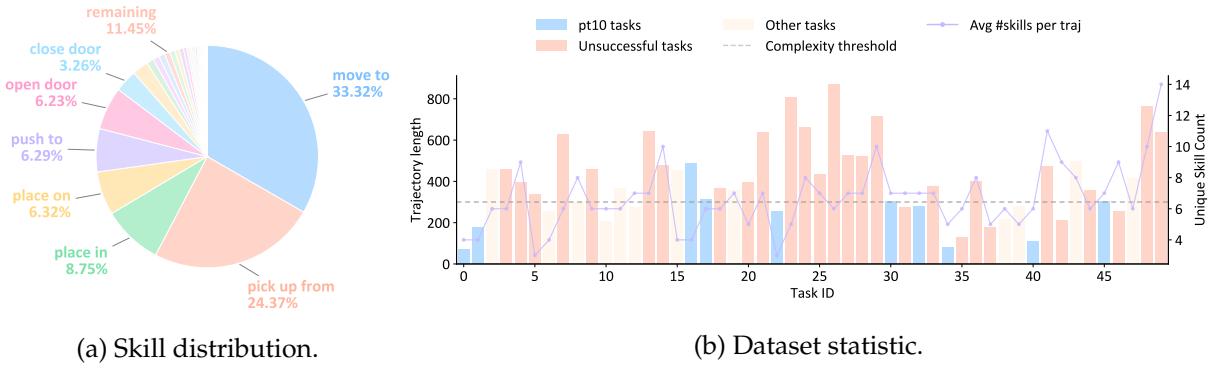


Figure 2: Dataset statistics and skill distribution for BEHAVIOR Challenge @ NeurIPS 2025. (a) Proportion of video frames occupied by each skill across the entire dataset. (b) Per-task distribution, showing the average trajectory length (in frames) and the average number of unique skills per trajectory.

tion sequences, while the RL rollouts introduce broader behavioral variability. Together, these additional trajectories substantially enrich the state–action coverage beyond human demonstrations.

To better understand the learning landscape, we analyze the demonstration dataset at the level of semantic skills and per-task complexity. As shown in Figure 2(a), the distribution is highly imbalanced: `move to` and `pick up from` dominate with roughly 33.3% and 24.4% of frames, followed by `place in` (8.8%) and a long tail of infrequent skills (11.5%). Figure 2(b) reports per-task statistics: many tasks require trajectories of several hundred frames and typically involve 5–10 distinct skills, with some exceeding 12. We treat tasks with an average length below 250 frames as relatively simple, since they involve shorter horizons and fewer skill compositions, and they are particularly useful for quickly bringing up our system and validating basic policy behavior in the simulator. In contrast, several tasks, including Task 48 and Task 49, fall into a clearly harder regime, characterized by long-horizon execution and rich mixtures of navigation and manipulation skills (e.g., moving between rooms, opening/closing containers, and rearranging multiple objects). These high-complexity tasks provide a stringent testbed for evaluating policies’ ability to handle extended temporal credit assignment and frequent skill switching.

## 4 Experiment

In this section, we present details of our implementations (Section 4.1), pre-training (Section 4.2) and post-training (Section 4.3). We report the results across different stages in Table 2 and present detailed ablations in Section 4.4.

### 4.1 Implementation details

We adopt  $\pi_{0.5}$  implemented in JAX as our policy, and all experiments are conducted on NVIDIA H200 GPUs with a per-device batch size of 64, using cosine decay scheduler. For single task SFT, we train for 15k–20k steps. For all multi-task pre-training, we use 50k training steps, with a learning rate of  $2.5 \times 10^{-5}$ . During both SFT and the subsequent RFT adaptation, we use a reduced learning rate of  $2.5 \times 10^{-6}$  on 8 GPUs to ensure stable fine-tuning.

To accelerate offline rollout generation, we parallelize evaluation across the 10 test instances by distributing them over multiple GPUs, which mitigates the extremely slow simulation rate of BEHAVIOR—where evaluating a single task can otherwise take from one hour to nearly a full day, making online RL potentially inefficient under this simulator Yu et al. (2025).

### 4.2 Pre-training

To study how pre-training task coverage and diversity affect long-horizon performance, we compare single-task pre-training `pt1` with three multi-task pre-training settings on the BEHAVIOR Challenge: `#pt7`, `#pt10`, and `#pt50`. Each setting trains a single VLA policy on demonstrations from 7, 10, or all 50 challenge tasks. For each configuration, we train the model end-to-end on the corresponding task subset and then evaluate on the same 50-task challenge protocol. This design allows us to isolate how increasing the number and heterogeneity of pre-training tasks influences task success, while keeping the overall training recipe fixed.

As shown in Figure 3, pt1 trains the policy only on demonstrations from a single target task; this single-task finetuning regime yields the lowest average success and only produces successful rollouts on 2 tasks, indicating that purely task-specific adaptation is insufficient for robust long-horizon control. Building on this baseline, pt7 pre-trains on a small subset of relatively short-horizon BEHAVIOR Challenge tasks such as bringing water, cook hot dogs, and make microwave popcorn, which share similar interactions with common household objects. As the pre-training set expands from pt7 to pt10 and finally pt50, more tasks begin to exhibit successful rollouts, showing that broader task coverage improves the model’s ability to generalize. In particular, pt10 augments pt7 with additional, slightly more complex tasks (e.g., moving boxes to storage, hanging pictures), while pt50 uses demonstrations from all 50 challenge tasks, including rare and highly compositional activities such as rearranging kitchen furniture and setting the fire, thus exposing the policy to the full long-horizon distribution during pre-training.

After pre-training, we reach a validation Q-score of 0.19 on the validation set, as shown in Table 2.

### 4.3 Post-training

Despite the recent progress in online RL (Chen et al., 2025), its low sample efficiency makes it impractical for the BEHAVIOR challenge. Moreover, online RL requires a heterogeneous compute setup: GPUs with RT Cores are needed for simulation, while GPUs with Tensor Cores are needed for model training and rollouts. Given these constraints, we instead adopt rejection sampling finetuning (RFT), a technique shown to be effective in both LLMs and VLMs (Ahn et al., 2024; Touvron et al., 2023; Azzolini et al., 2025). An overview of our RFT setup is provided in Algorithm 1. Starting from all the provided train and validation human demonstrations for each scene, we randomly perturb the robot’s initial pose and use our pre-trained policy to perform rollouts under these perturbed configurations. Using both train and validation set helps avoid overfitting to the validation set and provides better signals of the model quality. Successful rollouts are retained as additional demonstrations. We perform  $N = 3$  rounds of RFT in total, collecting on average  $T = 8500$  trajectories per round, and eventually selected approximately 2500 trajectories for training after de-duplication and task balancing.

After post-training, our suite of models achieves a validation Q-score of 0.22. In addition to the effectiveness, RFT provides an estimate of the theoretical upper bound on model performance by computing the union of success instances across all past model checkpoints. As shown in Figure 4 and Table 2, the theoretical best validation Q-score we could achieve is 0.31. Due to time constraints, we were unable to fully close this gap, leaving additional headroom for further post-training. Yet, the theoretical best Q-score provides a tangible target of additional improvements. Please refer to Figure 5 for representative visualizations of policy executions results.

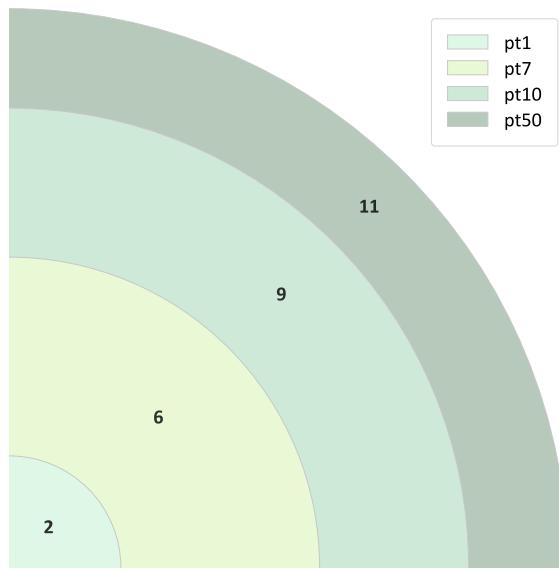


Figure 3: Training subset coverage and the number of successful tasks.

---

#### Algorithm 1 The RFT Algorithm

---

```

1: Initialize  $\mathcal{D} \leftarrow$  human demos.
2: Initialize  $\pi_1$  to pre-trained  $\pi_{pt}$ .
3: for  $i = 1$  to  $N$  do
4:    $\mathcal{D}_i \leftarrow \emptyset$ 
5:   for  $t = 1$  to  $T$  do
6:     Sample initial state  $s_0$  from  $\mathcal{D}$ .
7:      $s'_0 \leftarrow s_0 + \epsilon$ .
8:     Rollout  $\tau$  from  $s'_0$  using  $\pi_i$ .
9:      $\mathcal{D}_i \leftarrow \mathcal{D}_i \cup \mathcal{D}_\tau$  if successful.
10:  end for
11:   $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_i$ .
12:  Train  $\pi_{i+1}$  on  $\mathcal{D}$ .
13: end for
14: return best  $\pi_i$  on validation.

```

---

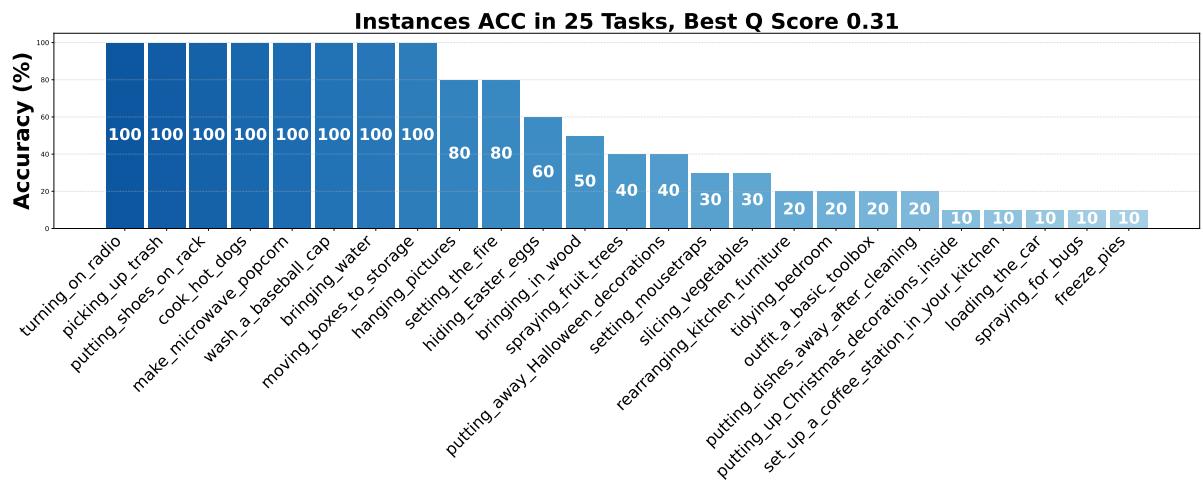


Figure 4: By selecting the best post-trained policy model *per task instance*, we can achieve an aggregated validation Q-score of 0.31, serving as the upper bound of our model’s performance.

Table 3: Ablation study of control mode, action horizon, input modality, and image resolution on the turning on radio task.

Settings	Control Mode	Action Horizon	Input Modality	Image Resolution	Success Rate
#1	Temporal Ensemble	50	RGB	Head:224	0.00
	Receding Temporal			Wrist:224	0.00
	Receding Horizon				<b>0.25</b>
#2	Receding Horizon	8	RGB		0.00
		16		Head:224	0.10
		50		Wrist:224	0.25
		32			<b>0.30</b>
#3	Receding Horizon	32	RGB+Depth Image	Head:224	0.20
			RGB+Point Cloud	Wrist:224	0.30
			RGB		<b>0.30</b>
#4	Receding Horizon	32	RGB	Head:224	0.30
				Wrist:224	
				Head:720	
				Wrist:480	<b>0.60</b>

#### 4.4 Ablations

Beyond pre-training and post-training, we find that low-level design decisions in training and inference have a substantial impact on the performance. Table 3 reports a set of controlled ablations along four such axes: Action Horizon and Input Modality for the training procedure, and Control Mode and Image Resolution for the inference strategy.

**Control mode (#1).** Temporal Ensemble and Receding Temporal fail to produce stable closed-loop behavior and result in near-zero success rates. In contrast, the Receding Horizon scheme, which executes all the predicted action segment and performs re-planning after finishing manipulation, significantly improves performance. This result highlights the necessity of continuous feedback for long-horizon manipulation and shows that smoothing or averaging open-loop predictions quickly accumulates error.

**Action horizon (#2).** Varying the action horizon reveals a non-monotonic relationship between prediction length and downstream control performance. Moderate horizons strengthen long-horizon manipulation by allowing the model to anticipate multi-stage behaviors, whereas excessively long horizons introduce conflicting temporal dependencies that compromise stable control. These findings indicate that the horizon length must be chosen to balance future awareness with the reliability of short-term control. We empirically find that setting the receding horizon to 32 gives us the best results.

**Input modality (#3).** We find that reconstructed point clouds improve performance compared to depth map as an input modality, which suggests that explicit geometric structure provides more informative cues for object-centric manipulation. However, the improvement over RGB-only inputs is limited, while the computational and latency overhead is considerable. Overall, explicit 3D geometry offers benefits in



Figure 5: Rollout Examples on BEHAVIOR-1K. Our policy successfully handles long-horizon, multi-stage household tasks involving navigation, fine manipulation, and tool use, demonstrating robust execution across diverse household activities.

Table 4: Ablation study of action representation, sampling frequency, proprioceptive state inputs, and skill weighting under default training settings in Table 3.

Settings	Action Representation	Action Sampling	State Input	Skill weighting	Success Rate
#5.1	Delta Joint Absolute Joint	30 Hz	✓	No Weighting	0.00 <b>0.30</b>
#5.2	Absolute Joint	15 Hz 30 Hz	✓	No Weighting	0.00 <b>0.30</b>
#5.3	Absolute Joint	30 Hz	✗ ✓	No Weighting	0.00 <b>0.30</b>
#5.4	Absolute Joint	30 Hz	✓	No Weighting Manip:Nav=2:1	<b>0.30</b> <b>0.30</b>

cluttered scenes, but the gains are not consistently large enough to justify the increased system complexity.

**Image resolution (#4).** Increasing the spatial resolution of both head and wrist views leads to substantial gains in performance. Low-resolution inputs of 224 by 224 pixels provide only coarse visual information, whereas high-resolution inputs more than double the success rate. These results indicate that precise visual cues are critical for reliable manipulation in visually complex environments, and that high-fidelity perception plays an essential role in long-horizon tasks.

**Data process (#5).** We further evaluate several data-processing strategies, including action representation, action subsampling, removal of proprioceptive state inputs, and skill weighting about Manipulation and Navigation in Table 4. We use default training setting as the folling: Receding Horizon, 32 action horizon, RGB input, and 224x224 image resolution. Results shows that relative-action parameterization and the removal of state both lead to worse results, suggesting that absolute action anchoring and explicit system observability are essential for stable closed-loop optimization. Action subsampling at 15 Hz accelerates execution but reduces temporal precision, leading to degraded manipulation accuracy. Finally, although reweighting the dataset toward manipulation segments shifts the empirical distribution, it yields no measurable improvement, indicating that simple resampling is insufficient to compensate for the structural difficulty and heterogeneous dynamics of long-horizon manipulation.

## 5 Conclusion

In this report, we presented our solution to the 2025 BEHAVIOR Challenge, adapting the publicly available  $\pi_{0.5}$  backbone to a demanding long-horizon household benchmark and systematically studying how pre-training task coverage, post-training, and inference-time design choices affect performance on all tasks. Our experiments reveal that scaling pre-training over more numerous and diverse BEHAVIOR tasks significantly generates and unlocks success on rare, compositional activities. Additionally, We show that RFT as a post-training technique avoids the infrastructure issues from online RL yet yield trackable targets as well as significant performance boost.

Despite the promising results, we acknowledge that our Q-score is still far from perfect. We find that despite the effectiveness of RFT, the sampling efficiency is still too low. Paradigms such as DAgger or on-policy distillation where an expert policy potentially using privileged data could greatly improve the sampling efficiency. In addition, RL approaches that provide both positive and negative rewards could balance the learning.

We believe that combining strong VLA backbones with more structured long-horizon reasoning, richer post-training objectives, and better curriculum design will further close the gap between synthetic benchmarks and real-world deployment, and we hope our empirical findings provide practical guidance for scaling foundation policies to complex, human-centric environments.

## 6 Authors

**Core Contributors:** Delin Qu\*, Qizhi Chen\*, Shangkun Sun\*, Zhaoshuo Li, Yu-Wei Chao, Xiaohui Zeng, Xuan Li, Junjie Bai, Tsung-Yi Lin, Ming-Yu Liu

**Contributors:** Kaichun Mo, Jinwei Gu, Moo Jin Kim, Fangyin Wei, Hongchi Xia, Nic Ma

\*First authors with random order

---

## References

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.
- Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.
- Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al.  $\pi_0$ : A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.
- 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.
- Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- Kang Chen, Zhihao Liu, Tonghe Zhang, Zhen Guo, Si Xu, Hao Lin, Hongzhi Zang, Quanlu Zhang, Zhaofei Yu, Guoliang Fan, et al. pirl: Online rl fine-tuning for flow-based vision-language-action models. *arXiv preprint arXiv:2510.25889*, 2025.
- Zixuan Chen, Ze Ji, Jing Huo, and Yang Gao. Scar: Refining skill chaining for long-horizon robotic manipulation via dual regularization. *Advances in Neural Information Processing Systems*, 37:111679–111714, 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.
- George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. *Advances in neural information processing systems*, 22, 2009.
- Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024.
- Xingyu Lin, Zhiao Huang, Yunzhu Li, Joshua B Tenenbaum, David Held, and Chuang Gan. Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools. *arXiv preprint arXiv:2203.17275*, 2022.
- Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.
- Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.
- Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023.
- Chenrui Tie, Shengxiang Sun, Jinxuan Zhu, Yiwei Liu, Jingxiang Guo, Yue Hu, Haonan Chen, Junting Chen, Ruihai Wu, and Lin Shao. Manual2skill: Learning to read manuals and acquire robotic skills for furniture assembly using vision-language models. *arXiv preprint arXiv:2502.10090*, 2025.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Chao Yu, Yuanqing Wang, Zhen Guo, Hao Lin, Si Xu, Hongzhi Zang, Quanlu Zhang, Yongji Wu, Chunyang Zhu, Junhao Hu, et al. Rlinf: Flexible and efficient large-scale reinforcement learning via macro-to-micro flow transformation. *arXiv preprint arXiv:2509.15965*, 2025.

---

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1702–1713, 2025.