Efficient and Robust VLA Models for Real Time Robotic Manipulation

Shriyadita Medavaram, Lam Nguyen, Yuzhang Shang
University of Central Florida, Orlando, FL

Abstract

- $\pi0.5$ is an open source VLA model developed by Physical Intelligence.
- Problem: $\pi0.5$ can be quite slow depending on the task, and can take up a lot of memory
- Goal: Reduce latency and dynamic memory usage of $\pi0.5$ model
- Contribution: Performed timestep distillation on diffusion process within $\pi0.5$ to decrease latency and memory usage during inference

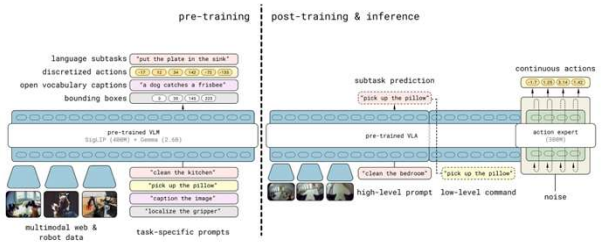

Figure 1: Original $\pi0.5$ model architecture [1]

Background

- While running latency and memory profile, it was found that the forward diffusion process takes 72.1% of the runtime, and 74.6% of the dynamic CPU usage during inference.
- This remained relatively constant across simulations and real-time manipulation

Section	Count	Avg time	Total time	Total CPU delta	Total GPU delta
Forward Diffusion	10	851.33 ms	8513.33 ms	50.21 MB	0.80 MB
Prefix Embedding	1	1919.79 ms	1919.79 ms	12.79 MB	0.40 MB
Key-Value Cache	1	1276.69 ms	1276.69 ms	4.25 MB	0.20 MB
Action Embedding	10	4.95 ms	49.47 ms	0.004 MB	0.00 MB
Diffusion Position Masking	10	2.85 ms	28.48 ms	0.00 MB	0.00 MB
Projection Update	10	1.14 ms	11.39 ms	0.002 MB	0.00 MB
Prefix Position Masking	1	1.80 ms	1.80 ms	0.00 MB	0.00 MB
Noise Initialization	1	0.26 ms	0.26 ms	0.007 MB	0.00 MB
Observation Preprocessing	1	10.02 ms	10.02 ms	0.015 MB	0.00 MB
Output Transform	1	1.59 ms	1.59 ms	0.002 MB	0.00 MB

Table 1: Latency and Memory profile results from $\pi0.5$ model running in simulation on LIBERO Goal

Process

- As the diffusion process took the most time and memory, timestep distillation on the $\pi0.5$ model was attempted
- The paper "Improved Distribution Matching Distillation for Fast Image Synthesis" was referenced for the timestep distillation process, and was slightly altered to be applicable to VLAs [2] [3]
- Over the course of various experiments, it was determined that a 3-step distilled diffusion model seemed to perform best

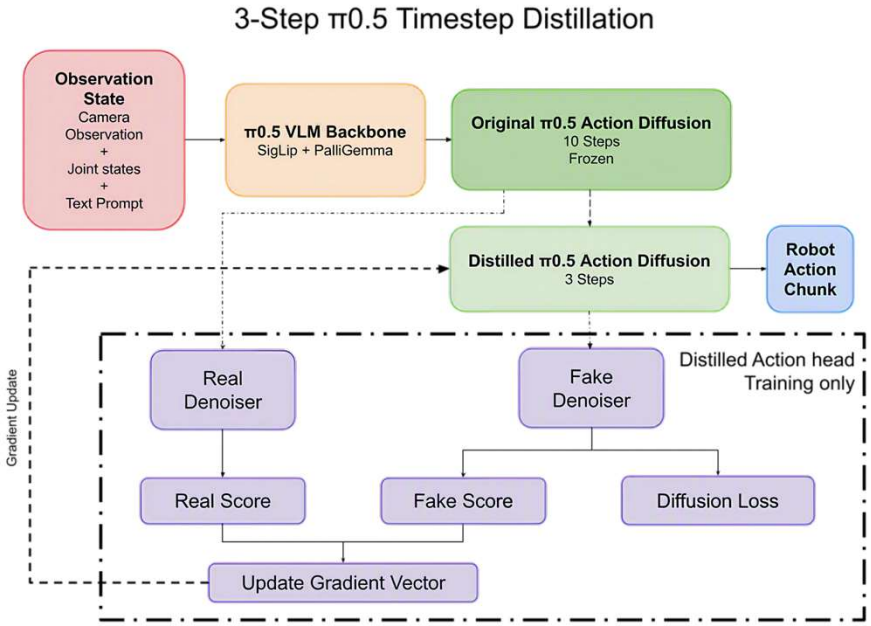

Figure 2: Diagram of how 3-step distillation of Pi0.5 model works. Describes training architecture in purple, and inference architecture

Latency and Memory Usage Comparison

- Compared to the original $\pi0.5$ model, the 3-step distilled model is faster and has a lighter memory usage

Metric	Original π0.5	3-step Distilled π0.5	Improvement
Full inference time	11.81 s	5.69 s	2.08x faster
Total CPU delta	672.73 MB	324.68 MB	2.07x lower
Total GPU delta	14 MB	8 MB	1.75x lower

Table 2: Latency and Memory profile comparison. Original $\pi0.5$ vs 3-step distilled model

Model Results

- The 3-step timestep distilled model was evaluated across various LIBERO suites
- The results are shown in the diagram below, compared to the original model accuracies

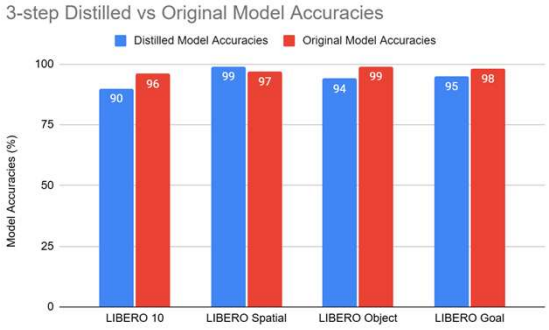

Figure 3: 3-step timestep distillation model accuracies vs original model accuracies across various LIBERO Suites

Future Work

- Apply technique to $\pi0.5$ fine-tuned for actual robot arm use
- Further increase model accuracy to get closer to original $\pi0.5$ model
- Distill other parts of $\pi0.5$ architecture to decrease the static/baseline memory needed to run the model

References

[1] K. Black *et al.*, "$\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization," Physical Intelligence, Tech. Rep., 2024. [Online]. Available: https://www.pi.website/download/pi05.pdf

[2] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park, "One-step diffusion with distribution matching distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 6613–6623. [Online]. Available: https://arxiv.org/abs/2311.18828

[3] T. Yin, M. Gharbi, T. Park, R. Zhang, E. Shechtman, F. Durand, and W. T. Freeman, "Improved distribution matching distillation for fast image synthesis," *arXiv preprint arXiv:2405.14867*, 2024. [Online]. Available: https://arxiv.org/abs/2405.14867

Acknowledgements

The author would like to thank and acknowledge the UCF Computer Vision REU program for all their support and the opportunity. She would also like to thank the faculty and students at the Institute of AI for their help. She would like to acknowledge Dr. Mubarak Shah and Dr. Niels de Victoria Lobo for their guidance.