Efficient and Robust VLAs for Real Time Robotic Manipulation

Shriyadita Medavaram, Lam Nguyen, Dr. Yuzhang Shang

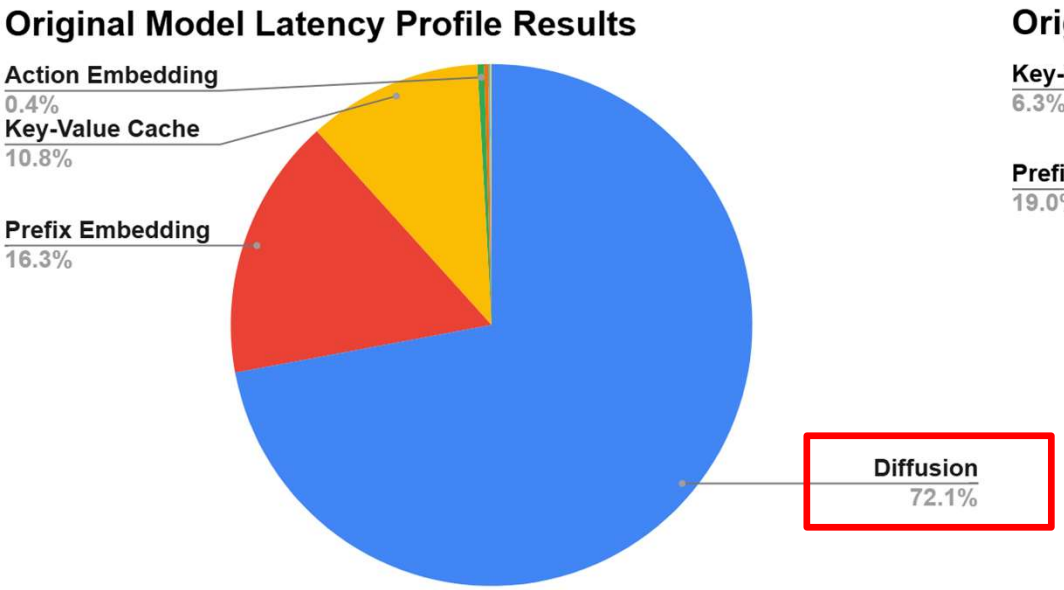
Project Overview

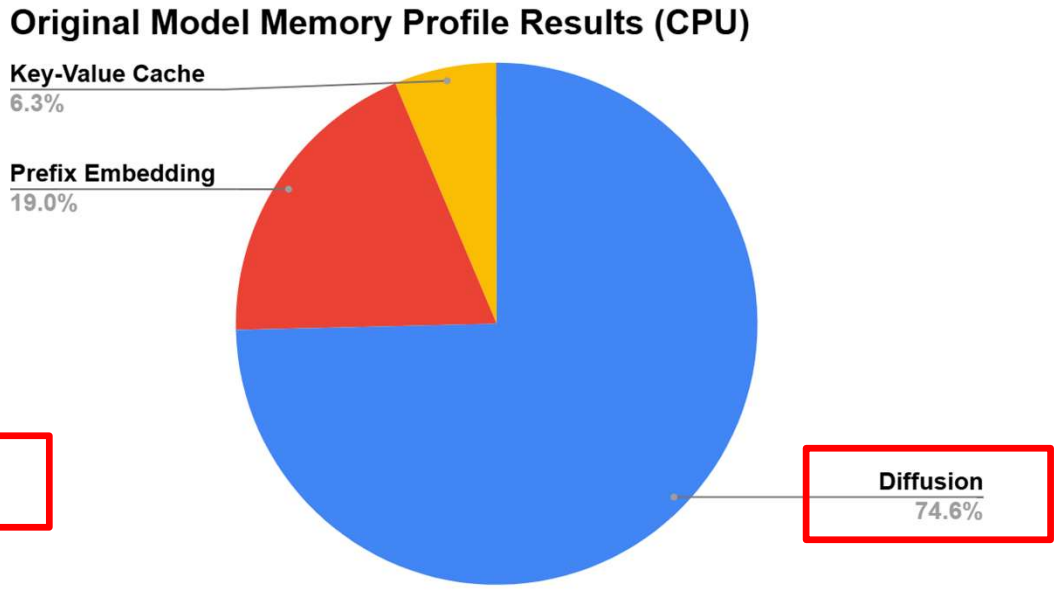
- Problem: Current VLA (Visual Language Action) models are too slow and take up too much memory
  - Causes interference when deployed on real robot arms
- Goal: Speed up VLA models by distilling the 10-step diffusion process in $\pi 0.5$ into a 3-step process



Our 3-step distilled model performance in simulation

Initial Findings

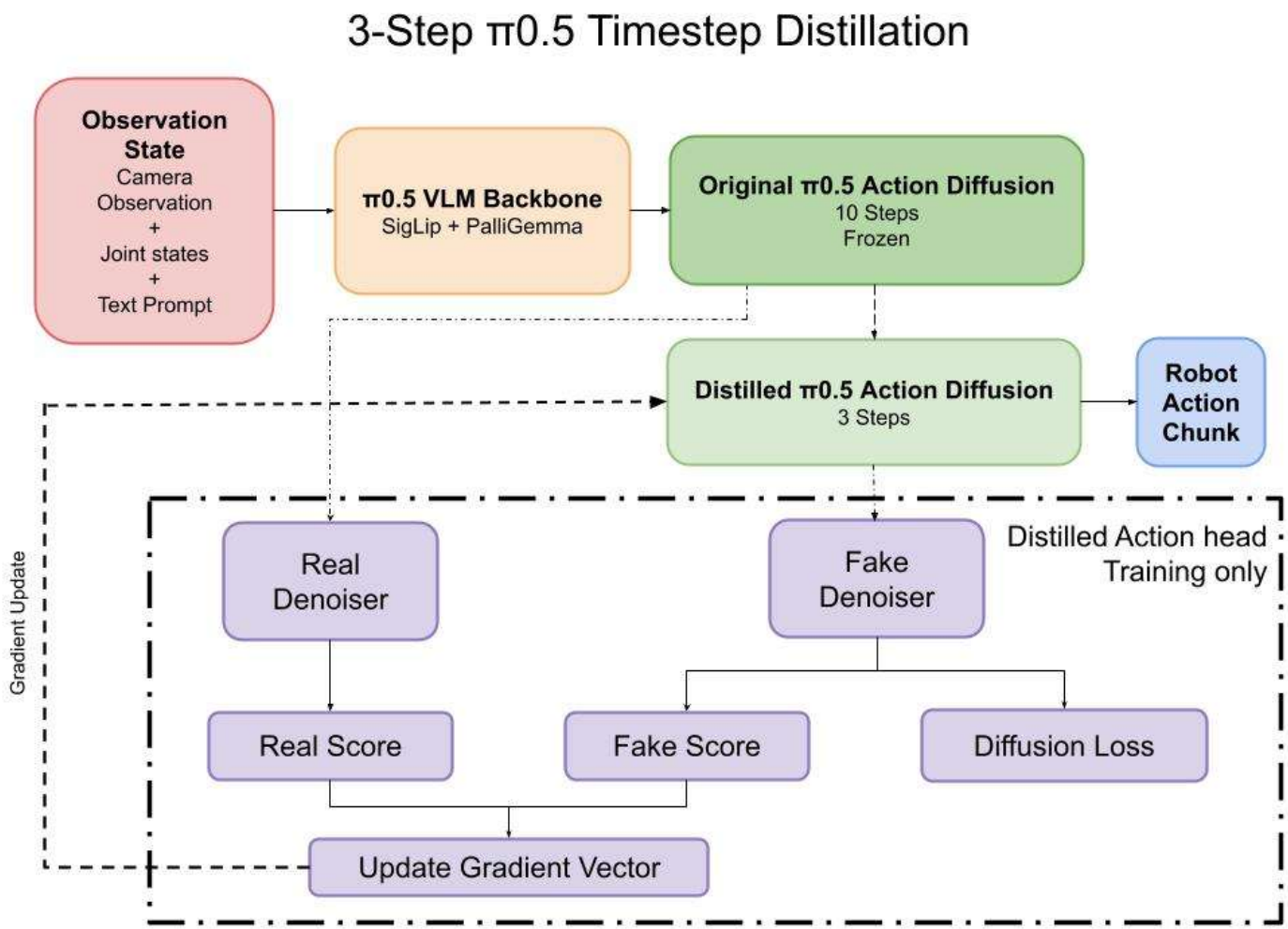




Method

- Time step distillation based on paper
  - “Improved Distribution Matching Distillation for Fast Image Synthesis” [3]
- Keeping overall $\pi 0.5$ architecture, distilling diffusion action head
- Using 3-step distillation, and methods described in paper to improve accuracy while reducing runtime

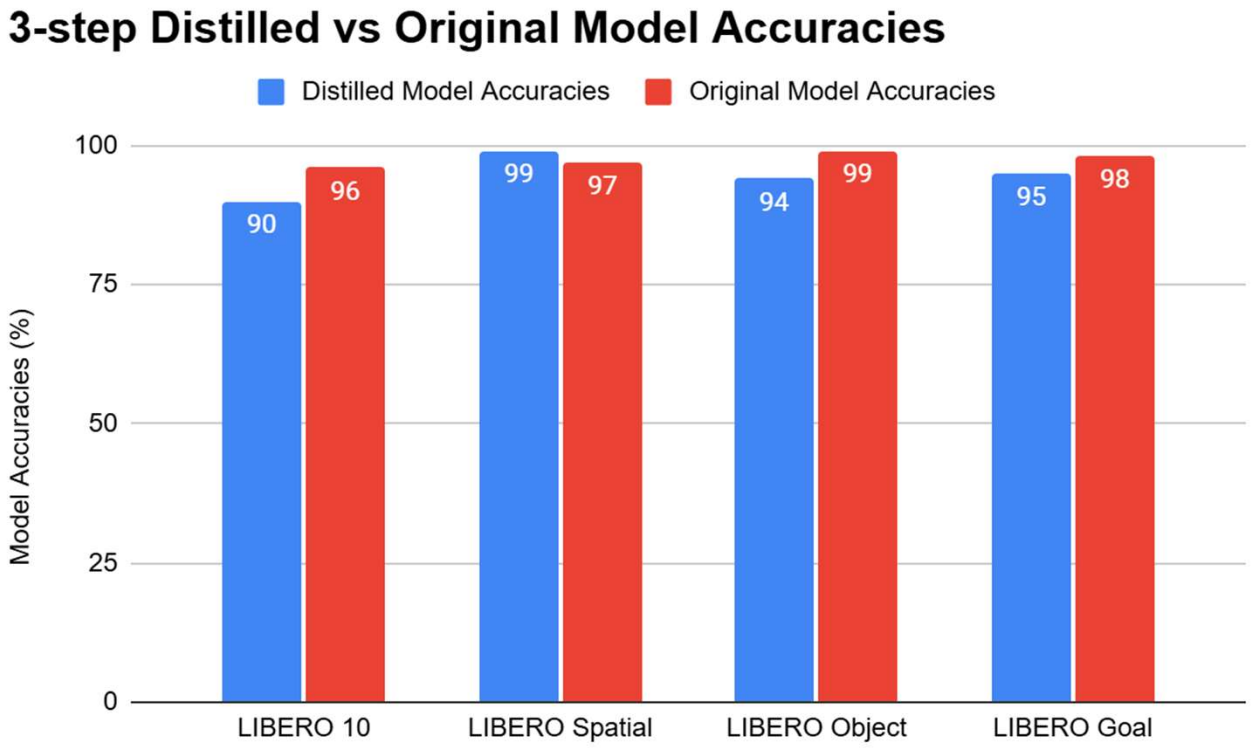
Method


3-Step π0.5 Timestep Distillation
Observation State
Camera Observation
+
Joint states
+
Text Prompt
π0.5 VLM Backbone
SigLip + PalliGemma
Original π0.5 Action Diffusion
10 Steps
Frozen
Distilled π0.5 Action Diffusion
3 Steps
Robot Action Chunk
Gradient Update
Real Denoiser
Fake Denoiser
Distilled Action head
Training only
Real Score
Fake Score
Diffusion Loss
Update Gradient Vector

Results

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

Results


3-step Distilled vs Original Model Accuracies
Distilled Model Accuracies
Original Model Accuracies
Model Accuracies (%)
100
75
50
25
0
90
96
99
97
94
99
95
98
LIBERO 10
LIBERO Spatial
LIBERO Object
LIBERO Goal

Sources

[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