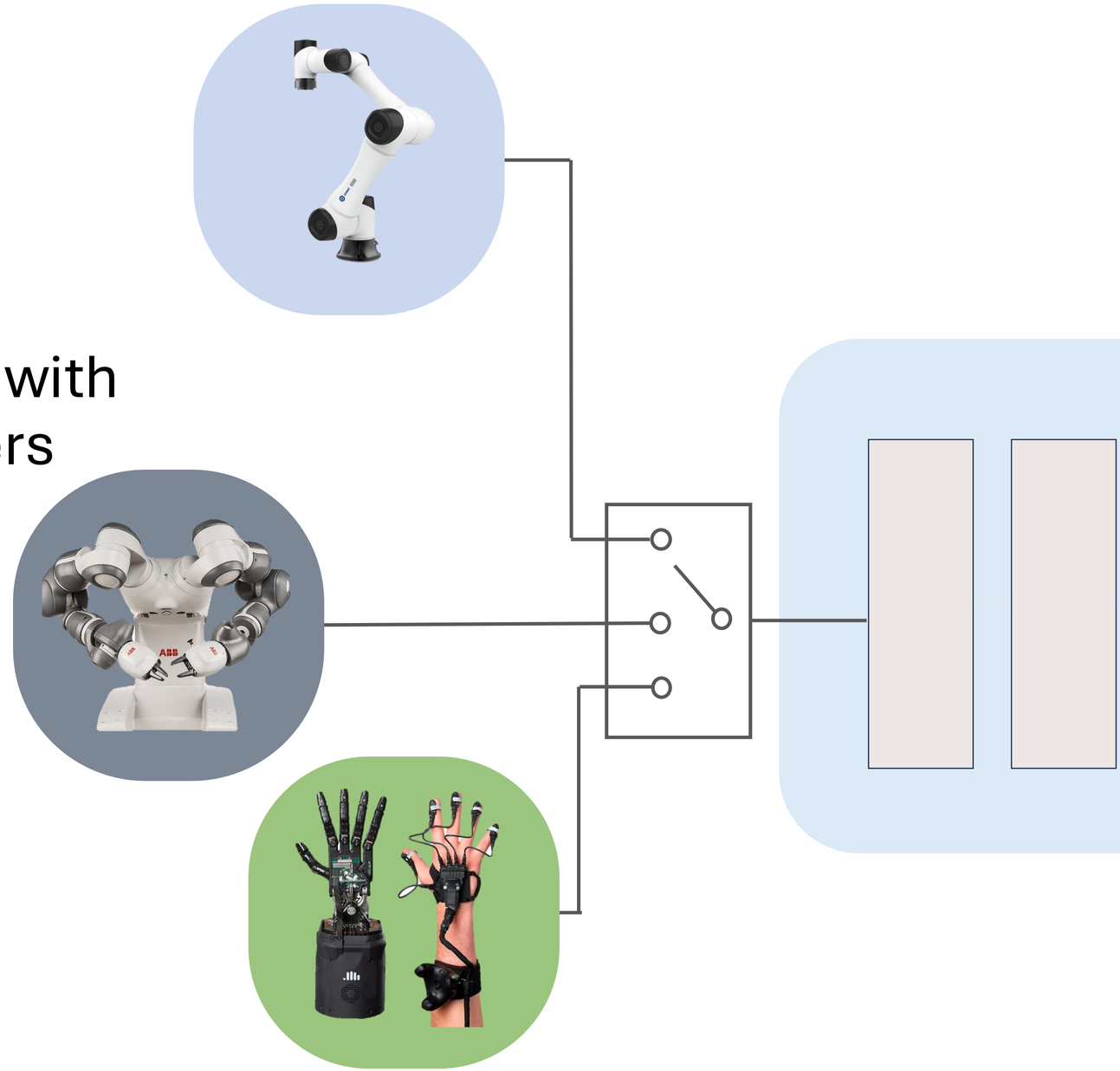
HPT

Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers

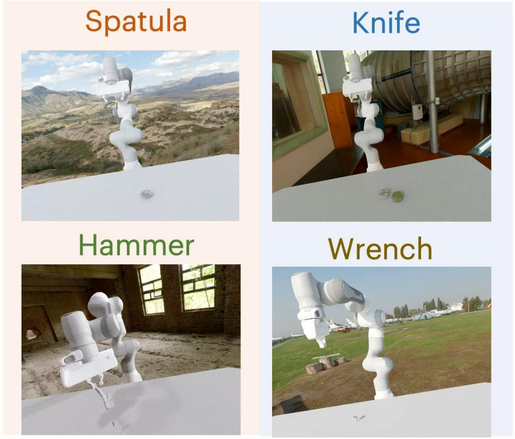
Lirui Wang, Xinlei Chen, Jialiang Zhao, Kaiming He
MIT CSAIL and Meta FAIR

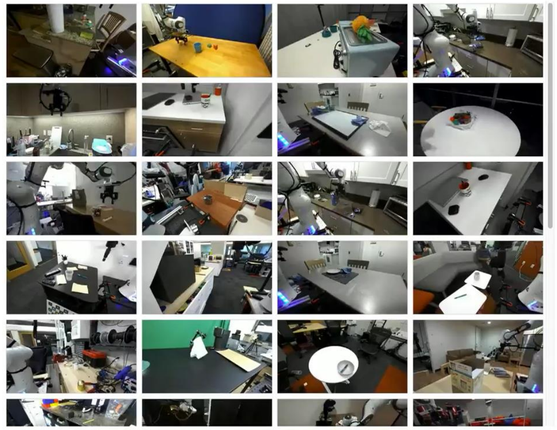
NeurIPS 2024 *Spotlight*



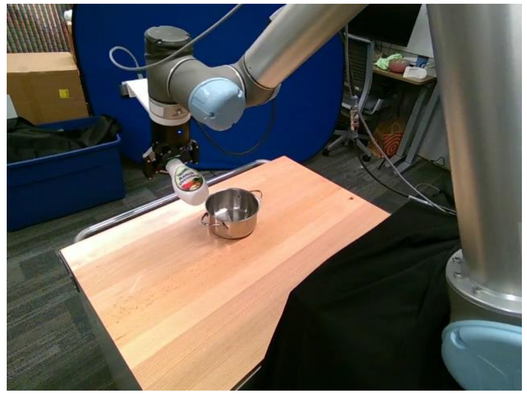
The Challenge Towards Robotic Foundation Models

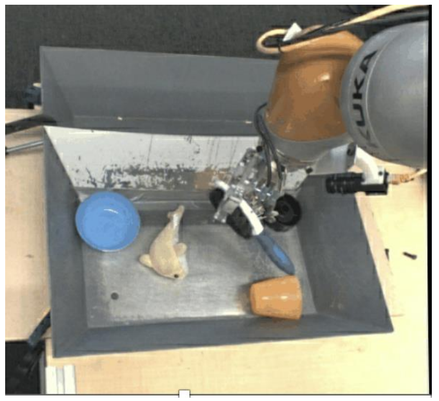
Different Domains / Tasks

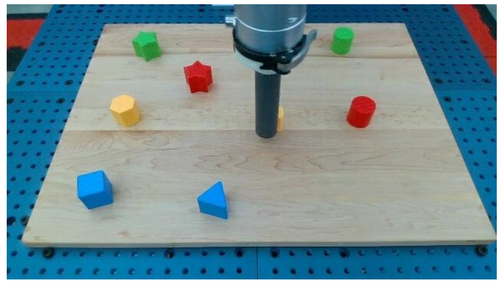




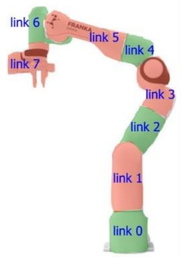
Different Embodiments



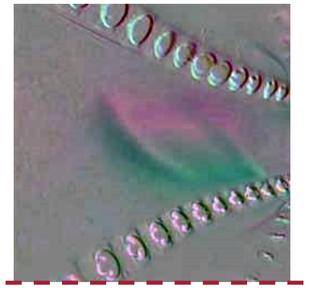




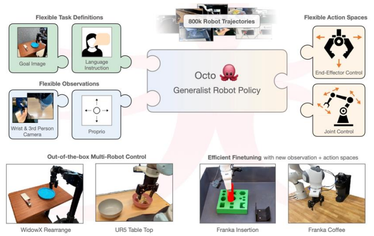
Different Sensor Modalities



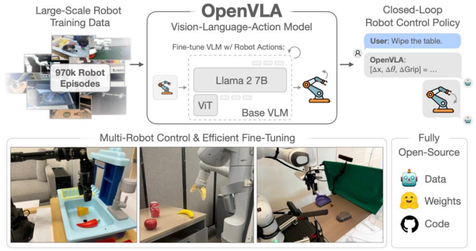
"...Use the hammer to hit the pin..."



Heterogeneous Pre-training

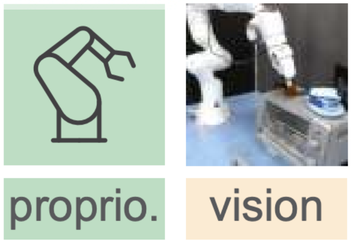


Octo, 2023



OpenVLA, 2024

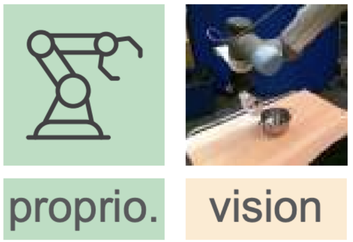
HALLO



proprio. | vision

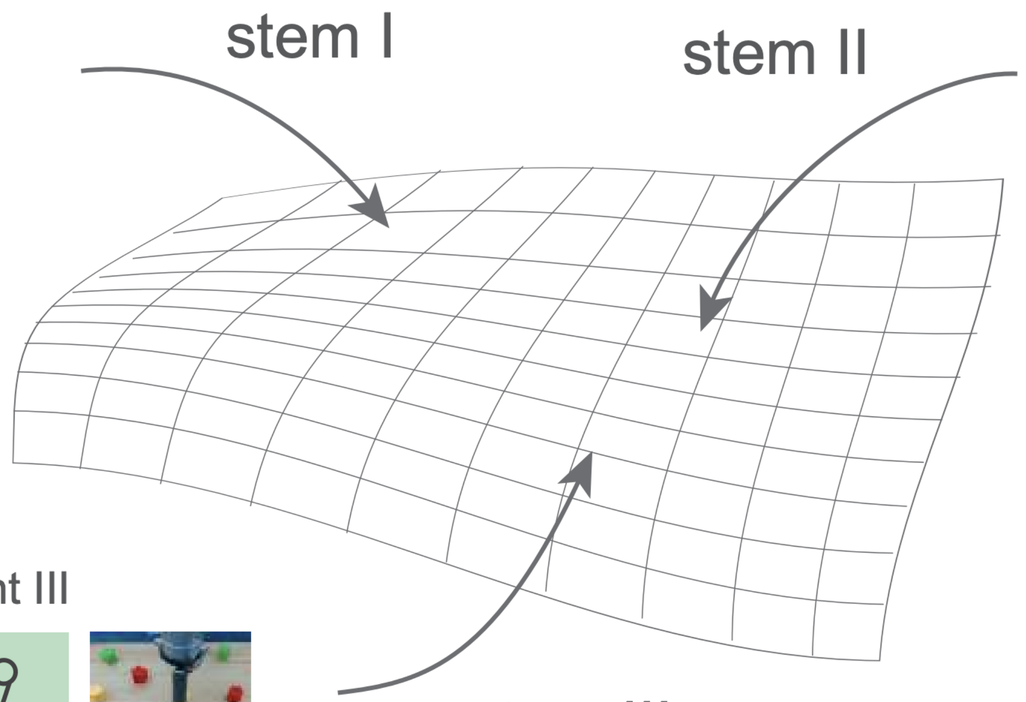
embodiment I

BONJOUR

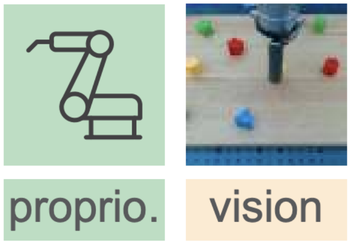


proprio. | vision

embodiment II



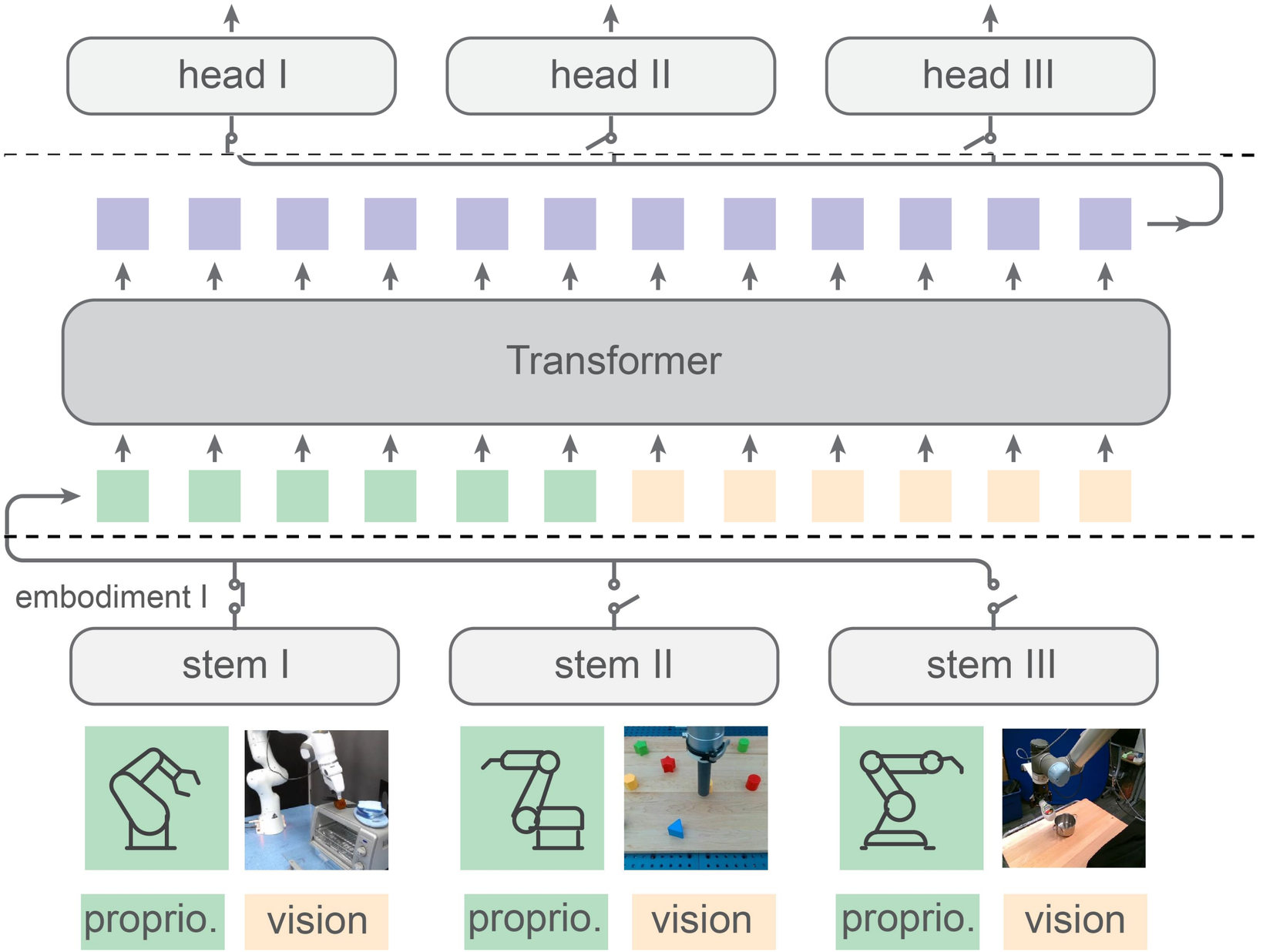
embodiment III



proprio. | vision

こんにちは

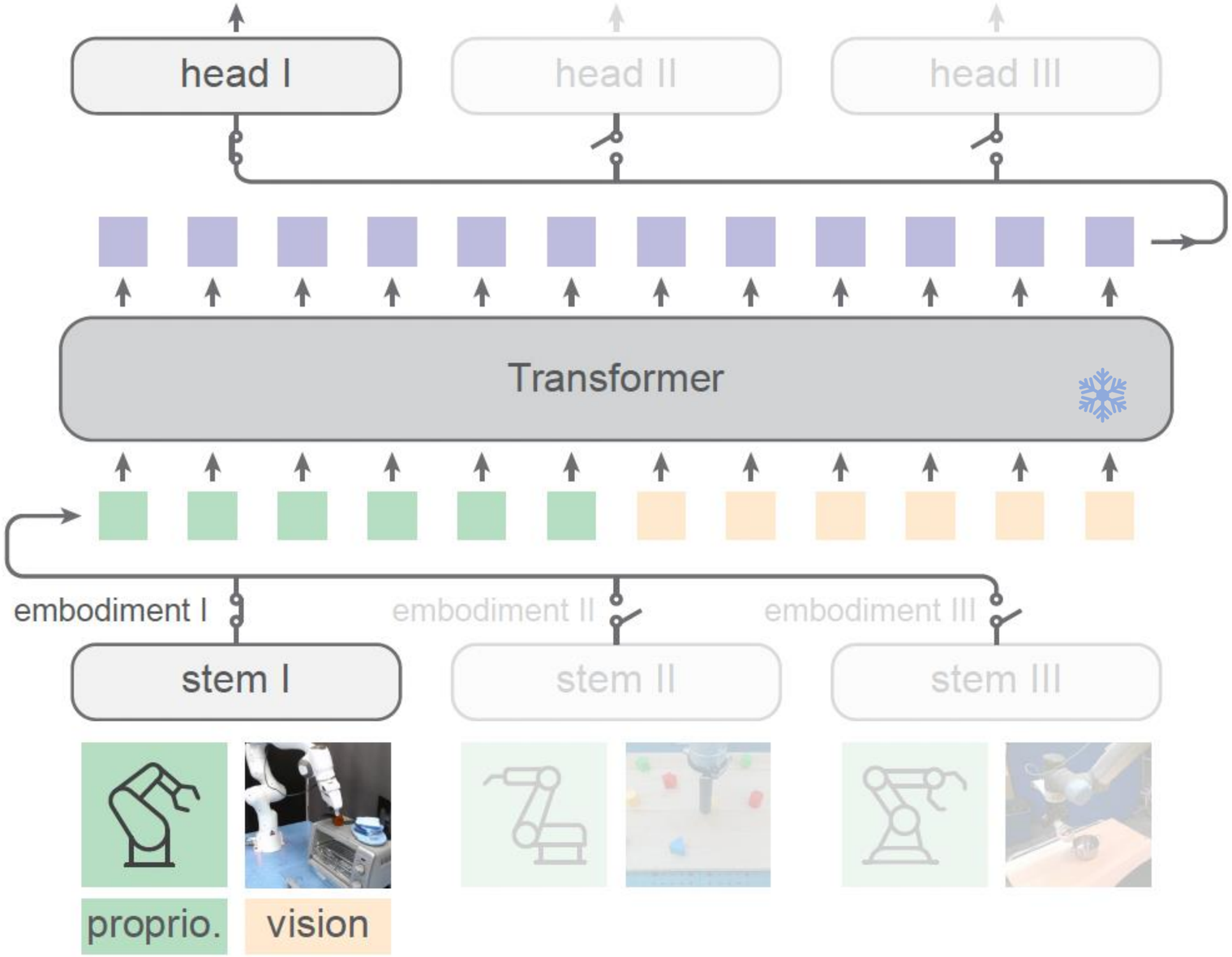
Heterogeneous Pre-trained Transformer (HPT)



Objective:

$$\min_{\theta} \frac{1}{K} \sum_{k=1}^{K} L(\theta; \mathcal{D}_k)$$

Heterogeneous Pre-trained Transformer (HPT)

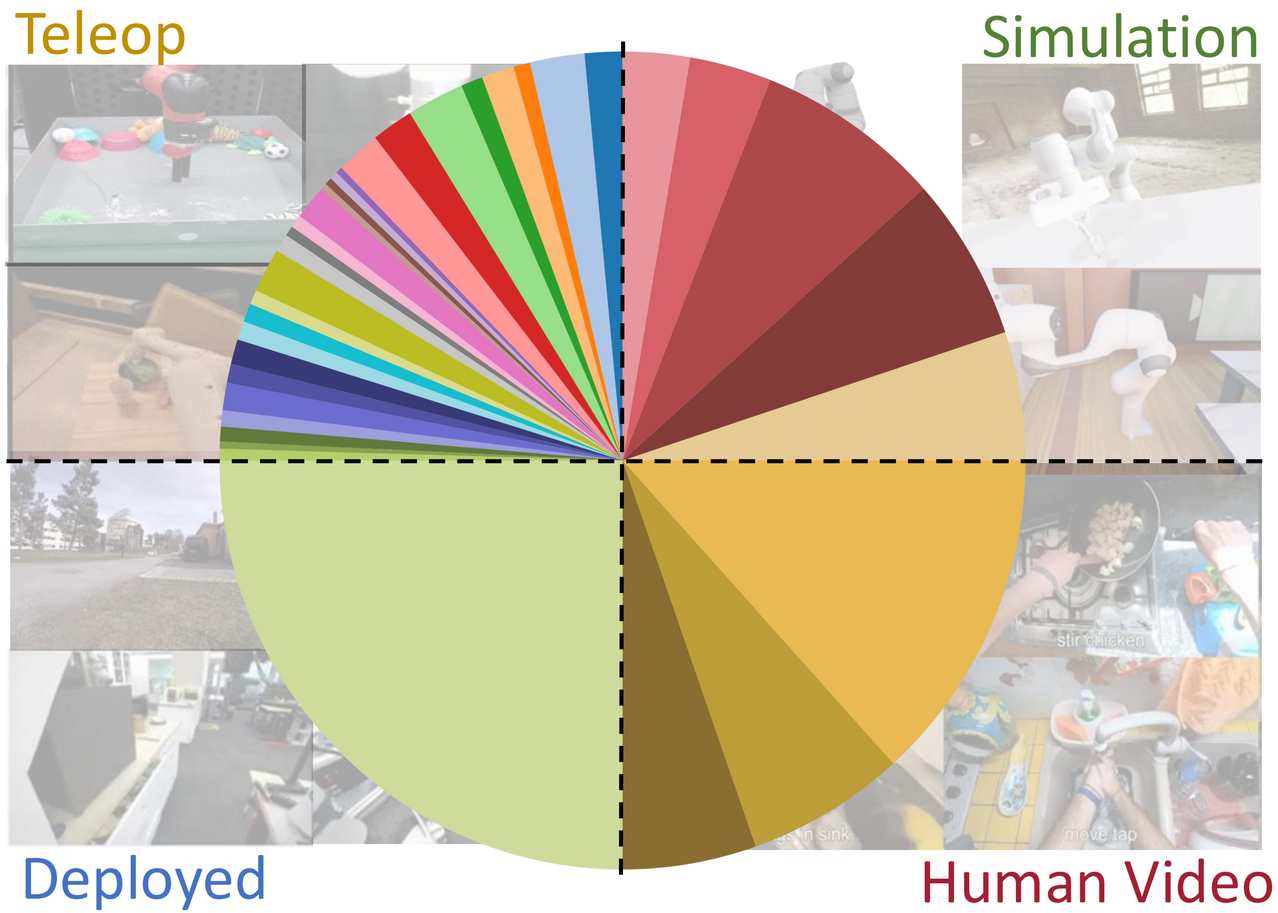


Reinit.

Frozen

Reinit.

Data Mixture



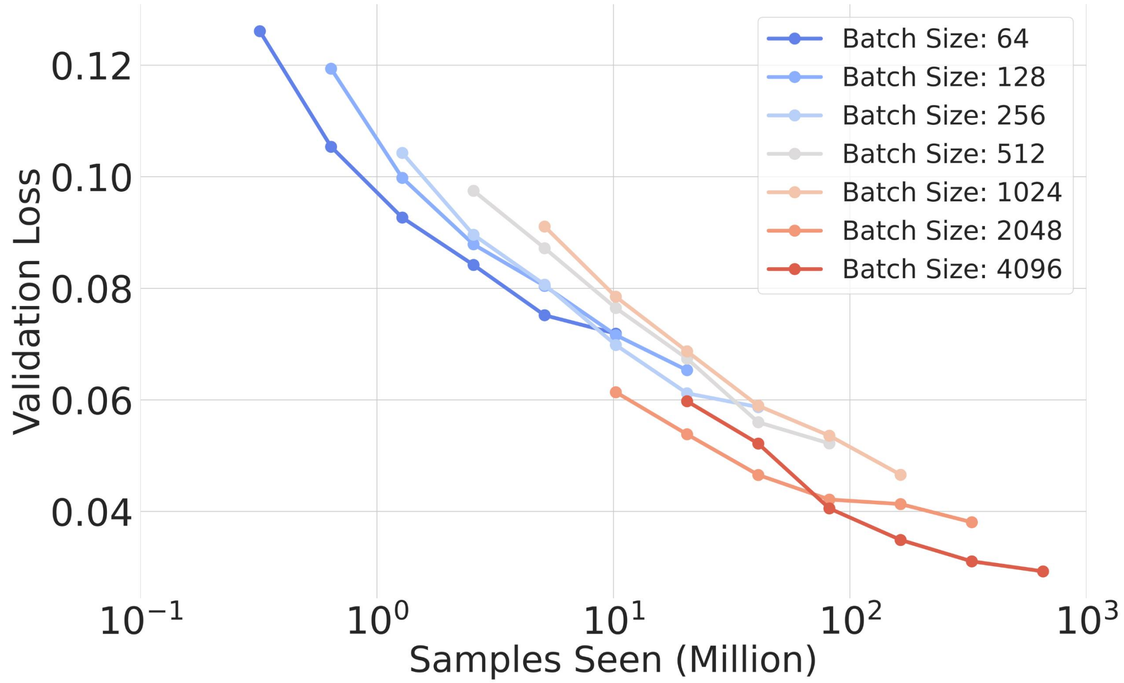
Data

- 52 Datasets
- 300k Trajectories
- 5B Tokens

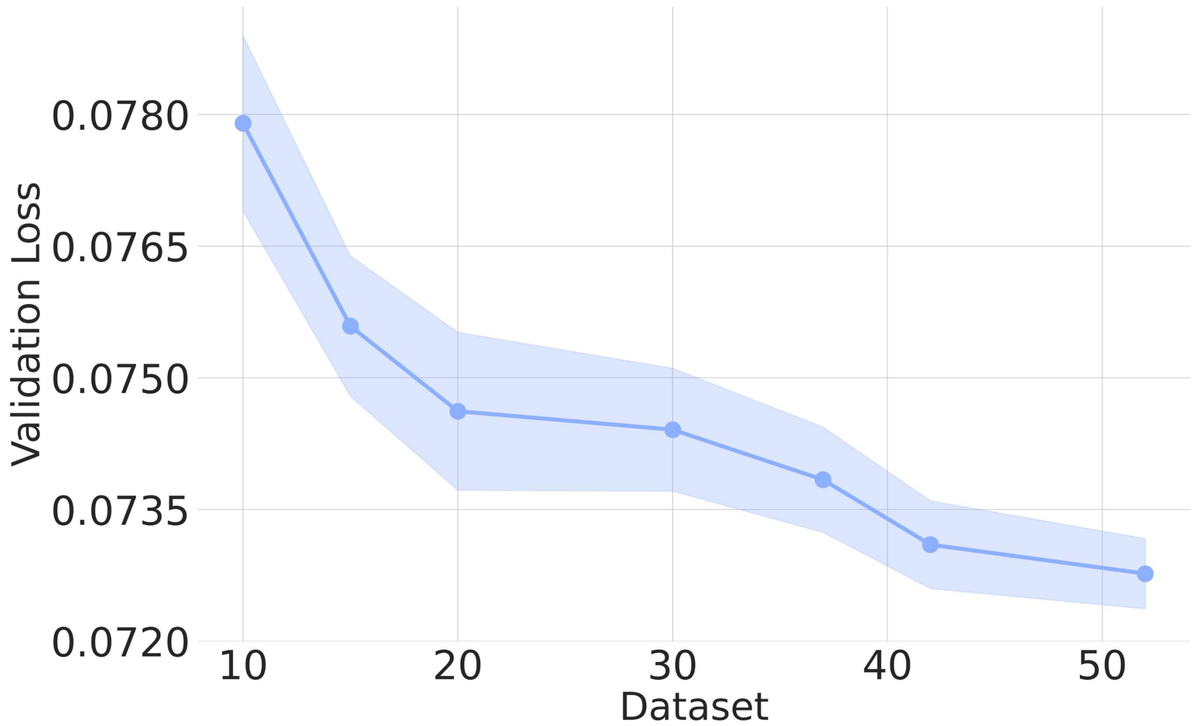
Compute

- 1B Params
- 31 GFLOPs
- 128 GPUs

Pre-training: Scaling Data Quantity and Diversity



Higher performance w/ more data

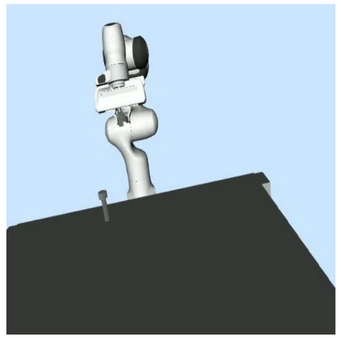


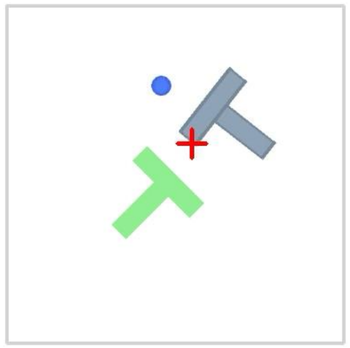
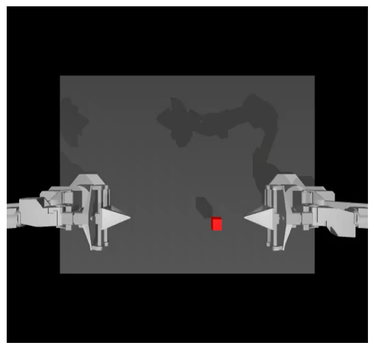
Higher performance w/ more diversity

Transfer to Embodiments in Simulation

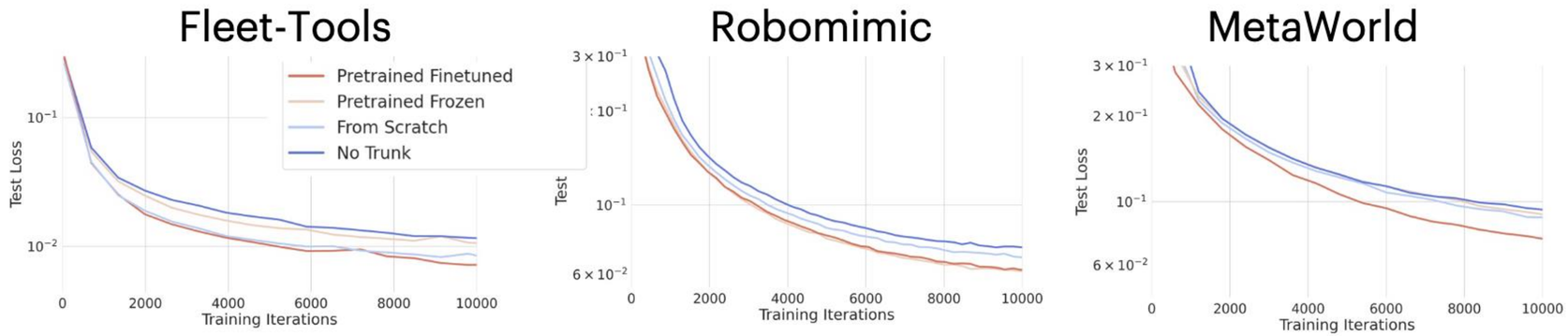
Visualization



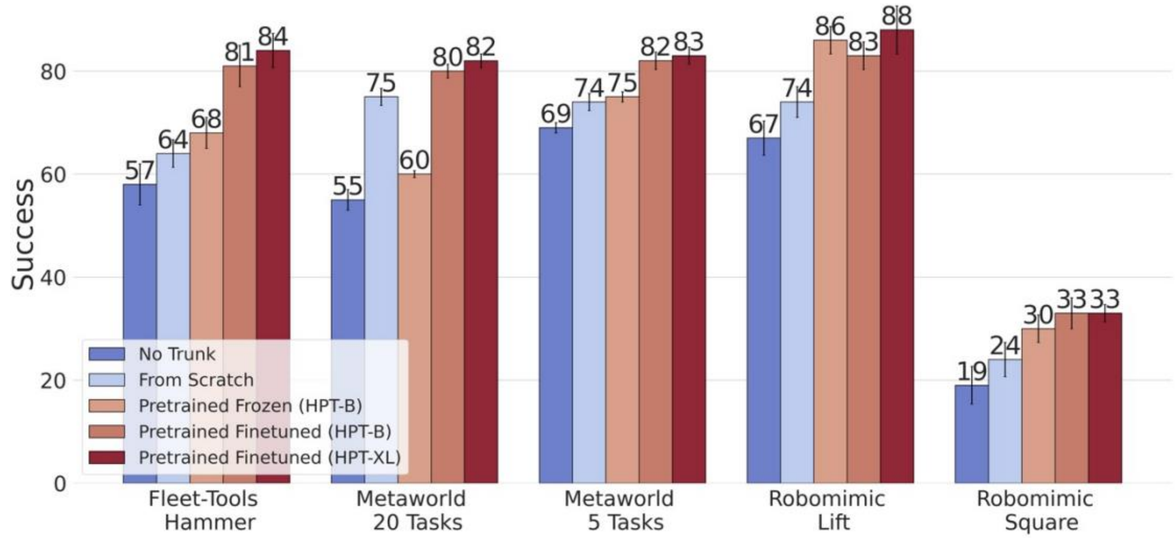


Transfer Learning Loss

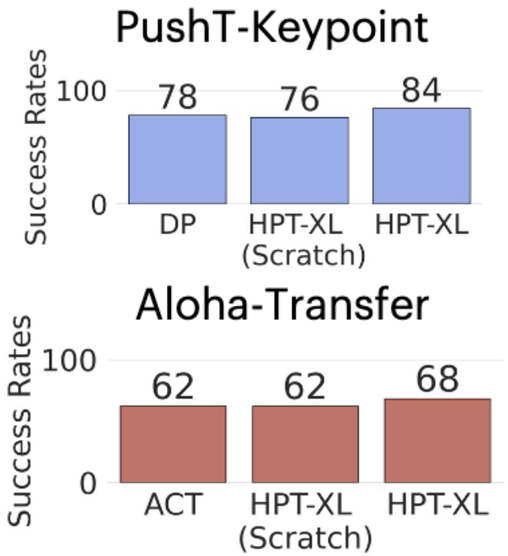


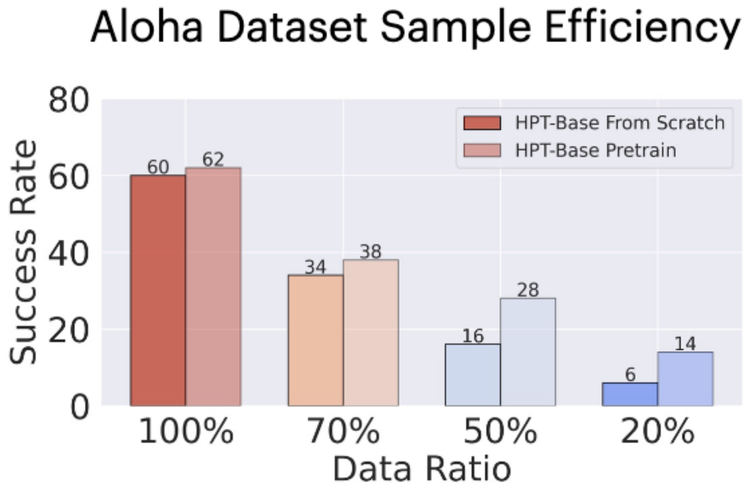
Transfer Learning Success Rates

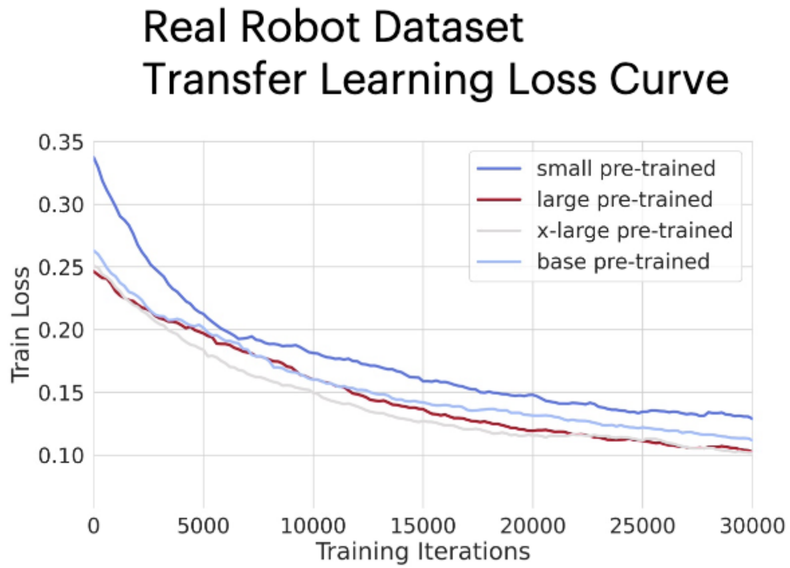


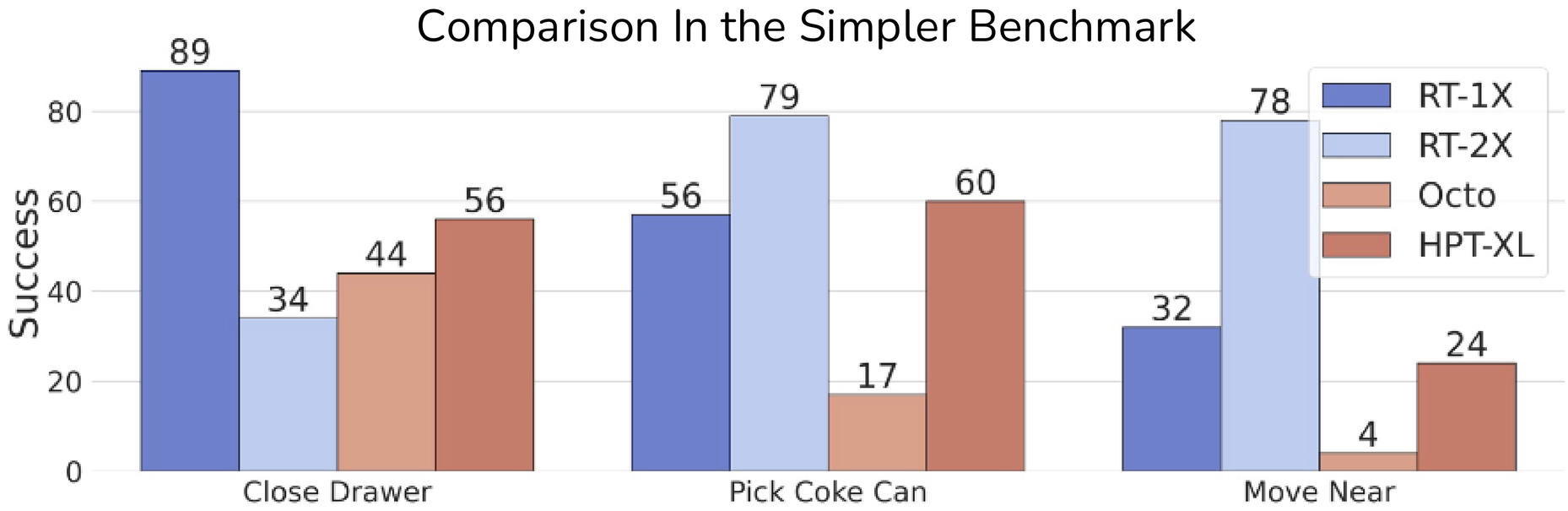
Minimal finetuning: 2% of params (Stem + Head) to generalize outside of the pre-training datasets

More Experiments with State-of-The-Art Methods



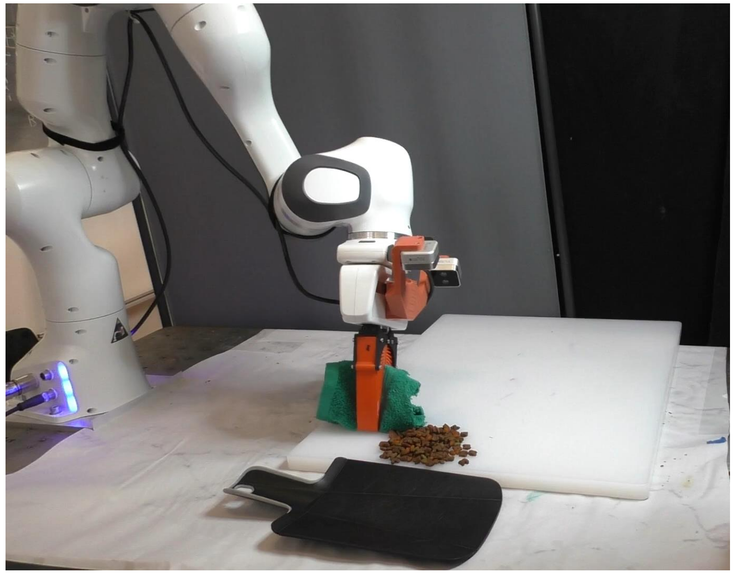






Transfer to Embodiments in the Real World

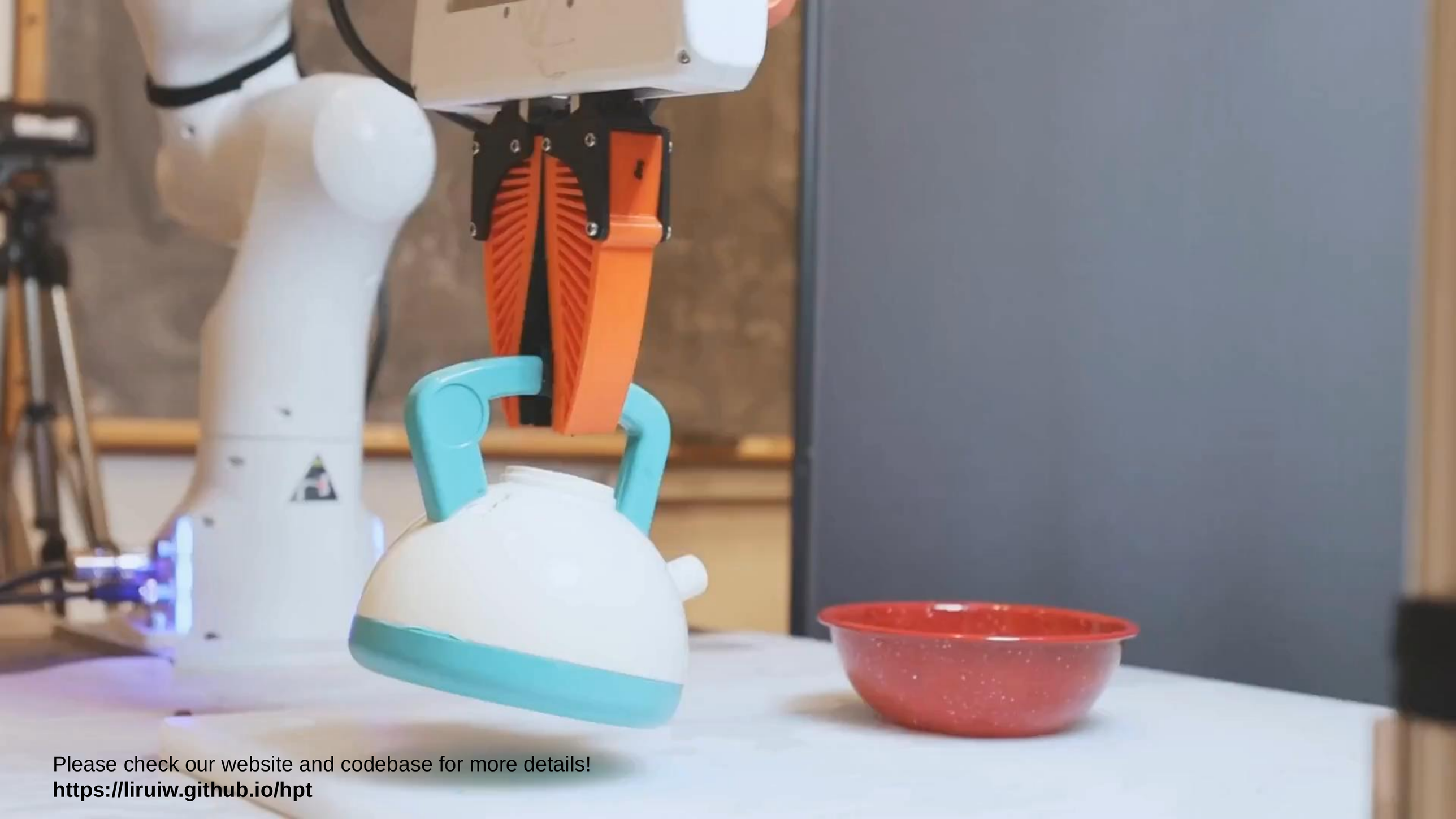
Embodiment 1: Pet Care



Embodiment 2: Insertion

Method	Success (%)	
From Scratch No Prop.	26.7±3.3	> train-from-scratch
From Scratch	43.3±3.8	
R3M [44]	50.0±3.0	does better than methods that pre-trains only vision
Voltron [28]	46.7±3.8	
VC-1 [40]	53.3±2.6	
HPT-B Finetuned	70.0±3.0	Improves with scale
HPT-XL Finetuned	**76.7**±3.3	

Robust to disturbance, new objects, and camera movements



Please check our website and codebase for more details!
https://liruiw.github.io/hpt