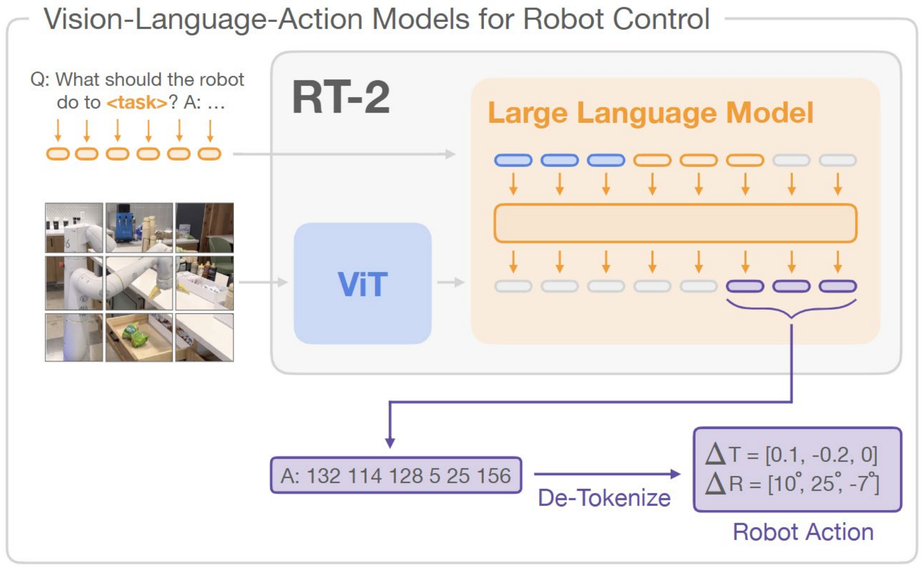
TwinVLA

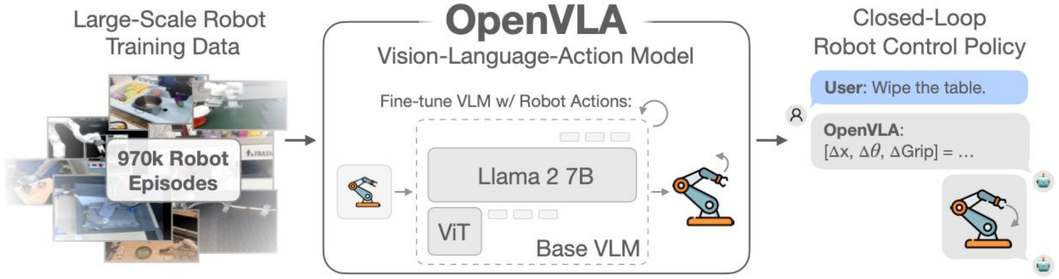
Data-Efficient Bimanual Policy with Twin Single-Arm Vision-Language-Action Models

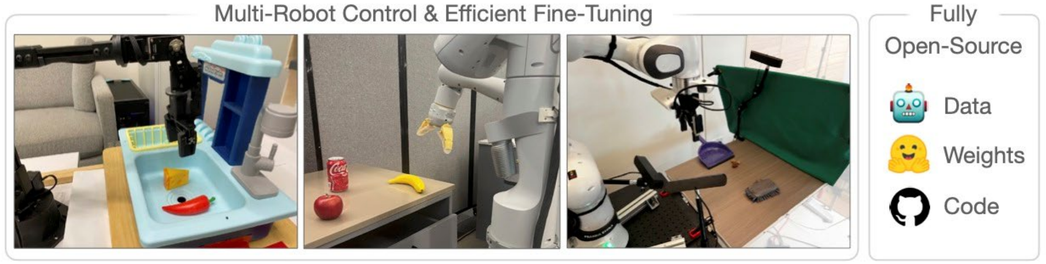
Hokyun Im[1,2], Euijin Jeong[1], Andrey Kolobov[2], Jianlong Fu[2], Youngwoon Lee[1]

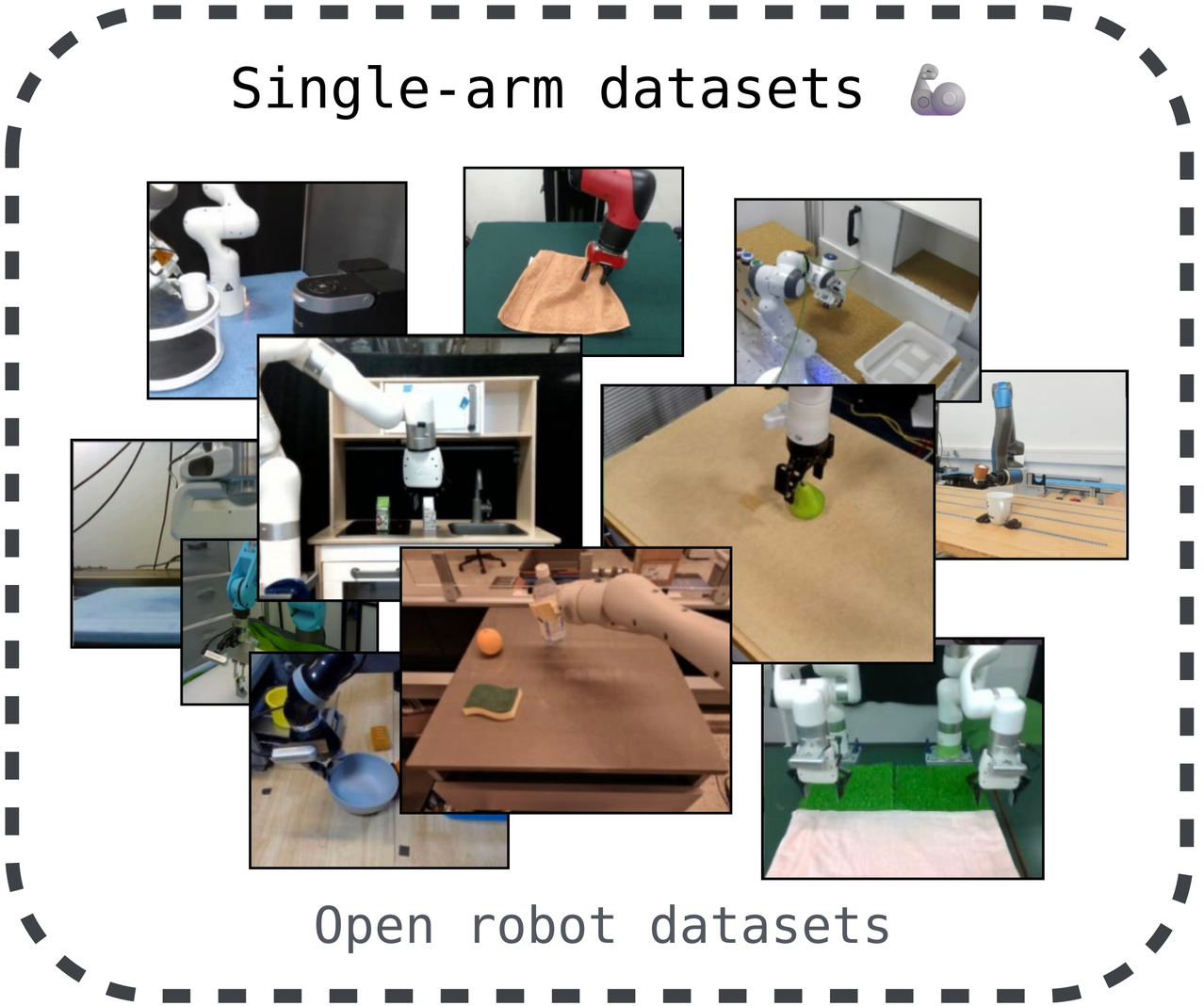
[1]Yonsei University, [2]Microsoft Research

Large, Diverse Robot Datasets Enable VLAs







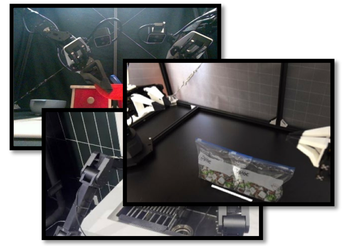


[1] Brohan et. al.,"RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control", arXiv 2023
[2] Kim et. al.,"OpenVLA: An Open – Source Vision – Language – Action Model", CoRL 2024
[3] OXE collaboration,"Open X – Embodiment: Robotic Learning Datasets and RT – X Models", ICRA 2024

Training Bimanual VLAs Remains Challenging

Bimanual VLA

Bimanual datasets

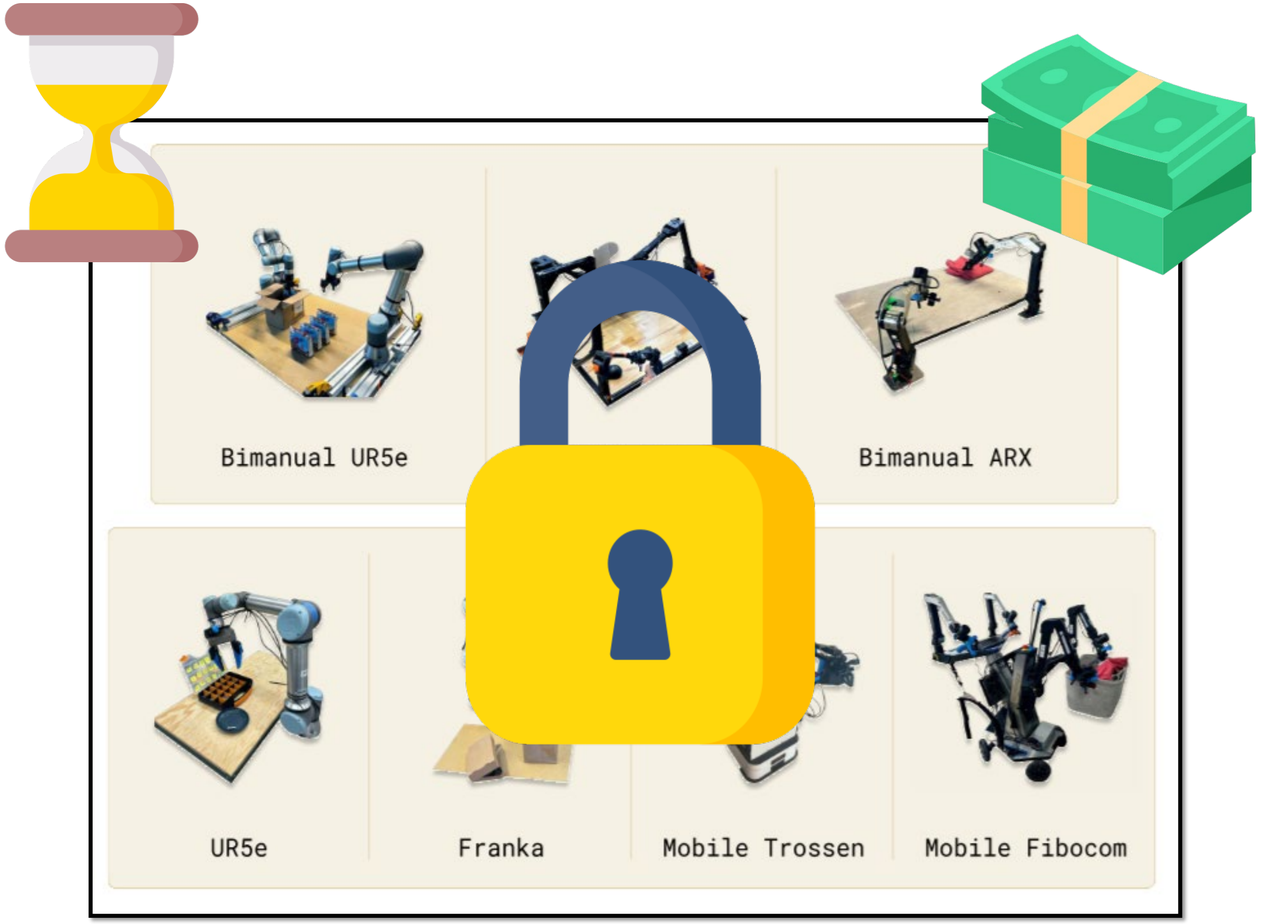


Scarce!

[4] Kevin Black et. al.,"π_0: A Vision-Language-Action Flow model for General Robot Control", RSS 2025

Training Bimanual VLAs Remains Challenging

Bimanual VLA



Proprietary datasets

[4] Kevin Black et. al.,"π_0: A Vision-Language-Action Flow model for General Robot Control", RSS 2025

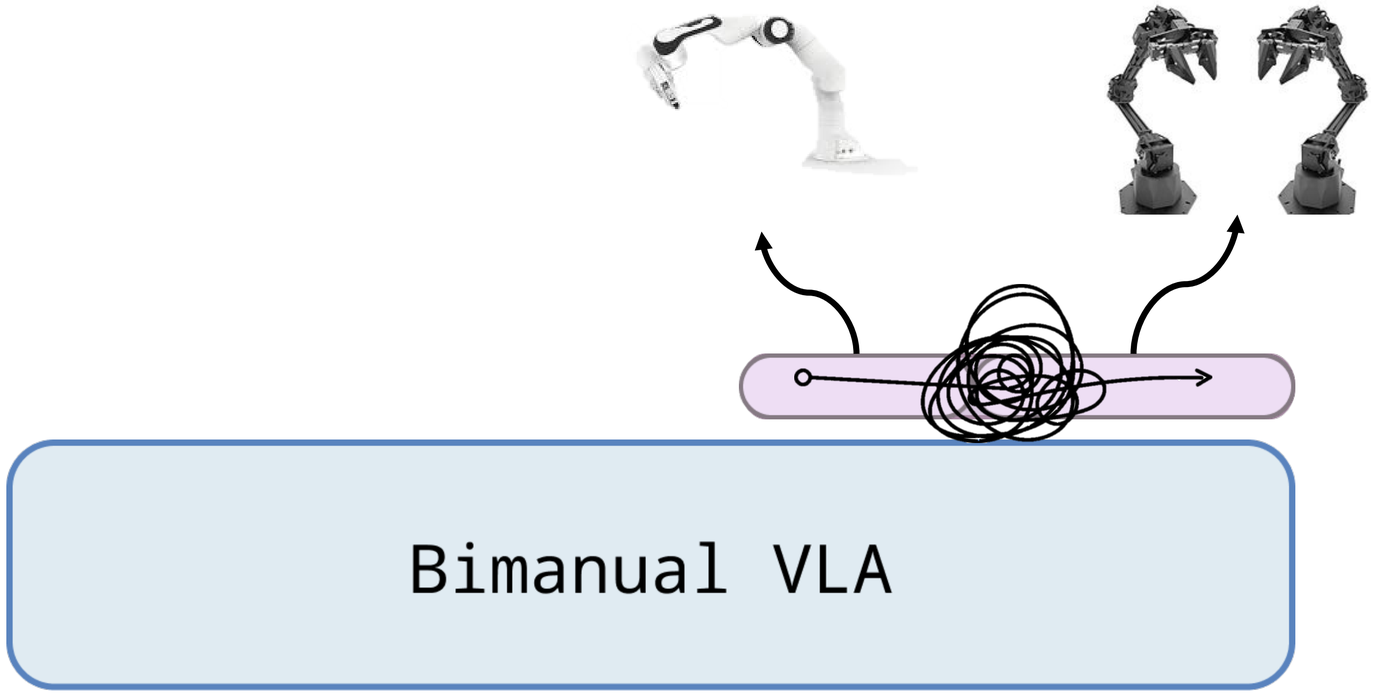
Training Bimanual VLAs Remains Challenging

Bimanual VLA

Not reproducible ❌

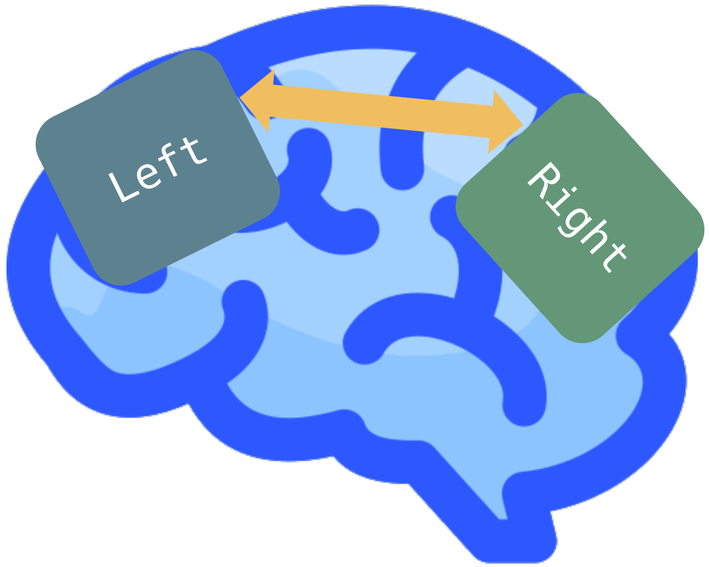
[4] Kevin Black et. al.,“π_0: A Vision-Language-Action Flow model for General Robot Control”, RSS 2025

Training Bimanual VLAs Remains Challenging



Monolithic models

Our Intuition :
Bimanual Manipulation is a Coordination of Single-Arm Skills

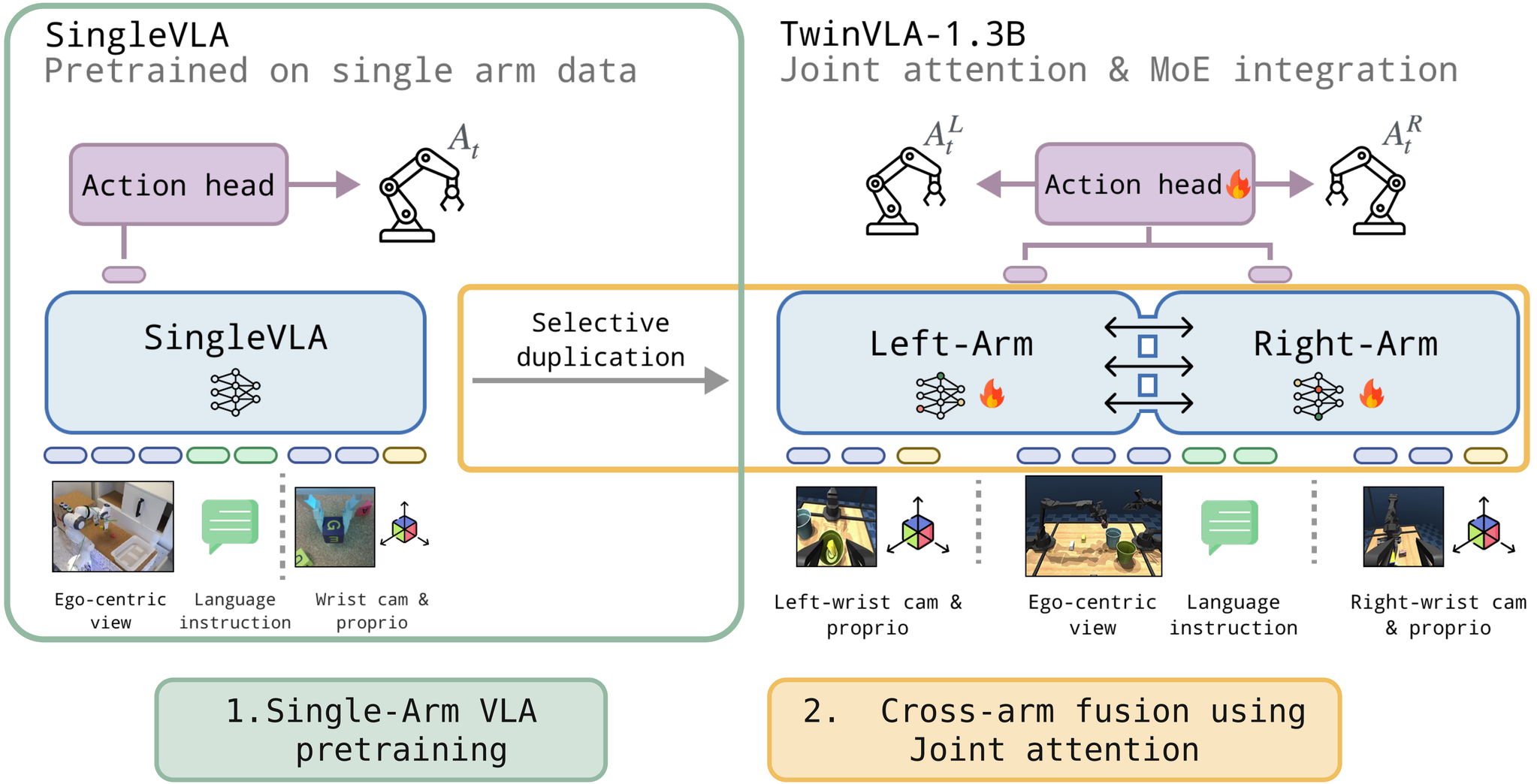


Human Brain

Bimanual Manipulation

Modular perspective on bimanual manipulation

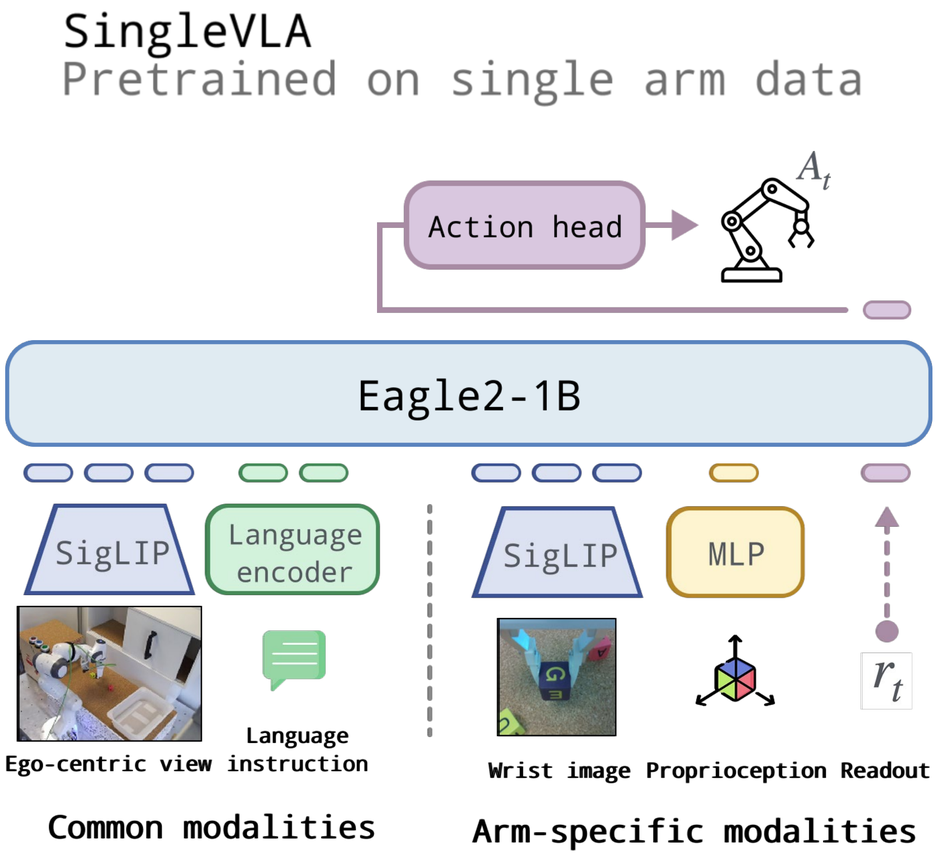
TwinVLA : Data-Efficient Bimanual VLA



1. Single-Arm VLA pretraining

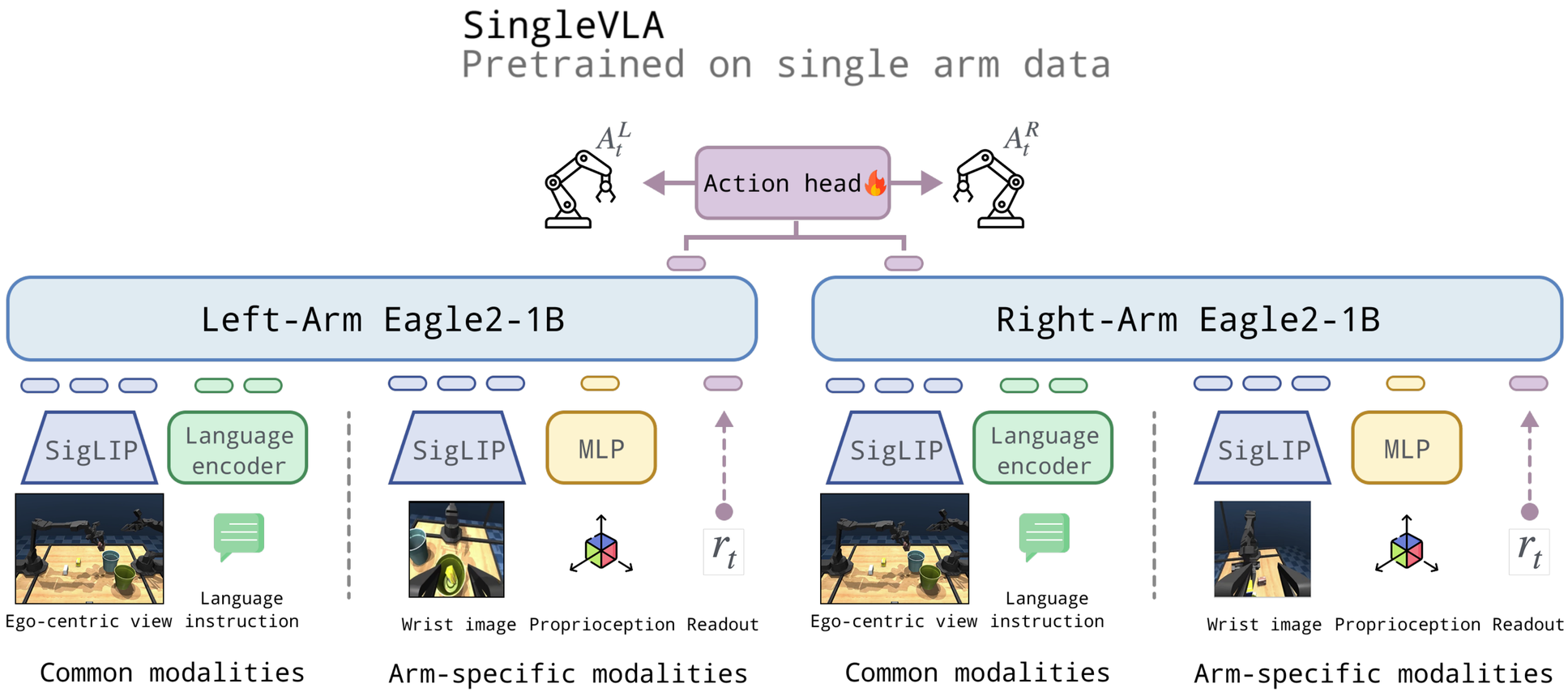
2. Cross-arm fusion using Joint attention

Cross-arm fusion using Joint Attention



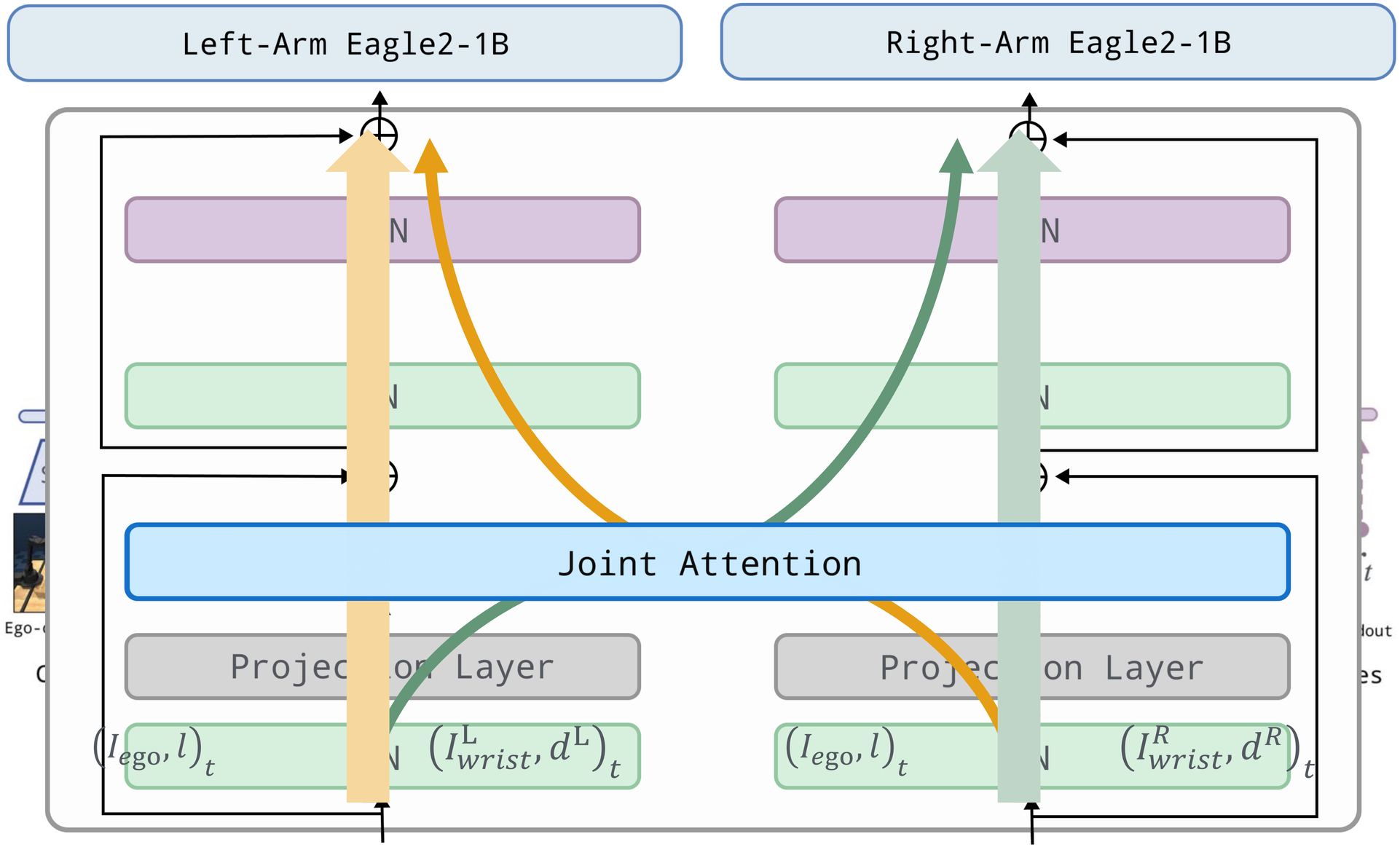
$$o_t^{\text{twin}} = \left(I_{\text{ego}}, l\right)_t \quad \left(I_{wrist}^{\text{L}}, d^{\text{L}}\right)_t \quad \left(I_{\text{ego}}, l\right)_t \quad \left(I_{wrist}^{R}, d^{R}\right)_t$$

Cross-arm fusion using Joint Attention



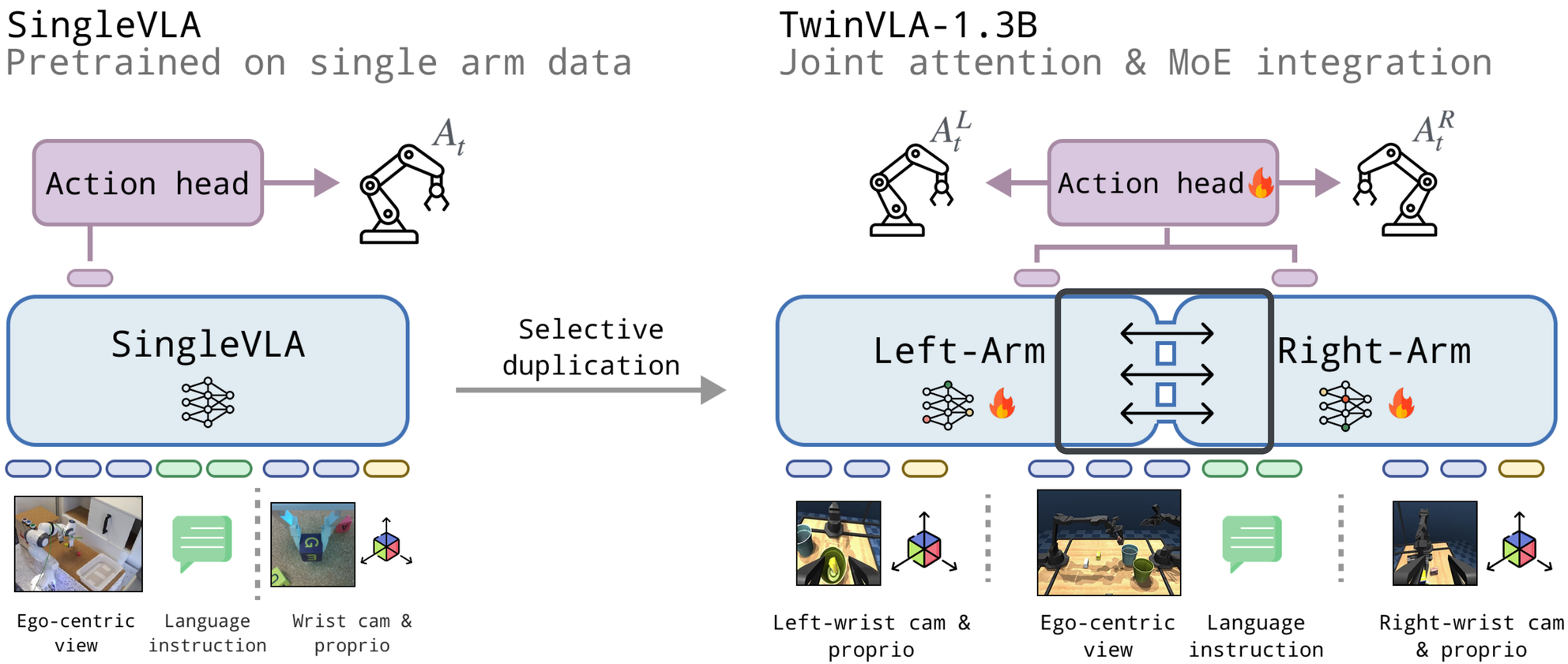
$$o_t^{\text{twin}} = \left(I_{ego}, l\right)_t \quad \left(I_{wrist}^{\text{L}}, d^{\text{L}}\right)_t \quad \left(I_{ego}, l\right)_t \quad \left(I_{wrist}^{R}, d^{R}\right)_t$$

Cross-arm fusion using Joint Attention



$o_t^{\text{twin}} =$

TwinVLA : Data-Efficient Twin Single-Arm VLA

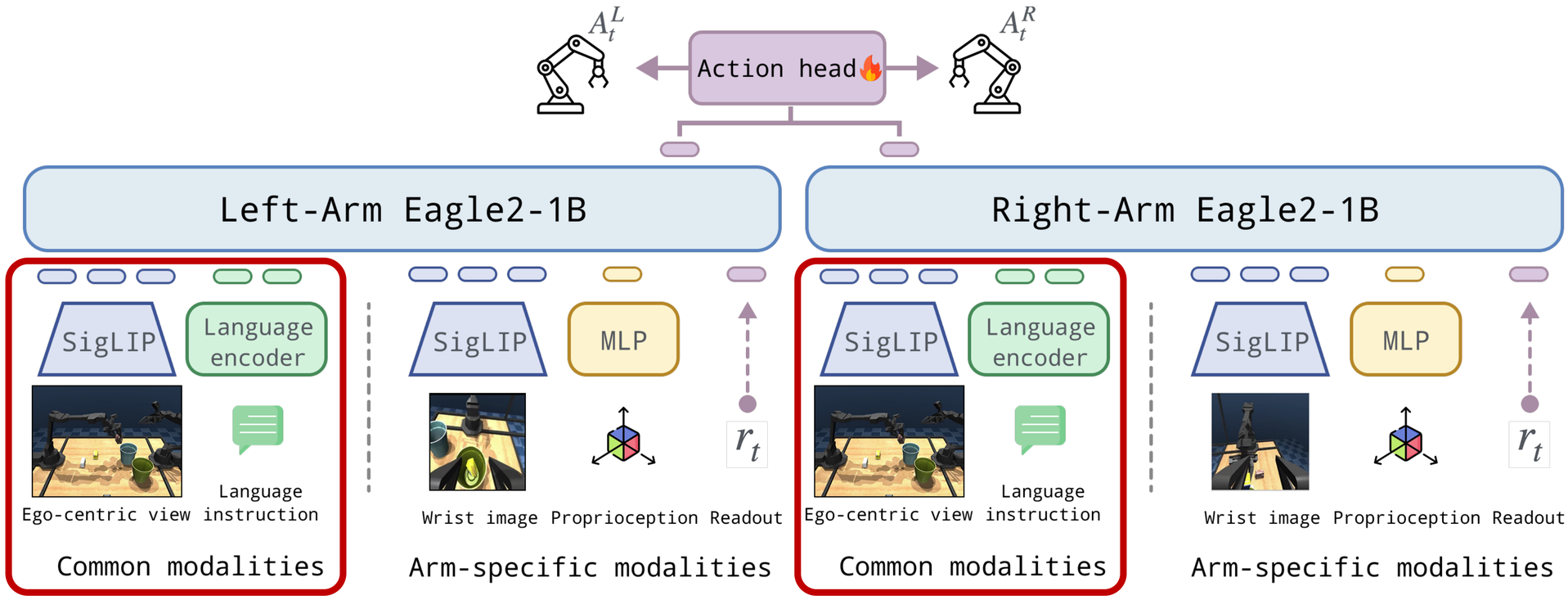


1. Single-Arm VLA pretraining

2. Cross-arm fusion using Joint attention

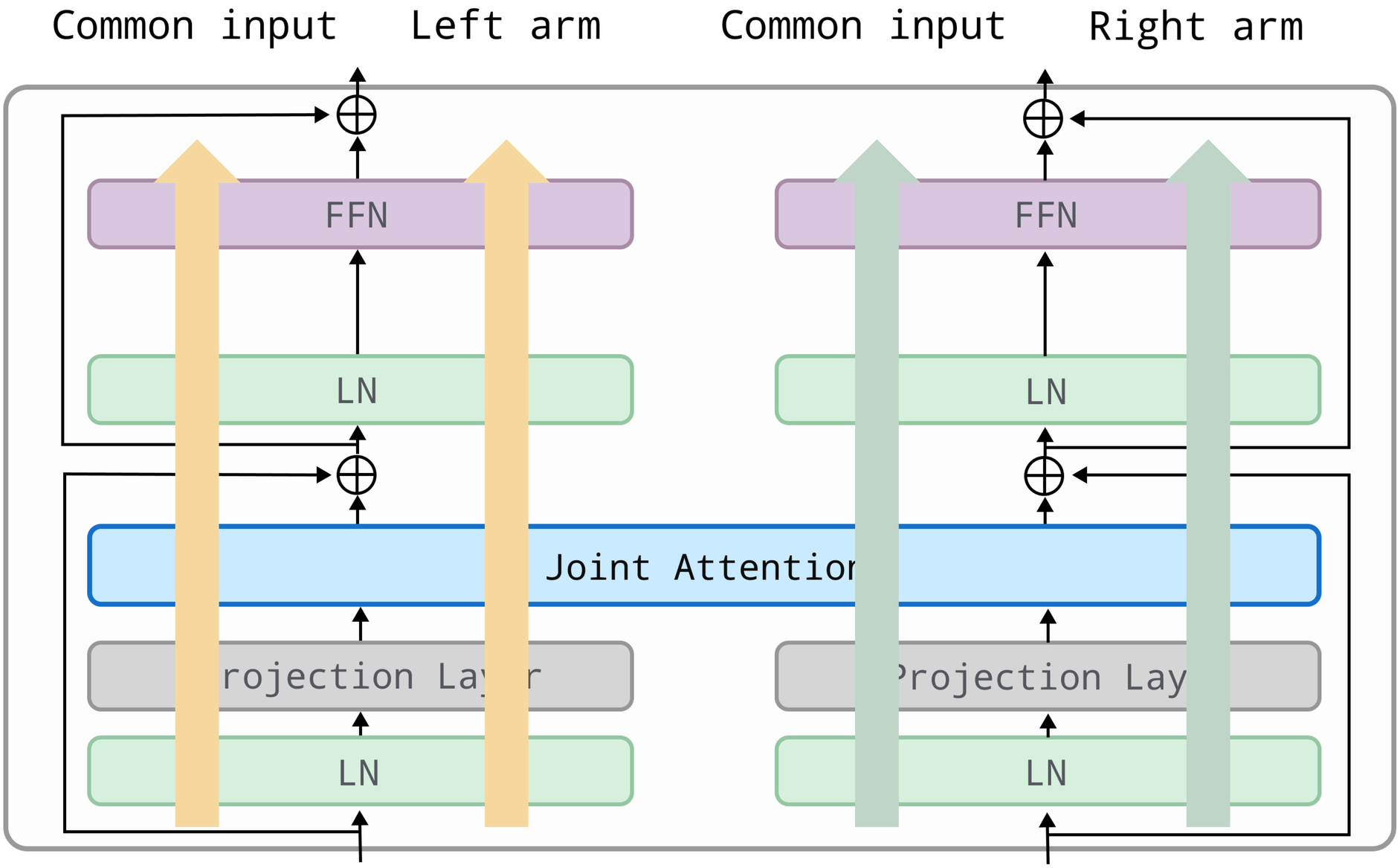
3. MoE for compute efficiency

MoE for compute efficiency

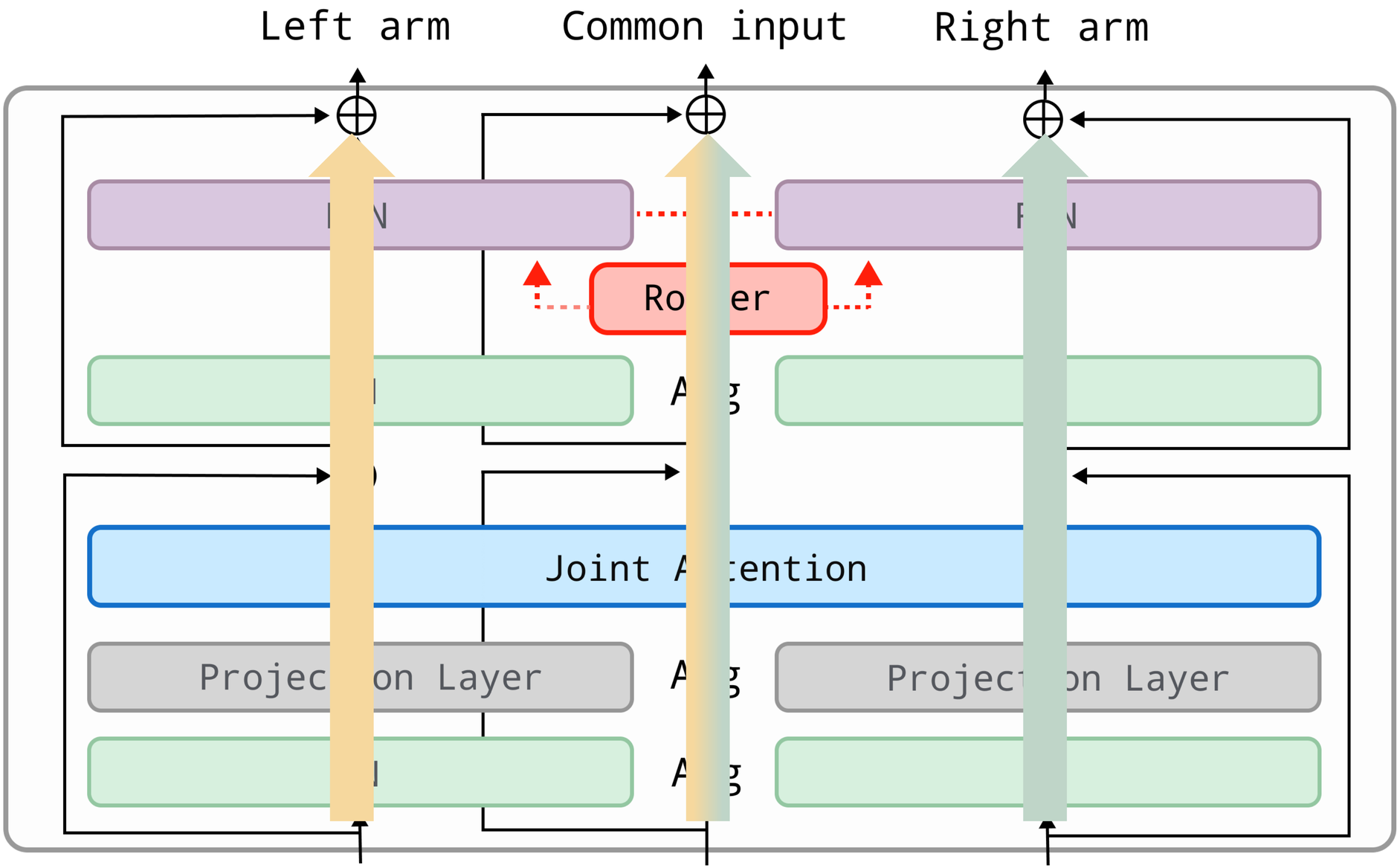


Increased input length → Increased VRAM usage

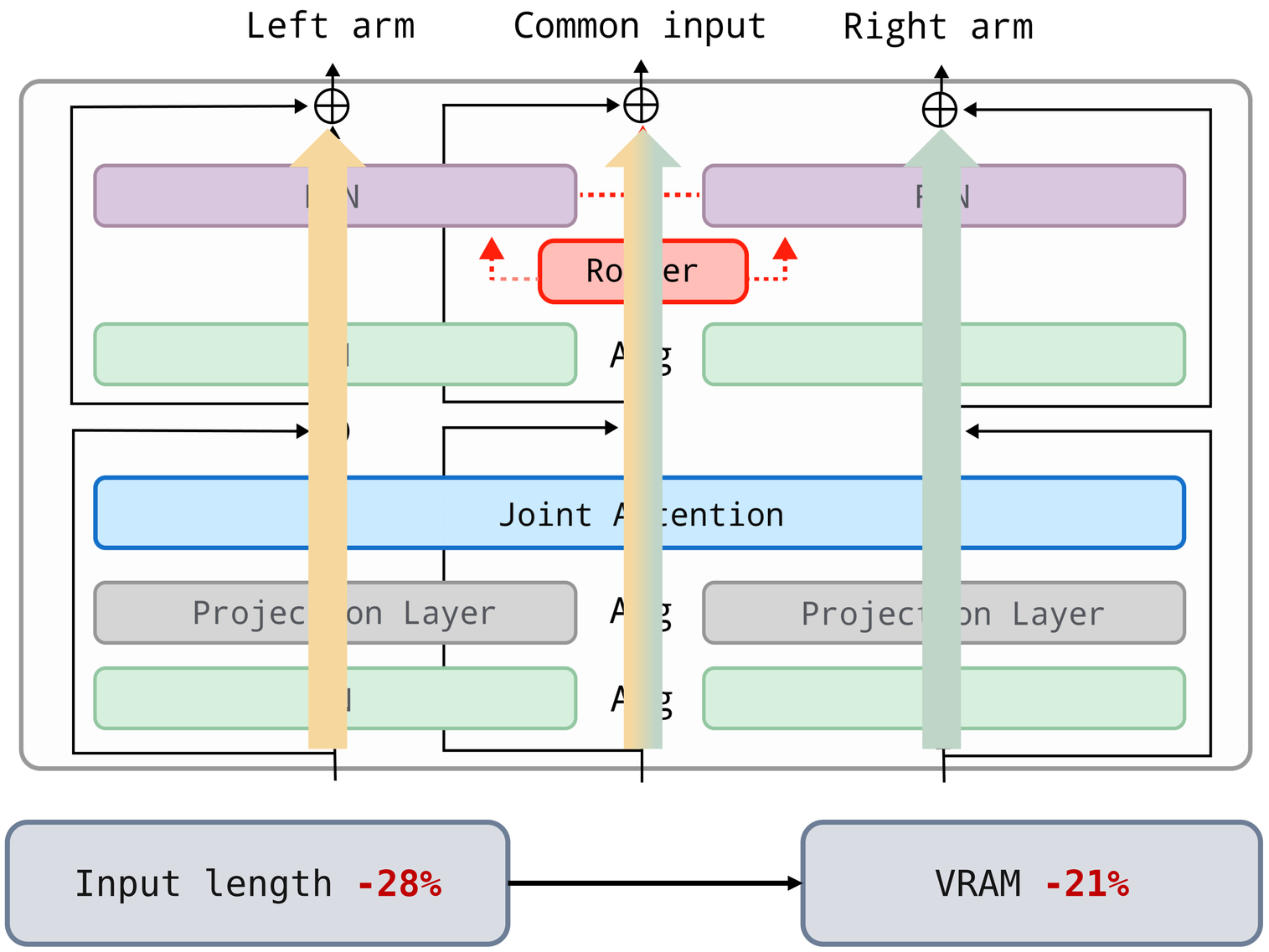
MoE for compute efficiency



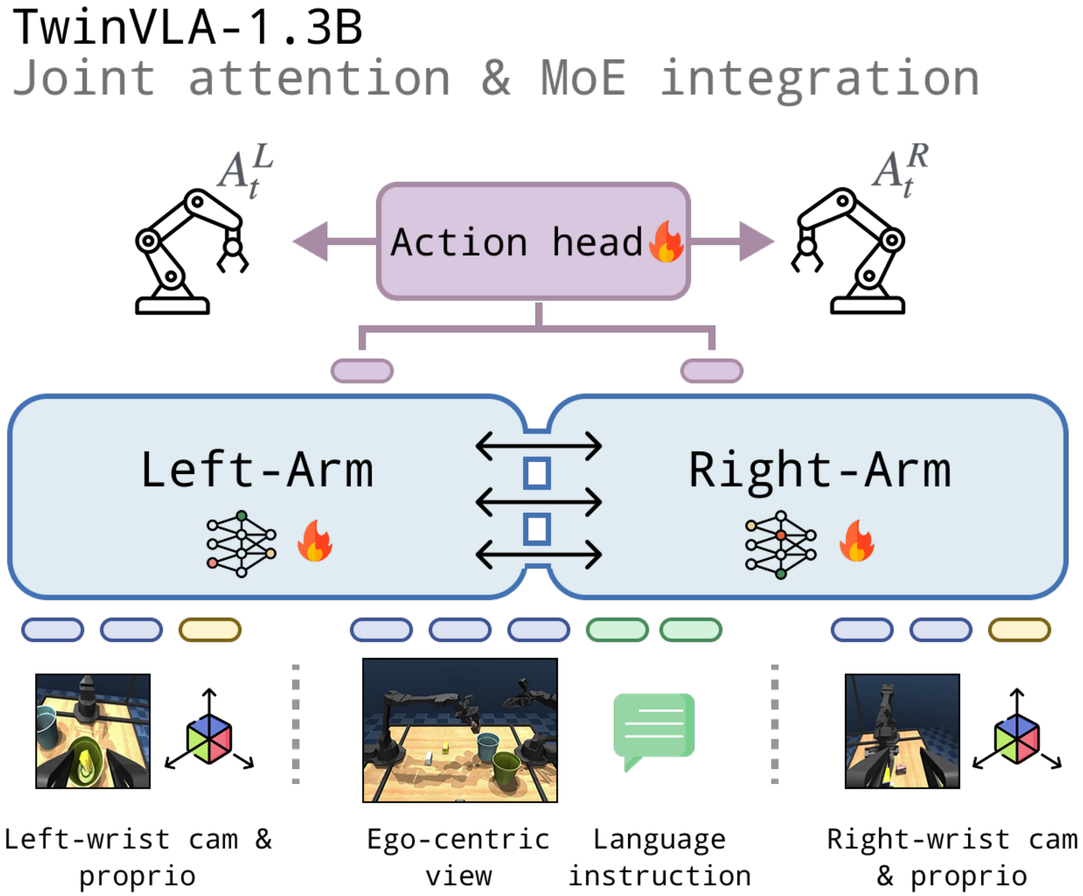
MoE for compute efficiency



MoE for compute efficiency

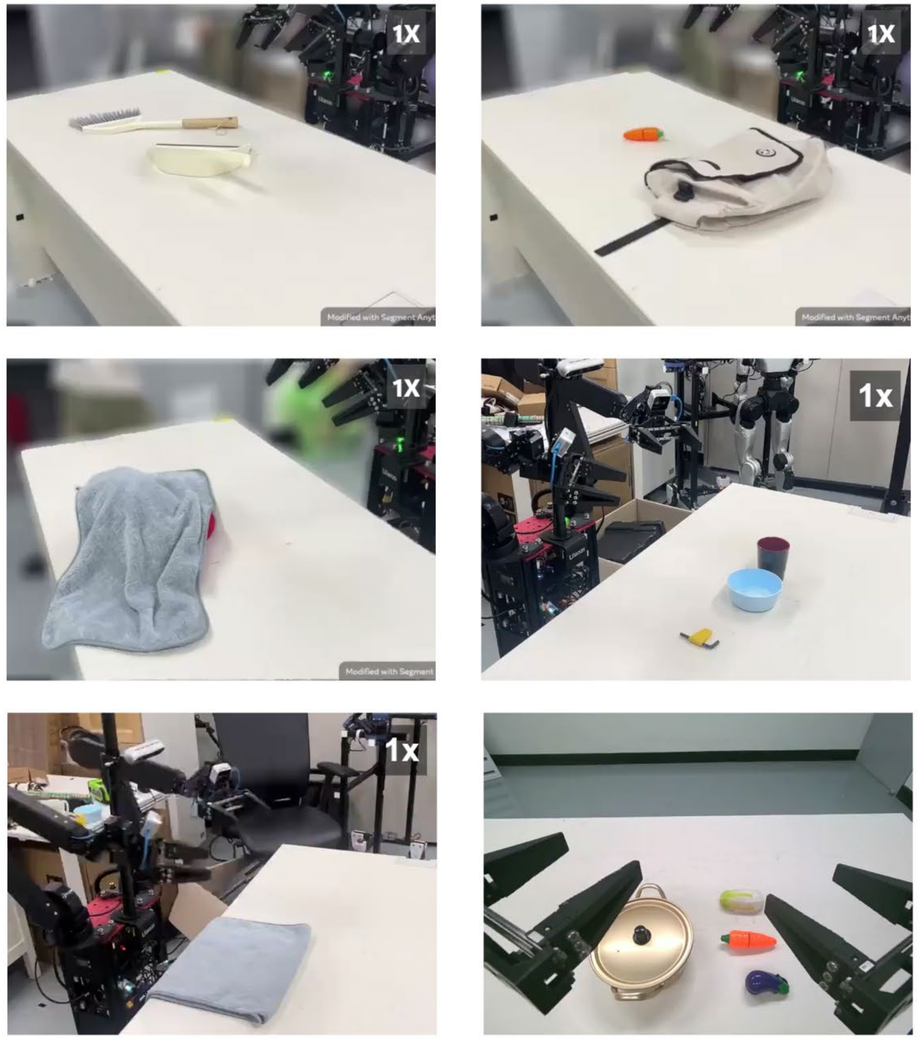


MoE for compute efficiency

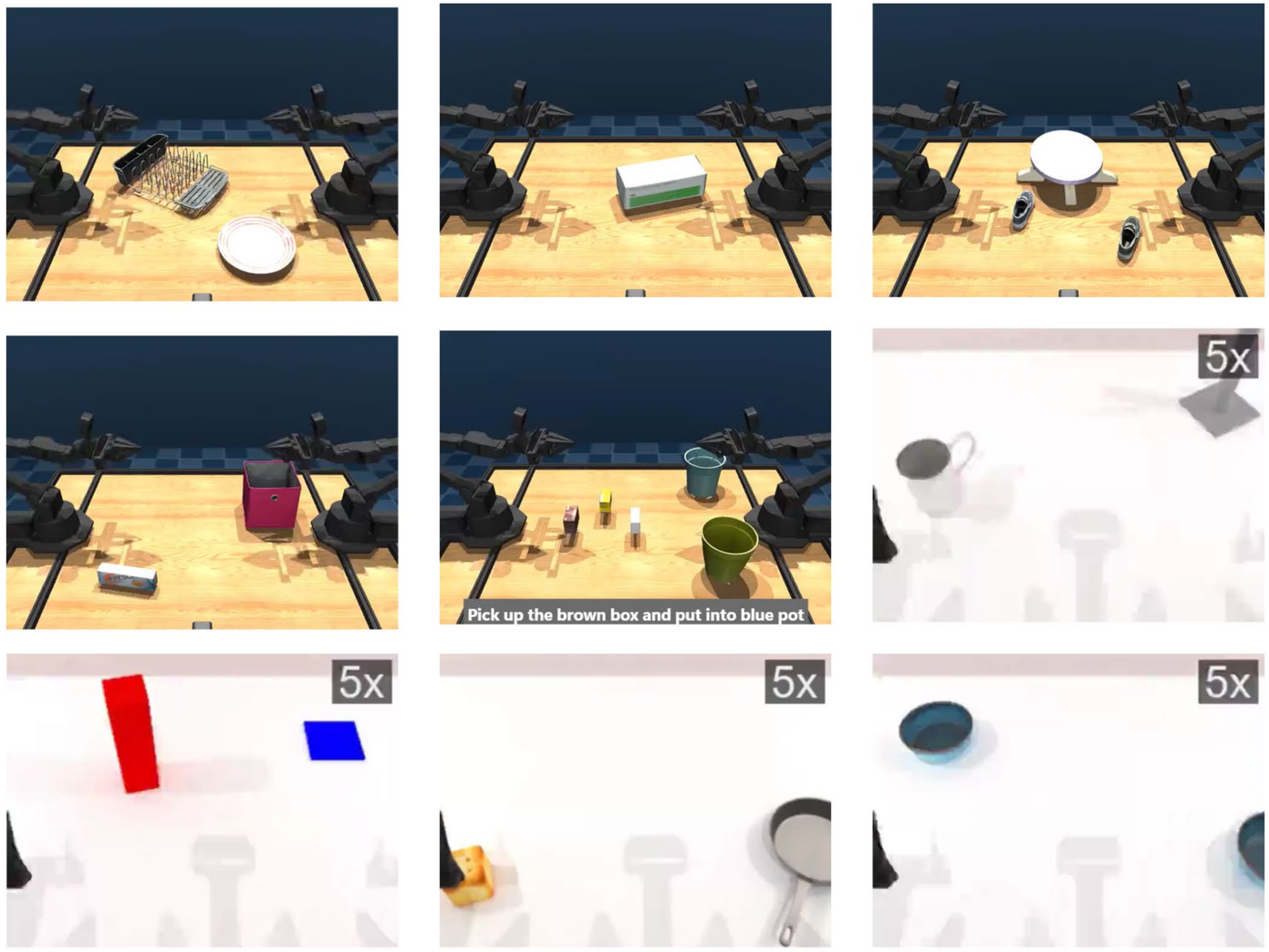


$$o_t^{\text{twin}} = \left(I_{wrist}^{L}, d^{L}\right)_t \quad \left(I_{ego}, l\right)_t \quad \left(I_{wrist}^{R}, d^{R}\right)_t$$

Experiments



Real World (Anubis Bimanual Robot)

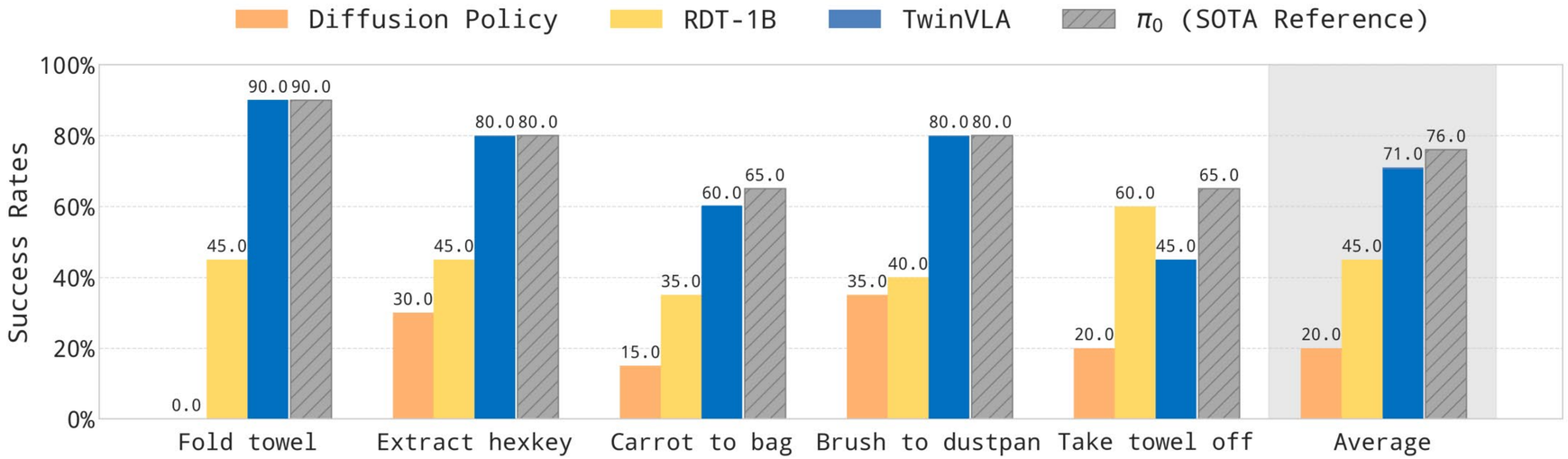


Simulation (Tabletop-Sim & RoboTwin2.0)

Experiments – Real World

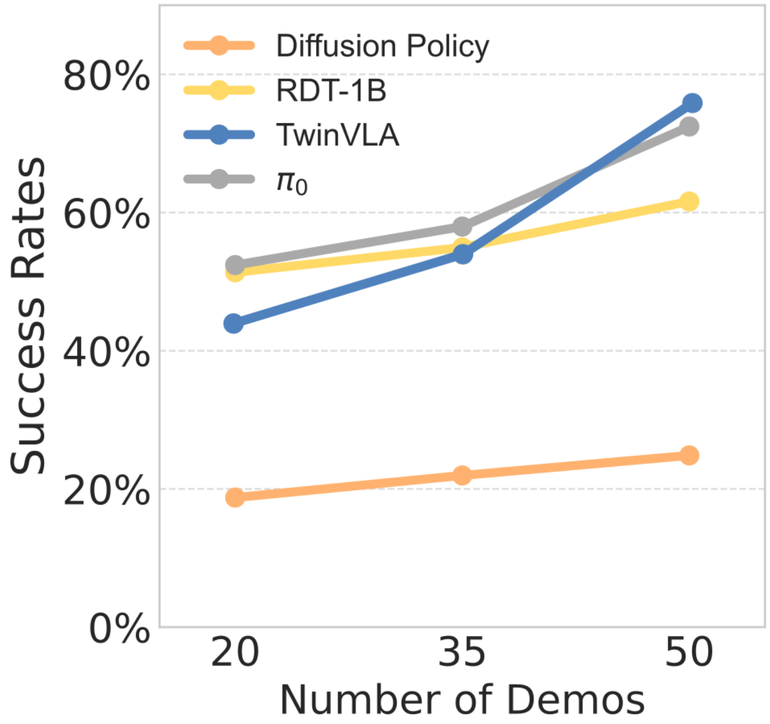
Long-Horizon

Arm Coordination



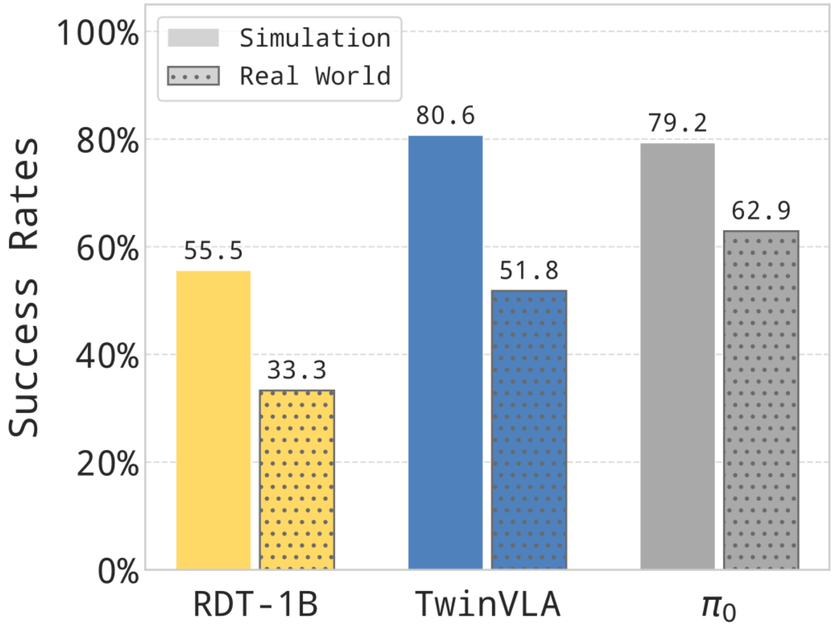
TwinVLA **outperforms** RDT, while comparable to PI0

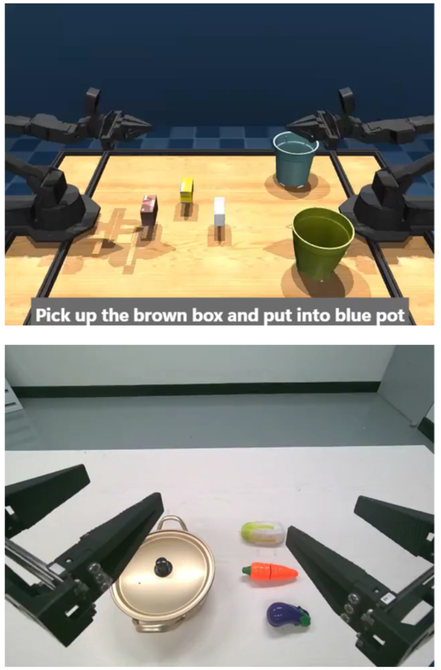
Experiments – Data efficiency & Language following



TwinVLA exhibits exceptional

data efficiency





Robust performance on

language-following tasks

TwinVLA : Data-Efficient Bimanual VLA

Summary

1. **Zero Bimanual Data for Pretraining:**
   We can make a bimanual VLA only from single-arm data

2. **Modular Architecture:**
   Single-arm VLAs + Joint Attention + MoE

3. **Outperforms Monolithic Baselines**