北京大学
PEKING UNIVERSITY

π_0: A Vision-Language-Action Flow Model for General Robot Control

Physical Intelligence

Presenter: Shaofan Sun
2025.2.14

Directory

北京大学 PEKING UNIVERSITY

Background: Vision-Language-Action Models (VLA)

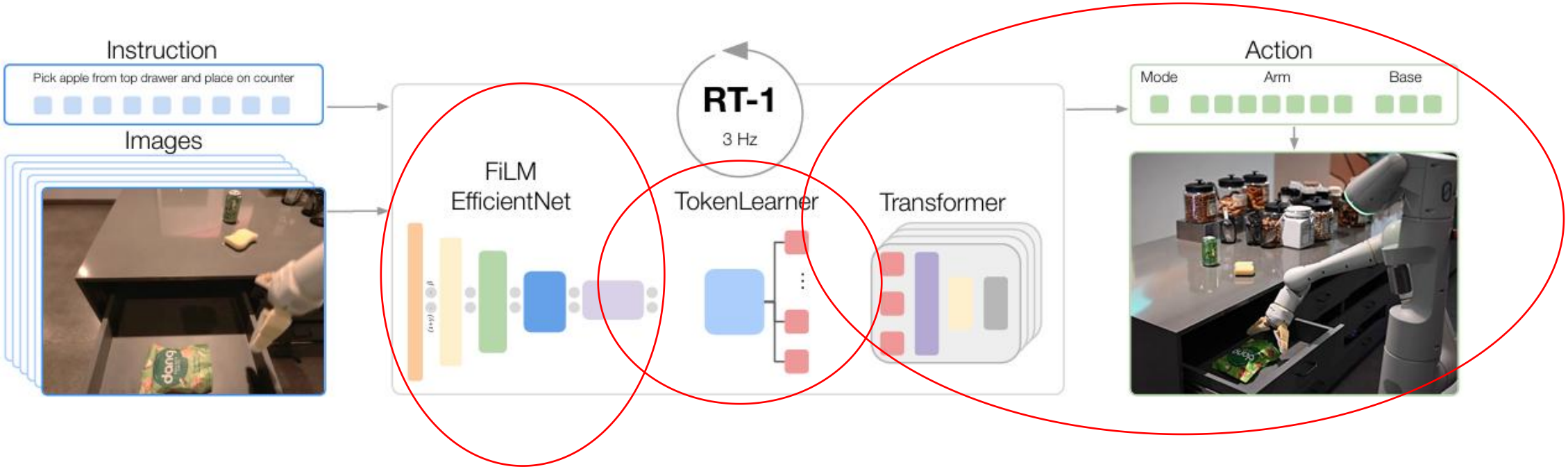
北京大学 PEKING UNIVERSITY

- Do planning for robot manipulation tasks (with large model)

- Main idea: π: $\mathcal{O} \times \mathcal{G} \rightarrow \mathcal{A}$
  - $\mathcal{O}$ is the set of observations, e.g., images or videos
  - $\mathcal{G}$ is the set of language descriptions, e.g., task descriptions
  - $\mathcal{A}$ is the set of robot actions
  - π is the mapping policy (the model)

- Challenges:
  - Multi-modal knowledge fusion
  - Accurate action prediction
  - Generalization ability

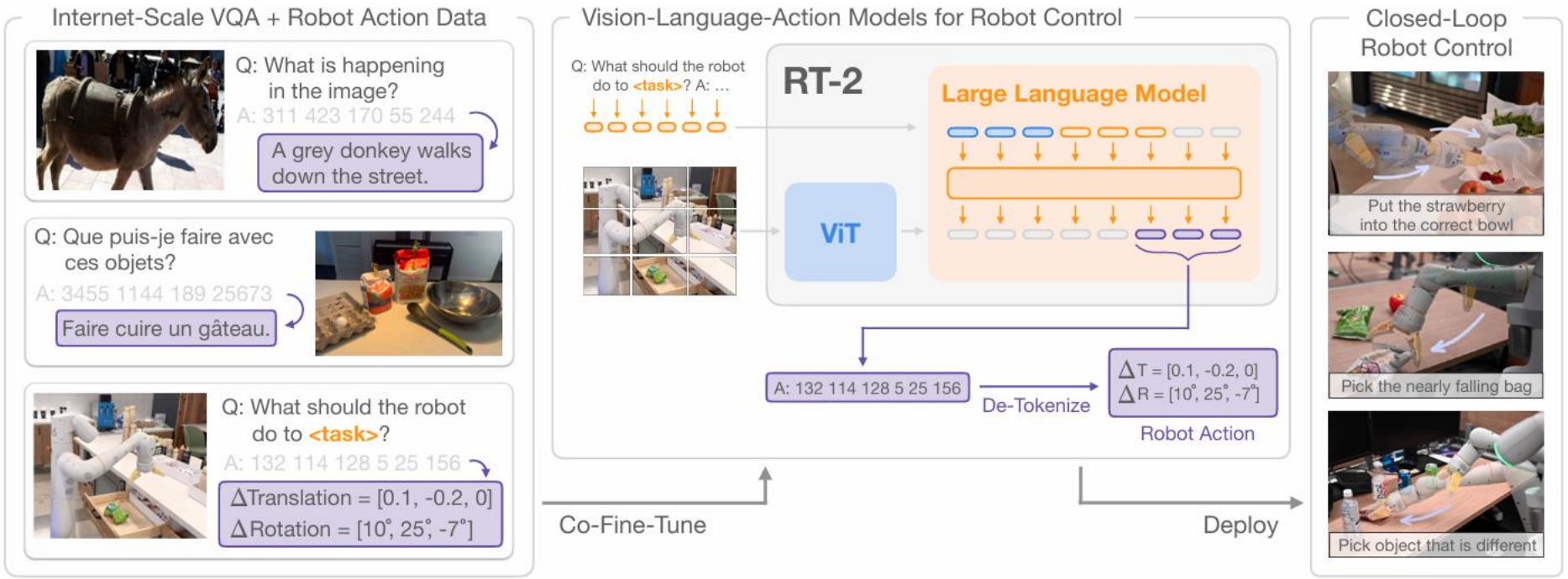
Background: RT-1



- **FiLM** + **EfficientNet** for fusing the knowledge from images and language.
- **TokenLearner** for decreasing the number of tokens passed to Transformer (81 to 8).
- **Transformer** for predicting robot actions.

RT-1: Robotics Transformer for Real-World Control at Scale, Anthony Brohan, et al., arXiv 2022

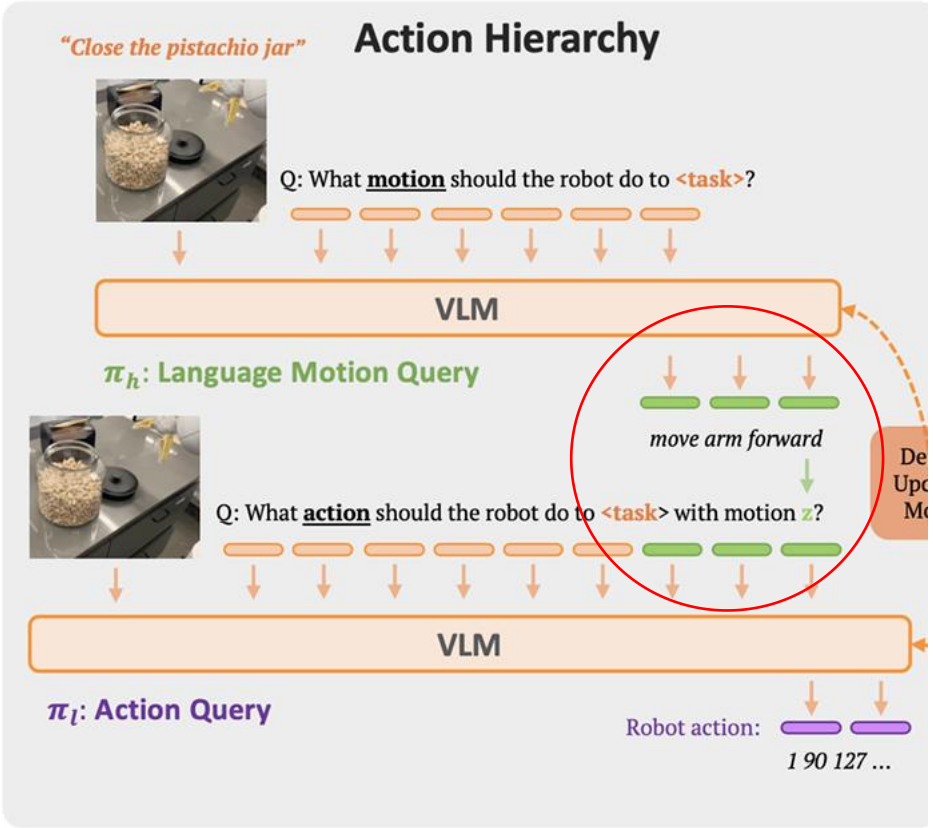
Background: RT-2



- LLM + large-scale data for better generalization ability.
- Generalizable in “understanding”, but not in “predicting”.

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, Anthony Brohan, et al., arXiv 2023

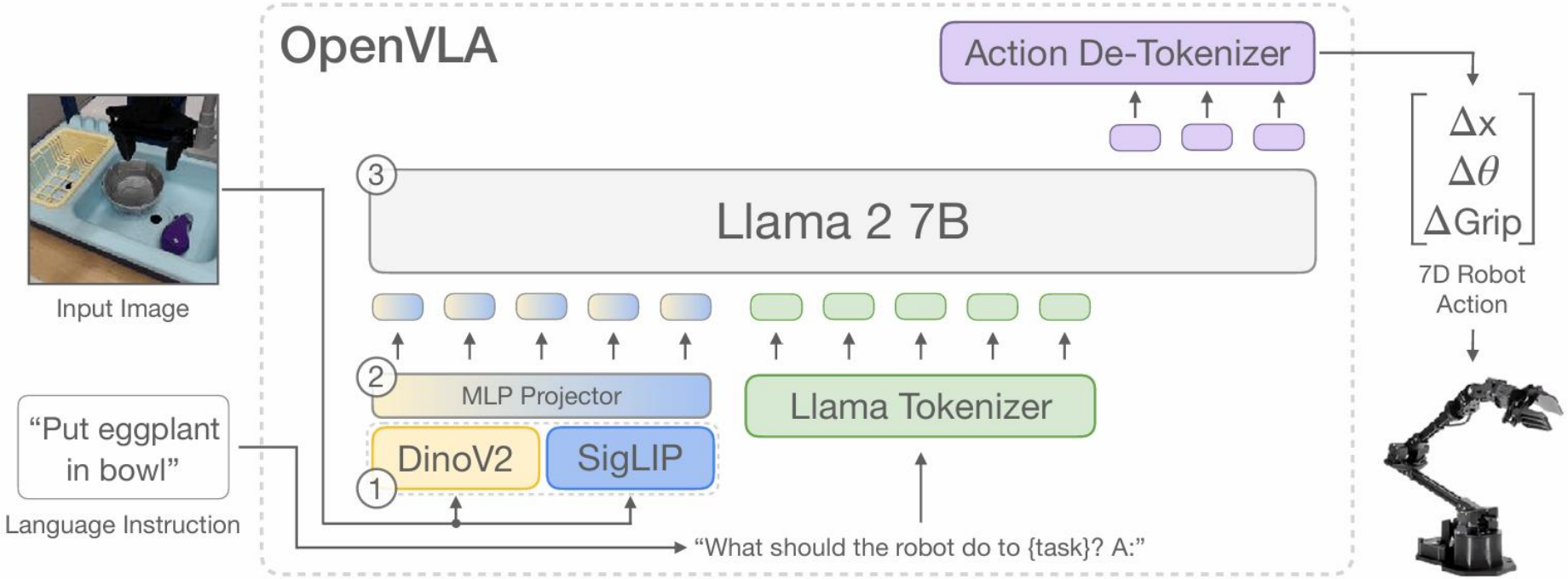
Background: RT-H



- Predict **intermediate action descriptions** and then accurate actions.
- Coarse-grained action descriptions are generalizable.

RT-H: Action Hierarchies Using Language, Suneel Belkhale, et al., arXiv 2024

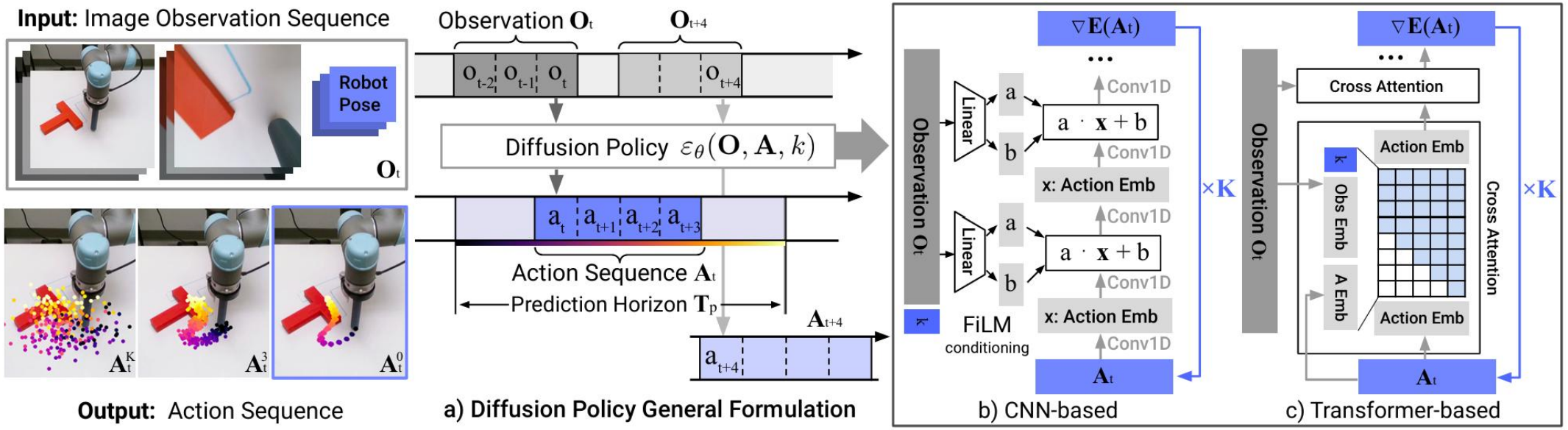
Background: OpenVLA



- The first open-source VLA model.
- Simple and direct, but currently no one has successfully replicated it.

OpenVLA: An Open-Source Vision-Language-Action Model, Moo Jin Kim, et al., CoRL 2024

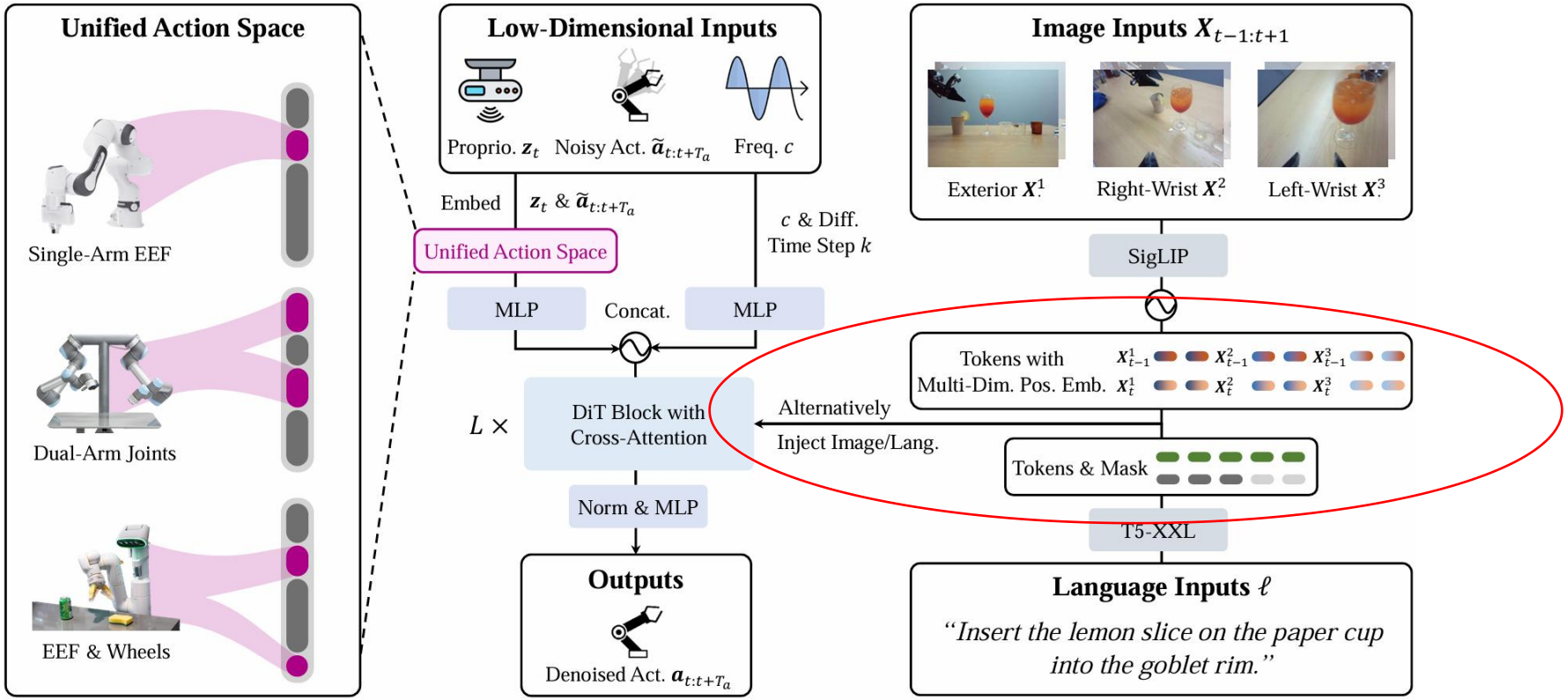
Background: Diffusion Policy



- Input image/video observations as the conditions.
- No language guidance, each model can only handle a single task.

Diffusion Policy: Visuomotor Policy Learning via Action Diffusion, Cheng Chi, et al., RSS 2023

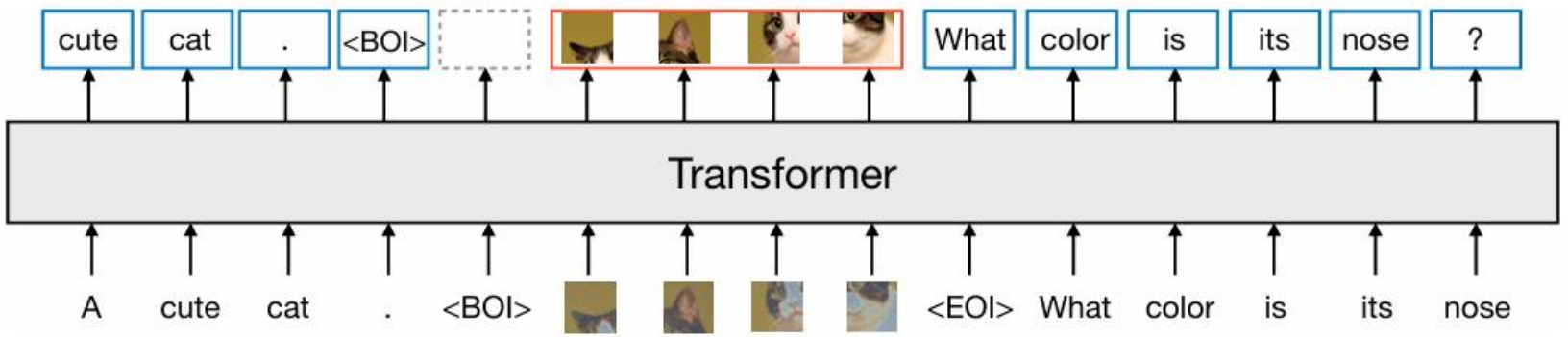
Background: RDT-1B

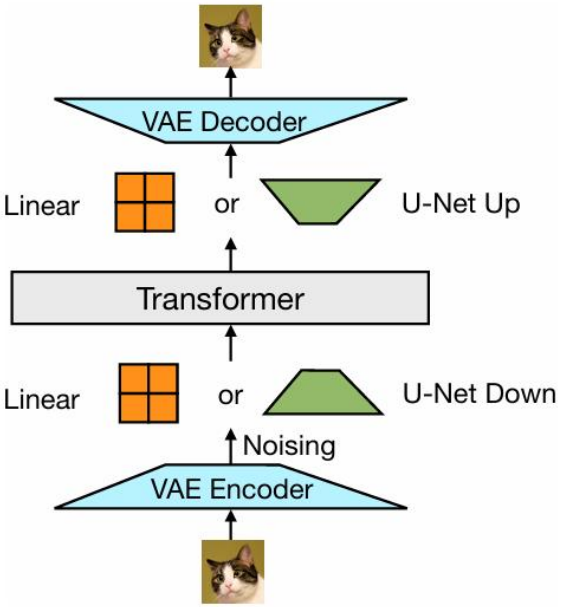


- Largest diffusion model for robot currently.
- Combine image and language tokens and then inject them as the condition.

RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation, Songming Liu, et al., ICLR 2025

Background: Transfusion





- Jointly optimize token and diffusion objectives.

Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model, Chunting Zhou, et al., arXiv 2024

Background: Flow matching

- Training CNFs based on regressing vector fields of fixed conditional probability paths.
- Flow Matching (FM) objective:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,p_t(x)} \left\| v_t(x) - u_t(x) \right\|^2$$

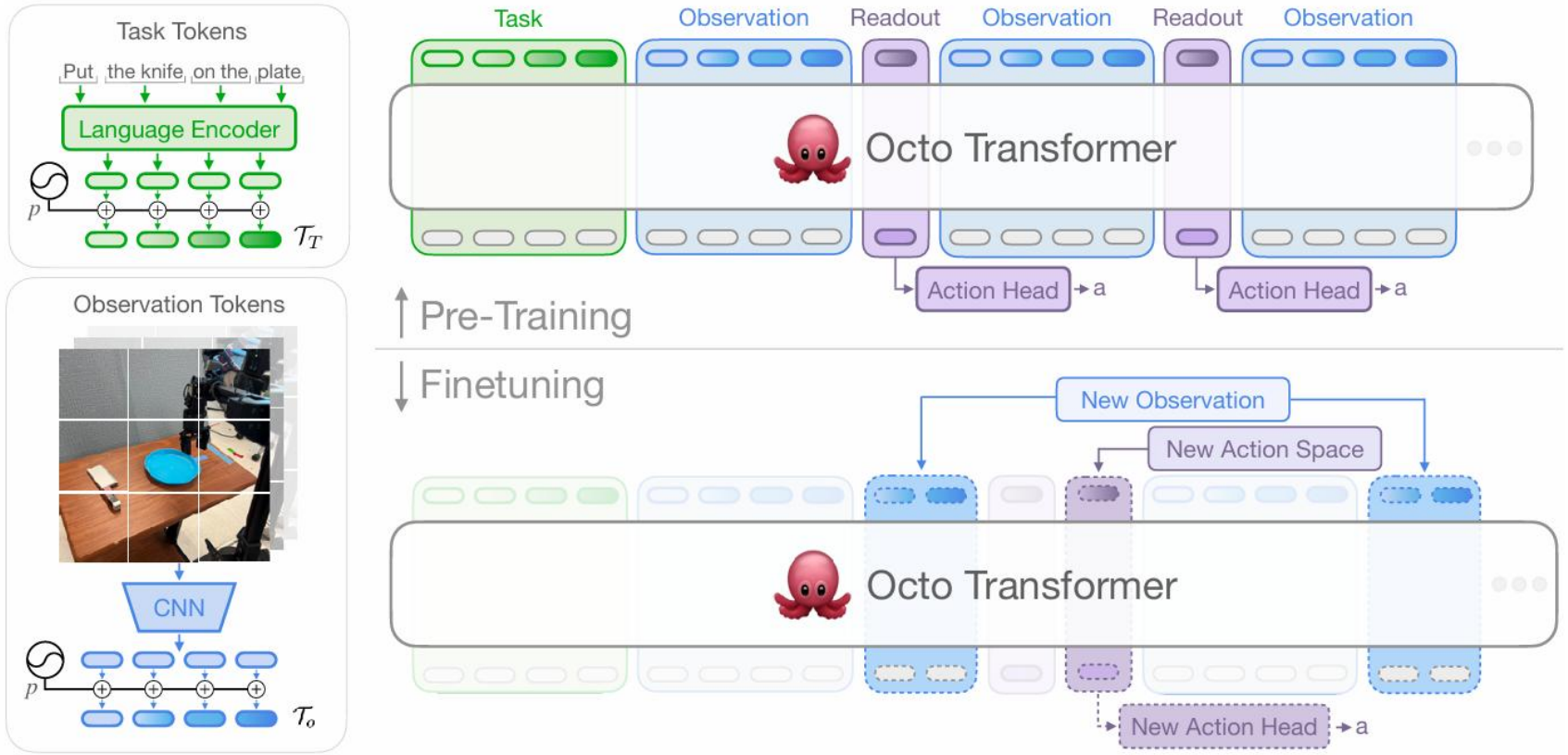
- Have no prior knowledge for p_t and u_t, intractable to compute u_t
- Conditional Flow Matching (CFM) objective:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)} \left\| v_t(x) - u_t(x|x_1) \right\|^2$$

- The FM and CFM objectives have **identical gradients**.

Flow Matching for Generative Modeling, Yaron Lipman, et al., ICLR 2023

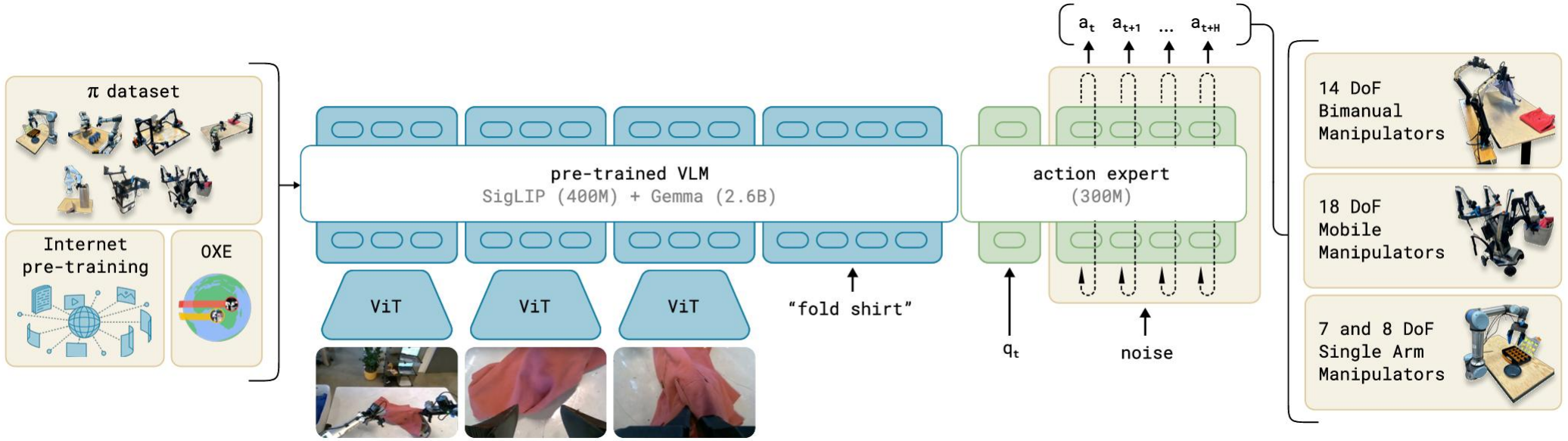
Background: Octo



- Transformer for token-based learning, diffusion head for action decoding.
- The predecessor work of π_0.

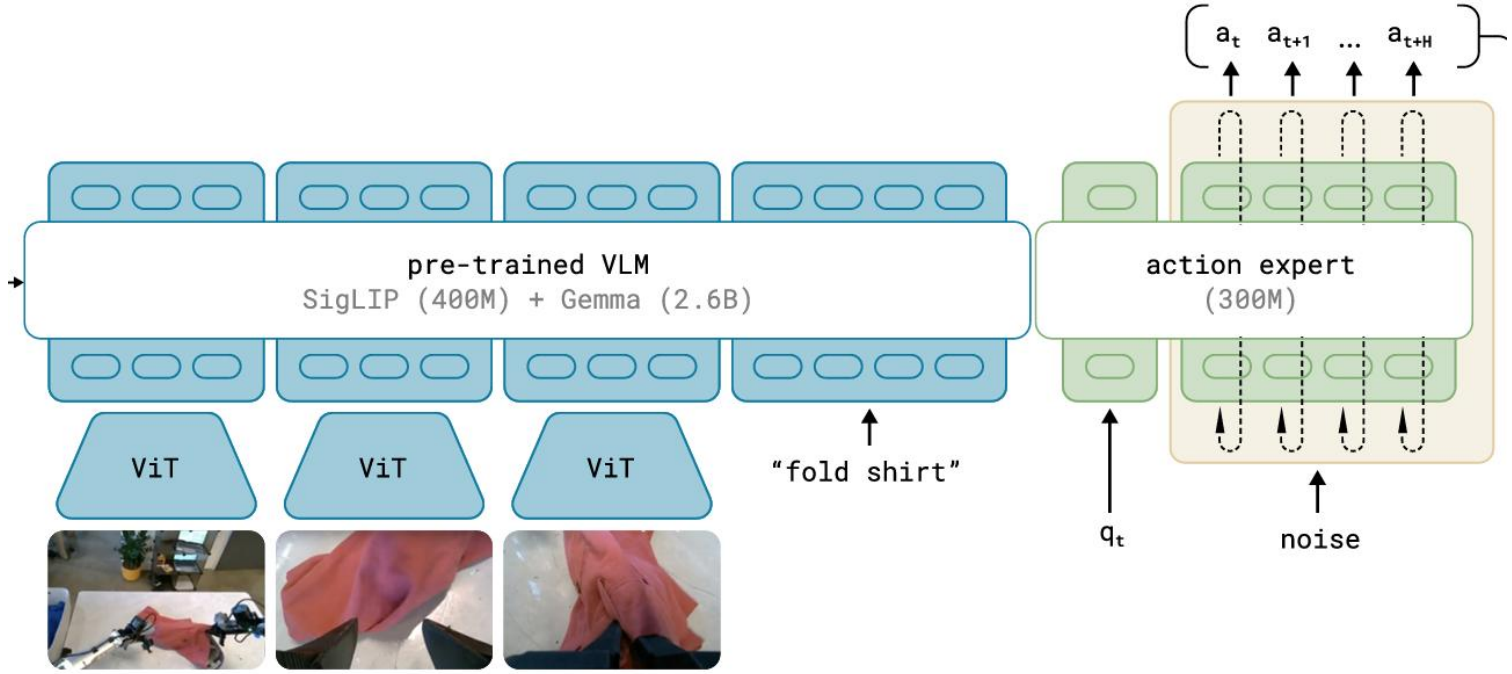
Octo: An Open-Source Generalist Robot Policy, Dibya Ghosh, et al., RSS 2024

Method: Overview



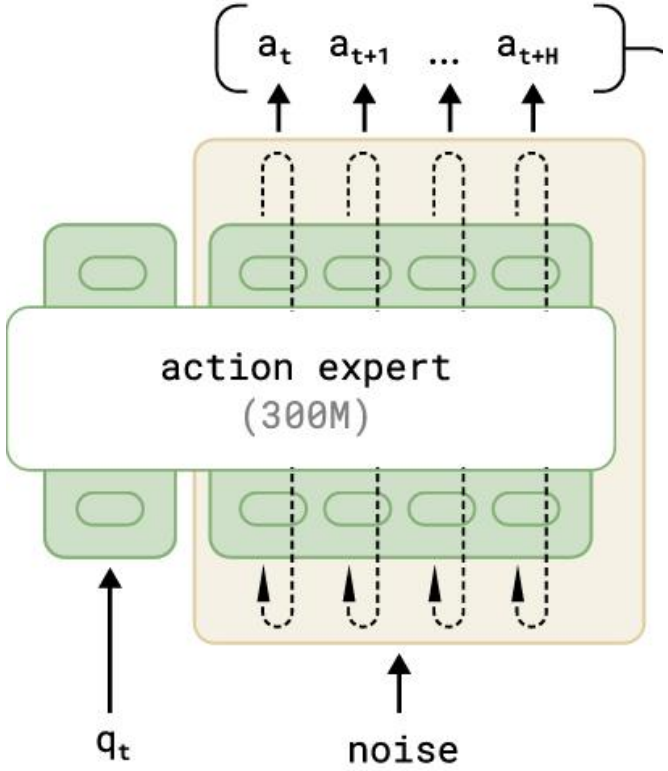
- Data: internet-scale data, large robotics data (OXE), π dataset
- Architecture: VLM for multi-modal understanding, flow matching for action decoding
- Deployment: multiple kinds of robots

Method: Architecture



- VLM backbone: initialized from PaliGemma, SigLIP + Gemma
- Action expert: a separate set of weights for the action modality, analogous to a mixture of experts

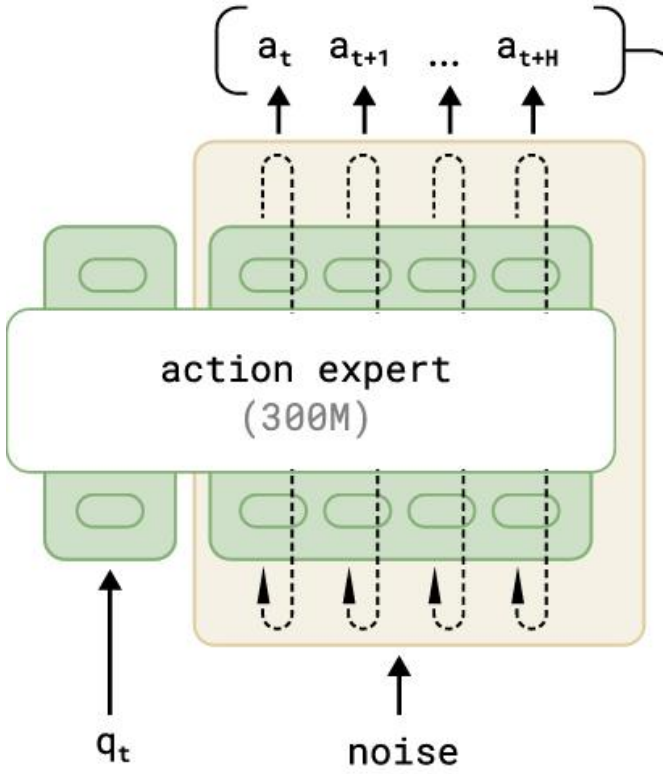
Method: Architecture



- Action chunk, H=50
  - Efficiency
  - Semantics

- Conditional flow matching
  - Modeling continuous action distributions
  - Efficiency

Method: Architecture



- Train:

$$L^\tau(\theta) = \mathbb{E}_{p(\mathbf{A}_t|\mathbf{o}_t),q(\mathbf{A}_t^\tau|\mathbf{A}_t)}||\mathbf{v}_\theta(\mathbf{A}_t^\tau, \mathbf{o}_t) - \mathbf{u}(\mathbf{A}_t^\tau|\mathbf{A}_t)||^2$$

$$q(\mathbf{A}_t^\tau|\mathbf{A}_t) = \mathcal{N}(\tau\mathbf{A}_t, (1-\tau)\mathbf{I})$$

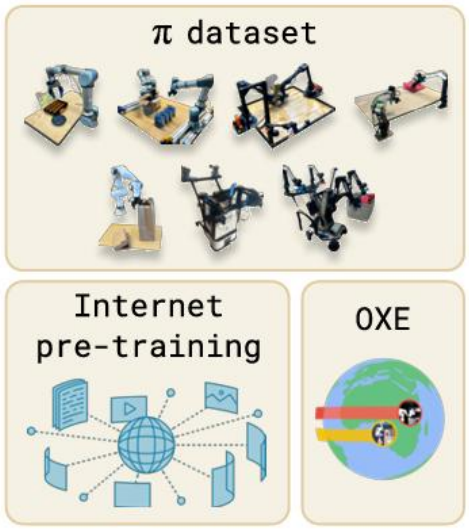
$$\mathbf{A}_t^\tau = \tau\mathbf{A}_t + (1-\tau)\epsilon$$

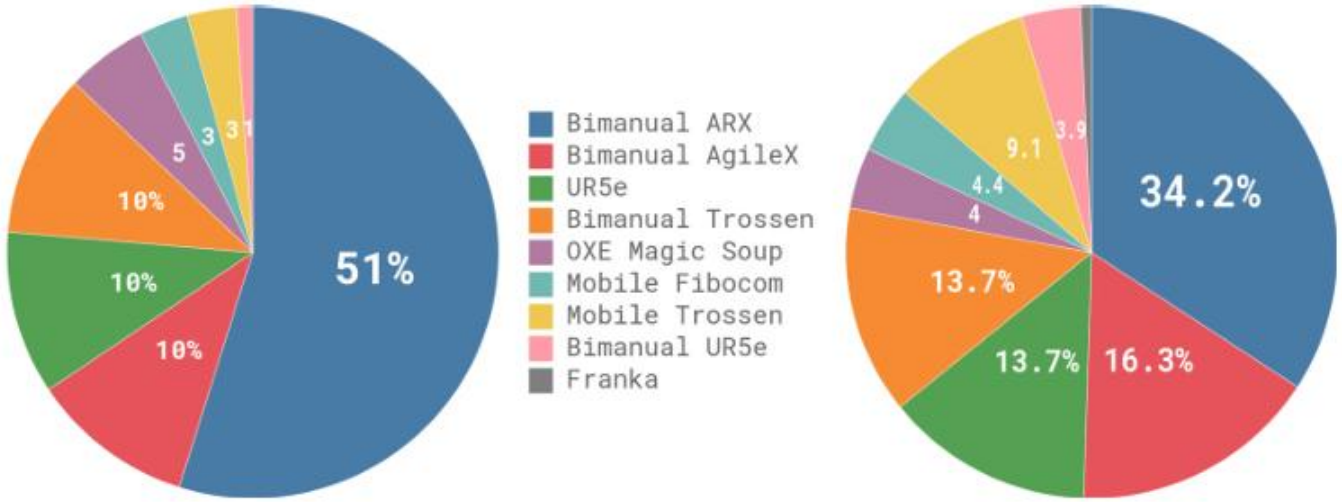
$$\mathbf{u}(\mathbf{A}_t^\tau|\mathbf{A}_t) = \epsilon - \mathbf{A}_t$$

- Inference:

$$\mathbf{A}_t^{\tau+\delta} = \mathbf{A}_t^\tau + \delta\mathbf{v}_\theta(\mathbf{A}_t^\tau, \mathbf{o}_t)$$

Method: Data





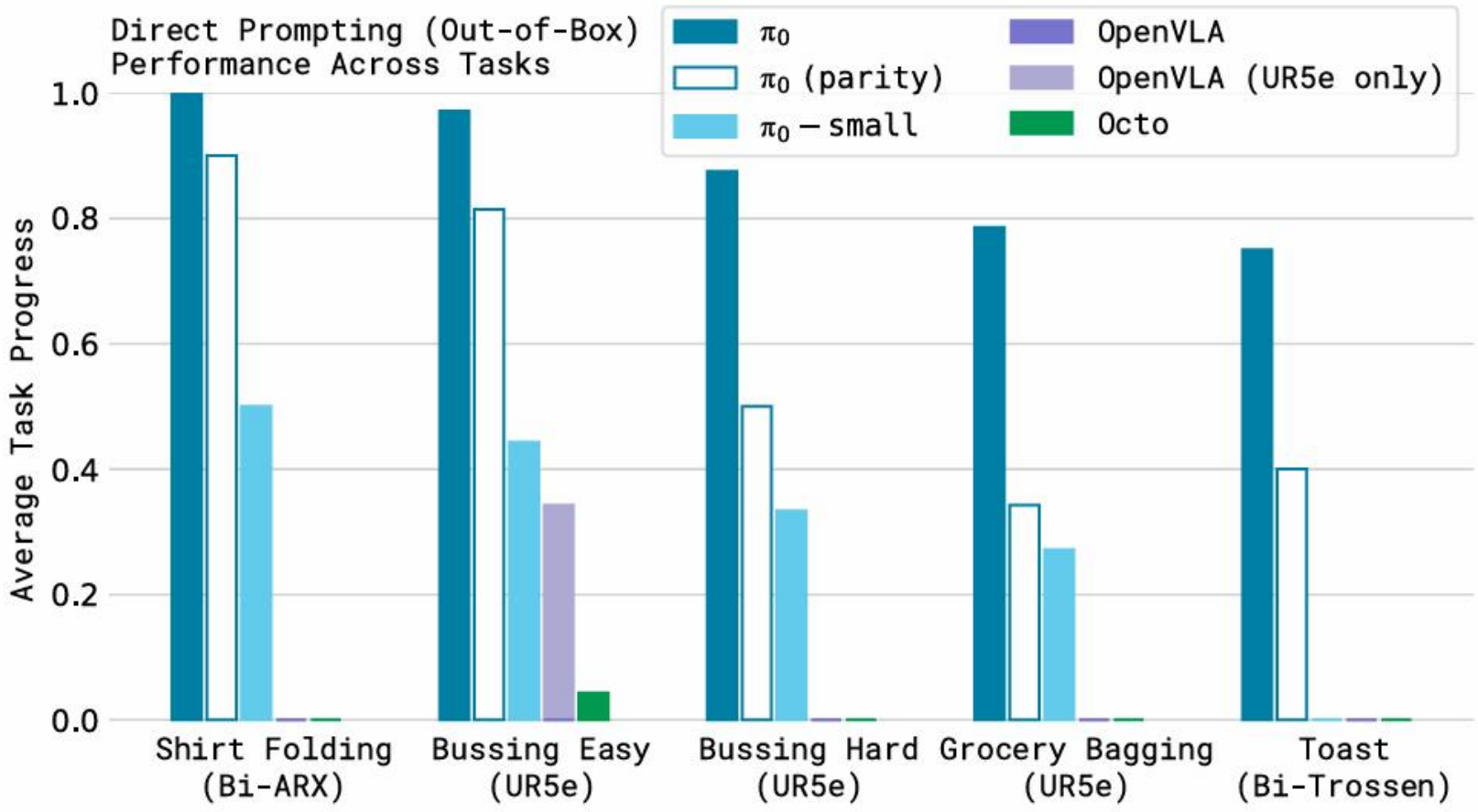
- Internet-scale data: inherited from pretrained VLM
- Open-X-Embodiment: a large-scale robot manipulation dataset
- π dataset: collected by the authors

Method: Language commands

北京大学 PEKING UNIVERSITY

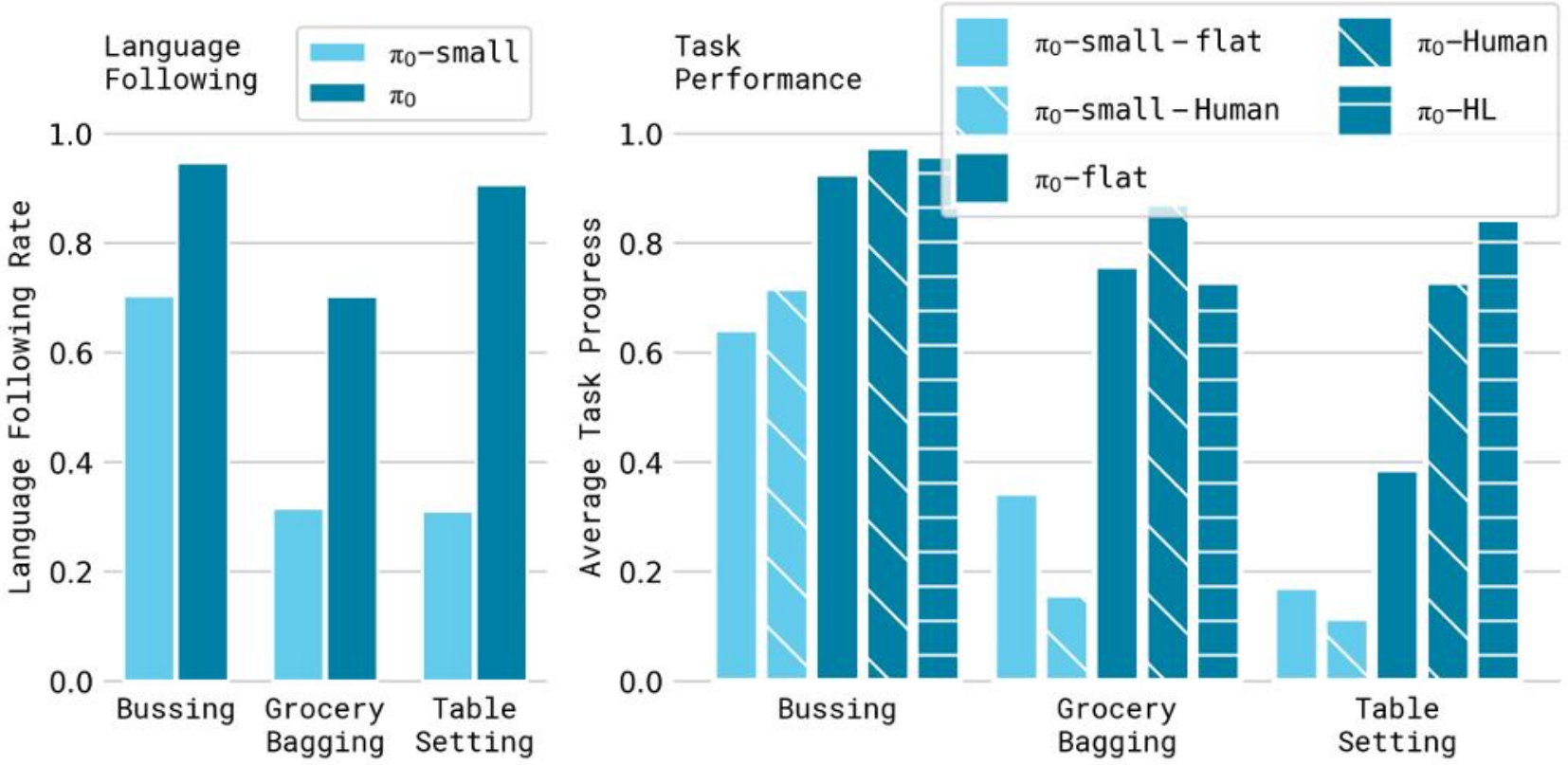
- Three kinds of commands:
  - high-level task commands like “bus the table”
  - intermediate subtasks like “pick up the napkin” and “throw the napkin into the trash”
  - use VLM to make semantic inferences

Experiments: Without post-training

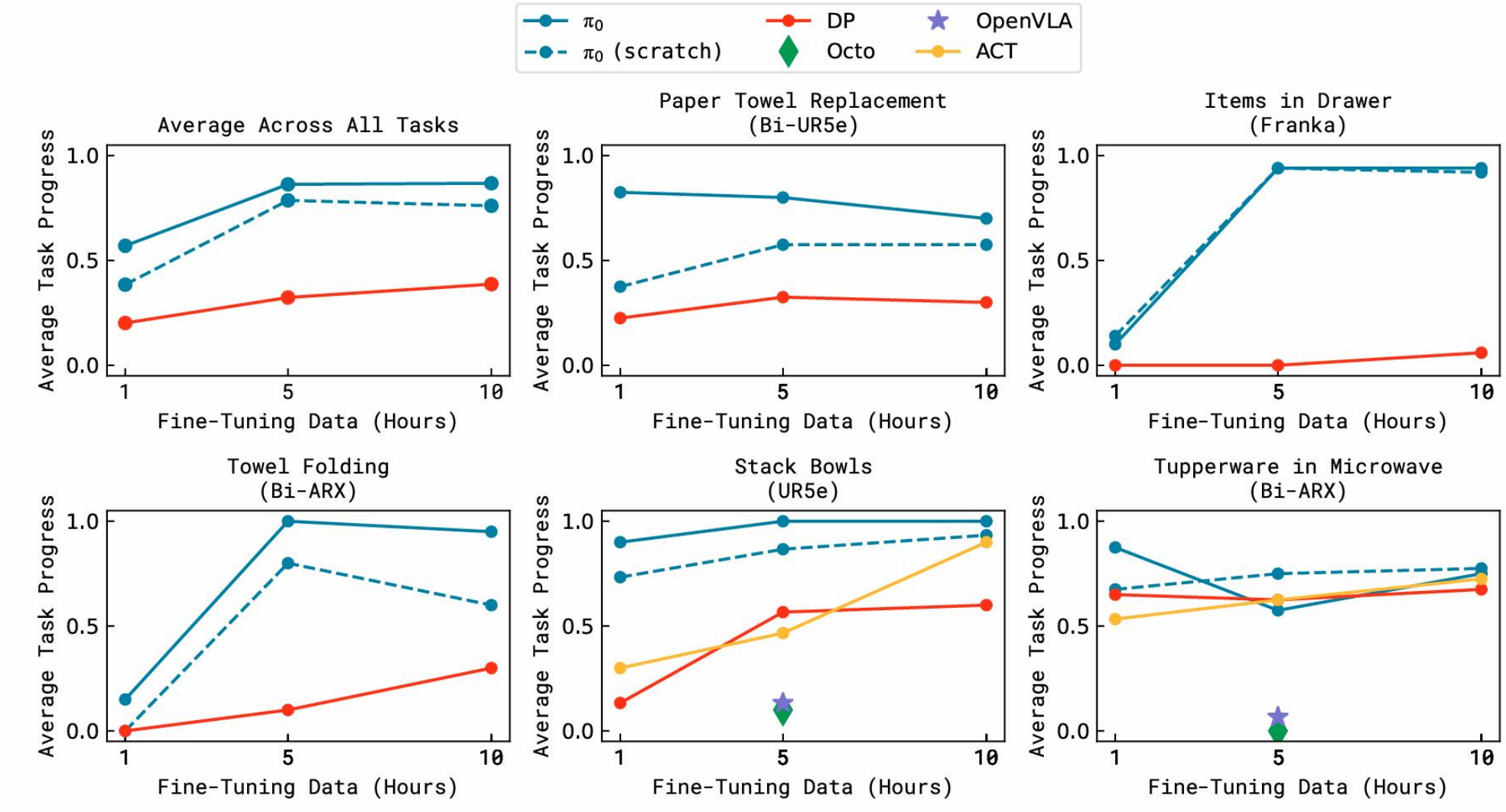


Experiments: Intermediate commands

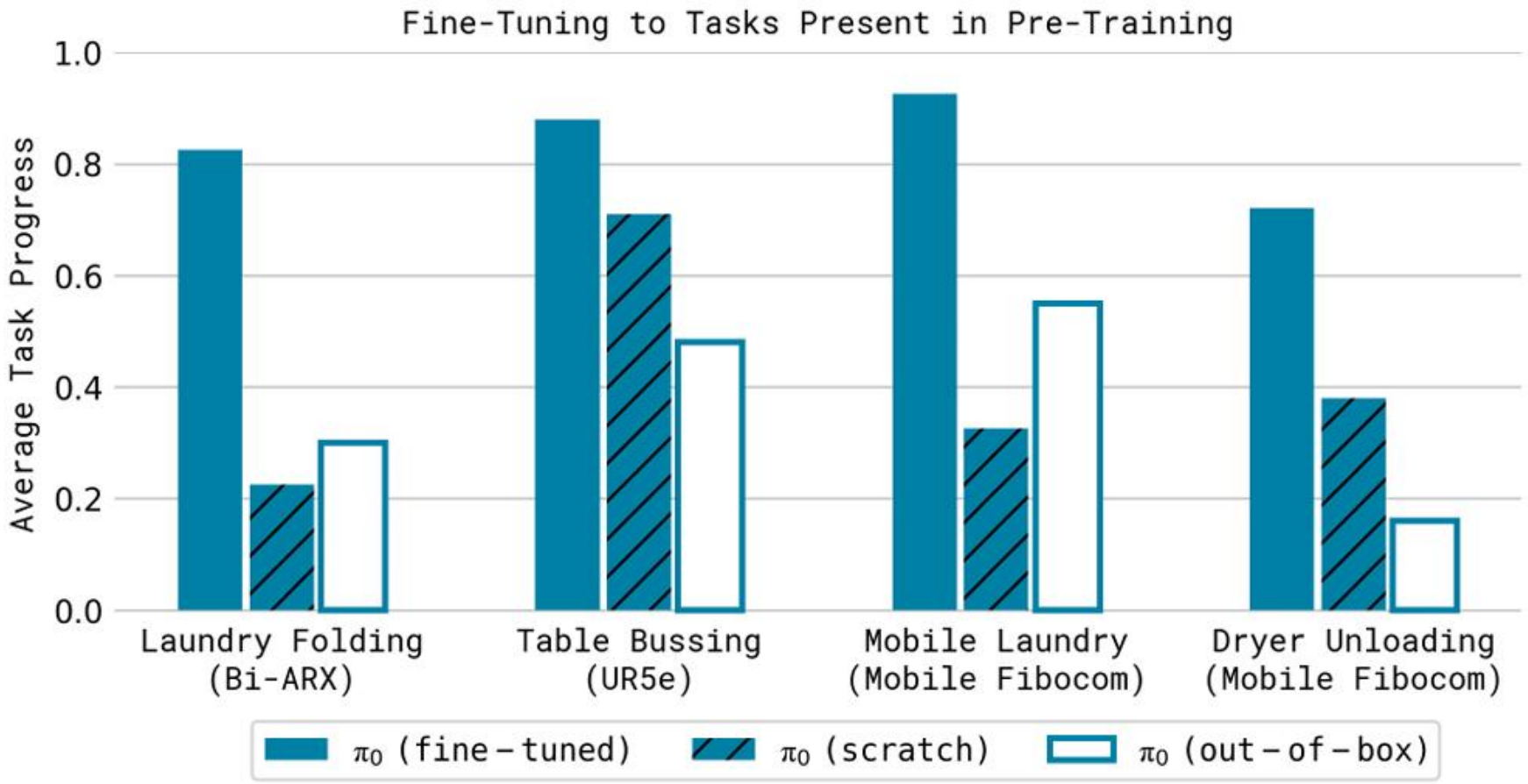
- Benefit more from intermediate commands with initialized VLM



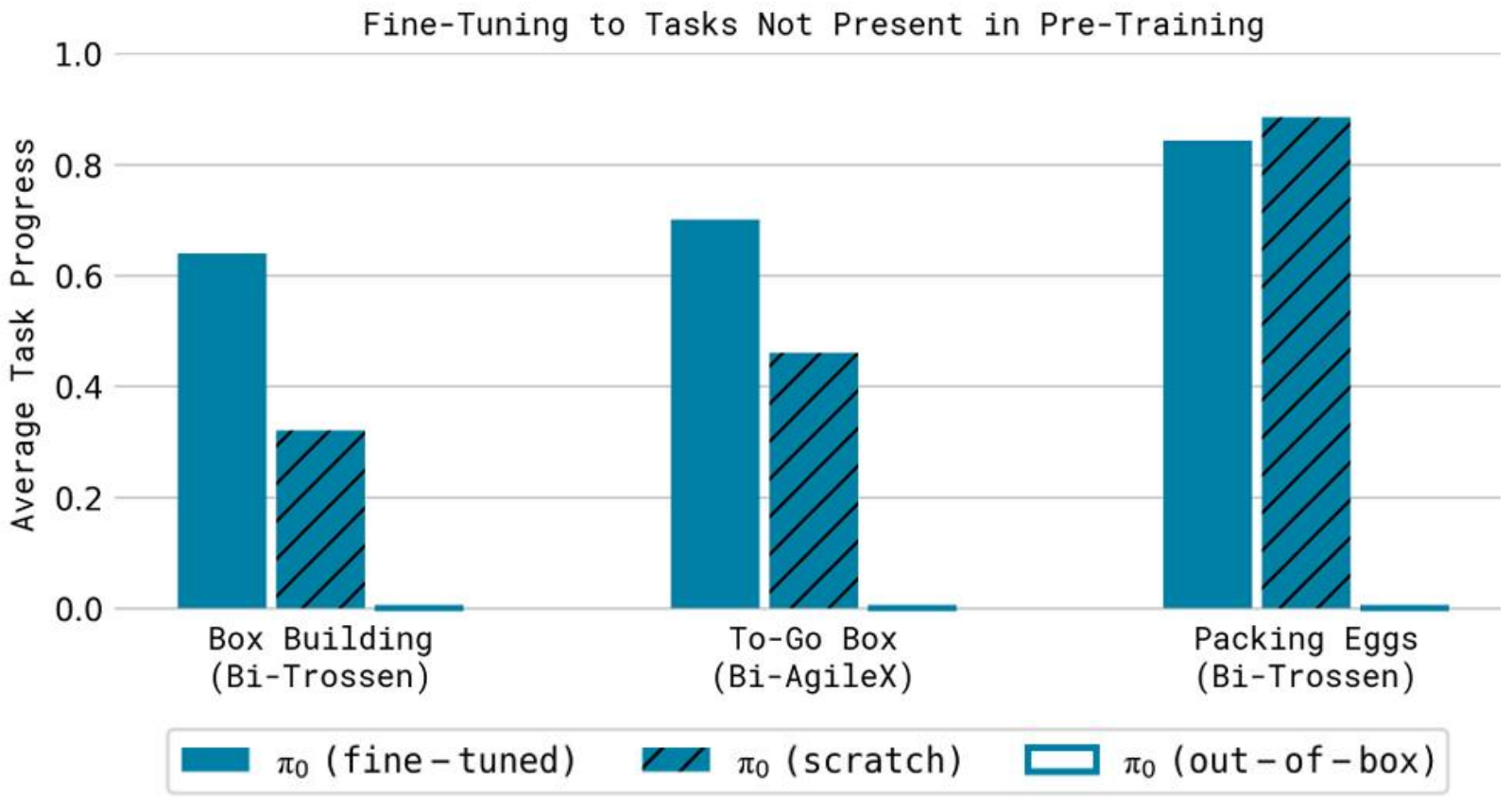
Experiments: New tasks



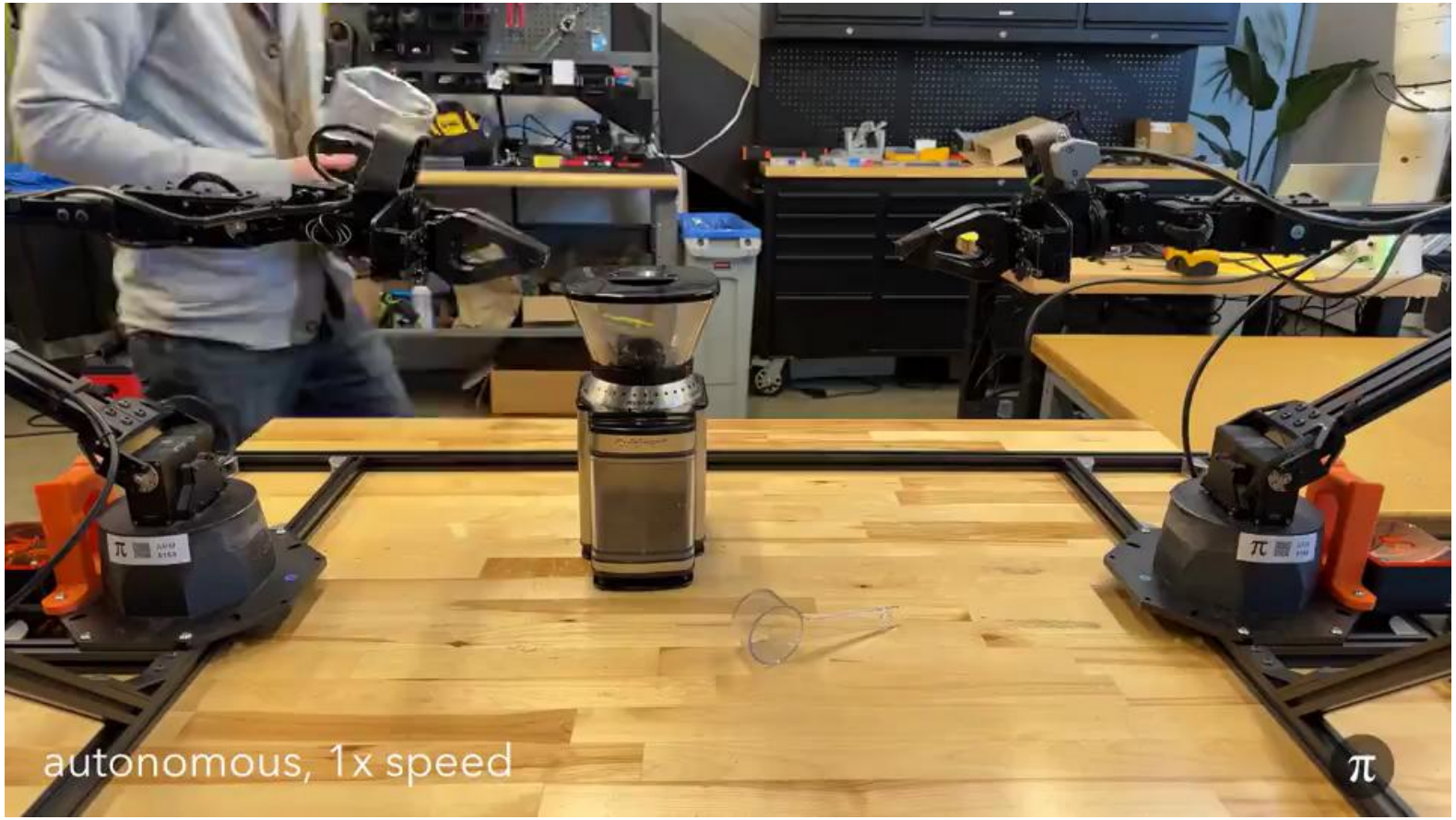
Experiments: Complex tasks



Experiments: Complex tasks



Experiments: videos



- https://www.physicalintelligence.company/blog/pi0

Conclusion

北京大学 PEKING UNIVERSITY

- Contributions
  - A VLA model that draws on the strengths of various approaches.
  - Extensive experiments validated the importance of the knowledge acquired through pretraining.

- Limitations
  - Data combination strategy.
  - Cannot directly generalize to unseen scenarios.
  - Unfair comparisons with the non-VLM model.

Conclusion

北京大学 PEKING UNIVERSITY

- Future work:
  - Data: How to appropriately expand/select pretraining data.
  - LLM/VLM: How to more effectively analyze/utilize the prior knowledge of large models.
  - Action prediction: How to balance prediction accuracy and generalization.

Thanks for listening!