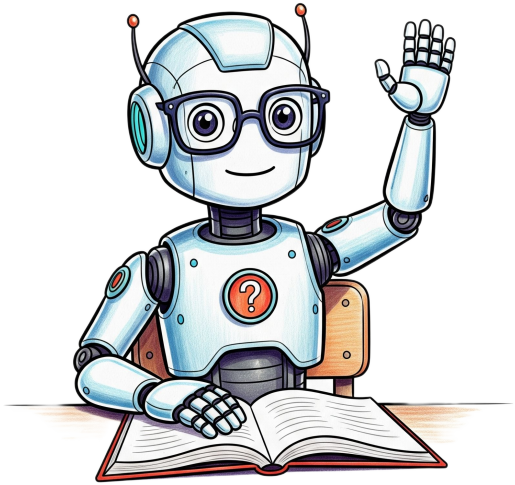


# VLA

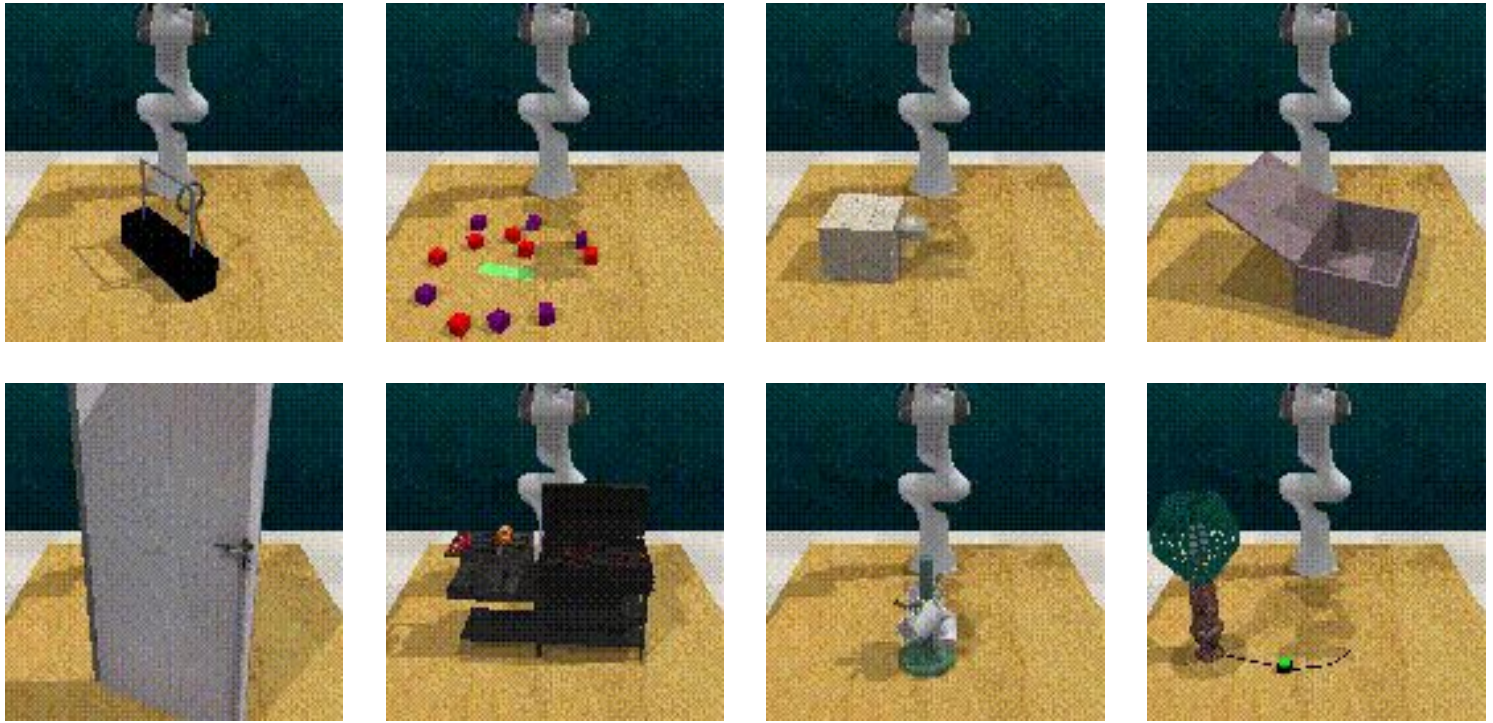
16 october 2025

# О чем сегодня поговорим?

- VLA
  - Зачем это нужно?
  - Как это работает?
- Pi0, Pi0.5
- OpenVLA
- GROOT

RLBench: The Robot Learning Benchmark & Learning Environment

# Задачи для VLA



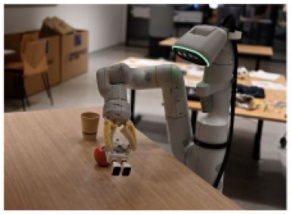
RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

# Real World задачи для VLA

put strawberry into the correct bowl

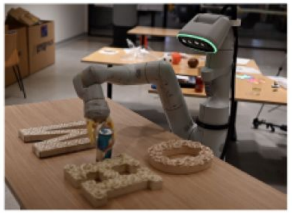
pick up the bag about to fall off the table

move apple to Denver Nuggets



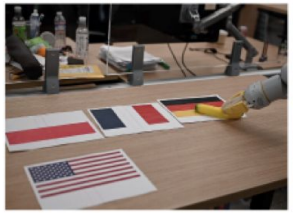
pick robot

place orange in matching bowl

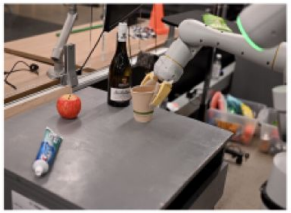


move redbull can to H

move soccer ball to basketball



move banana to Germany



move cup to the wine bottle

pick animal with different colour

RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control

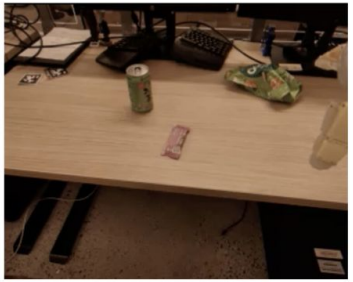
# Real World задачи для VLA

Prompt:
Given <img> Instruction: Bring me a drink.
Prediction:
Plan: pick 7up can. Action: 1 143 129 123 145 114 115 127

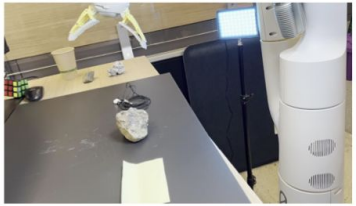
Prompt:
Given <img> Instruction: Move all the objects together.
Prediction:
Plan: move green can near green rice chip bag. Action: 1 128 126 127 135 123 119 127

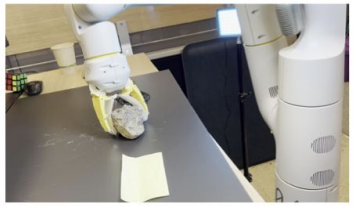
Prompt:
Given <img> Instruction: Pick the object that is different from all other objects
Prediction:
Plan: pick rxbar chocolate. Action: 1 128 129 125 131 125 128 127

Prompt:
Given <img> Instruction: Move the green objects together.
Prediction:
Plan: move green can near green rice chip bag. Action: 1 130 129 121 131 127 128 127



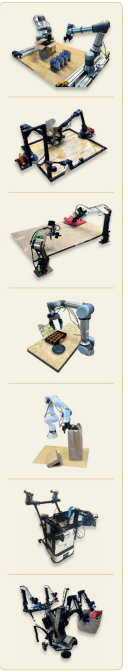
Prompt:
Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127

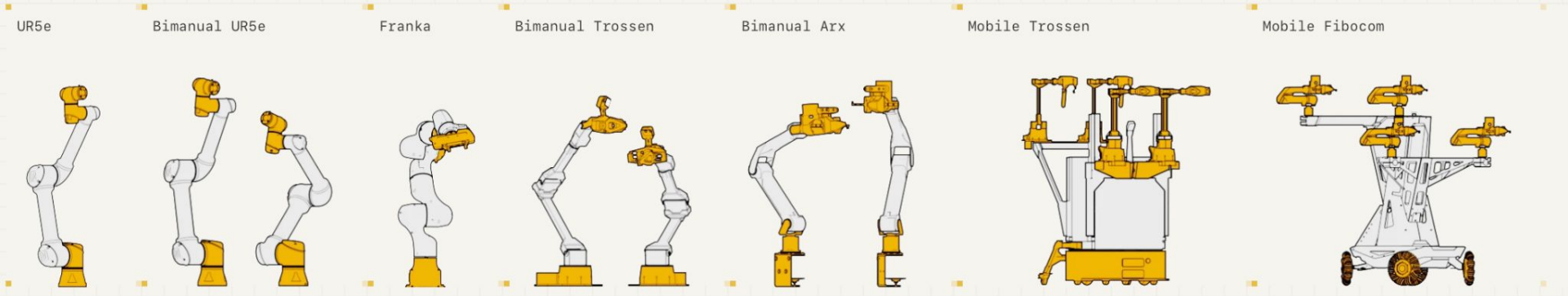




# Проблемы традиционной робототехники

- **Узкая специализация:** Каждый робот или задача требовали создания и обучения отдельных моделей, оптимизированных под конкретные условия.
- **Необходимость переобучения:** Для каждой новой задачи или изменения среды требовалось трудоёмкое и дорогостоящее переобучение всей системы.
- **Изолированные решения:** Отсутствие единой архитектуры приводило к созданию разрозненных, несовместимых решений для разных типов роботов и платформ.
- **Высокие затраты на разработку:** Кастомная разработка для каждого применения влекла за собой значительные временные и финансовые издержки.



# Foundational models

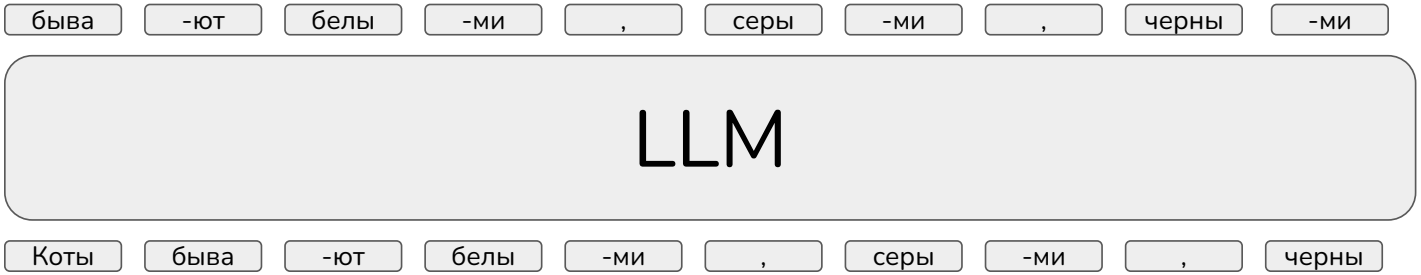
В контексте робототехники **foundation model** — это обобщающая политика (policy), предобученная на множестве гетерогенных данных: визуальных наблюдениях, действиях, обратной связи и состояниях различных роботов.

Такой подход решает три ключевые задачи:

1. **Инвариантность к воплощению (embodiment)** — одна и та же модель должна работать на разных роботах (рука, мобильная платформа, манипулятор и т.д.).
2. **Инвариантность к задаче** — способность выполнять новые задачи без специальной дообучения.
3. **Инвариантность к среде** — перенос между различными сценами, объектами, освещением, фоном.

**Примеры из других областей:**

# От текста к действию




**LLM**

Понимание языка и текстовые ответы

**VLM**

Интеграция визуального восприятия

**VLA**

Генерация физических действий в реальном мире

# От текста к действию


**LLM**

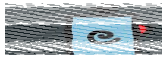
Понимание языка и текстовые ответы

**VLM**

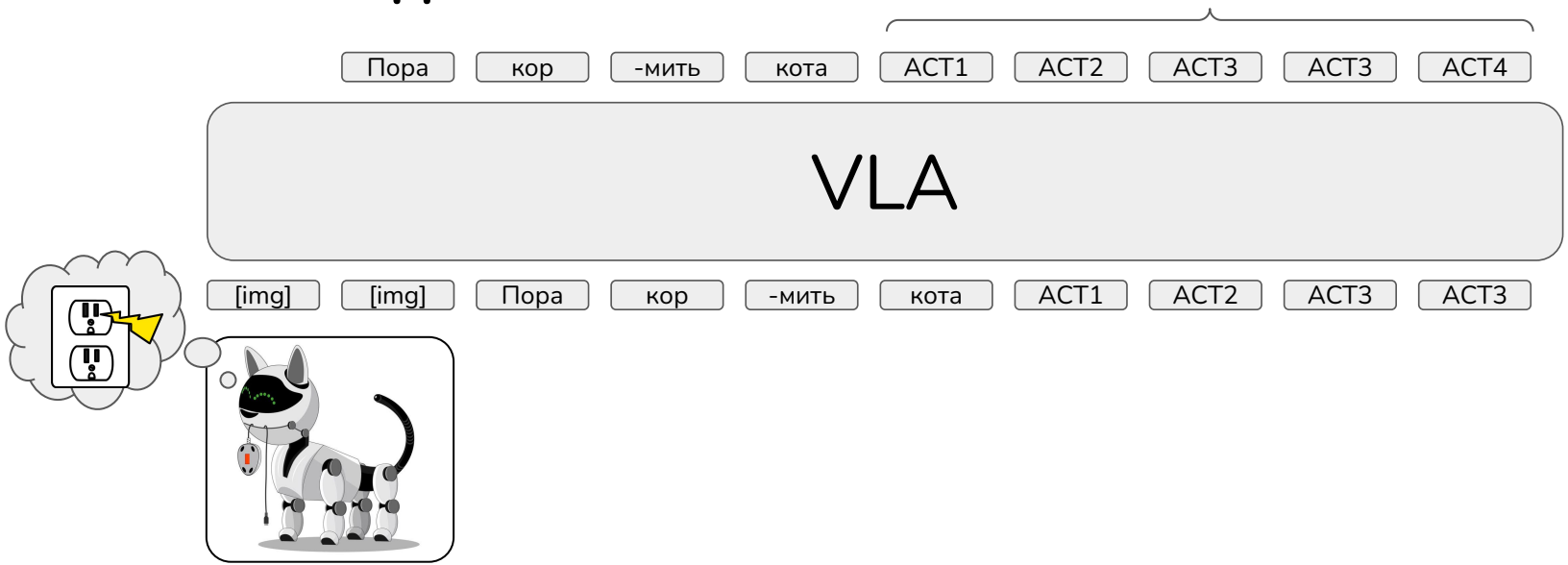
Интеграция визуального восприятия

**VLA**

Генерация физических действий в реальном мире



# От текста к действию




**LLM**

Понимание языка и текстовые ответы

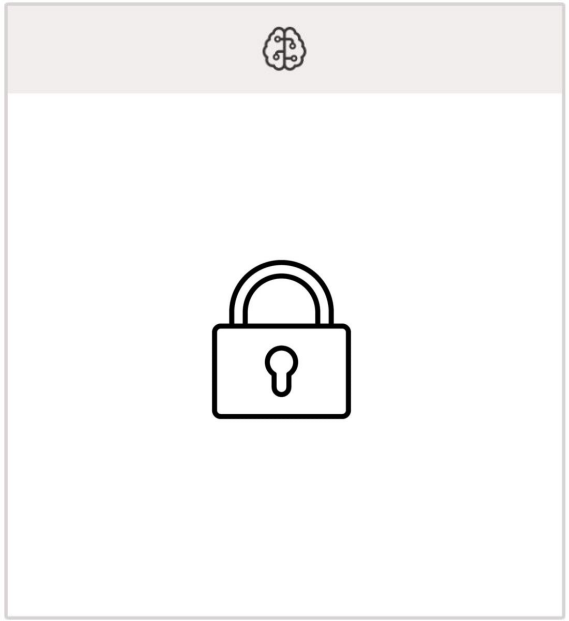
**VLM**

Интеграция визуального восприятия

**VLA**

Генерация физических действий в реальном мире

# 3 кита современных VLA



# 3 кита современных VLA

### π0 / π0.5

**Physical Intelligence**

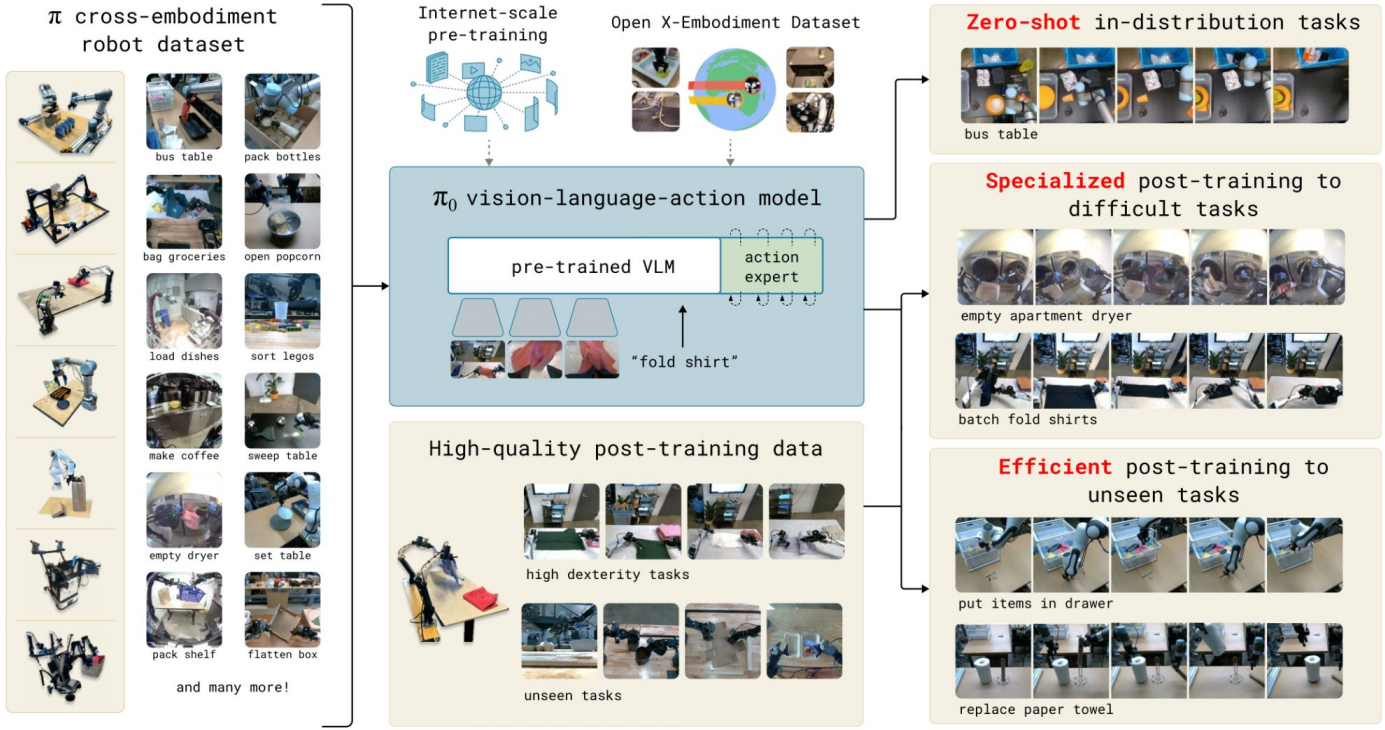
Масштабируемая архитектура с единым токенным пространством для языка и действий

- Более 10M эпизодов обучения
- Поддержка 3D-восприятия
- Впечатляющее zero-shot обобщение

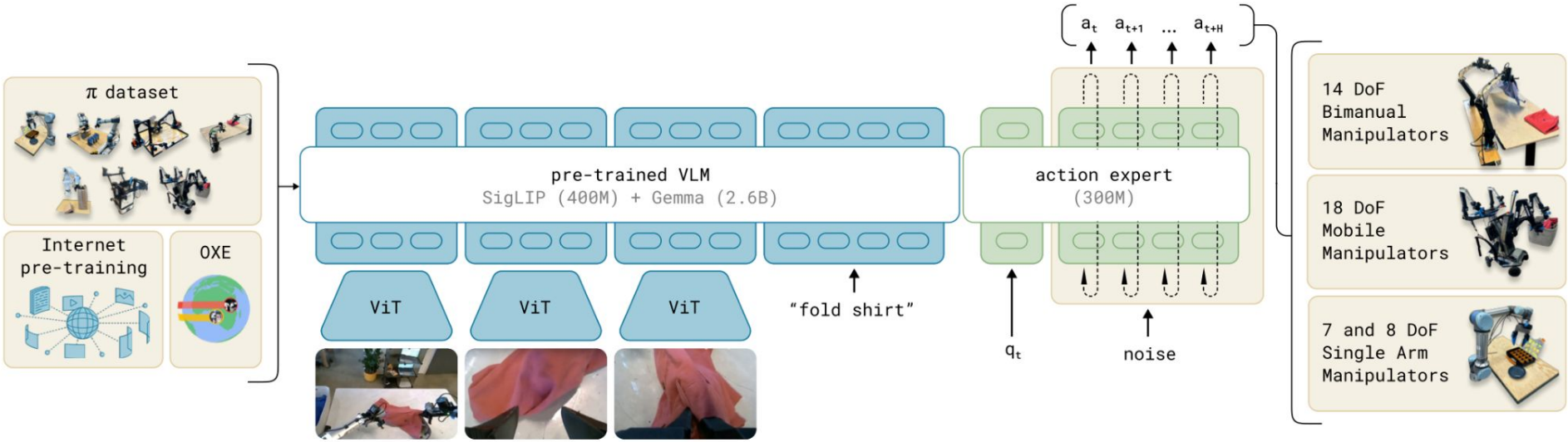
https://www.physicalintelligence.company/blog/pi0

# $\pi_0$

[CLS] возьми красный куб [ACTION_X+0.02] [ACTION_Y-0.01] [GRIP_CLOSE] [EOS]

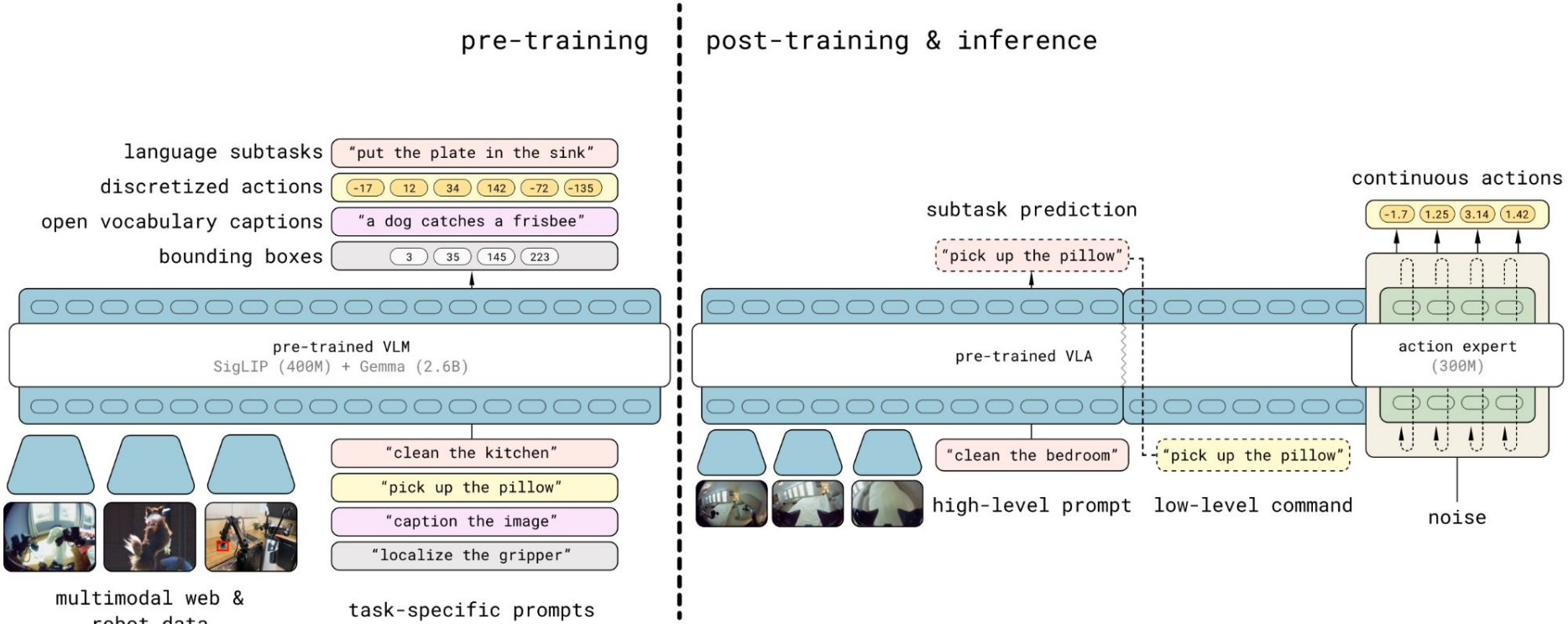


# $\pi_0$


π dataset
Internet pre-training
OXE
pre-trained VLM
SigLIP (400M) + Gemma (2.6B)
ViT
ViT
ViT
"fold shirt"
$a_t$ $a_{t+1}$ ... $a_{t+H}$
action expert
(300M)
$q_t$
noise
14 DoF Bimanual Manipulators
18 DoF Mobile Manipulators
7 and 8 DoF Single Arm Manipulators

https://www.physicalintelligence.company/blog/pi05

# $\pi_{0.5}$



# 3 кита современных VLA

## π0 / π0.5

**Physical Intelligence**

Масштабируемая архитектура с единым токенным пространством для языка и действий

- Более 10M эпизодов обучения
- Поддержка 3D-восприятия
- Впечатляющее zero-shot обобщение

## OpenVLA

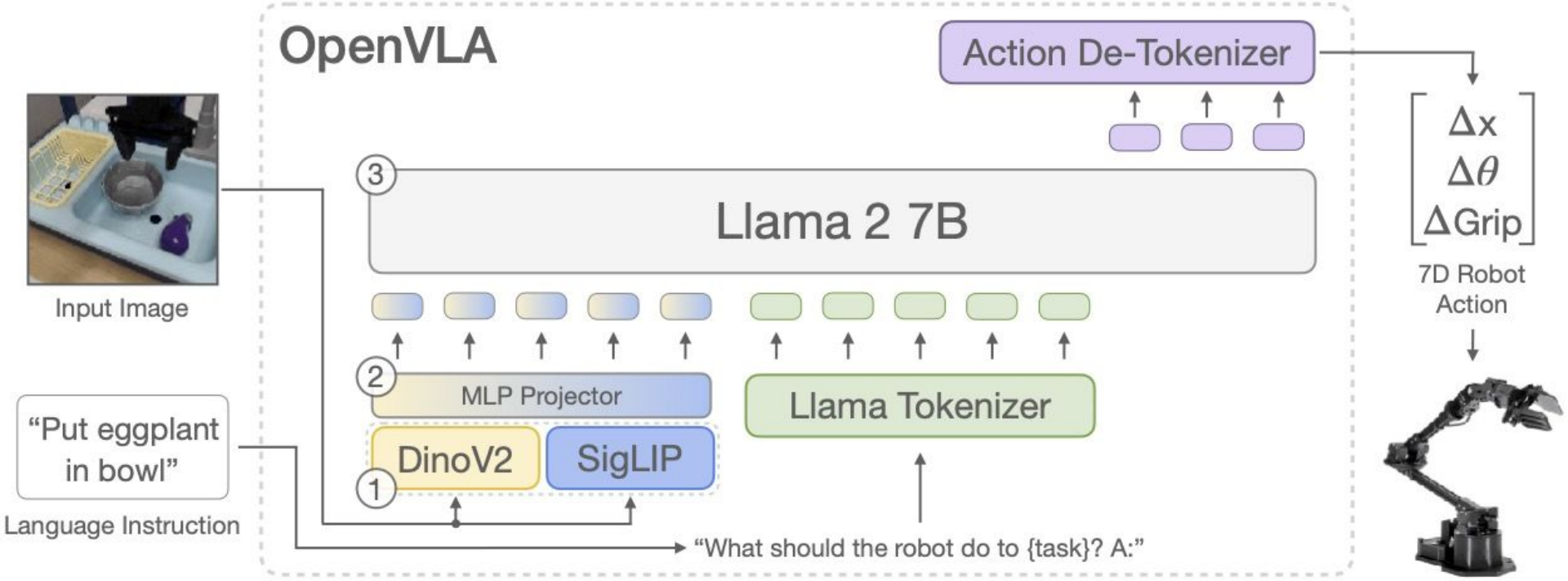
**Stanford / UC Berkeley**

Открытая платформа для исследований и прототипирования

- Полностью открытый код и веса
- Простой fine-tuning за 1 день
- Модульная архитектура

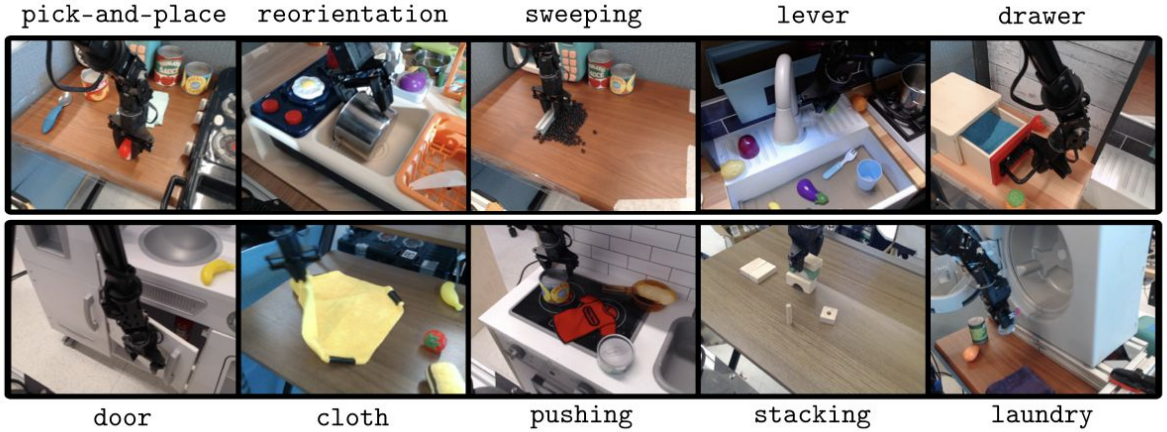
https://openvla.github.io/

# OpenVLA

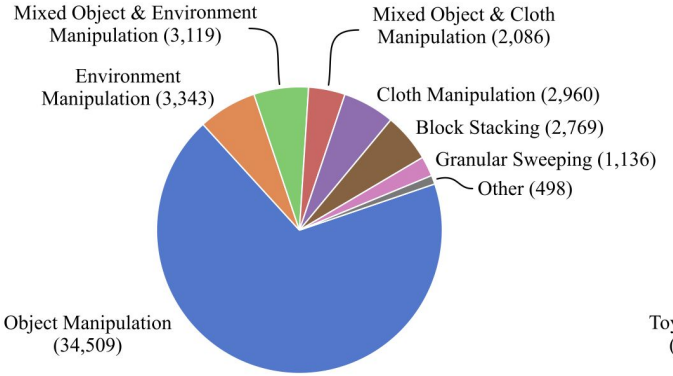


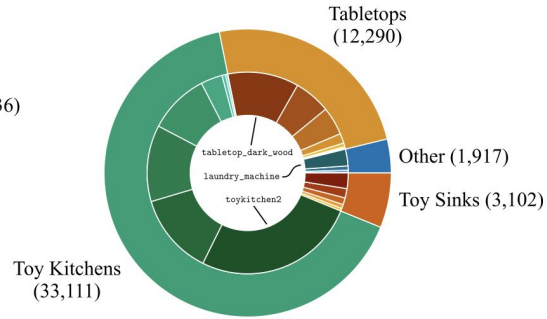
https://rail-berkeley.github.io/bridgedata/

# BridgeData V2



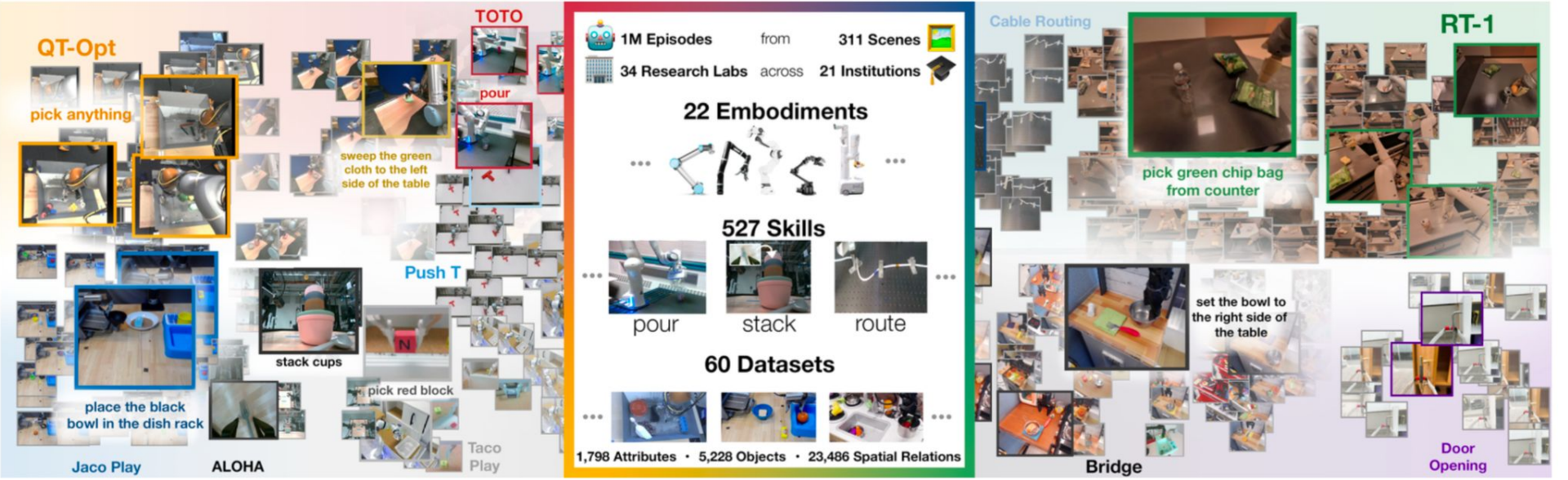




https://robotics-transformer-x.github.io/

# Open X-Embodiment



# 3 кита современных VLA

## π0 / π0.5

**Physical Intelligence**

Масштабируемая архитектура с единым токенным пространством для языка и действий

- Более 10M эпизодов обучения
- Поддержка 3D-восприятия
- Впечатляющее zero-shot обобщение

## OpenVLA

**Stanford / UC Berkeley**

Открытая платформа для исследований и прототипирования

- Полностью открытый код и веса
- Простой fine-tuning за 1 день
- Модульная архитектура

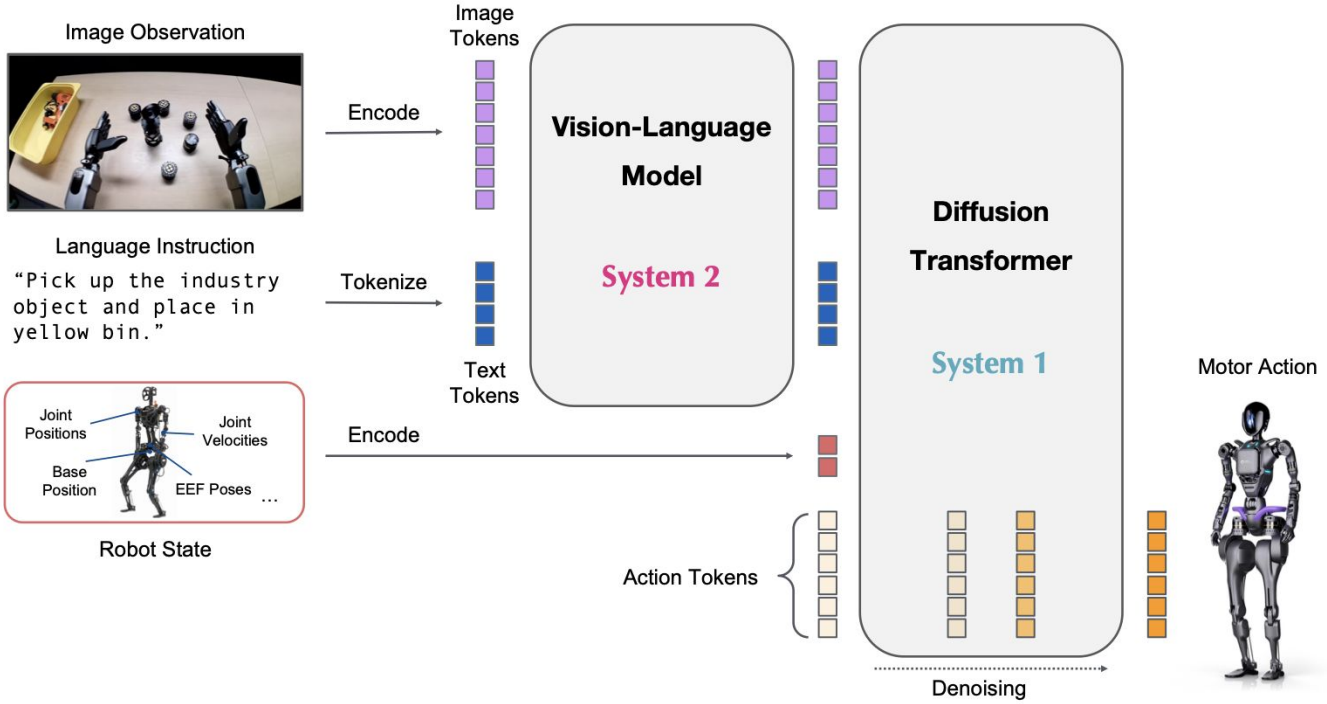
## GROOT

**NVIDIA**

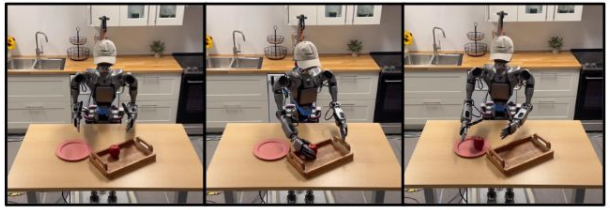
Промышленное решение для реального мира

- Задержка менее 200 мс
- Мультимодальное взаимодействие
- Интеграция с Isaac Sim

# GROOT N1: An Open Foundation Model for Generalist Humanoid Robots



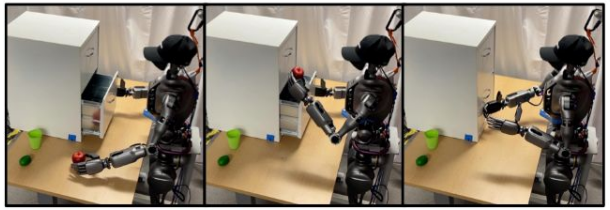
# Двуручные манипуляции роботом Fourier GR-1



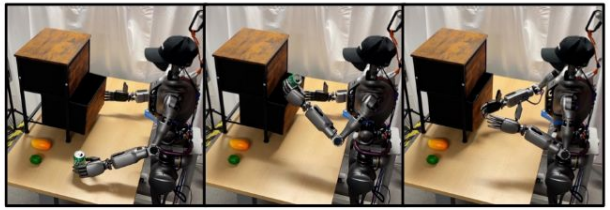
Pick-and-Place: Tray to Plate

Pick-and-Place: Placemat to Basket

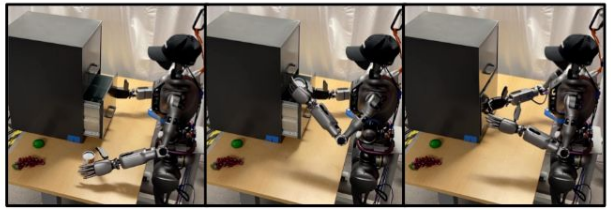
Pick-and-Place: Cutting Board to Pan



Articulated: White Drawer

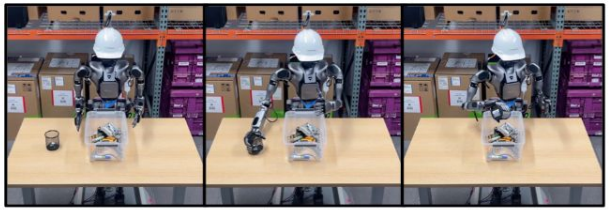


Articulated: Wooden Chest

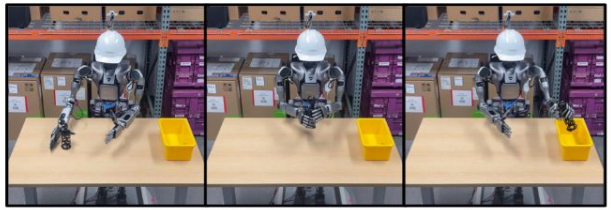


Articulated: Dark Cabinet

Industrial: Machinery Packing



Industrial: Mesh Cup Pouring



Industrial : Cylinder Handover

# Teleoperation Data Collection

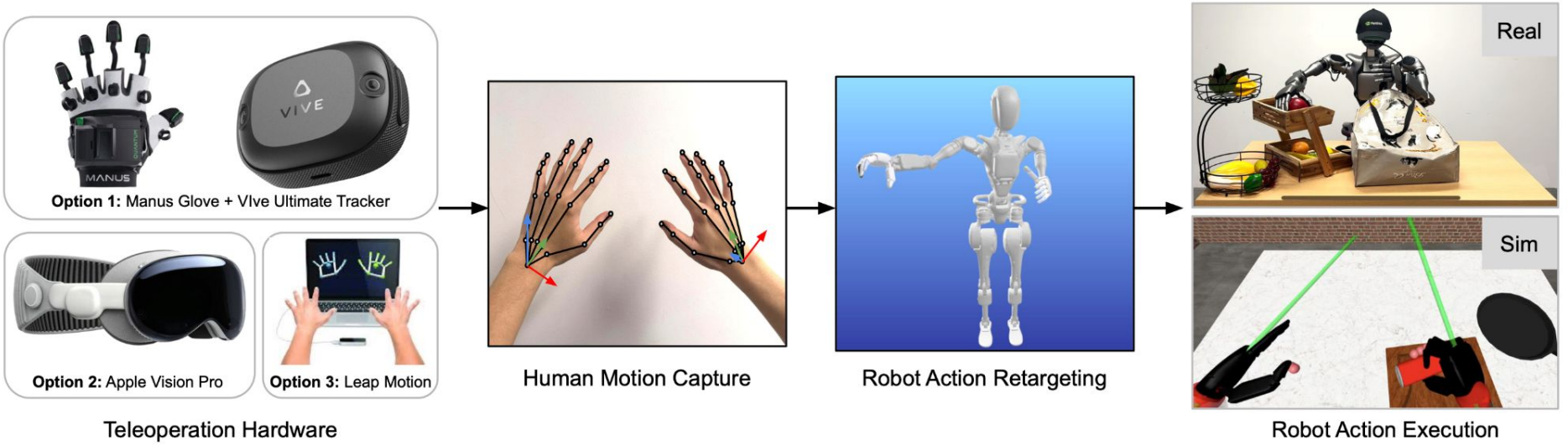



Figure 6: **Data Collection via Teleoperation.** Our teleoperation infrastructure supports multiple devices to capture human hand motion, including 6-DoF wrist poses and hand skeletons. Robot actions are produced through retargeting and executed on robots in real and simulation environments.

# Итоги

VLA – foundation модели для робототехники:

- ➔ Объединяют **понимание сцены и задач** с моторикой робота.
- ➔ Способны к **многозадачности и обобщению** на новые объекты, среды и роботы.
- ➔ Поддержка **разных типов роботов**: манипуляторы, мобильные платформы, гуманоиды.

1 **2024: Прорыв**

Появление π0, OpenVLA и GROOT — три подхода к одной цели

2 **2025: Масштабирование**

Тысячи лабораторий начинают эксперименты с VLA

3 **2026-2030: Внедрение**

VLA-роботы появляются на складах, в больницах, домах

4 **Будущее**

Универсальные роботизированные ассистенты становятся реальностью

# Вопросы?