LLMs ❄

VLMs ❄

VLAs ❄

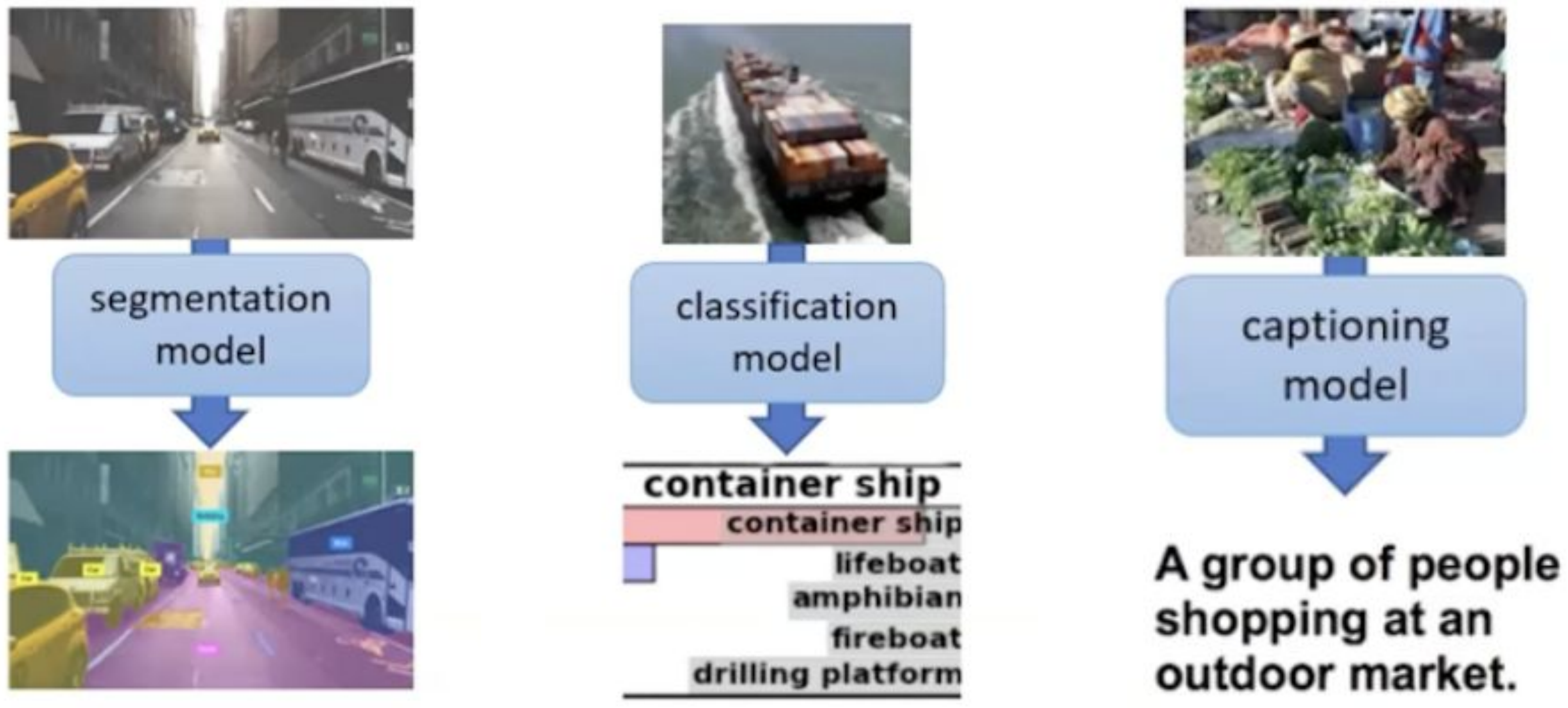
❄ Maybe Robot Model? 🔥

DeepRob

[Group 6] Lecture 8
Foundational Models and Robotics

by Chang, Franklin, Rohan
University of Minnesota

How AI used to work in the past？


segmentation model

classification model

container ship

container ship

lifeboat

amphibian

fireboat

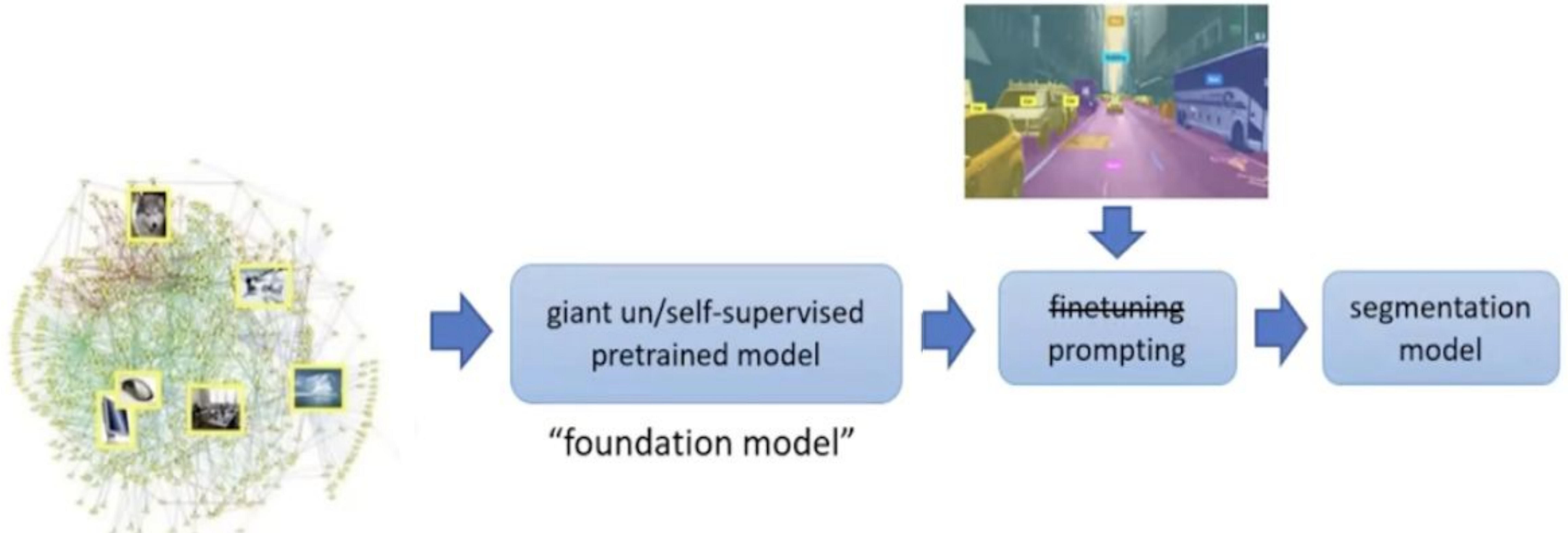
drilling platform

captioning model

A group of people shopping at an outdoor market.

....

How AI works now?


giant un/self-supervised pretrained model

“foundation model”

~~finetuning~~ prompting

segmentation model

Ref : https://www.youtube.com/watch?v=ET2HsfLrNMs

What is Foundation Model?

FOUNDATION models are pretrained on extensive internet-scale data and can be fine-tuned for adaptation to a wide range of downstream tasks.

Can you give me some examples?

GPT-4, BERT, LLaMA, Claude, PaLM, DALL-E, Stable Diffusion, Midjourney, SAM, Gemini, CLIP, Florence, Codex, StarCoder, AlphaCode, Copilot...

Do you guys think Transformer is a Foundation model?

Evolution of Foundation Models and Architectures

Key Milestones Leading to Foundation Models

2012 - AlexNet

Revolutionary CNN architecture that won ImageNet competition and kicked off the deep learning revolution in computer vision

2014 - VGG

Introduced deep convolutional networks with small filters, showing that network depth is crucial for performance

2015 - ResNet

Solved deep network training with residual connections, enabling networks with hundreds of layers

2017 - Transformer Architecture

Introduced "attention is all you need" architecture that became the foundation for modern language models

2018 - BERT

Bidirectional encoder representations from transformers, revolutionized NLP with pre-training and fine-tuning paradigm

2019 - GPT-2

Demonstrated impressive text generation capabilities and raised discussions about AI safety

2020 - DALL-E

Created by OpenAI, demonstrated the ability to generate images from text descriptions

2020 - GPT-3

Showed emergent abilities in few-shot learning and task adaptation without fine-tuning

2021 - CLIP

Zero-shot learning model that connects text and images, enabling flexible visual recognition tasks

2021 - Codex

GPT model fine-tuned on code, powering GitHub Copilot and demonstrating code generation capabilities

2022 - PaLM

Pathways Language Model demonstrated strong reasoning and multilingual capabilities

2022 - Stable Diffusion

Open-source text-to-image model that democratized AI image generation

2022 - ChatGPT

Fine-tuned GPT-3.5 that sparked widespread adoption of conversational AI

2023 - RT-2

Robotic Transformer 2 bridges vision-language models with robotic control for real-world tasks

2023 - GPT-4V

Multimodal model capable of understanding and reasoning about both images and text

2023 - DALL-E 3

Latest iteration of OpenAI's text-to-image generation model

Architectural Foundations (2017-2018)

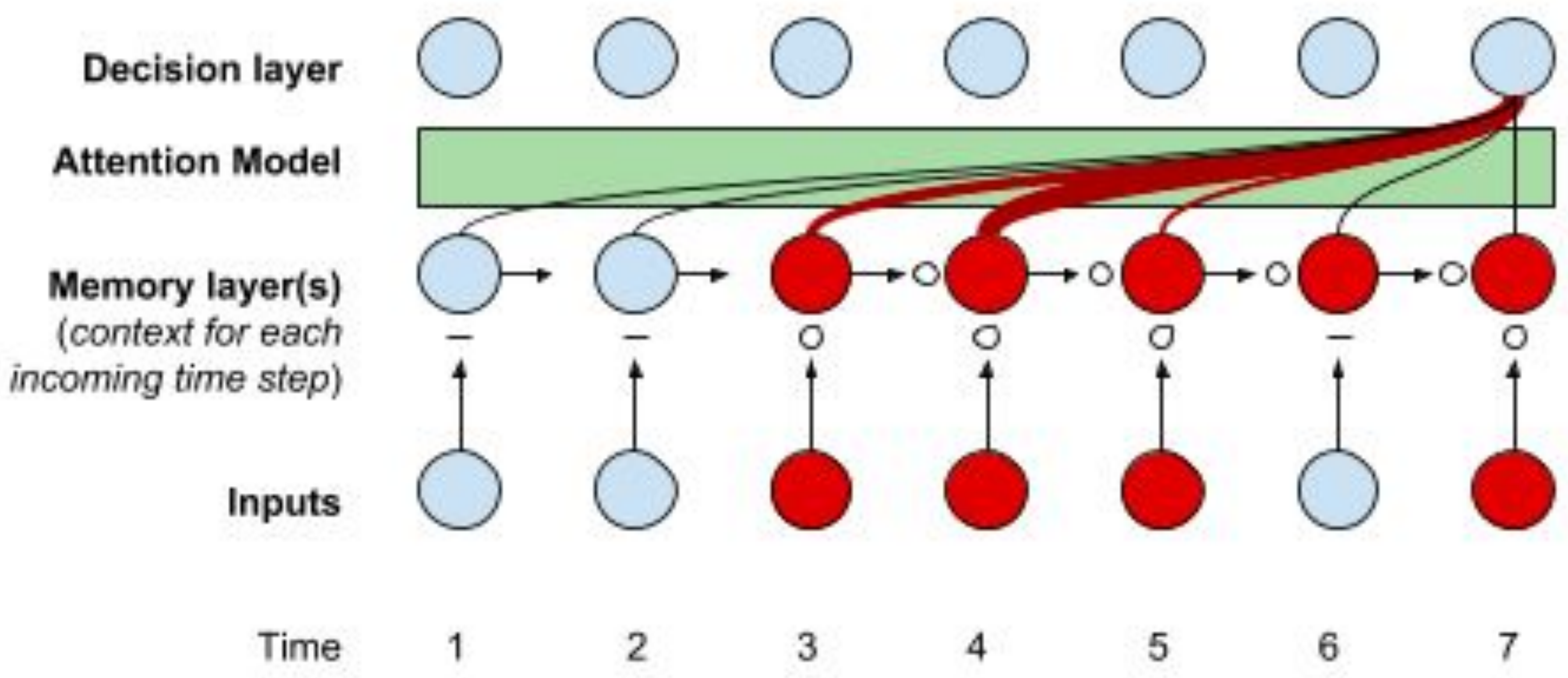
Key Architectural Innovations:

- Transformer architecture (2017)
- LSTM and attention mechanisms
- ResNet and deep residual learning
- Encoder-decoder architectures

Key Developments:

- Shift from CNNs/RNNs to attention-based models
- Breakthrough in sequence modeling
- Enabling deeper model architectures

Attention Mechanism



https://deepai.org/machine-learning-glossary-and-terms/attention-models

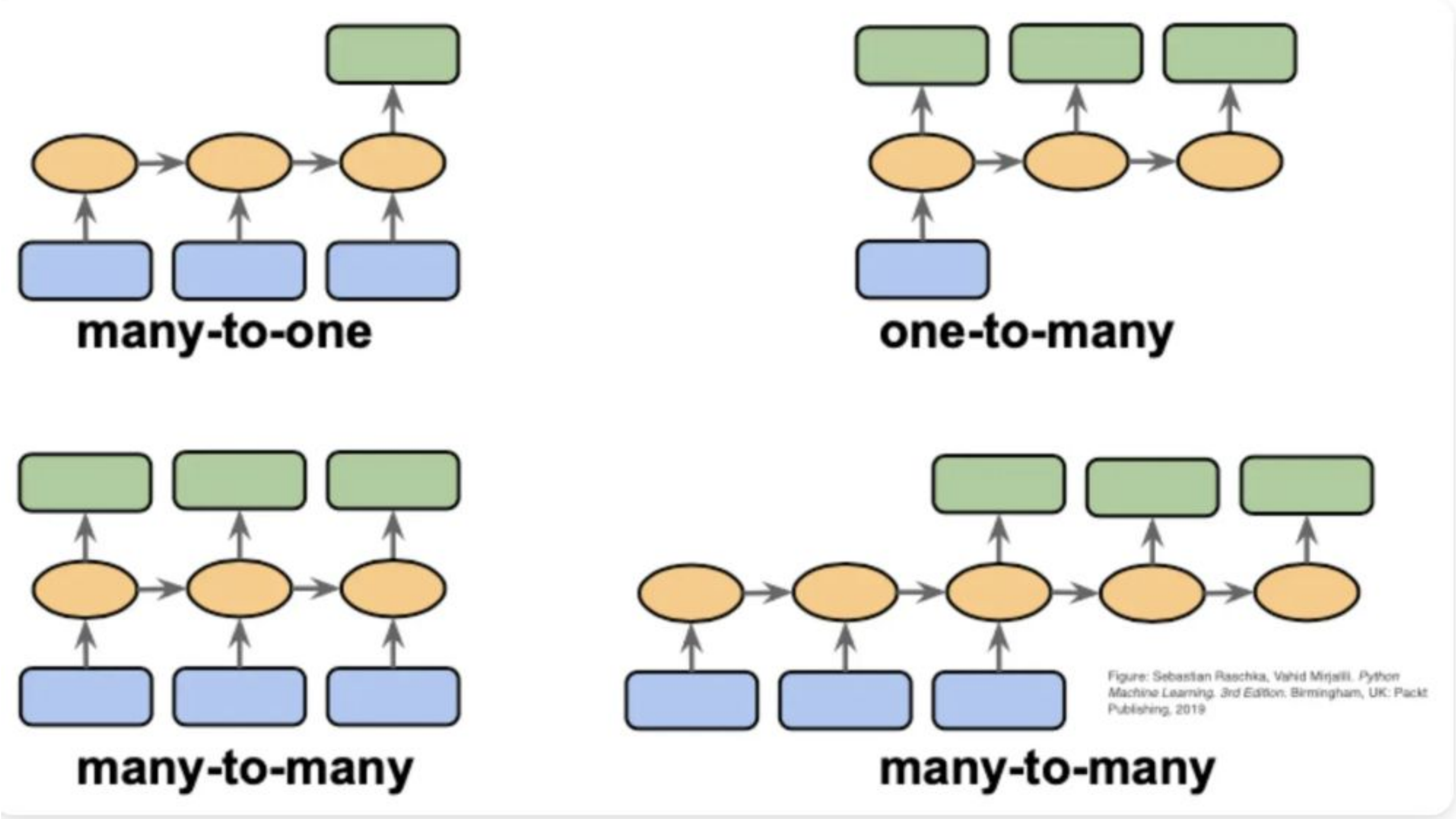


Figure: Sebastian Raschka, Vahid Mirjalili. *Python Machine Learning. 3rd Edition*. Birmingham, UK: Packt Publishing, 2019

https://botpenguin.com/glossary/sequence-modeling

Early Foundation Models (2018-2020)

Key Architectural Innovations:

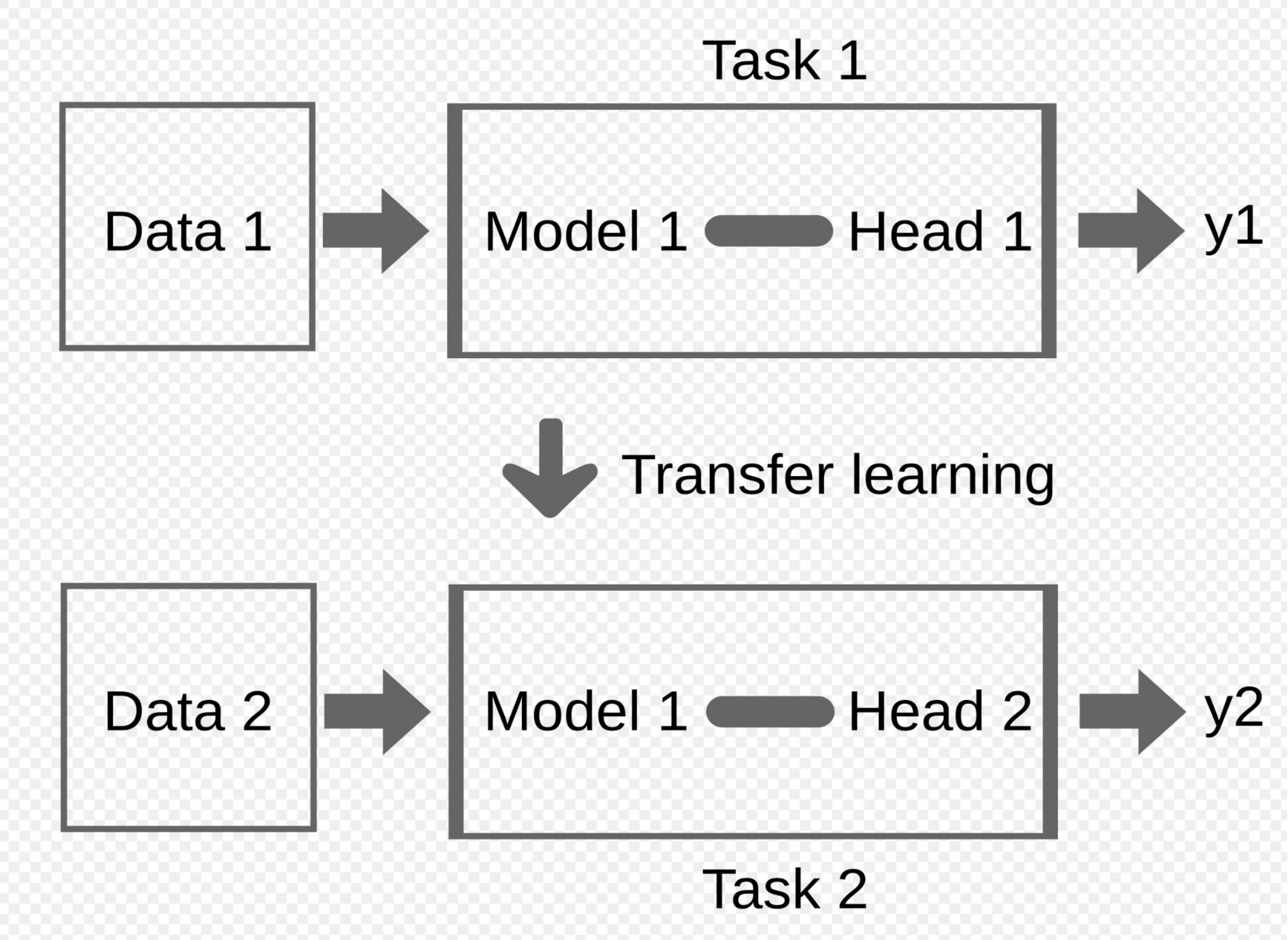
- BERT's bidirectional encoder architecture
- GPT's autoregressive architecture

Foundation Models:

- BERT: First large-scale bidirectional language model
- GPT-2: Demonstrated scaling potential
- GPT-3: First true foundation model with emergent abilities

Key Developments

- Introduction of pre-training and fine-tuning paradigm
- Discovery of transfer learning capabilities
- Emergence of few-shot learning abilities



https://en.wikipedia.org/wiki/Transfer_learning

Multimodal Evolution (2021-2022)

Key Architectural Innovations:

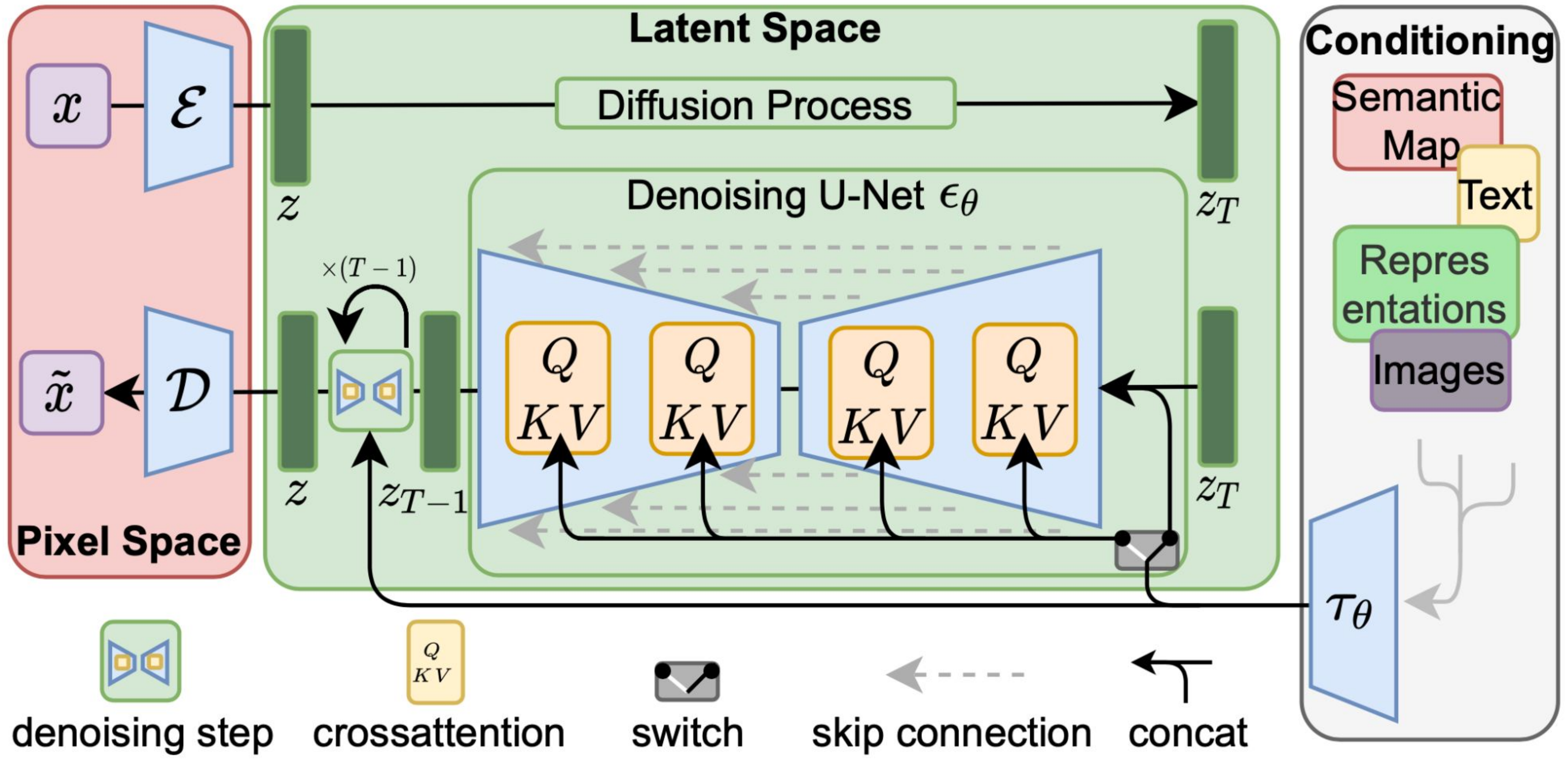
- Vision Transformer (ViT)
- Diffusion model architectures

Foundation Models:

- DALL-E: Text-to-image foundation model
- PaLM: Pathways Language Model
- Stable Diffusion: Open source image generation
- Gato: Multi-task foundation model

Key Developments:

- Extension to multiple modalities
- Emergence of specialized domain models
- Democratization through open-source models



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

Advanced Foundation Models (2023-2024)

Key Architectural Innovations:

- Mixture of Experts (MoE)
- Advanced attention mechanisms
- Multimodal architectures

Foundation Models

GPT-4: Advanced multimodal capabilities

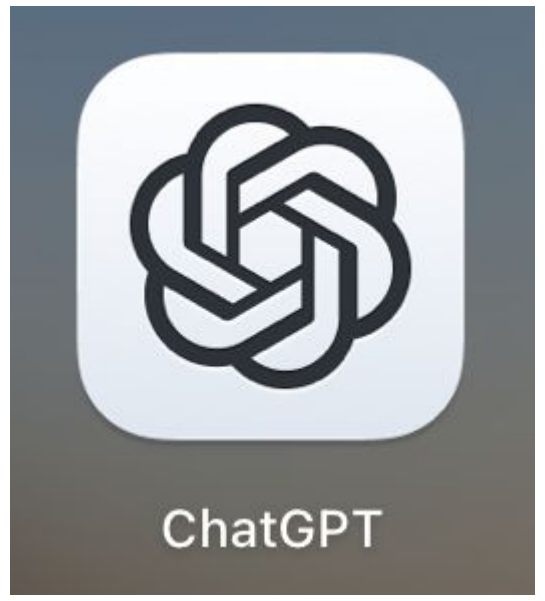
Claude: Advanced reasoning abilities

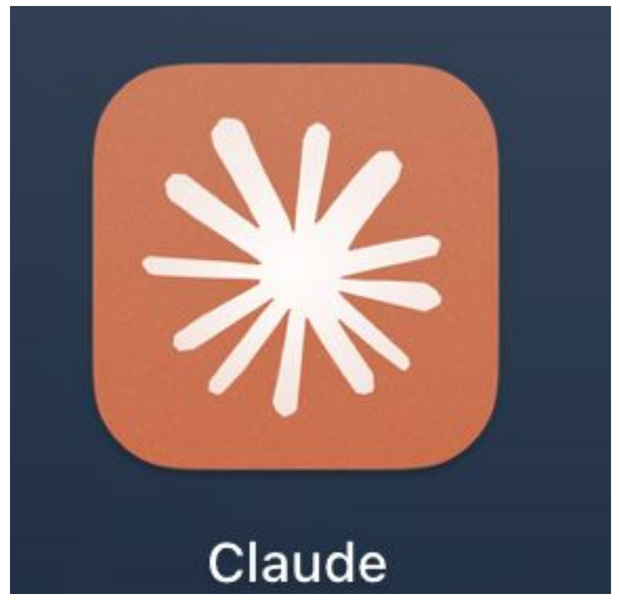
Gemini: Multimodal understanding

DALL-E 3: Enhanced image generation

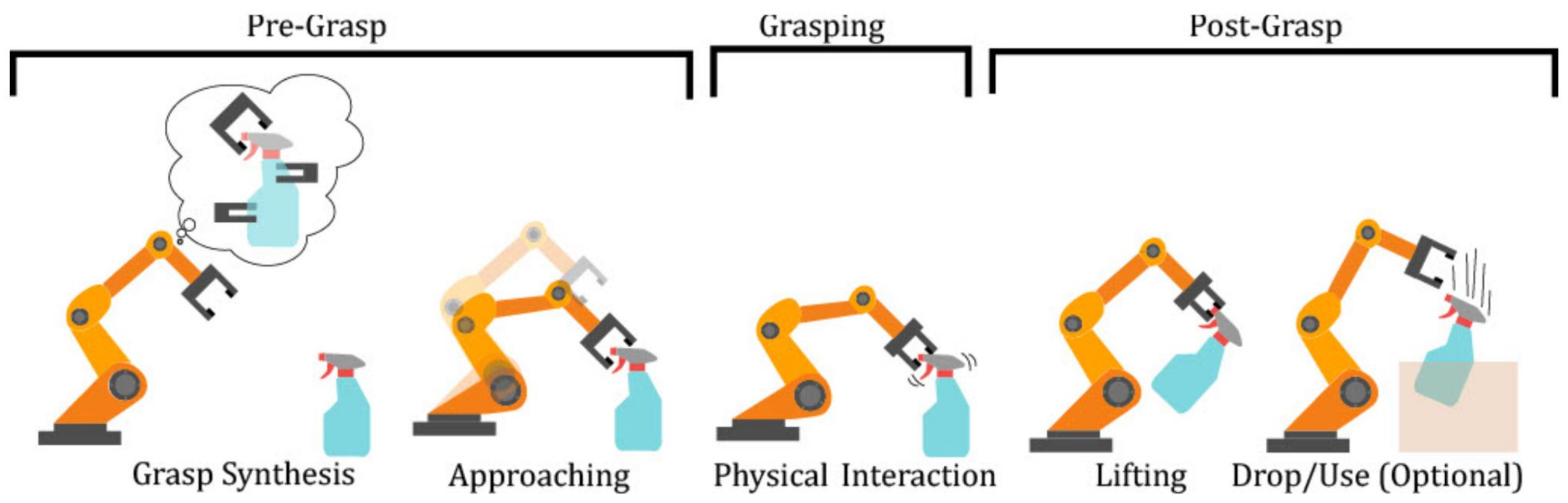
Key Developments

- Integration of multiple modalities
- Focus on model efficiency and scaling
- Improved reasoning and instruction following



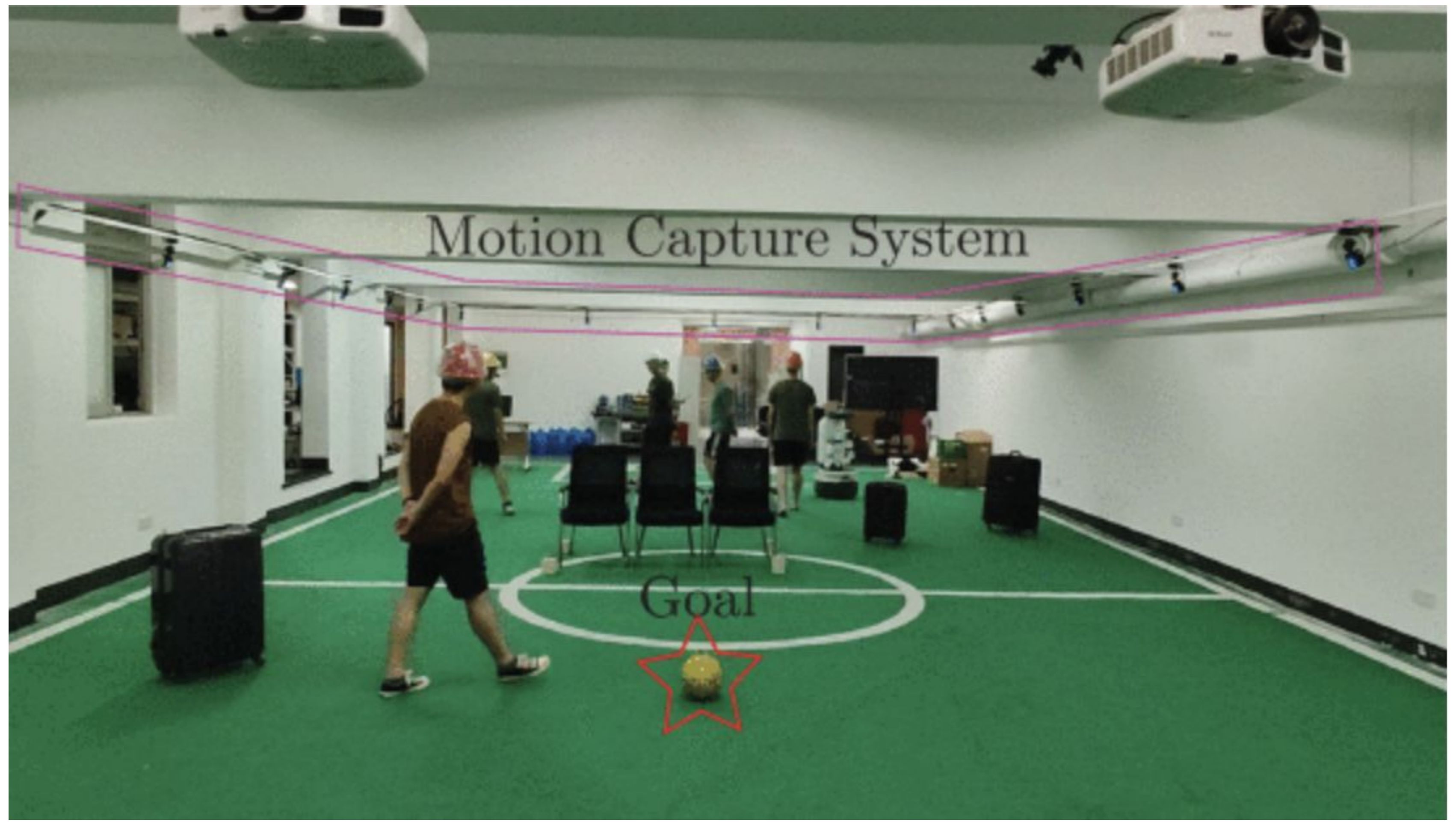


How Robotic learning works now？



https://rhys-newbury.github.io/projects/6dof/

Grasping



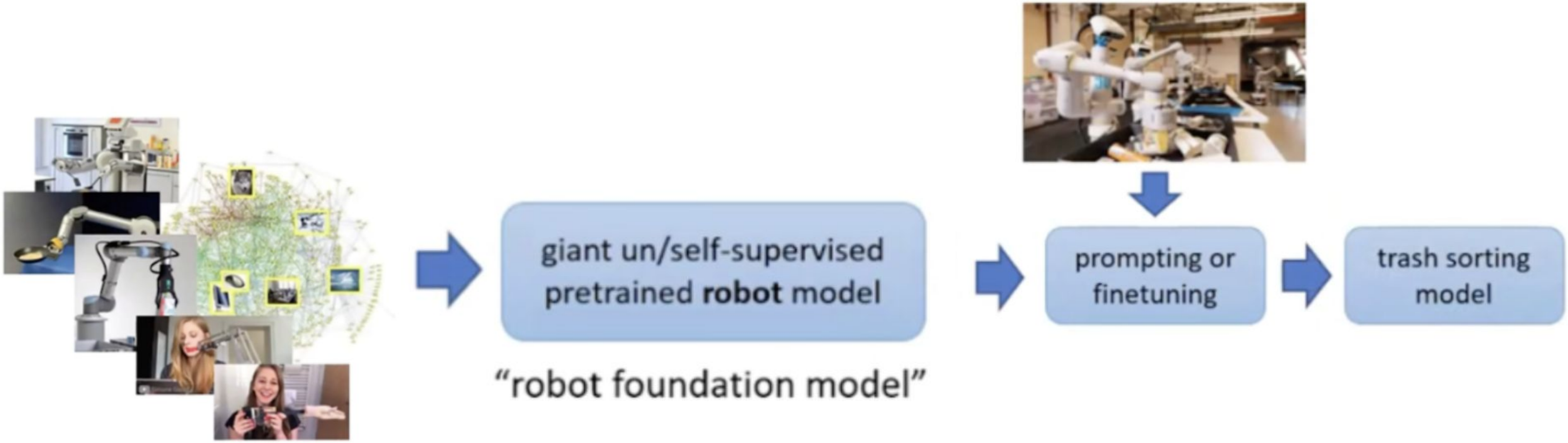
Zhou, Zhiqian, Zhiwen Zeng, Lin Lang, Weijia Yao, Huimin Lu, Zhiqiang Zheng, and Zongtan Zhou. "Navigating robots in dynamic environment with deep reinforcement learning." *IEEE Transactions on Intelligent Transportation Systems* 23, no. 12 (2022): 25201-25211.

Navigating

.
.
.

.
.
.

How Robotic learning will work in the future



giant un/self-supervised pretrained **robot** model

"robot foundation model"

prompting or finetuning

trash sorting model

Ref : https://www.youtube.com/watch?v=ET2HsfLrNMs

Foundation Models used in Robotics

Why Foundation Models Matter in Robotics

- Enable robots to process language, images, actions seamlessly

Improve understanding and performance in varying environments

Key Features

- Multimodal Learning:
  - Integrate vision, language, and action.
- Generalization:
  - Adapt to multiple tasks without the need for retraining.
- Human Interaction:
  - Understand and respond to natural language commands.

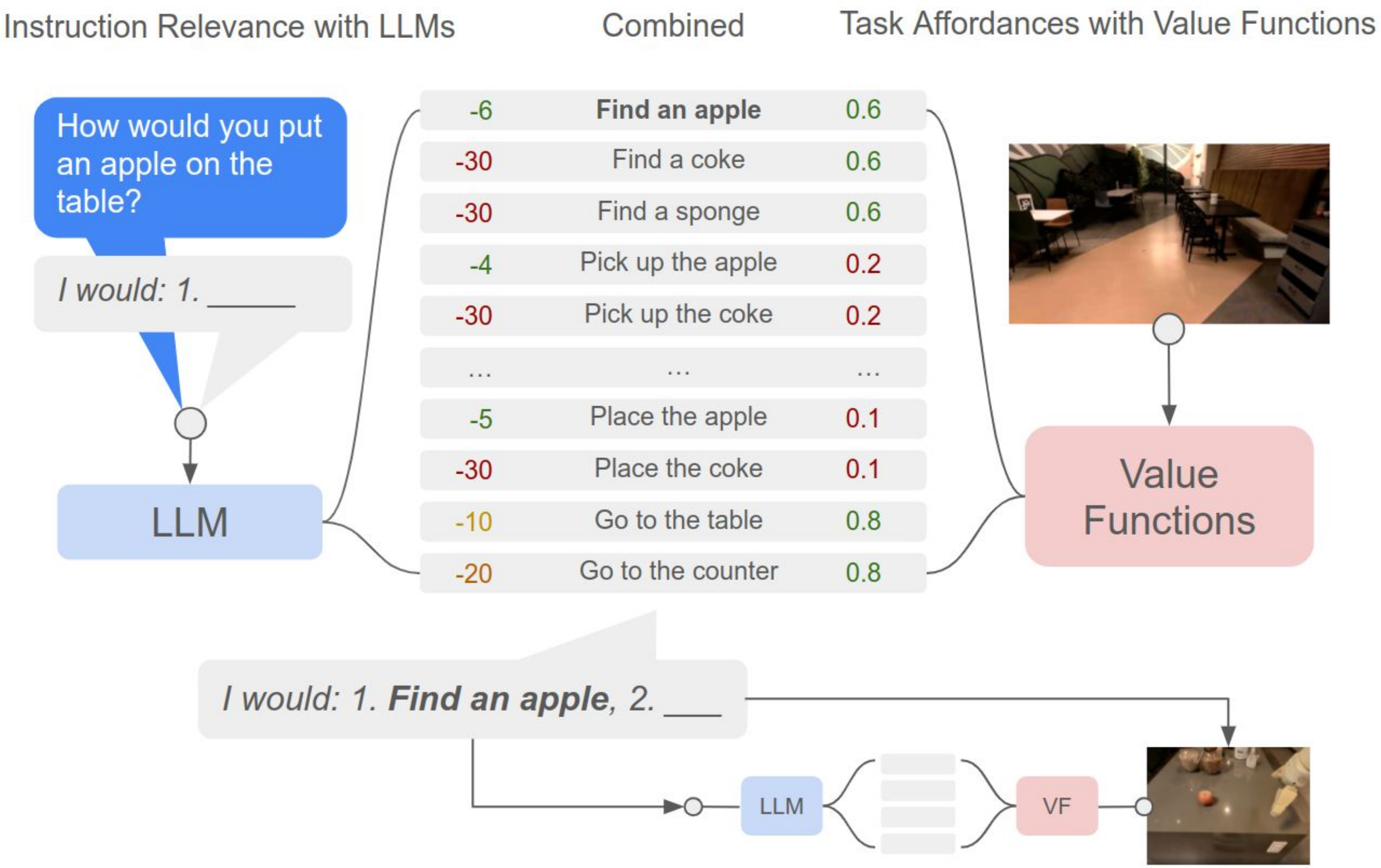
Some examples of Using Foundation Models in Robotics

- SayCan
- DALL-E-Bot
- CLIPort

SayCan

- A system by Google combining language models with robotic control.
- Uses a language model for interpreting tasks.
- Robot policy model executes physical actions.
- Use Cases:
  - Performing household tasks.
  - Assisting in warehouses.
- Language Model: Processes complex human commands.
- Action Model: Aligns tasks with robot capabilities.
  - Example: "Bring me a snack" → Identifies snack location → Fetches it.

Brohan, Anthony, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz et al. "Do as i can, not as i say: Grounding language in robotic affordances." In *Conference on robot learning*, pp. 287-318. PMLR, 2023.



Instruction Relevance with LLMs	Combined	Task Affordances with Value Functions
-6	**Find an apple**	0.6
-30	Find a coke	0.6
-30	Find a sponge	0.6
-4	Pick up the apple	0.2
-30	Pick up the coke	0.2
…	…	…
-5	Place the apple	0.1
-30	Place the coke	0.1
-10	Go to the table	0.8
-20	Go to the counter	0.8

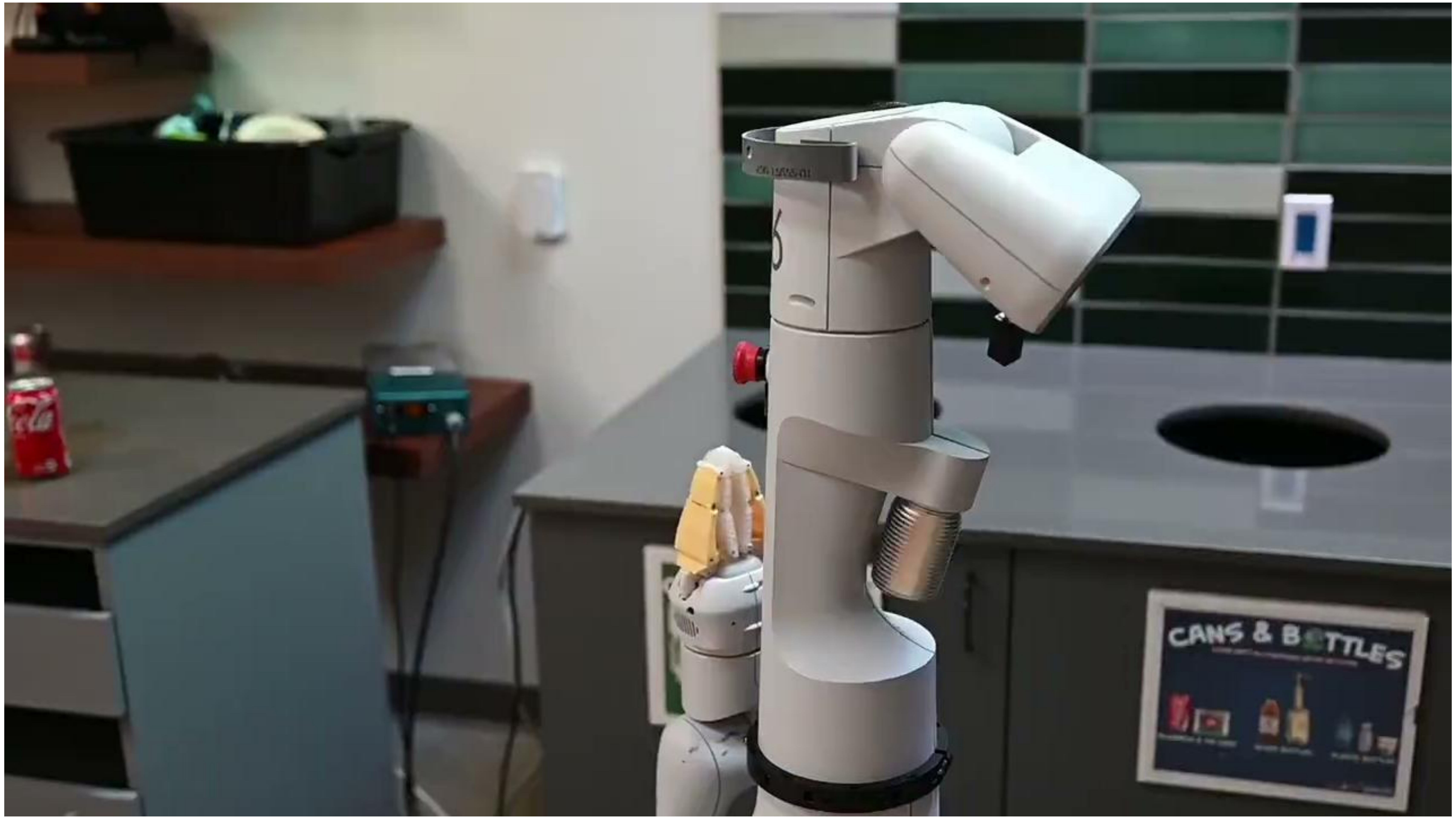
How would you put an apple on the table?

I would: 1. ______

I would: 1. ***Find an apple****, 2. ___*

Brohan, Anthony, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz et al. "Do as i can, not as i say: Grounding language in robotic affordances." In *Conference on robot learning*, pp. 287-318. PMLR, 2023.

SayCan Demo

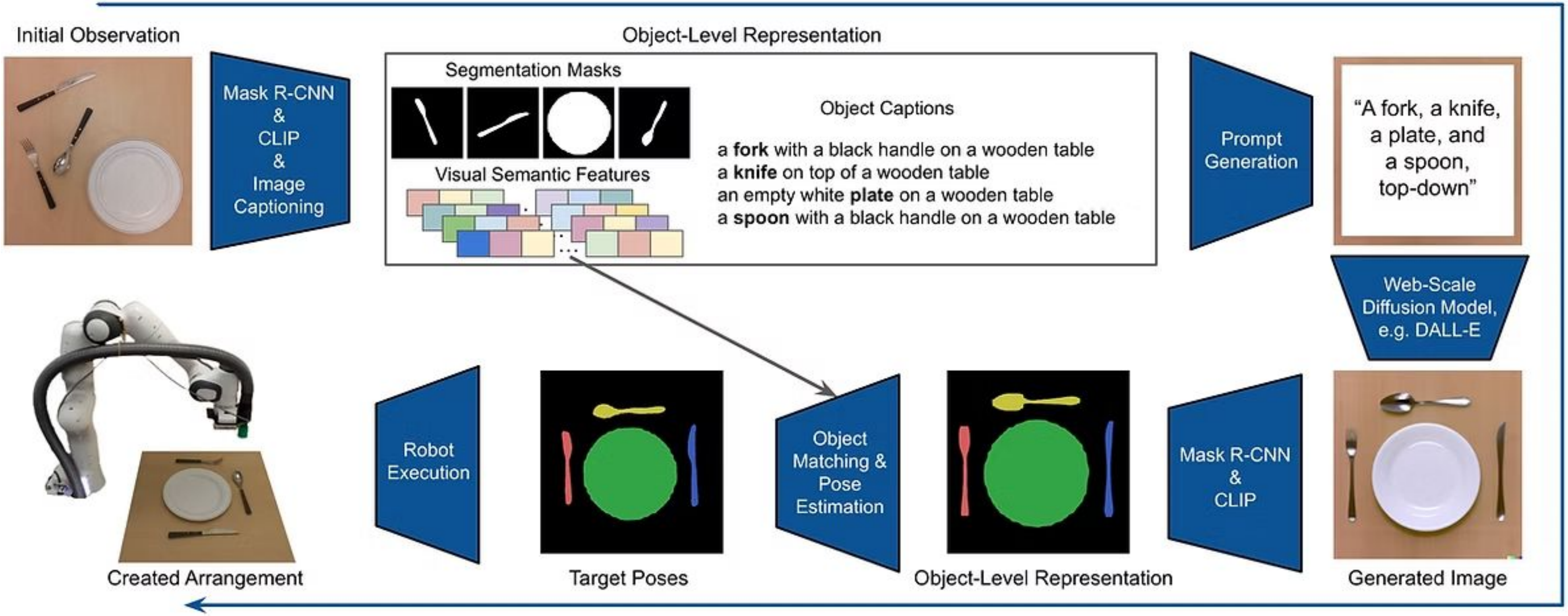


Brohan, Anthony, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz et al. "Do as i can, not as i say: Grounding language in robotic affordances." In *Conference on robot learning*, pp. 287-318. PMLR, 2023.

DALL-E-Bot

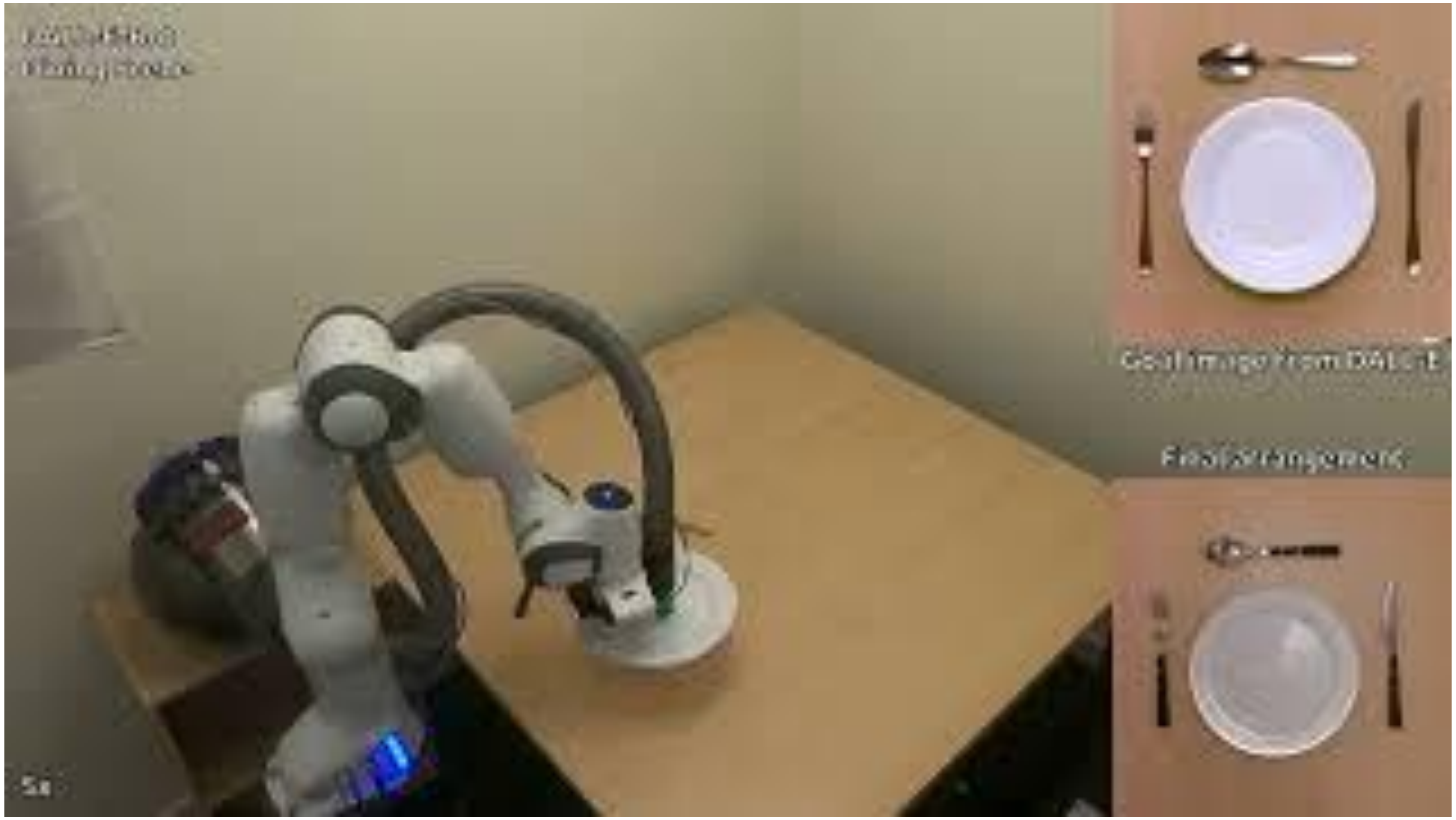
- DALL-E is an artificial intelligence model that can generate images based on fed prompts
- DALL-E-Bot using DALLE’s generative image capabilities to rearrange objects in a scene
  - Generates images to interpret visual tasks.
  - Enhances robot adaptability in creative scenarios.
- Primarily used for creative tasks and artistic problem-solving

Kapelyukh, Ivan, Vitalis Vosylius, and Edward Johns. "Dall-e-bot: Introducing web-scale diffusion models to robotics." *IEEE Robotics and Automation Letters* 8, no. 7 (2023): 3956-3963.


Initial Observation
Mask R-CNN & CLIP & Image Captioning
Object-Level Representation
Segmentation Masks
Visual Semantic Features
Object Captions
a **fork** with a black handle on a wooden table
a **knife** on top of a wooden table
an empty white **plate** on a wooden table
a **spoon** with a black handle on a wooden table
Prompt Generation
"A fork, a knife, a plate, and a spoon, top-down"
Web-Scale Diffusion Model, e.g. DALL-E
Generated Image
Mask R-CNN & CLIP
Object-Level Representation
Object Matching & Pose Estimation
Target Poses
Robot Execution
Created Arrangement

Kapelyukh, Ivan, Vitalis Vosylius, and Edward Johns. "Dall-e-bot: Introducing web-scale diffusion models to robotics." *IEEE Robotics and Automation Letters* 8, no. 7 (2023): 3956-3963.

DALL-E-Bot in Action



Ref : https://youtu.be/z2g6OdqDcGQ?si=XRozGQn5JVa9qV-p

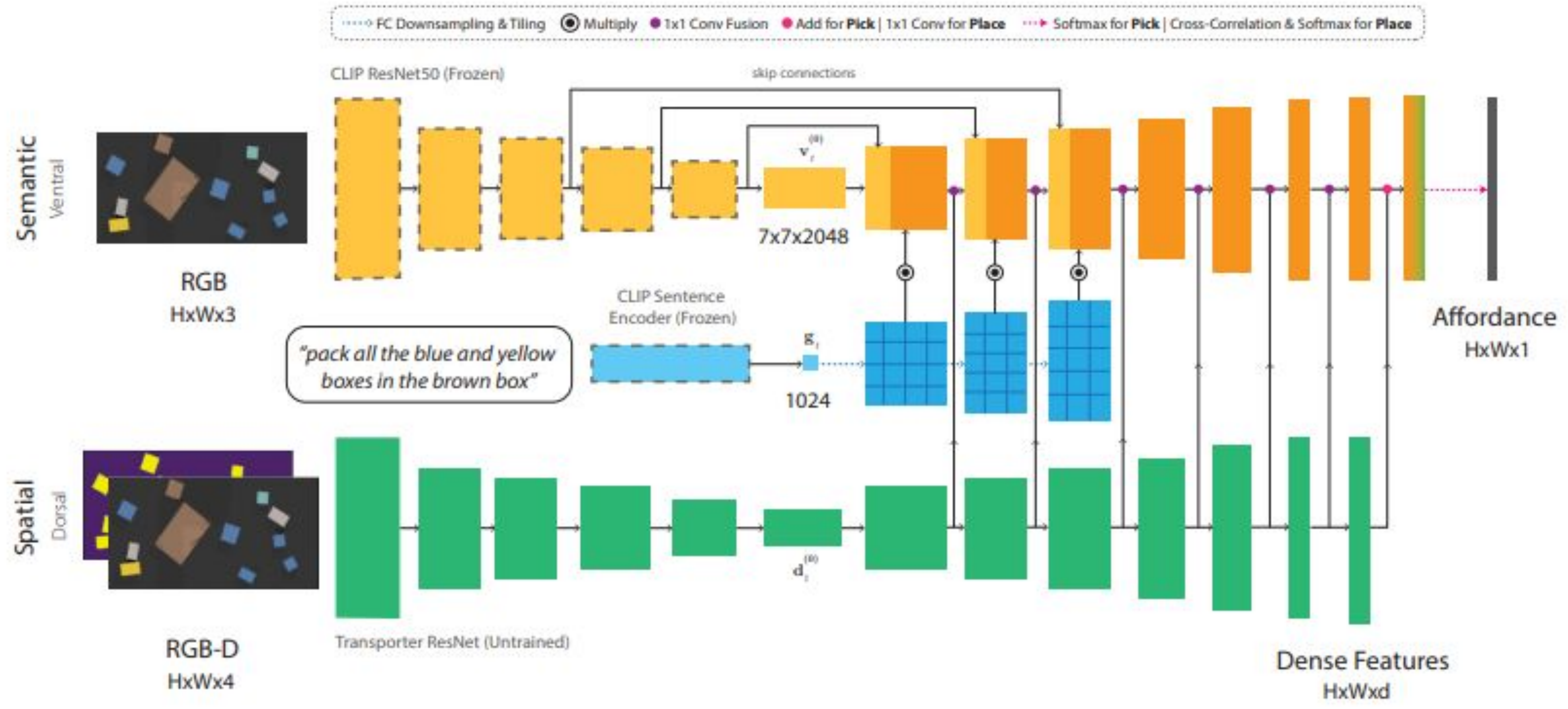
Kapelyukh, Ivan, Vitalis Vosylius, and Edward Johns. "Dall-e-bot: Introducing web-scale diffusion models to robotics." *IEEE Robotics and Automation Letters* 8, no. 7 (2023): 3956-3963.

CLIPort

- CLIP (Contrastive Language-Image Pre-Training) is a neural network trained on a variety of (image, text) pairs.
- CLIPort combines CLIP with robot manipulation.
- Matches language instructions to objects.
- Executes precise actions for object manipulation.
- Applications:
  - Warehouse automation
  - Object sorting

Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. "Cliport: What and where pathways for robotic manipulation." In *Conference on robot learning*, pp. 894-906. PMLR, 2022.

How CLIPort works



Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. "Cliport: What and where pathways for robotic manipulation." In *Conference on robot learning*, pp. 894-906. PMLR, 2022.

CLIPort Demo

Put in Bowl Task
" put the red blocks in the green bowl "
(unseen red-green goal combo)

Ref : https://youtu.be/UdzoagBgWTA?si=l6vsAVAzX-xfffFH

Shridhar, Mohit, Lucas Manuelli, and Dieter Fox. "Cliport: What and where pathways for robotic manipulation." In *Conference on robot learning*, pp. 894-906. PMLR, 2022.

Benefits of Using Foundation Models in Robotics

Improved Understanding: Robots interpret human commands more accurately.

Versatility: One model can be used to perform several tasks.

Challenges of Using Foundation Models in Robotics

- Computational Demands:
  - Larger models require significant resources.
- Safety Concerns:
  - Ensuring robots act reliably in all scenarios.
- Real Time Challenges:
  - Adaptability of dynamic environments
  - Integration with real world sensors

Future Direction for Foundation Models in Robotics

- Optimization:
  - Making models faster for real-time usage.
- Enhanced Multimodality:
  - Better integration of vision, language, and actions.
- Applications:
  - Currently limited to domains with a lot of data
  - Expanding use in agriculture, healthcare, education, etc.

Overview

Need for Robotic Foundation Models

- **Challenges**
- **Characteristics of a Robotic Foundation model**

Recent Attempts to Create Robotic Foundation Models

- Open X-Embodiment Dataset
- Foundation Models
  - Google DeepMind RT-1-X
  - Google DeepMind RT-2-X
  - OpenVLA
- Conclusion

Need for Robotic Foundation Models

Robots are great specialists, but poor generalists. Typically, you have to train a model for each task, robot, and environment. Changing even a single variable often requires starting from scratch.

Ref : https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/

Need for Robotic Foundation Models

Robots are great specialists, but poor generalists. Typically, you have to train a model for each task, robot, and environment. Changing even a single variable often requires starting from scratch.

This is due to the following key challenges:

- **Complex & Dynamic Environments:** Robots often operate in unpredictable, unstructured real-world spaces. Traditional models struggle to adapt beyond their narrow training scenarios.

Ref : https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/

Need for Robotic Foundation Models

Robots are great specialists, but poor generalists. Typically, you have to train a model for each task, robot, and environment. Changing even a single variable often requires starting from scratch.

This is due to the following key challenges:

- **Complex & Dynamic Environments:** Robots often operate in unpredictable, unstructured real-world spaces. Traditional models struggle to adapt beyond their narrow training scenarios.
- **Limited transferability:** Knowledge gained in one scenario rarely helps in another.

Ref : https://deepmind.google/discover/blog/scaling-up-learning-across-many-different-robot-types/

Need for Robotic Foundation Models

To solve these challenges we need a model that has:

- **Generalized Understanding:** Model trained from large, diverse datasets, capturing a broad range of skills and concepts which exhibits zero-shot and few-shot performance on novel tasks.

Need for Robotic Foundation Models

To solve these challenges we need a model that has:

- **Generalized Understanding:** Model trained from large, diverse datasets, capturing a broad range of skills and concepts which exhibits zero-shot and few-shot performance on novel tasks.

- **Bridge Between Vision, Language, and Action:** Integrates multimodal inputs (images, language instructions) to produce meaningful, goal-directed actions.

Overview

Need for Robotic Foundation Models

- **Challenges**
- **Characteristics of a Robotic Foundation model**

Recent Attempts to Create Robotic Foundation Models

- **Open X-Embodiment Dataset**
- Foundation Models
  - Google DeepMind RT-1-X
  - Google DeepMind RT-2-X
  - OpenVLA
- Conclusion

Recent Attempts to Create Robotic Foundation Models

Open X-Embodiment Dataset

QT-Opt
pick anything
TOTO
pour
sweep the green cloth to the left side of the table
Push T
stack cups
pick red block
place the black bowl in the dish rack
Jaco Play
ALOHA
Taco Play
1M Episodes from 311 Scenes
34 Research Labs across 21 Institutions
22 Embodiments
527 Skills
pour
stack
route
60 Datasets
1,798 Attributes • 5,228 Objects • 23,486 Spatial Relations
Cable Routing
RT-1
pick green chip bag from counter
set the bowl to the right side of the table
Bridge
Door Opening

O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Open X-Embodiment Dataset

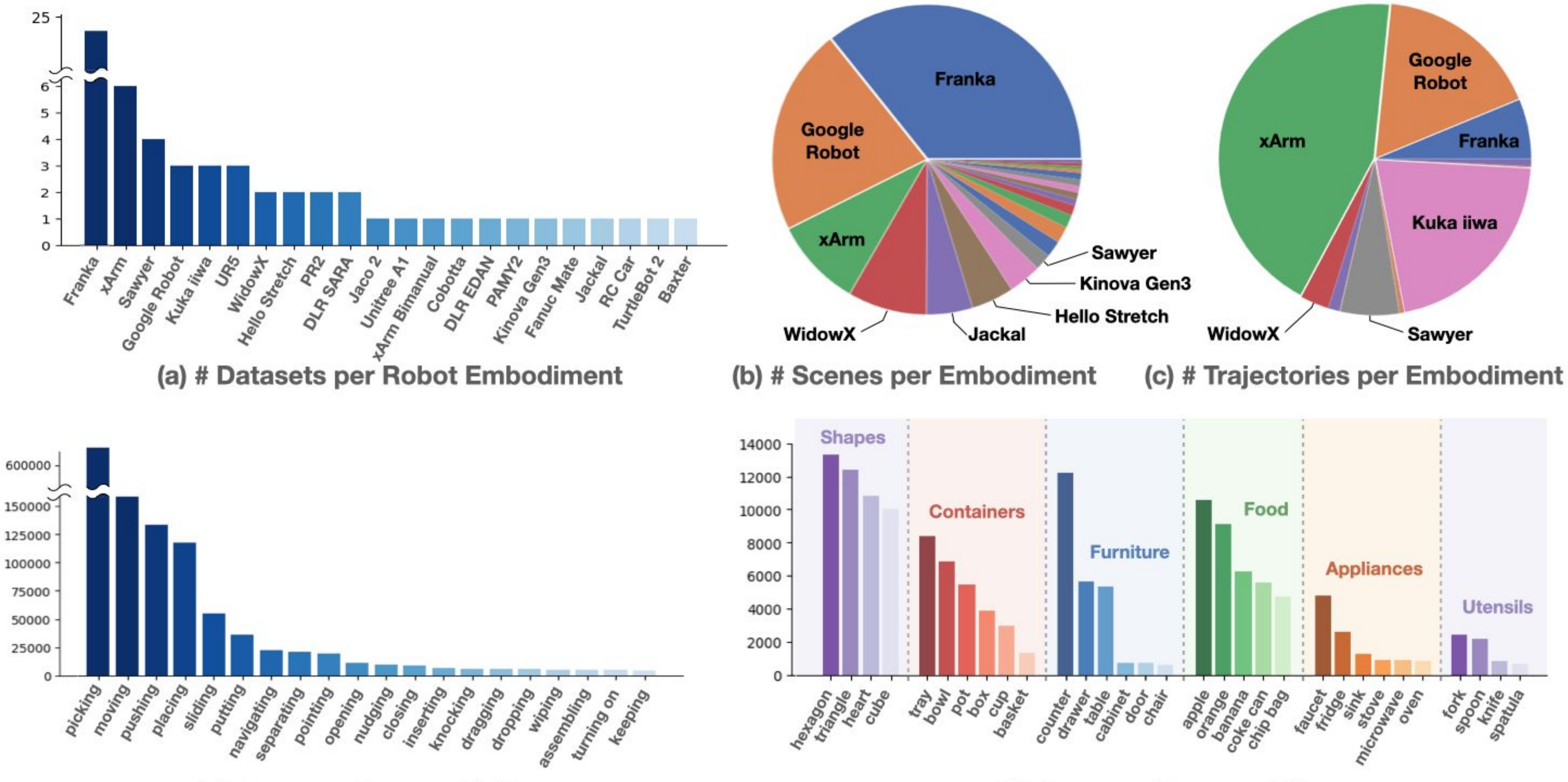
Dataset Scale:

- Over 1 million episodes spanning 22 robot embodiments.
- Unified from **60 individual datasets** across 34 global research institutions.

Dataset Features:

- **Robot Diversity:** Franka, Sawyer, xArm, Google Robot, WidowX, etc.
- **Skill Diversity:** Picking, placing, wiping, assembling, dragging, etc.
- **Object Variety:** Household items, appliances, utensils, and more.

O’Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.



(a) # Datasets per Robot Embodiment

(b) # Scenes per Embodiment

(c) # Trajectories per Embodiment

(d) Common Dataset Skills

(e) Common Dataset Objects

O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Overview

Need for Robotic Foundation Models

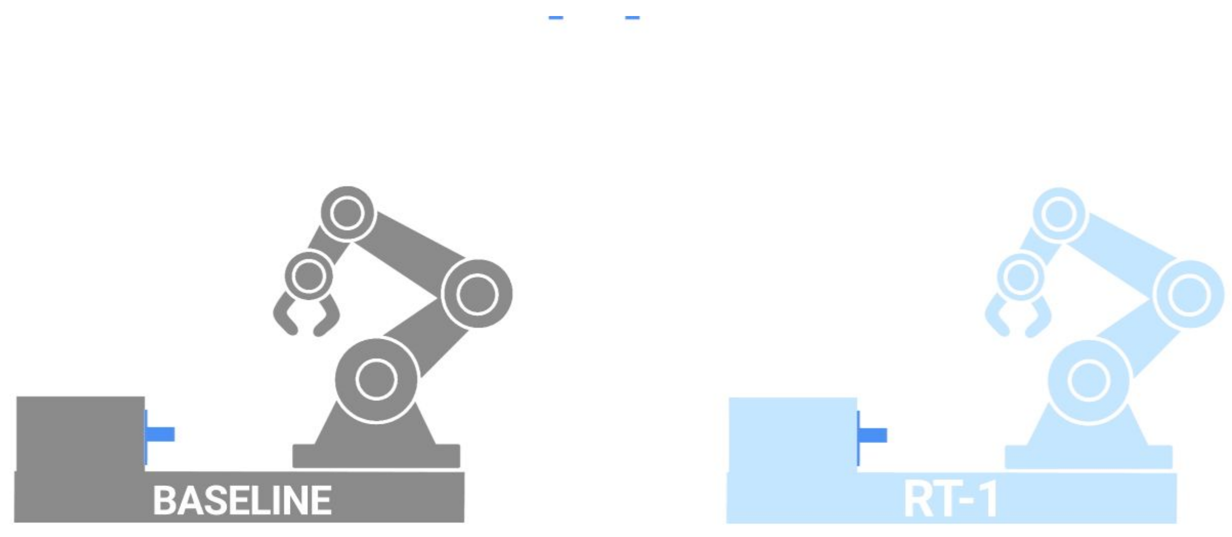
- **Challenges**
- **Characteristics of a Robotic Foundation model**

Recent Attempts to Create Robotic Foundation Models

- **Open X-Embodiment Dataset**
- **Foundation Models**
  - **Google DeepMind RT-1-X**
  - Google DeepMind RT-2-X
  - OpenVLA
- Conclusion

Google DeepMind RT-1-X

- **Released by Google DeepMind in 2023**: Built as an extension of the RT-1 model.
- **Trained on Open X-Embodiment Dataset**: Utilizes over 1 million episodes spanning 22 robotic embodiments for diverse task adaptability.



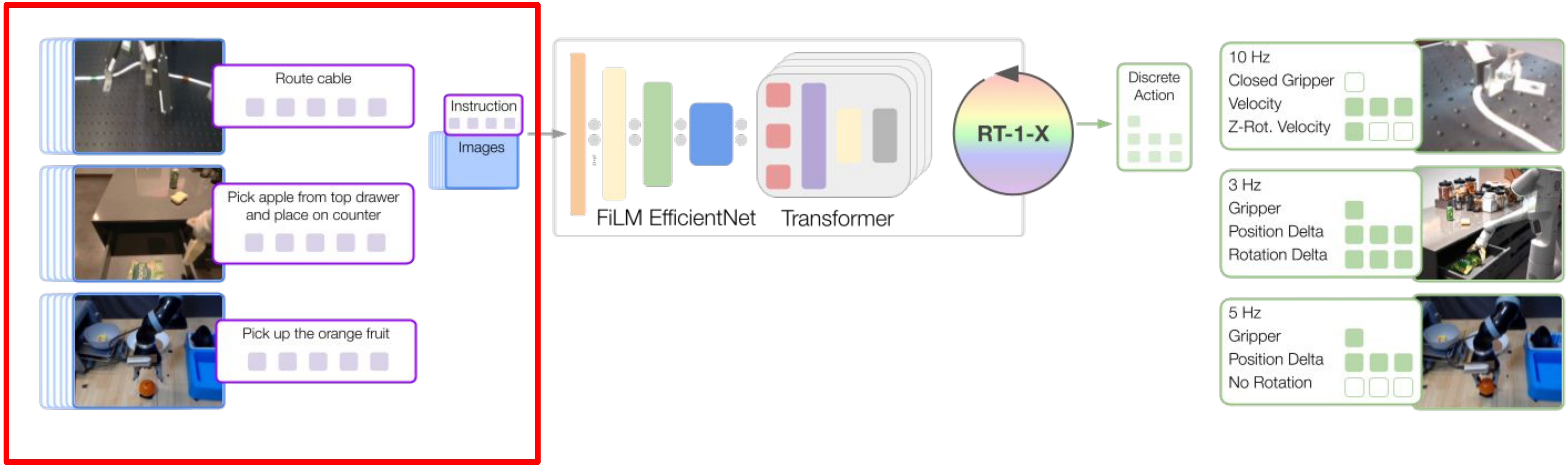
https://robotics-transformer1.github.io/

O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-1-X

Input:

- History of 15 RGB images from a single canonical camera view, resized to a standard resolution.
- Natural language instruction describing the task.

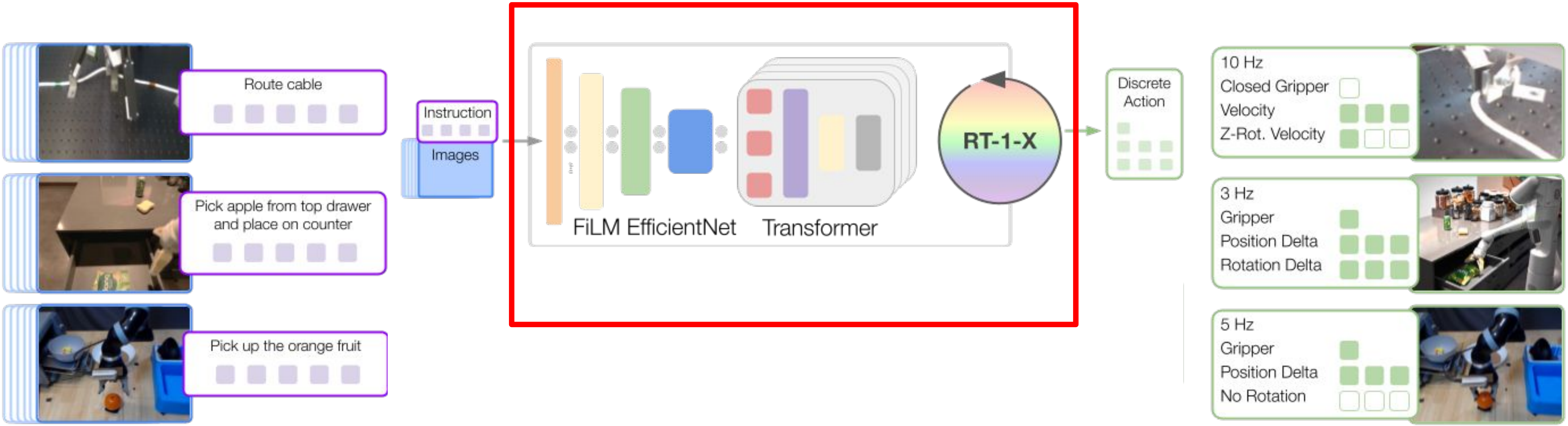


O’Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-1-X

Components:

1. **Vision Backbone: EfficientNet** pretrained on ImageNet for image feature extraction.

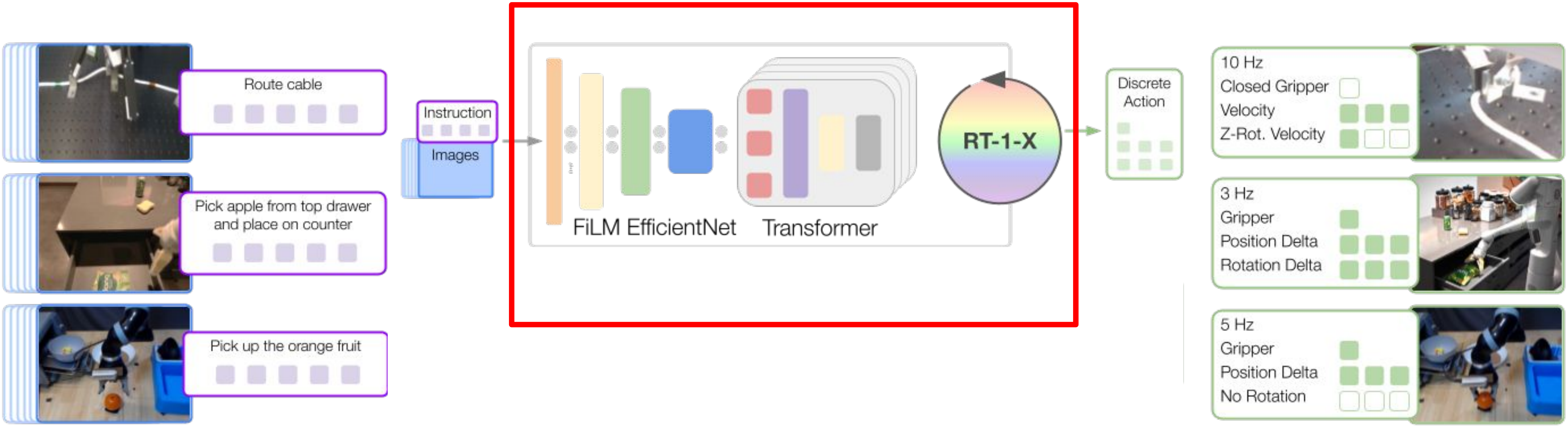


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-1-X

Components:

1. **Vision Backbone: EfficientNet** pretrained on ImageNet for image feature extraction.
2. **Language Encoder: Universal Sentence Encoder (USE)** for embedding natural language instructions.

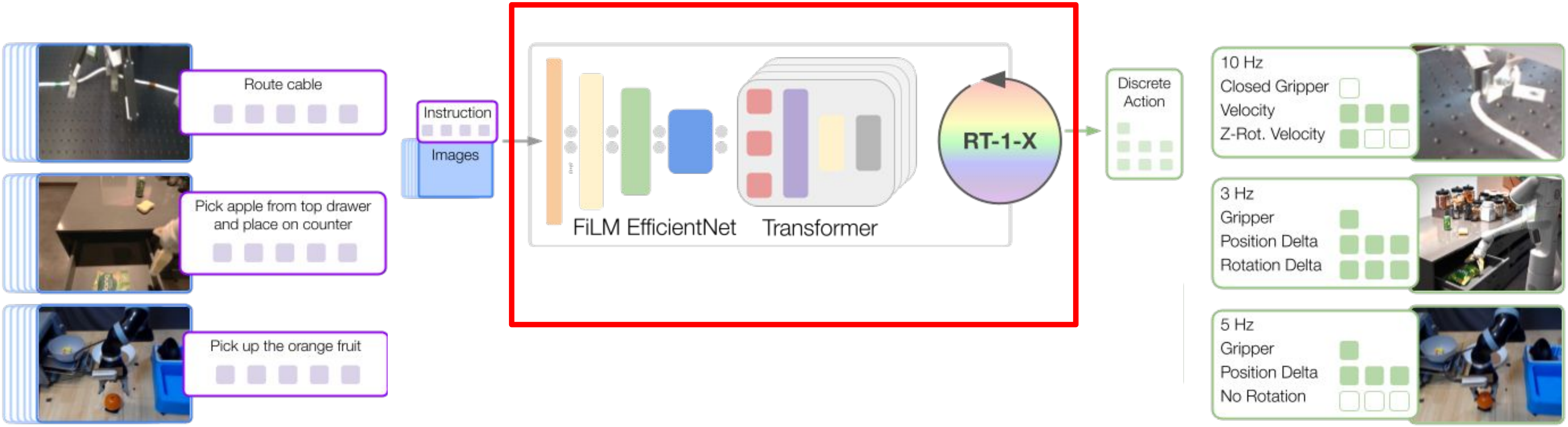


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-1-X

Components:

1. **Vision Backbone: EfficientNet** pretrained on ImageNet for image feature extraction.
2. **Language Encoder: Universal Sentence Encoder (USE)** for embedding natural language instructions.
3. **FiLM (Feature-wise Linear Modulation):** Integrates vision and language representations into **81 interwoven tokens**.

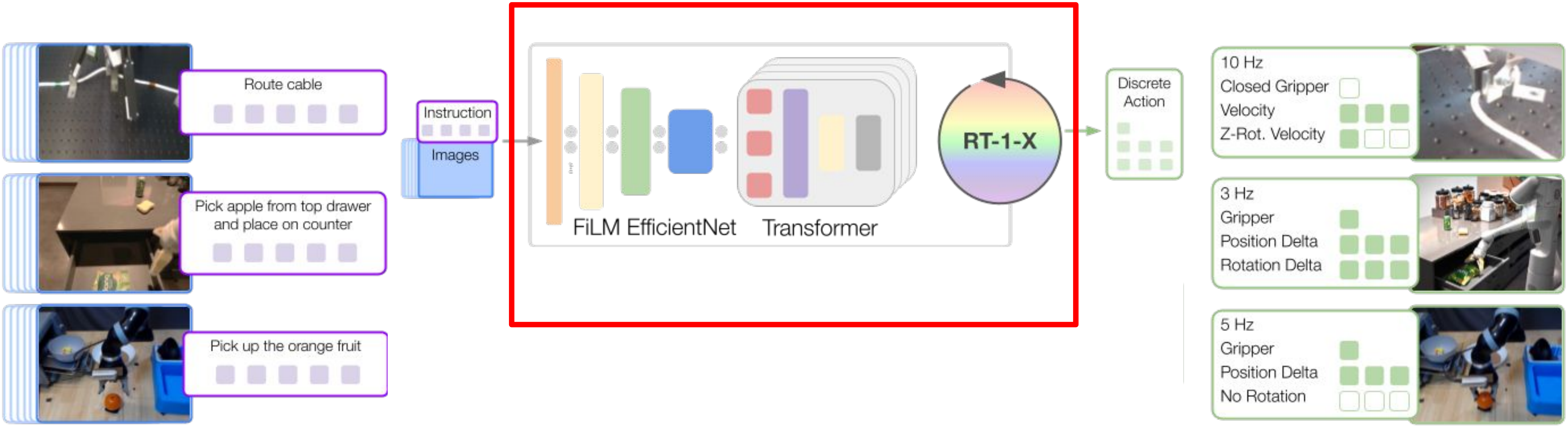


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-1-X

Components:

1. **Vision Backbone: EfficientNet** pretrained on ImageNet for image feature extraction.
2. **Language Encoder: Universal Sentence Encoder (USE)** for embedding natural language instructions.
3. **FiLM (Feature-wise Linear Modulation):** Integrates vision and language representations into **81 interwoven tokens**.
4. **Transformer Decoder:** Processes vision-language tokens to generate action tokens.

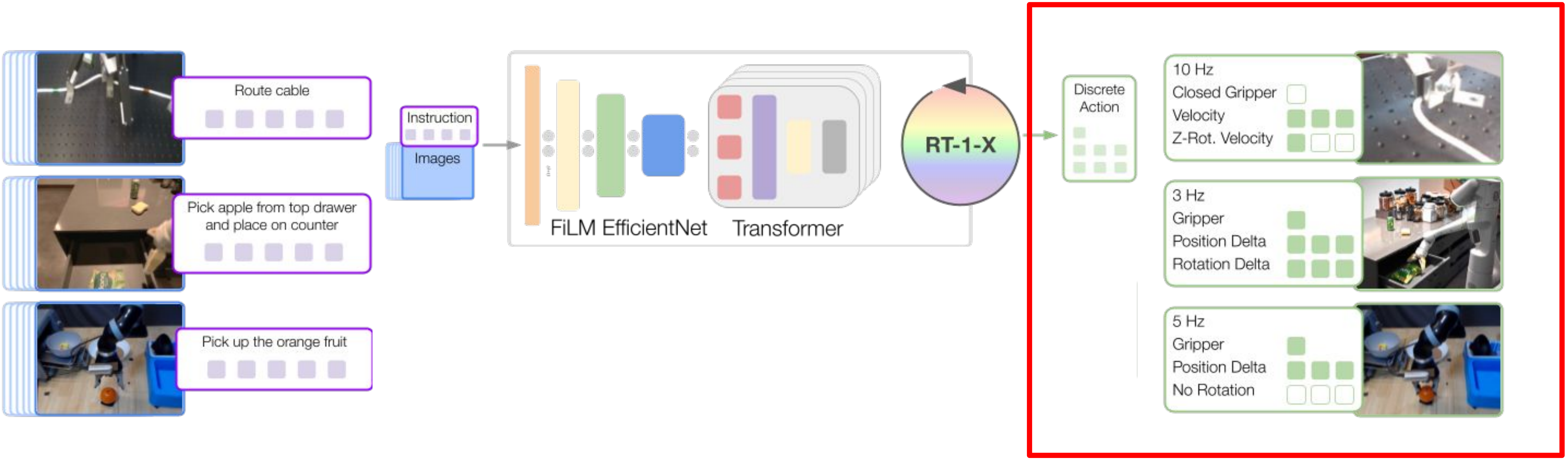


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-1-X

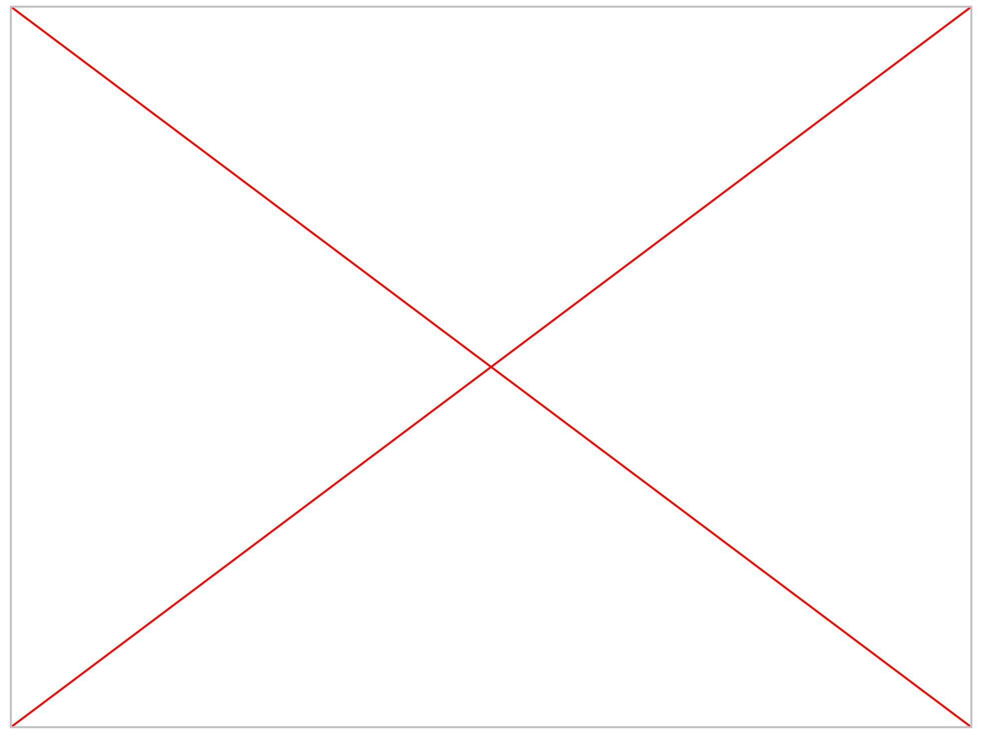
Output:

- **7 DoF Actions:** x, y, z, roll, pitch, yaw, gripper opening/closing.
- Discretized into 256 bins for efficient action generation.

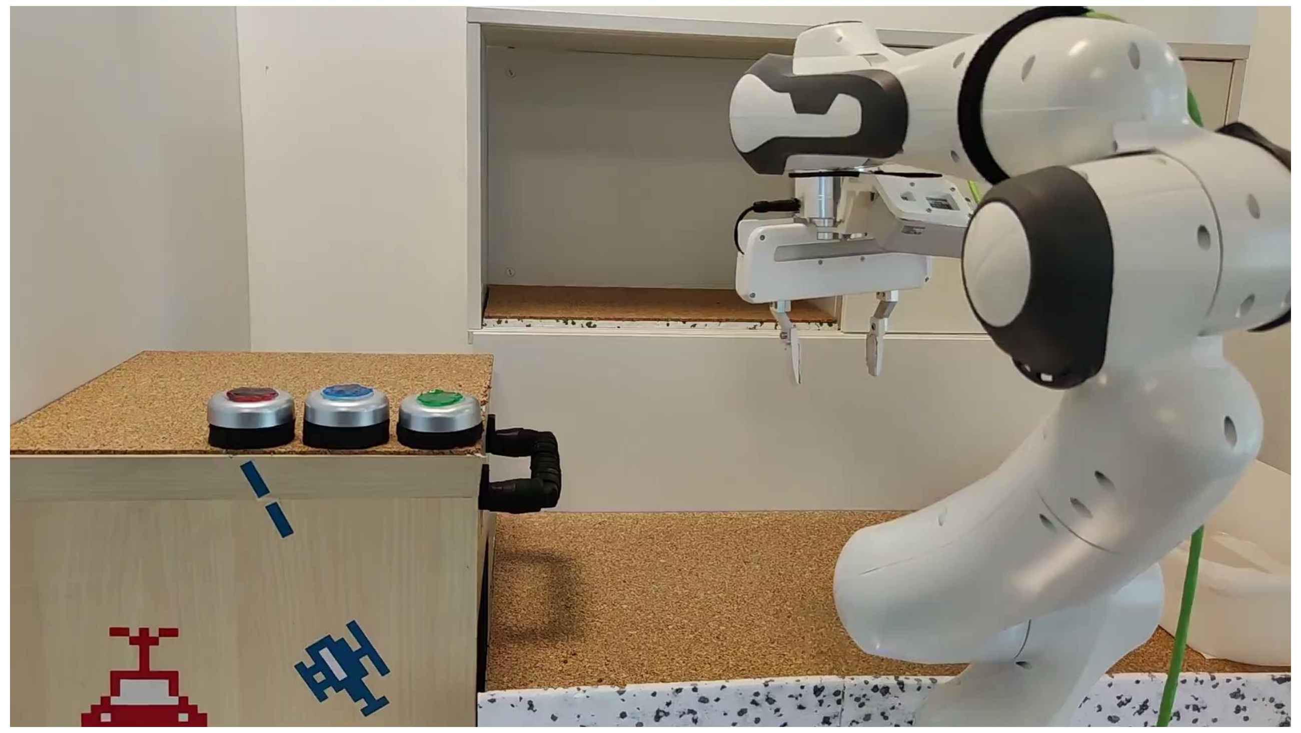


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

RT-1-X Evaluation

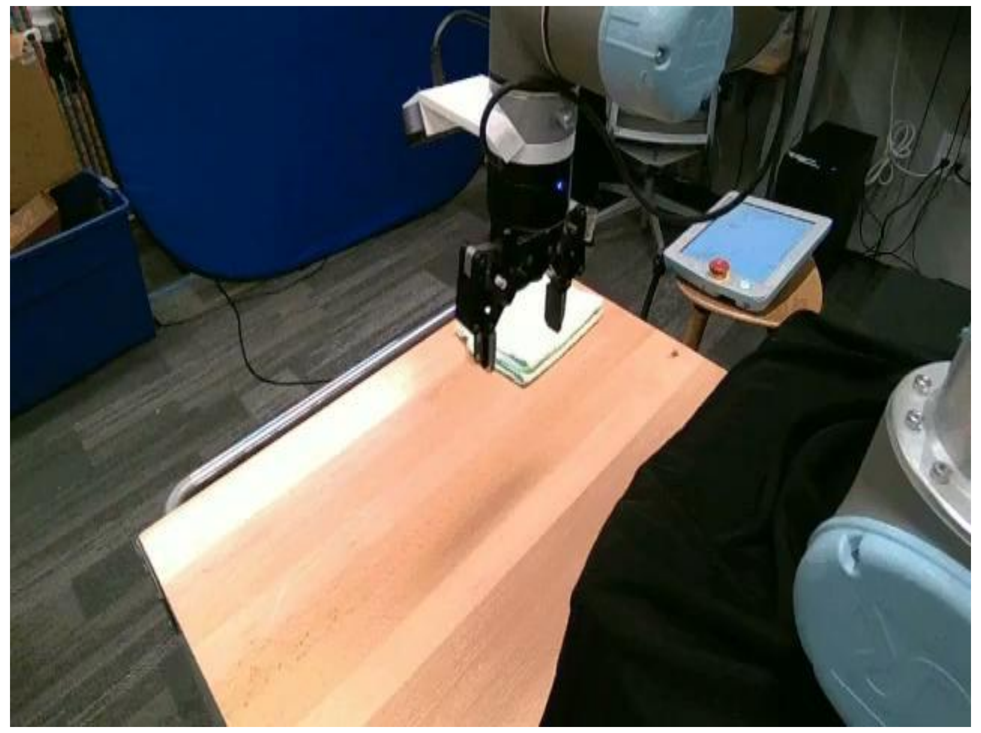


UC Berkeley(RAIL)

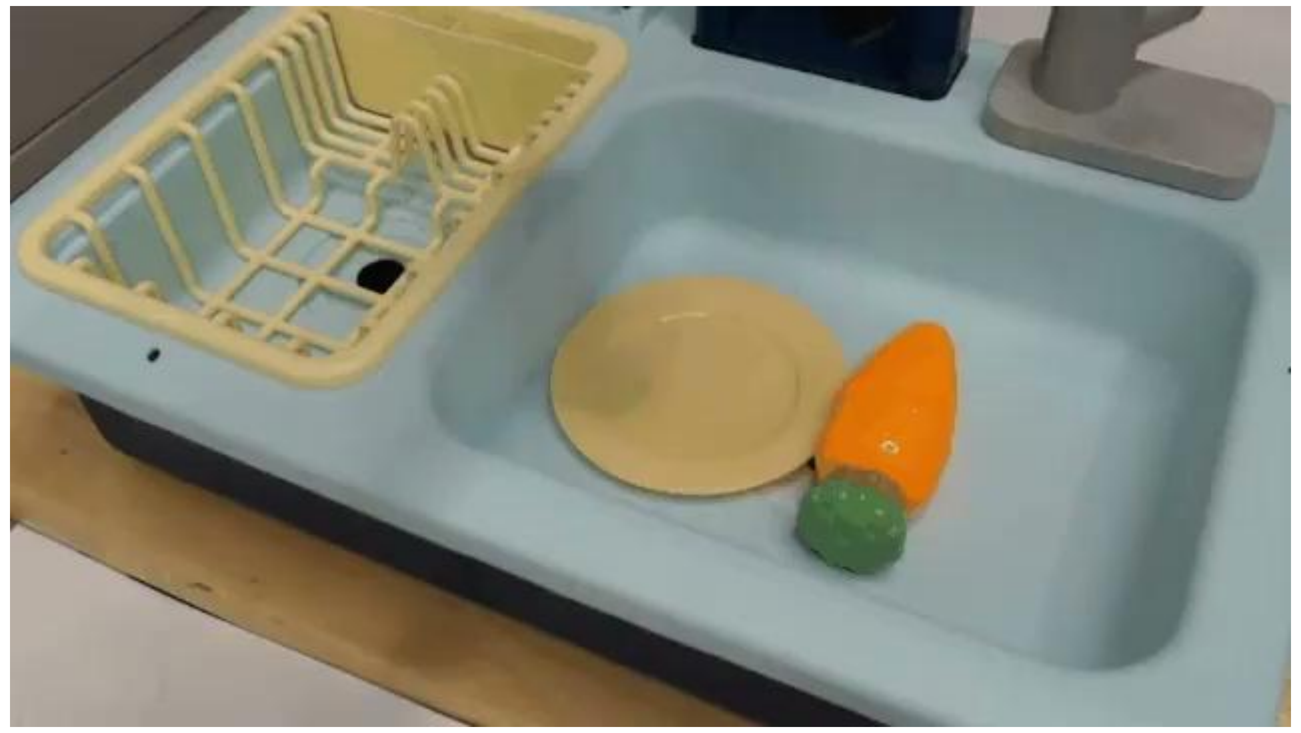


University of Freiburg(AiS)

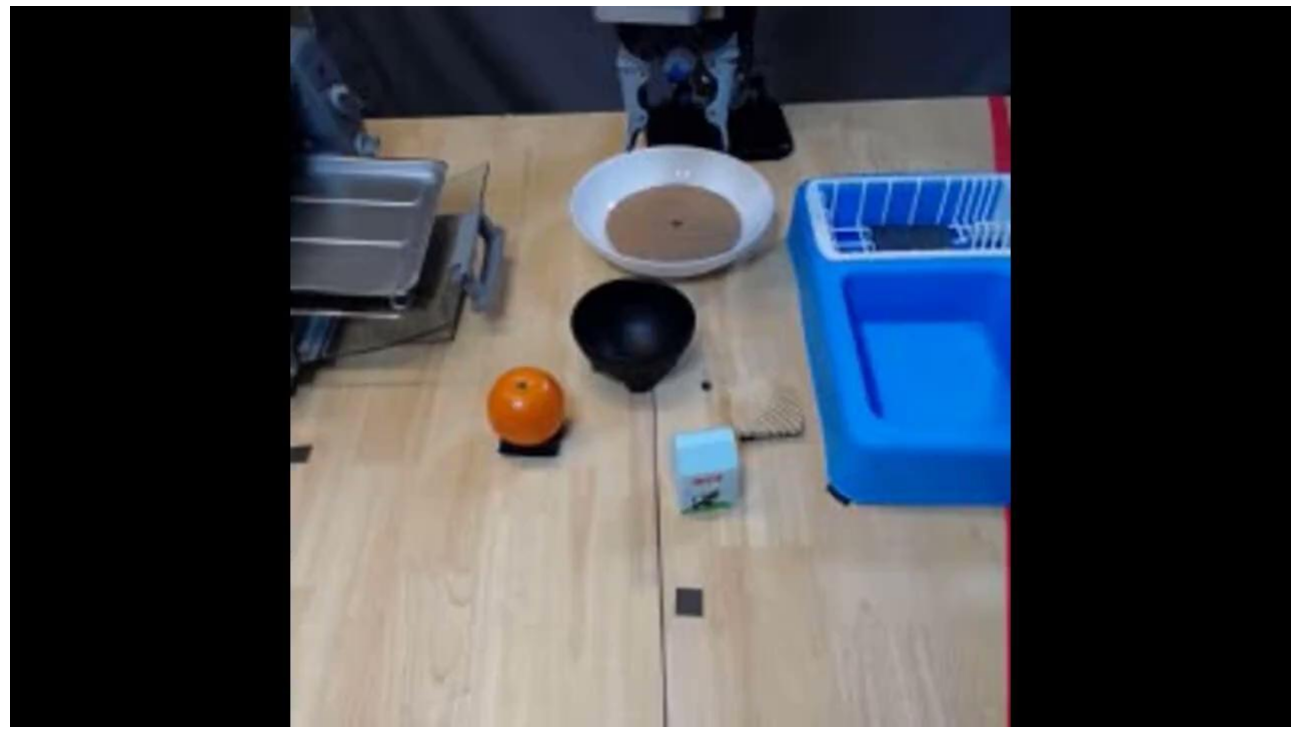
NYU (CILVR)



UC Berkeley(AUTOLab)



Stanford(IRIS)

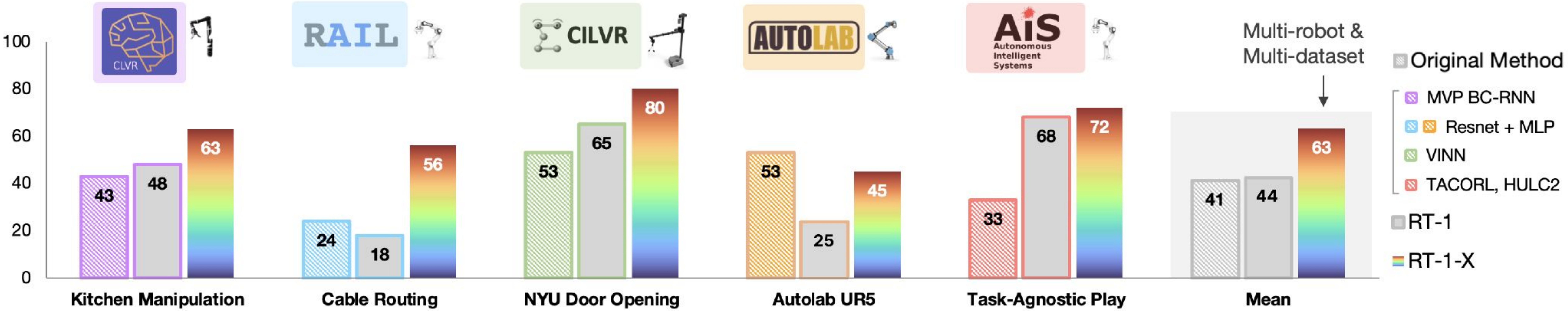


USC(CLVR)

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

O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

RT-1-X Evaluation



RT-1-X models outperform RT-1 or Original Methods trained on individual datasets by 50% in the small-data domain

O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Overview

Need for Robotic Foundation Models

- **Challenges**
- **Characteristics of a Robotic Foundation model**

Recent Attempts to Create Robotic Foundation Models

- **Open X-Embodiment Dataset**
- **Foundation Models**
  - **Google DeepMind RT-1-X**
  - **Google DeepMind RT-2-X**
  - OpenVLA
- Conclusion

Google DeepMind RT-2-X

- **Released by Google DeepMind in 2023**: RT-2-X builds on the RT-2 model, incorporating vision-language-action (VLA) capabilities for robotic control.

- **Trained on a Mix of Robotics and Web-scale Data**: Combines robotic datasets with web-scale vision-language data for enhanced generalization.

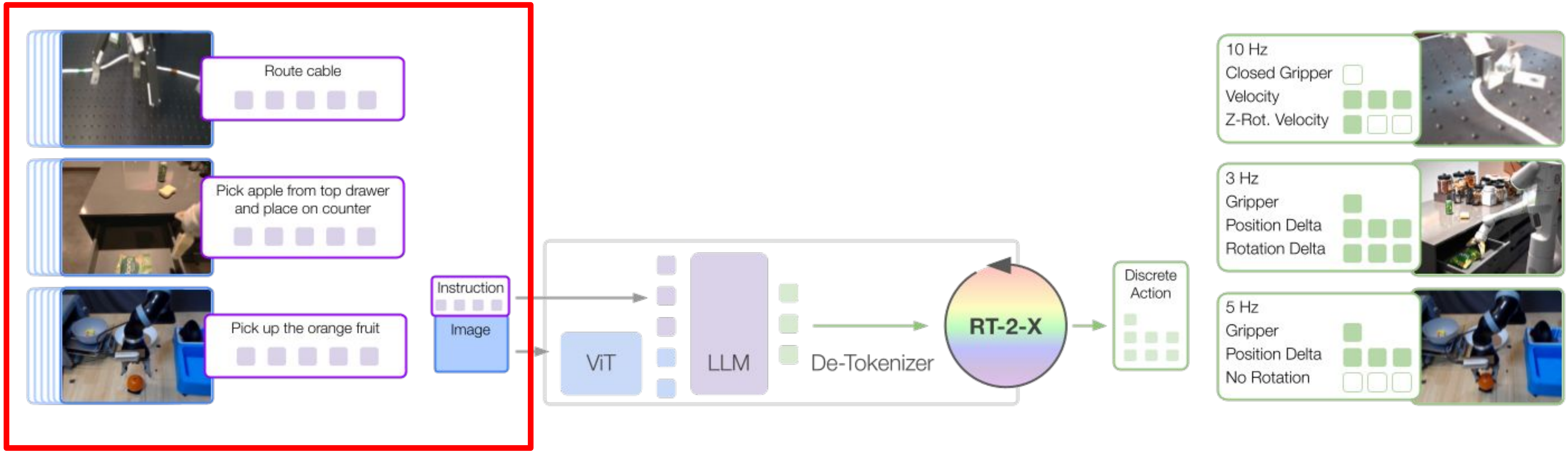
https://robotics-transformer2.github.io/

O’Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-2-X

Input:

- **Visual Input:** Single RGB image from a canonical camera view, resized to a standard resolution.
- **Language Input:** Natural language instruction describing the robotic task.

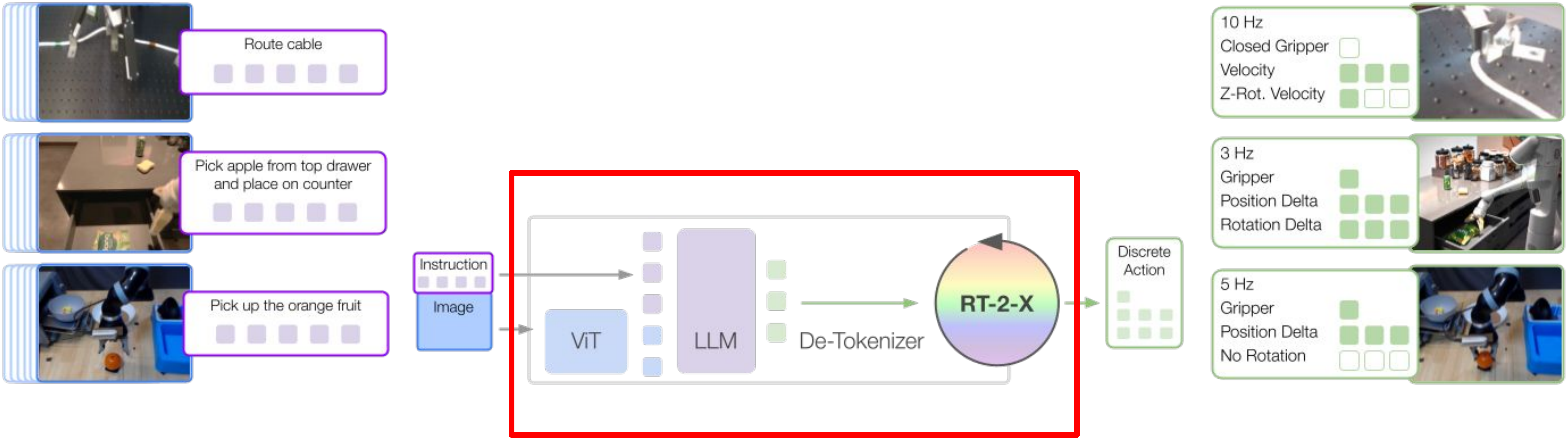


O’Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-2-X

Components:

1. **Vision Backbone: Vision Transformer (ViT):** Pretrained on large-scale vision datasets for general-purpose feature extraction. Processes input images to generate spatial feature embeddings.

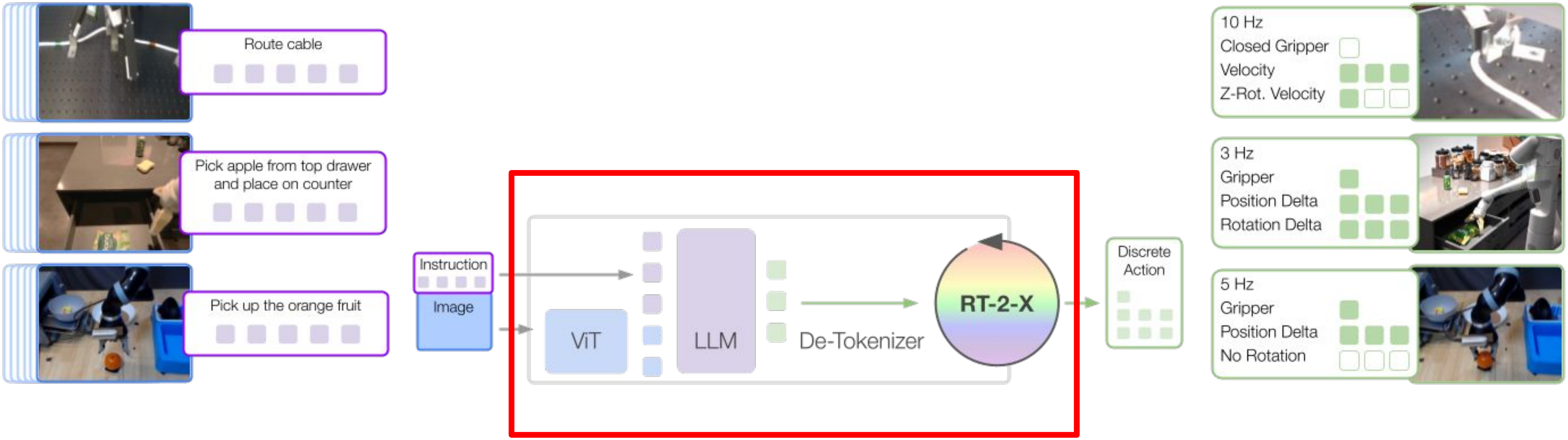


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-2-X

Components:

1. **Vision Backbone: Vision Transformer (ViT):** Pretrained on large-scale vision datasets for general-purpose feature extraction. Processes input images to generate spatial feature embeddings.
2. **Language Encoder: Unified Language Learner (UL2):** Converts task instructions into semantic embeddings. Pretrained primarily on WebLI dataset for extensive task understanding.

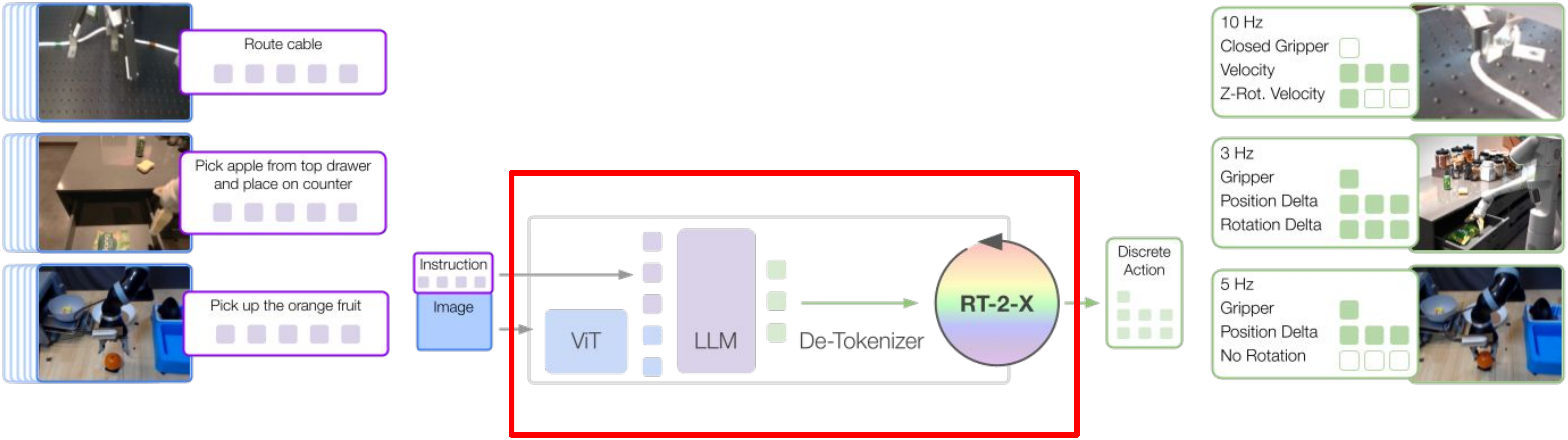


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-2-X

Components:

1. **Vision Backbone: Vision Transformer (ViT):** Pretrained on large-scale vision datasets for general-purpose feature extraction. Processes input images to generate spatial feature embeddings.
2. **Language Encoder: Unified Language Learner (UL2):** Converts task instructions into semantic embeddings. Pretrained primarily on WebLI dataset for extensive task understanding.
3. **Transformer Decoder:** Processes fused embeddings to generate sequential action tokens. Each layer incorporates attention mechanisms for understanding dependencies between modalities.

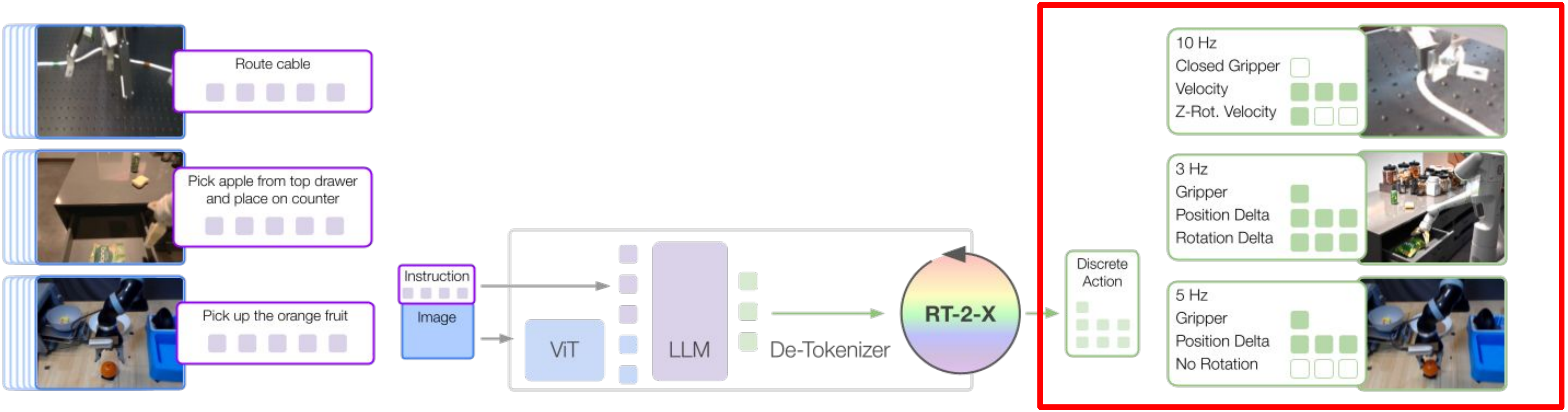


O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

Google DeepMind RT-2-X

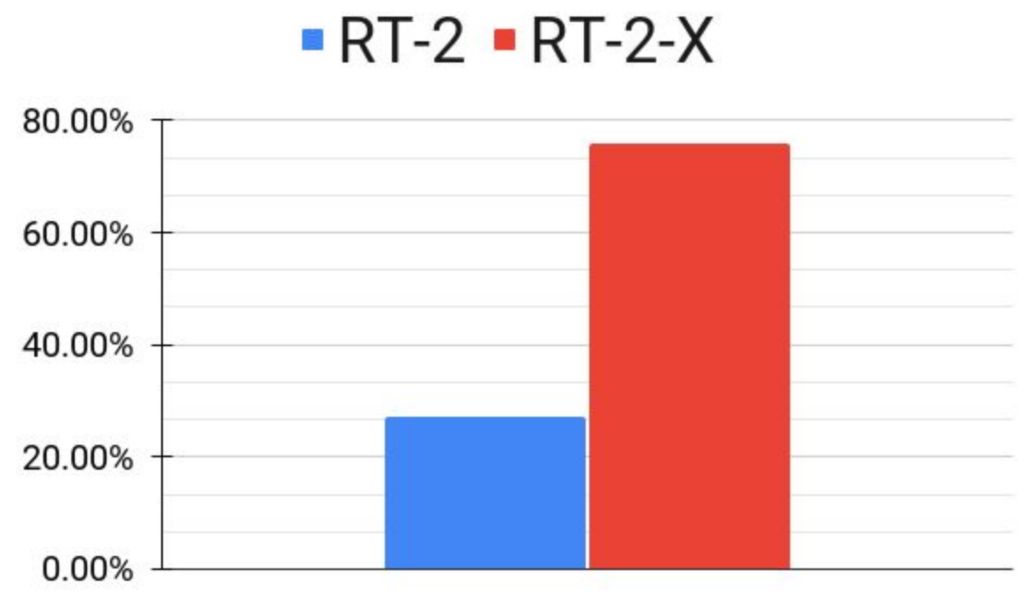
Output:

- **7 DoF Actions:** x, y, z, roll, pitch, yaw, gripper opening/closing.
- Discretized into 256 bins for efficient action generation.



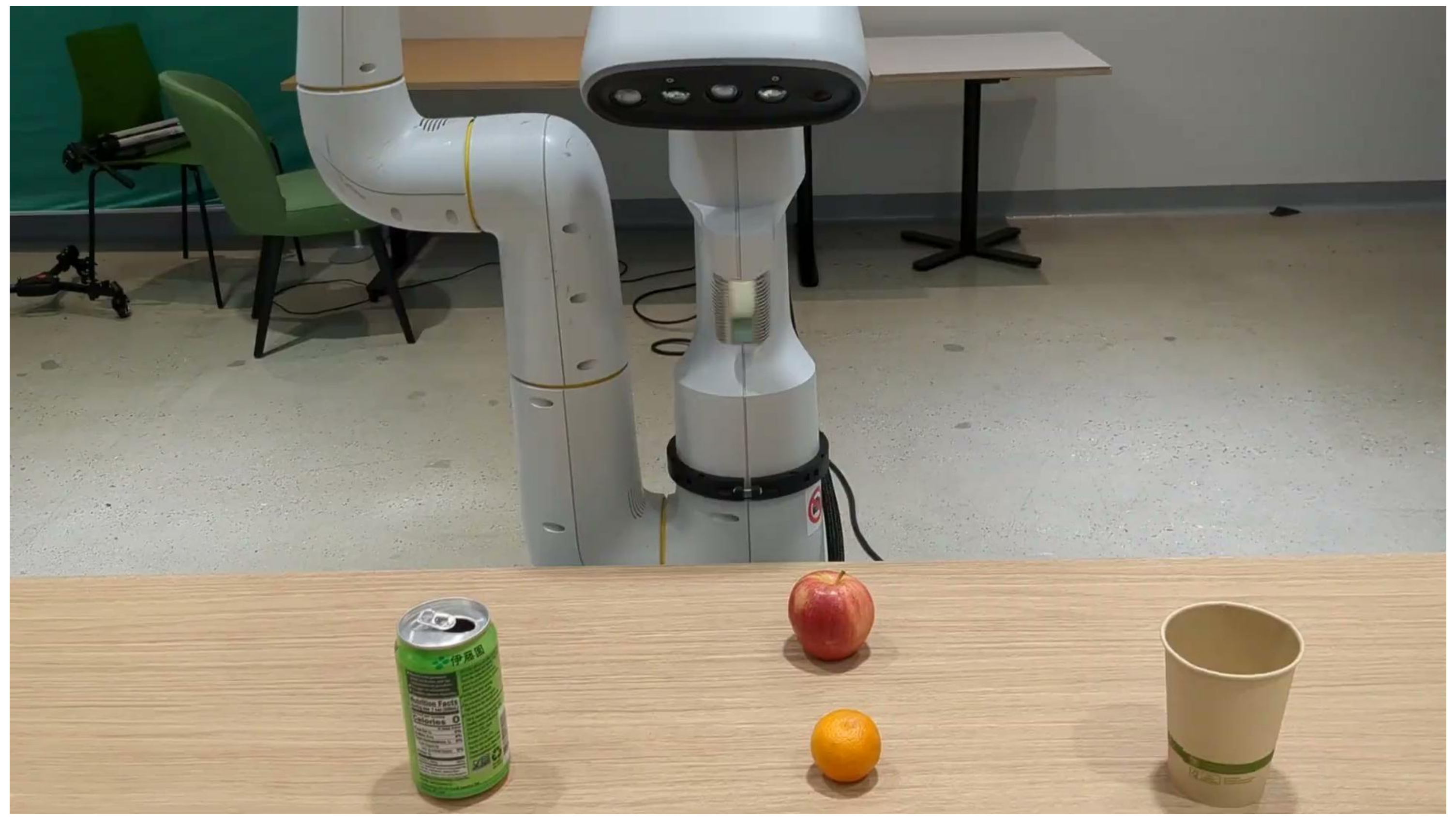
O'Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

RT-2-X Evaluation on Emergent Skills



move apple near cloth

move apple on cloth



move apple between can & orange

RT-2-X modulates low-level behaviors based on small changes in prepositions (see “on” vs “near” above) and demonstrates understanding of spatial relationships between objects

RT-2-X outperforms RT-2 by 3x in emergent skill evaluations

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

O’Neill, Abby, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley et al. "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0." In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892-6903. IEEE, 2024.

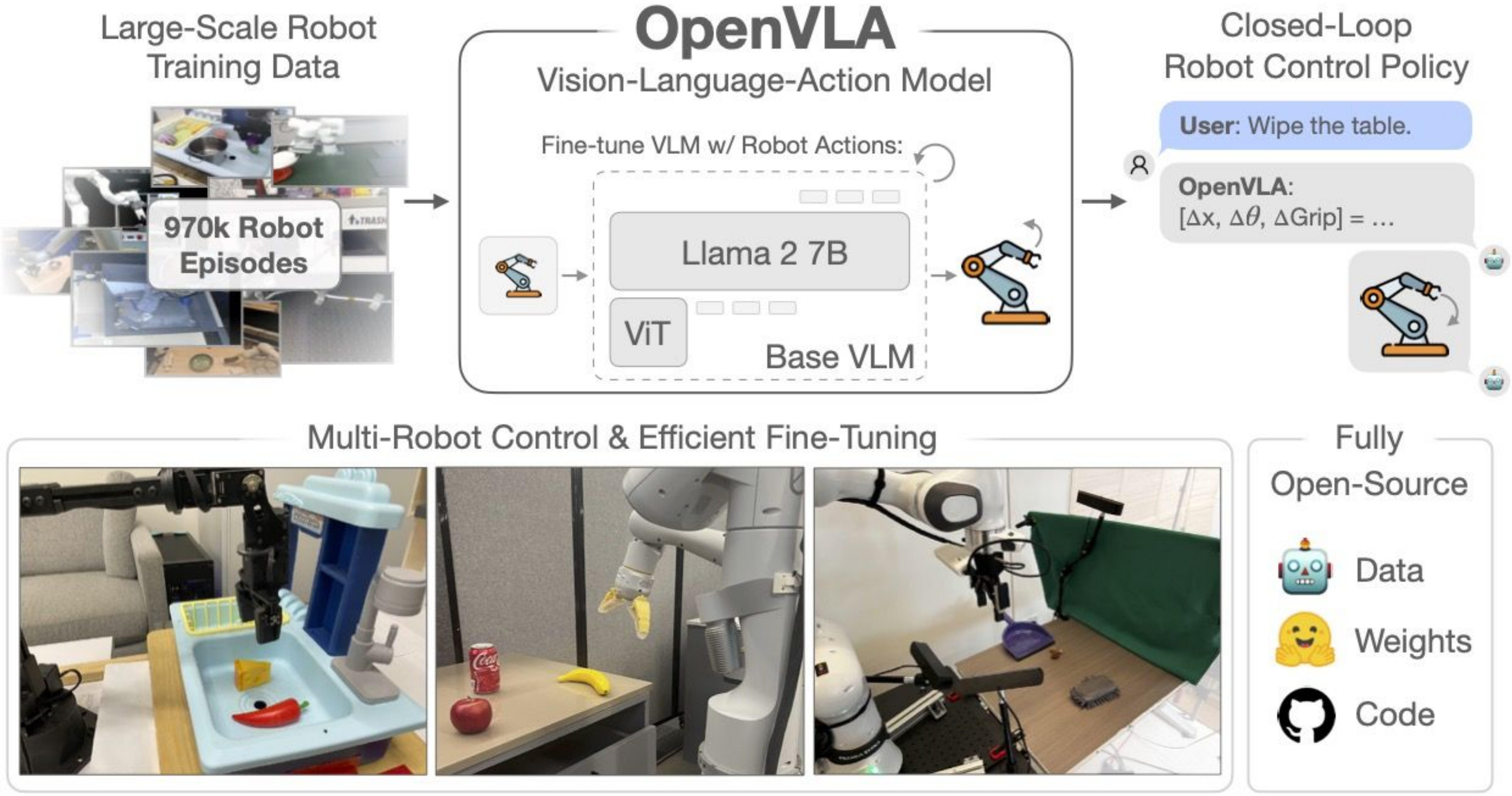
Overview

Need for Robotic Foundation Models

- Challenges
- Characteristics of a Robotic Foundation model

Recent Attempts to Create Robotic Foundation Models

- Open X-Embodiment Dataset
- Foundation Models
  - Google DeepMind RT-1-X
  - Google DeepMind RT-2-X
  - OpenVLA
- Conclusion



Large-Scale Robot Training Data

970k Robot Episodes

OpenVLA

Vision-Language-Action Model

Fine-tune VLM w/ Robot Actions:

Llama 2 7B

ViT

Base VLM

Closed-Loop Robot Control Policy

User: Wipe the table.

OpenVLA: [Δx, Δθ, ΔGrip] = …

Multi-Robot Control & Efficient Fine-Tuning

Fully Open-Source

Data

Weights

Code

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

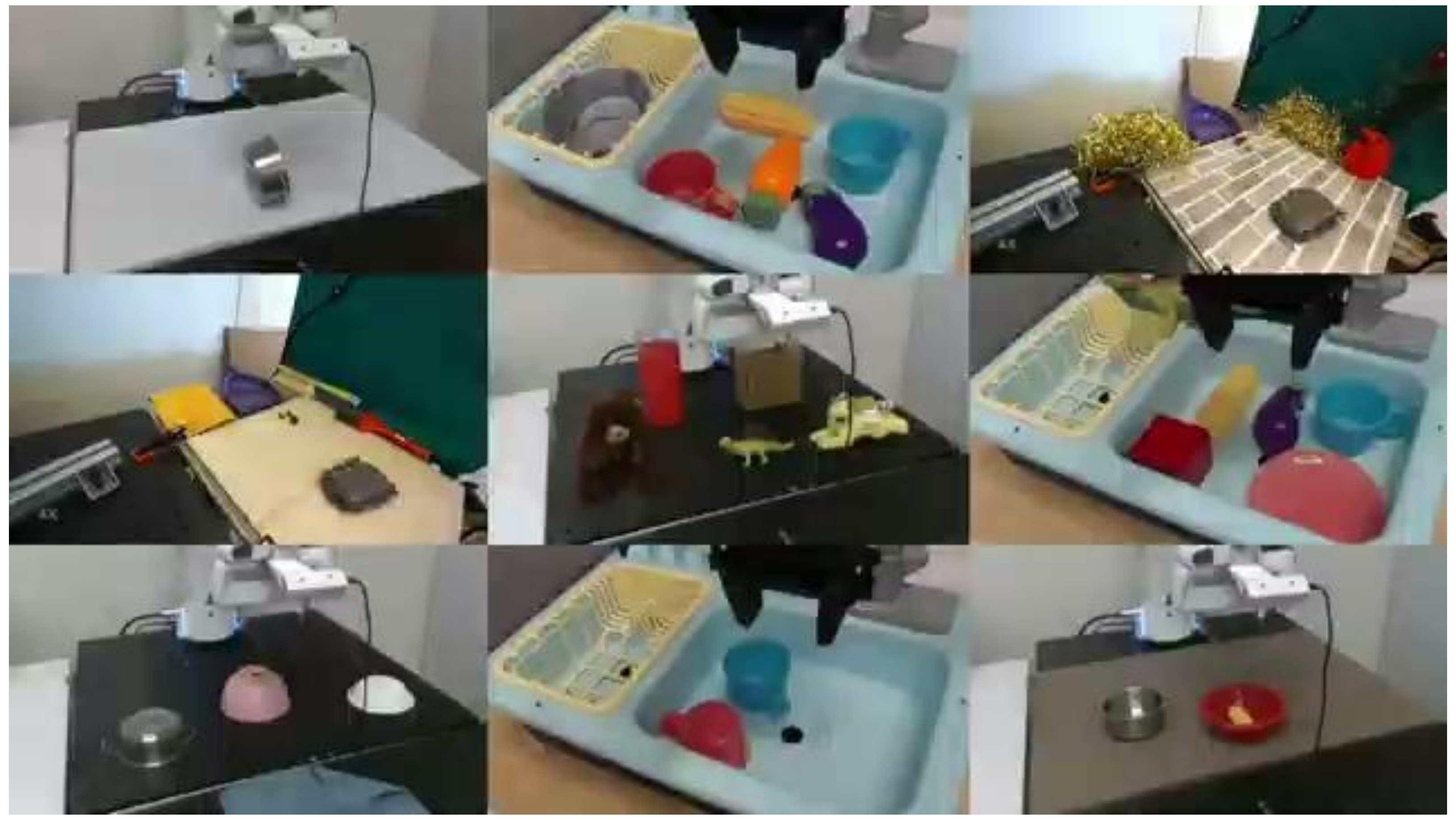
OpenVLA: An Open-Source Vision-Language-Action Model

- **Released in 2024 by Stanford University and Collaborators:** OpenVLA was developed through a collaboration between Stanford University, UC Berkeley, Toyota Research Institute, Google DeepMind, Physical Intelligence and MIT.

- **Built on Open X-Embodiment Dataset:** The model uses 970k robot episodes from the Open X-Embodiment dataset, encompassing a wide range of tasks and 22 diverse robotic embodiments, making it highly versatile for generalist robotic manipulation.

- **Architecture:** OpenVLA is a 7 billion parameter model, combining vision, language, and action modalities to enable seamless integration across perceptual, reasoning, and execution tasks.

- **Fine-Tuning for New Robots:** The model supports parameter-efficient fine-tuning, allowing it to adapt quickly to new robotic configurations and tasks with minimal computational overhead.



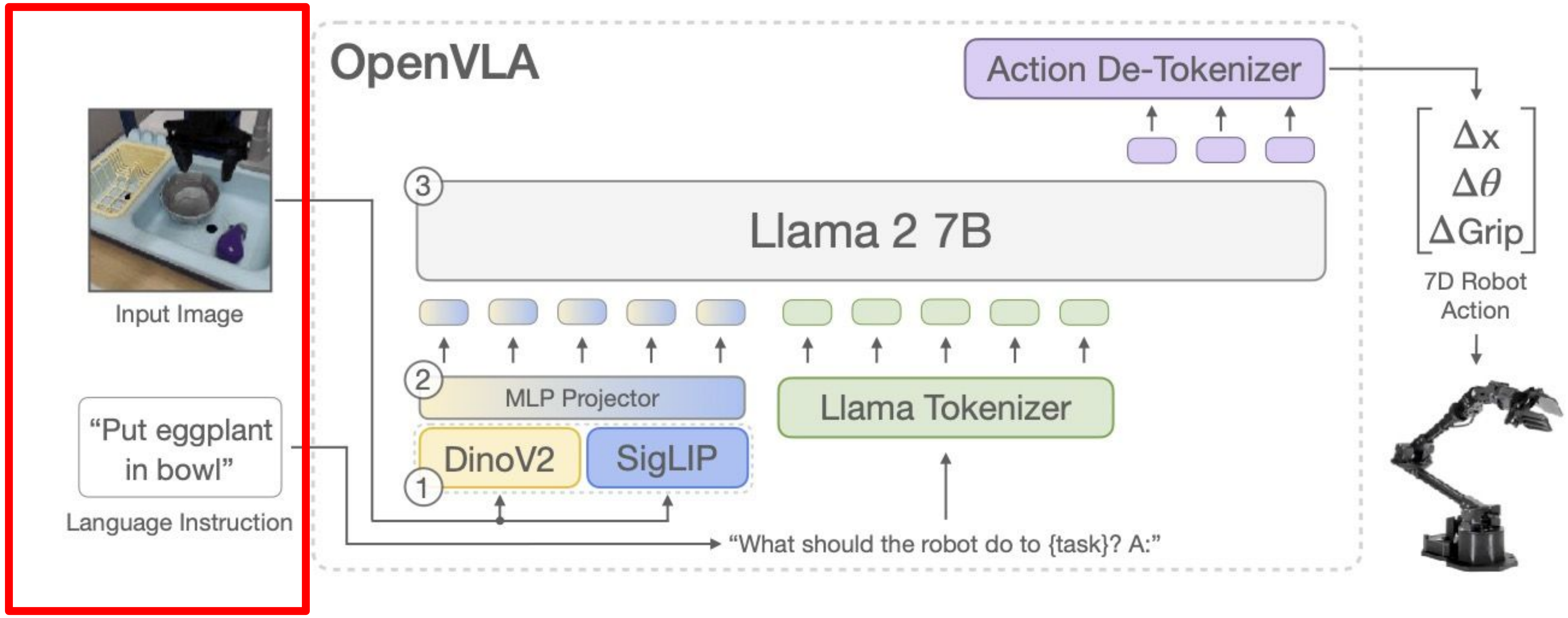
https://openvla.github.io/

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

Input:

- Takes an image and language instruction as input.

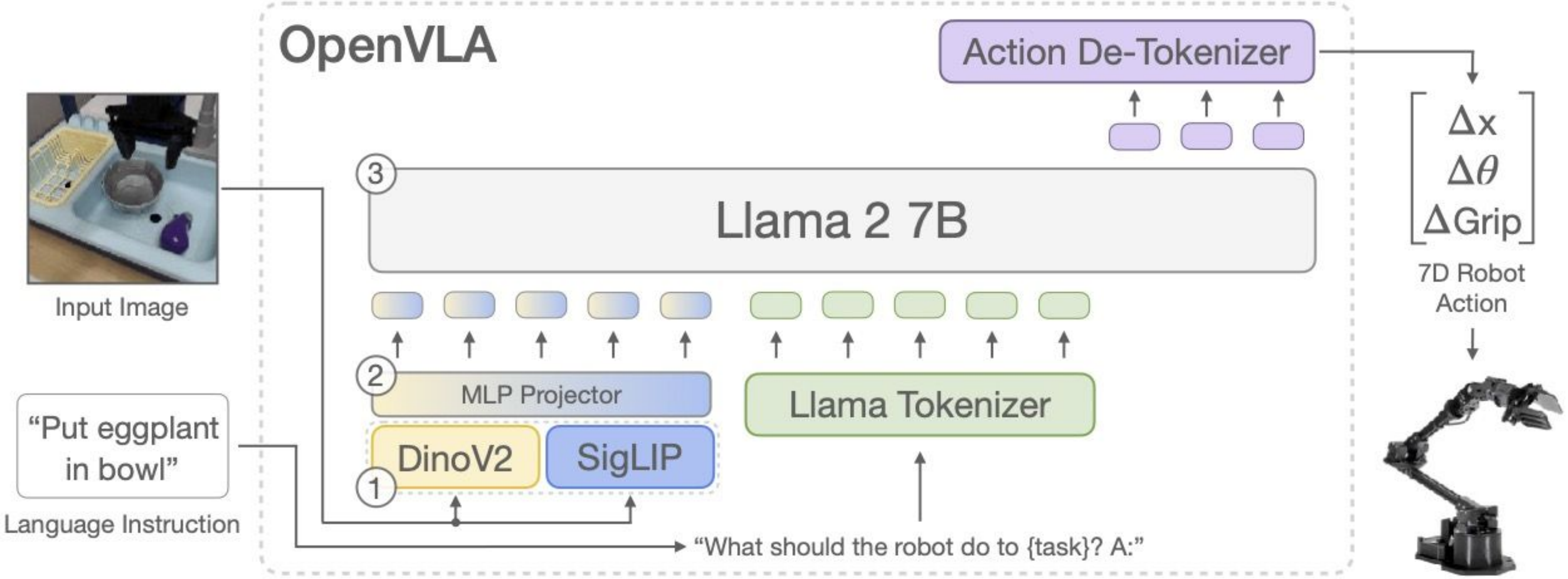


Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

Components:

1. **Fused Vision Encoder**: Consisting of a **SigLIP** and a **DinoV2** backbone, that maps image inputs to a number of image patch embeddings

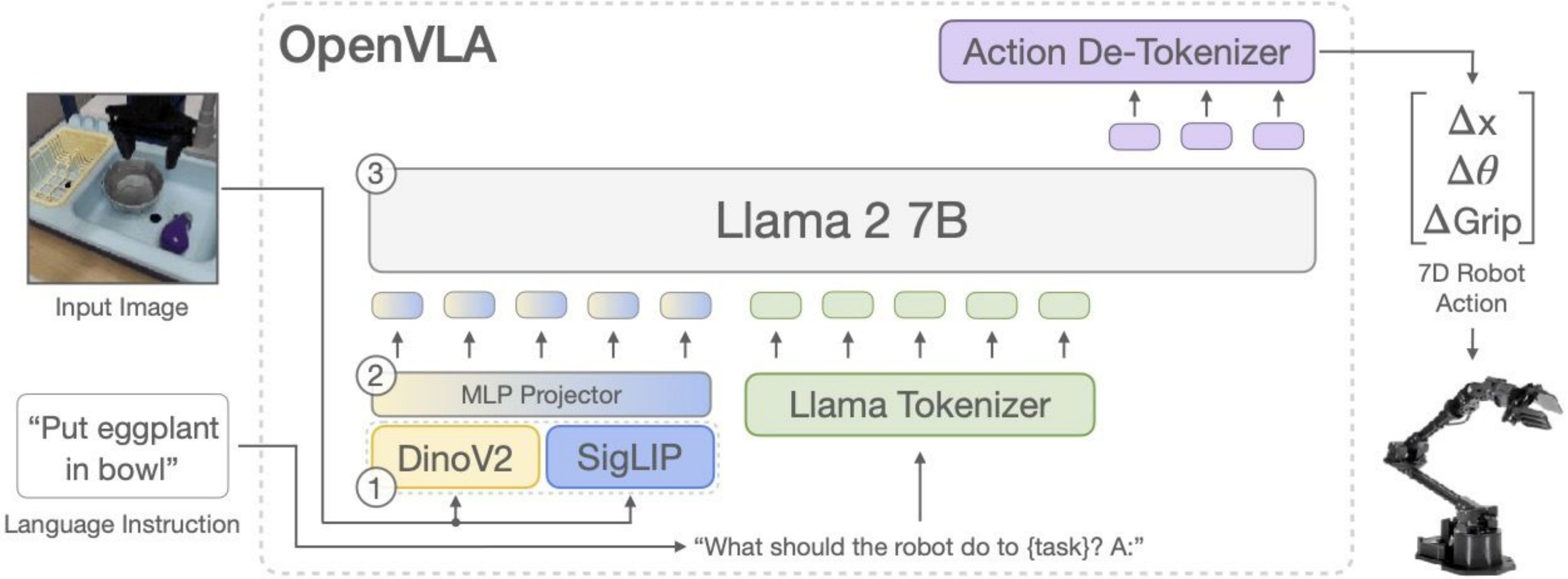


Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

Components:

1. **Fused Vision Encoder**: Consisting of a **SigLIP** and a **DinoV2** backbone, that maps image inputs to a number of image patch embeddings
2. **MLP Projector:** Takes the output embeddings of the fused visual encoder and maps them into the input space of a language model

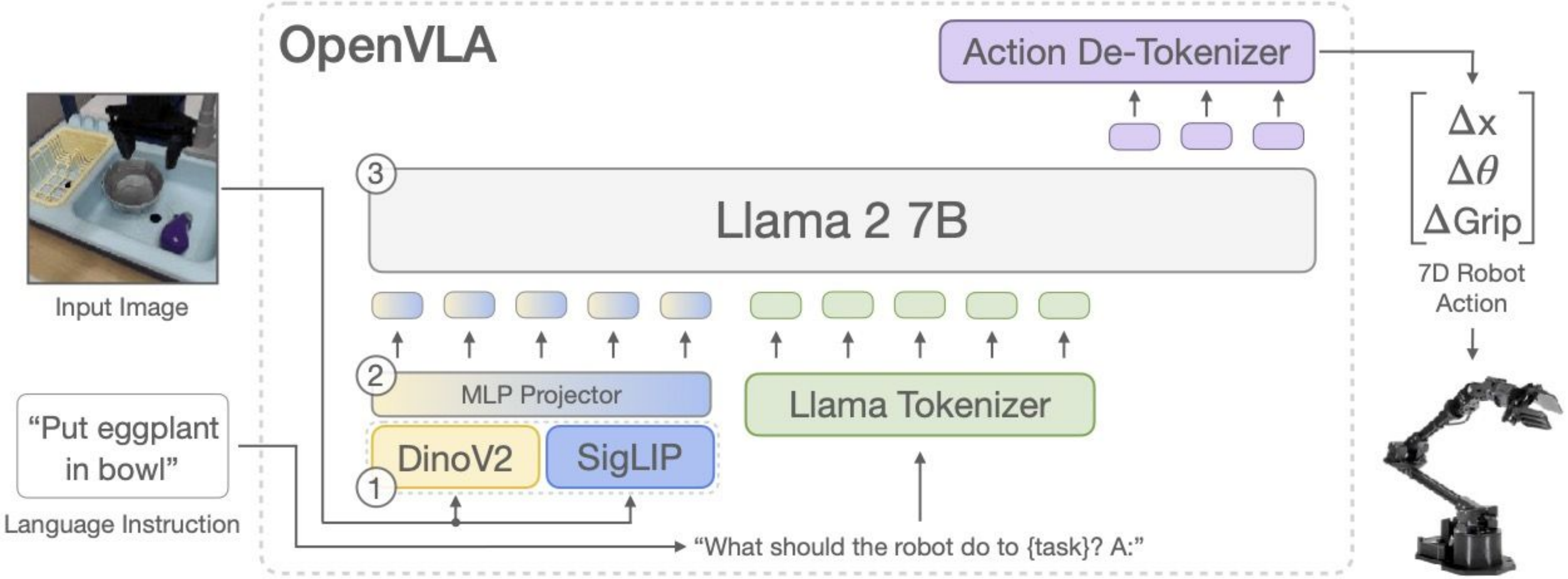


Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

Components:

1. **Fused Vision Encoder**: Consisting of a **SigLIP** and a **DinoV2** backbone, that maps image inputs to a number of image patch embeddings
2. **MLP Projector:** Takes the output embeddings of the fused visual encoder and maps them into the input space of a language model
3. **Llama 2 Language Model Backbone**: Predicts tokenized output actions.

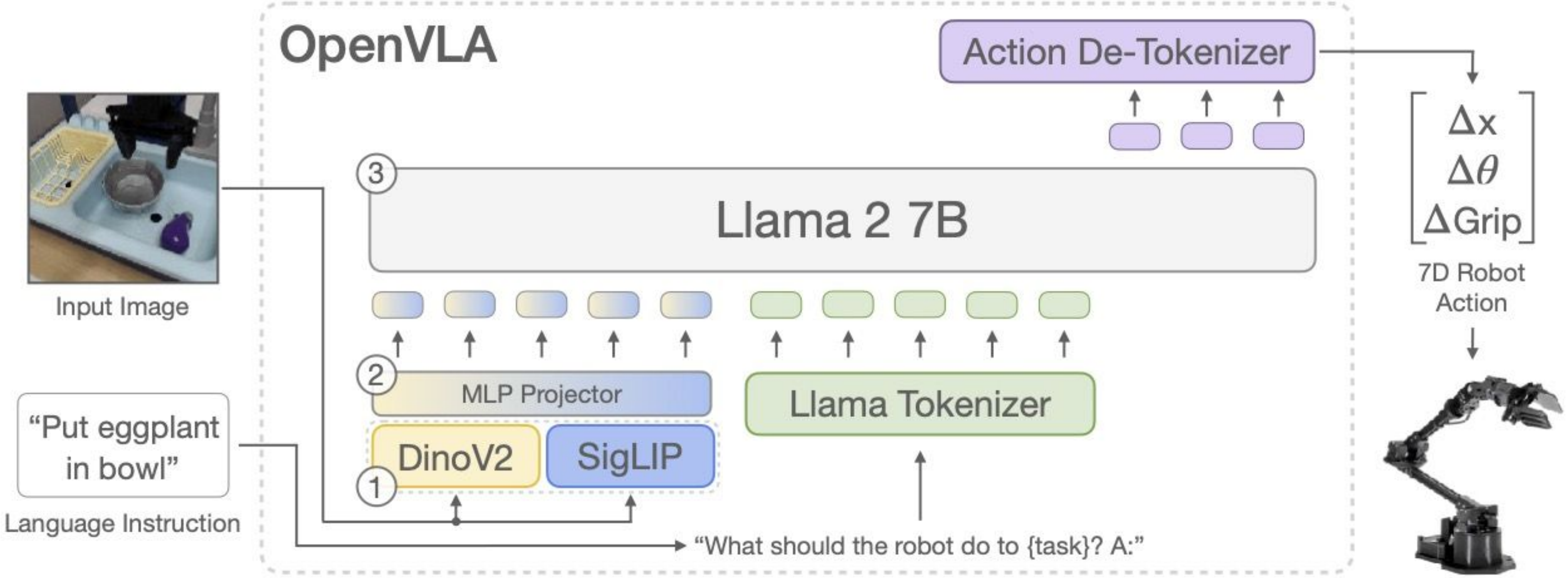


Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

Components:

1. **Fused Vision Encoder**: Consisting of a **SigLIP** and a **DinoV2** backbone, that maps image inputs to a number of image patch embeddings
2. **MLP Projector:** Takes the output embeddings of the fused visual encoder and maps them into the input space of a language model
3. **Llama 2 Language Model Backbone**: Predicts tokenized output actions.
4. **Action De-tokenizer**: Maps discrete tokens into robot control output.

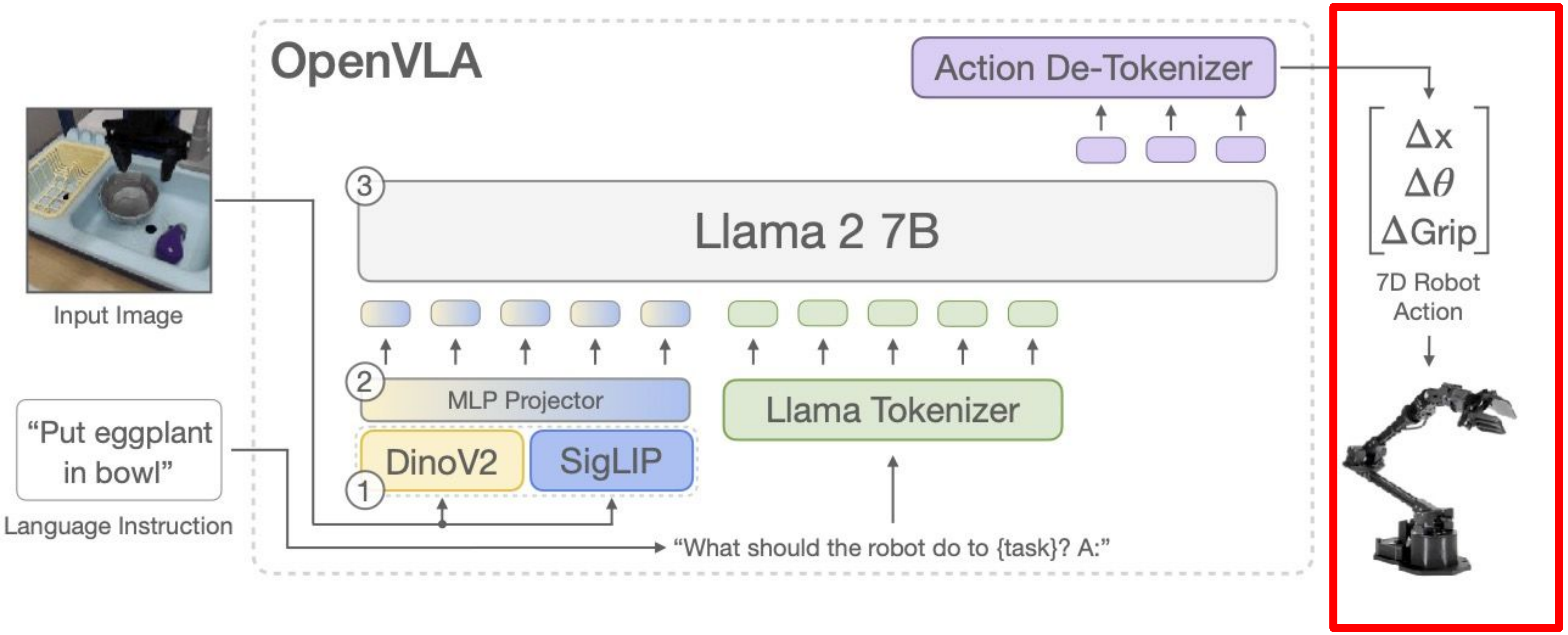


Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

OpenVLA

Output:

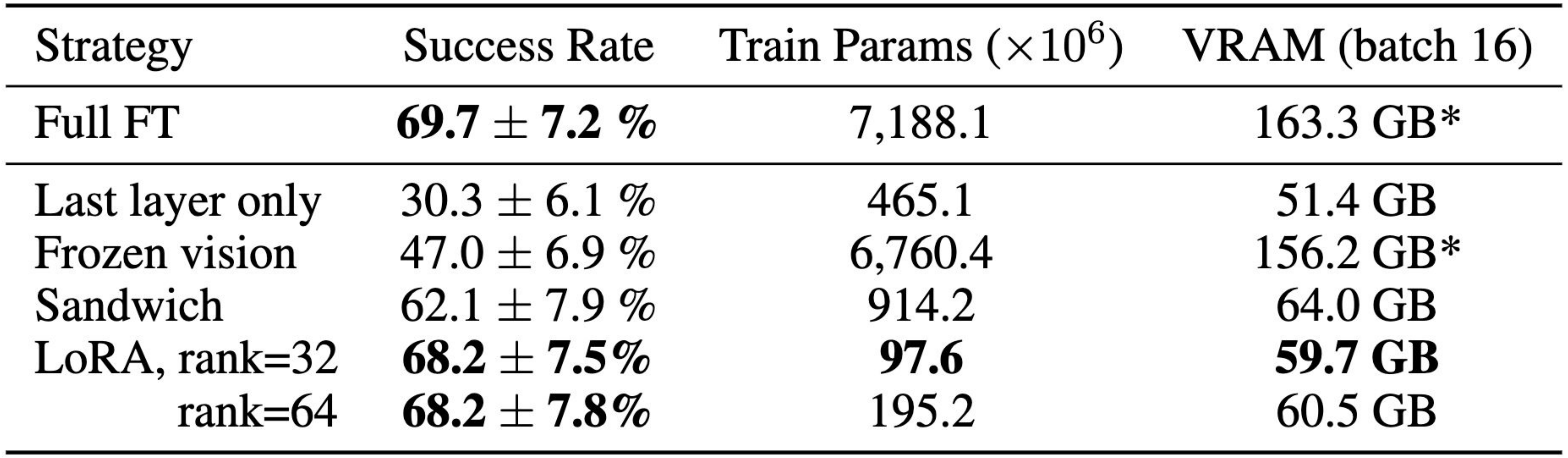
Produces robot control commands in a 7D format (e.g., x, y, z, rotation, grip).



Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Efficient Fine-Tuning

- **Techniques**:
  a. **Full Fine-Tuning**: Updates all parameters but is compute-intensive.
  b. **LoRA (Low-Rank Adaptation)**:
     i. Fine-tunes a subset of parameters.
     ii. Matches full fine-tuning performance while reducing memory and compute requirements (1.4% parameters tuned).
- **Inference**:
  a. Supports quantized inference (4-bit precision) for consumer GPUs, maintaining performance while reducing memory footprint.



Strategy	Success Rate	Train Params ($\times 10^6$)	VRAM (batch 16)
Full FT	**69.7 ± 7.2 %**	7,188.1	163.3 GB*
Last layer only	30.3 ± 6.1 %	465.1	51.4 GB
Frozen vision	47.0 ± 6.9 %	6,760.4	156.2 GB*
Sandwich	62.1 ± 7.9 %	914.2	64.0 GB
LoRA, rank=32	**68.2 ± 7.5%**	**97.6**	**59.7 GB**
rank=64	**68.2 ± 7.8%**	195.2	60.5 GB

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Performance Evaluation

Benchmarks:

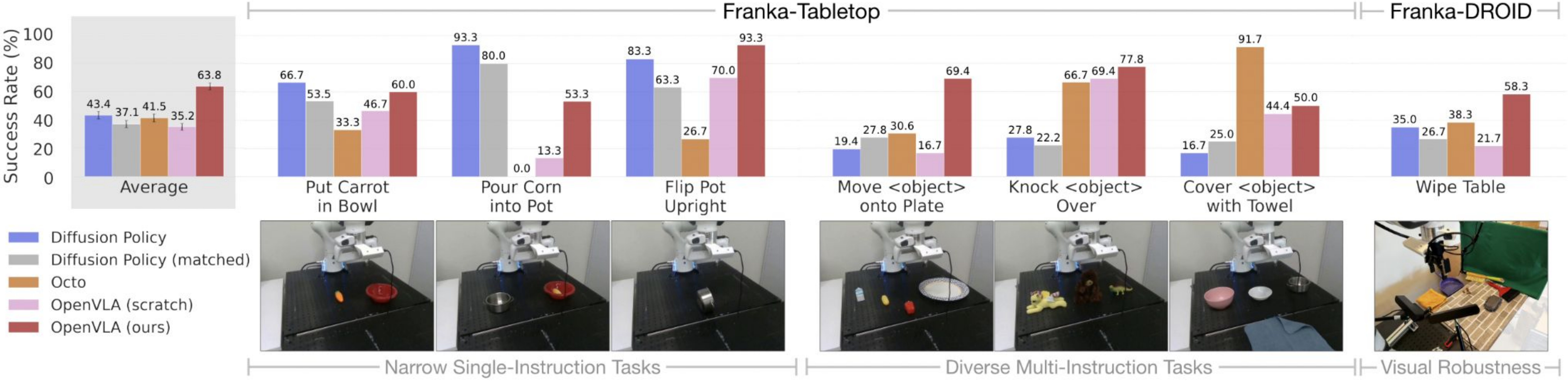
- Evaluated on tasks involving unseen objects, distractors, and language-specific instructions.
- Outperforms RT-2-X in multi-task environments, achieving a 16.5% higher success rate across 29 tasks.

Robot Platforms:

- Tested on WidowX and Google Robots.
- Adapts to new setups like the Franka-Tabletop robot with minimal fine-tuning data.

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Data-Efficient Adaptation to New Robot Setups



OpenVLA (scratch) Model - where we directly finetune the underlying base Prismatic VLM on the target robot setup.
OpenVLA (Ours) Model – where we finetune the OpenX-pretrained OpenVLA model on the target robot setup.

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Comparison with State-of-the-Art Models

RT-2-X:
Place Banana on Plate
(OOD: unseen target object & instruction)

OpenVLA:
Place Banana on Plate
(OOD: unseen target object & instruction)

Both RT-2-X (closed-source 55B-parameter model) and OpenVLA perform reliably on in-distribution and basic out-of-distribution (OOD) generalization tasks.

Ref : https://openvla.github.io/

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Comparison with State-of-the-Art Models

RT-2-X:
Move Coke Can near Taylor Swift
(OOD: unseen concept from Internet)

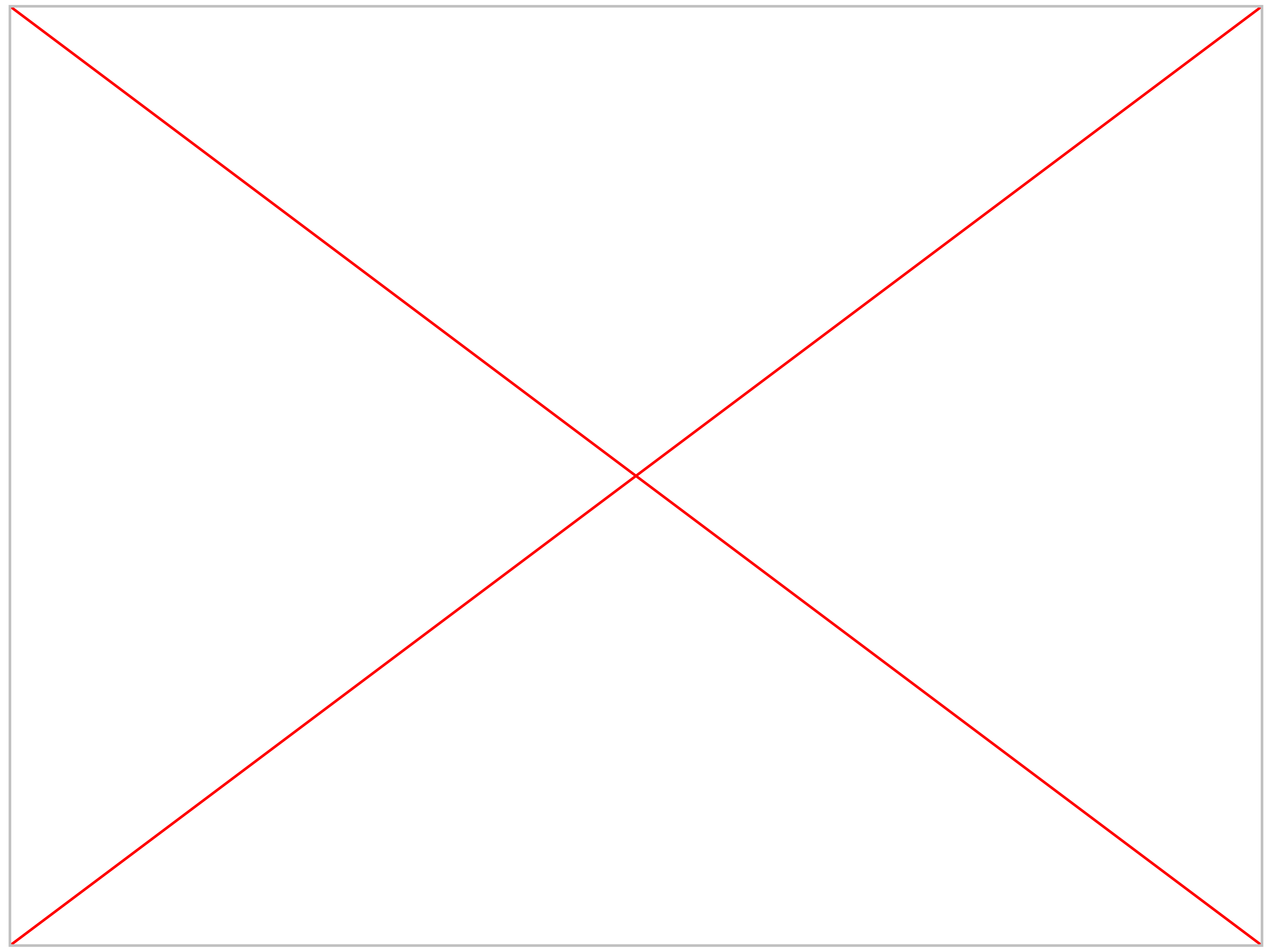
OpenVLA:
Move Coke Can near Taylor Swift
(OOD: unseen concept from Internet)

- However, RT-2-X performs better than OpenVLA on difficult semantic generalization tasks, i.e., tasks that require knowledge of concepts from the Internet that do not appear in the robot action training data, such as Taylor Swift in the video.

Ref : https://openvla.github.io/

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Comparison with State-of-the-Art Models



RT-2-X:
Move Coke Can near Taylor Swift
(OOD: unseen concept from Internet)

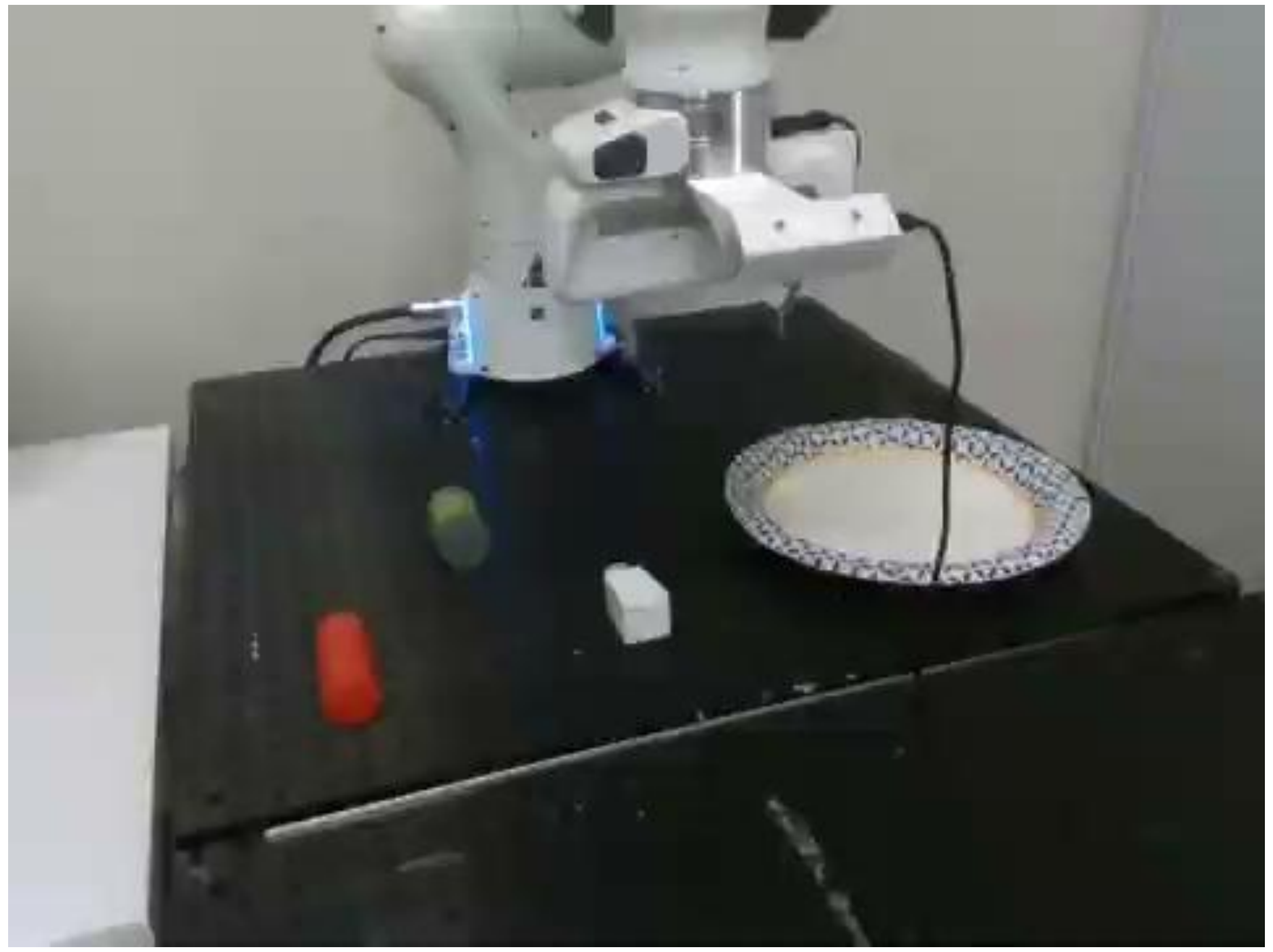
OpenVLA:
Move Coke Can near Taylor Swift
(OOD: unseen concept from Internet)

- However, RT-2-X performs better than OpenVLA on difficult semantic generalization tasks, i.e., tasks that require knowledge of concepts from the Internet that do not appear in the robot action training data, such as Taylor Swift in the video.
- This is expected given that RT-2-X uses larger-scale Internet pretraining data and is co-fine-tuned with both robot action data and Internet pretraining data to better preserve the pretraining knowledge (for OpenVLA, they fine-tune the pretrained vision-language model solely on robot action data for simplicity).

Ref : https://openvla.github.io/

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Comparison between OpenVLA(Scratch) and OpenVLA(Ours)



OpenVLA (Scratch):
Move White Salt Shaker onto Plate

OpenVLA (Ours):
Move White Salt Shaker onto Plate

OpenVLA(Ours) model exhibits much more reliable behaviors than OpenVLA (Scratch).

OpenVLA (scratch) Model - where we directly finetune the underlying base Prismatic VLM on the target robot setup.
OpenVLA (Ours) Model – where we finetune the OpenX-pretrained OpenVLA model on the target robot setup.

Ref : https://openvla.github.io/

Kim, Moo Jin, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov et al. "OpenVLA: An Open-Source Vision-Language-Action Model." *arXiv preprint arXiv:2406.09246* (2024).

Overview

Need for Robotic Foundation Models

- Challenges
- Characteristics of a Robotic Foundation model

Recent Attempts to Create Robotic Foundation Models

- Open X-Embodiment Dataset
- Foundation Models
  - Google DeepMind RT-1-X
  - Google DeepMind RT-2-X
  - OpenVLA
- Conclusion

Conclusion

- **OpenVLA: A Promising Choice**
  OpenVLA stands out as a highly efficient option among recent robotic foundation models due to its **high performance**, **smaller parameter size**, enabling local execution without heavy computational demands. This makes it an attractive solution for applications requiring portability and real-time processing.

- **Exploring Beyond the Present**
  While we focused on a few models, it’s important to recognize the broader landscape, including innovative models like **Crossformer** and cutting-edge foundation models under developments such as **NVIDIA Project GR00T**. These ongoing advancements continue to redefine the capabilities and accessibility of robotic foundation models.

https://crossformer-model.github.io/

https://developer.nvidia.com/project-gr00t

Doshi, Ria, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. "Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation." *arXiv preprint arXiv:2408.11812* (2024).

Ref : https://youtu.be/h0R5aumX_Uo?si=vgcHQ-OJ0VrT5vt5

Thank you!

Student Rating of Teaching

Please do them!

DeepRob

[Group 6] Lecture 8
Foundational Models and Robotics
by Chang, Franklin, Rohan
University of Minnesota

LLMs ❄

VLMs ❄

VLAs ❄

❄ Maybe Robot Model? 🔥