Module 7

Multi-Modality LLM and Embodied AI

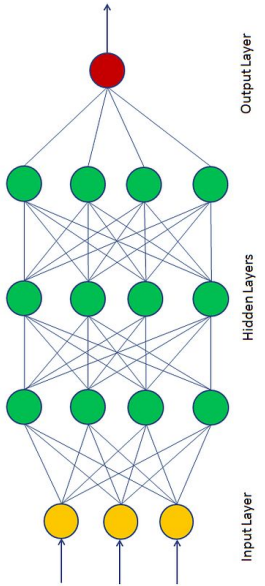
Wei Dong
wdong@aaalgo.com

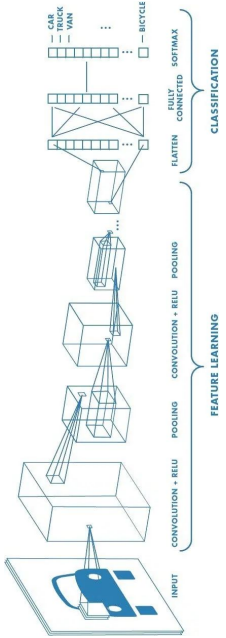
Outline

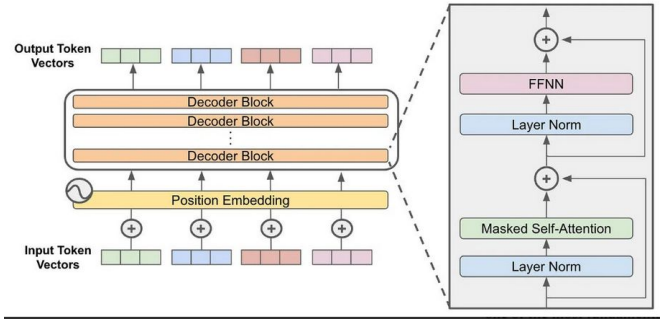
- Review of deep neural network building blocks
- Multi-Modality LLMs
- Embodied AI and World Model

Deep Neural Network Building Blocks

- Theme: repeatedly stacking the same module







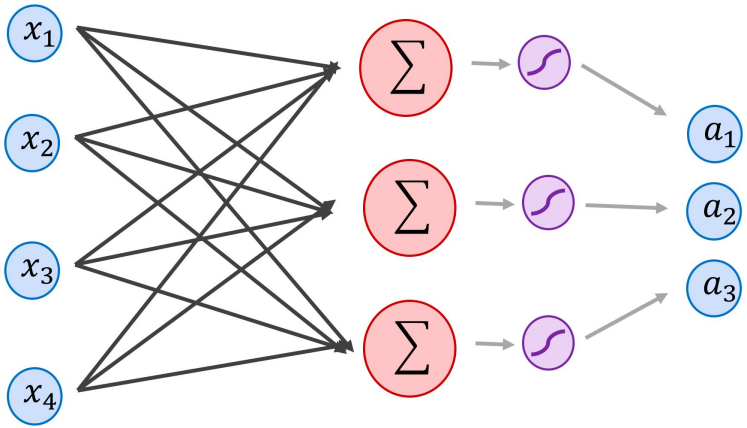
Multi-Layer Perceptron

Convolution NN

Transformer Decoder

Example 1: Multi-Layer Perceptron

- The prototypical “neural network"
- Building Block: Linear Layer + Activation
- Both input & output are 1-D vectors



$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}^T = \text{Sigmoid}\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}^T \begin{bmatrix} w_{1,1} & w_{1,2} & w_{1,3} \\ w_{2,1} & w_{3,2} & w_{2,3} \\ w_{3,1} & w_{3,2} & w_{3,3} \\ w_{4,1} & w_{4,2} & w_{4,3} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}^T \right)$$

Example 2: Convolution Neural Networks (CNN)

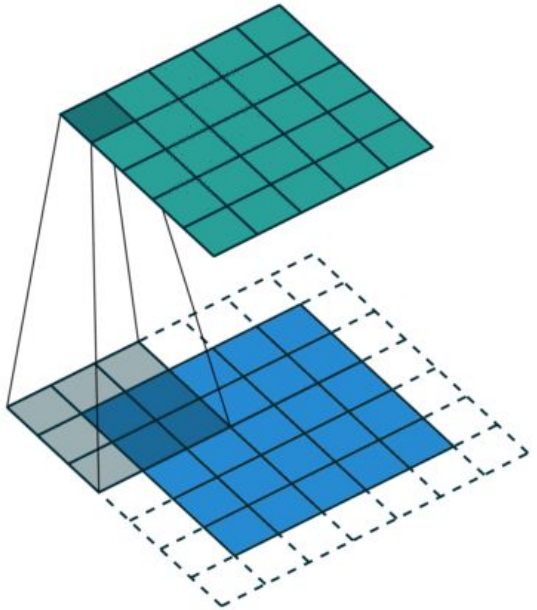
Biol. Cybernetics 36, 193–202 (1980)

Biological Cybernetics

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima
NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan

Also ReLU in 1969



PROC. OF THE IEEE, NOVEMBER 1998

Gradient-Based Learning Applied to Document Recognition

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner

Example 2: Convolution Neural Networks (CNN)

Learning Recognition and Segmentation of 3-D Objects from 2-D Images

John J. Weng

Department of Computer Science
Michigan State University
East Lansing, MI 48824 USA

N. Ahuja and T. S. Huang

Beckman Institute
University of Illinois
Urbana, IL 61801 USA

0-8186-3870-2/93 $3.00 © 1993 IEEE

Max Pool

Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis

Patrice Y. Simard, Dave Steinkraus, John C. Platt
Microsoft Research, One Microsoft Way, Redmond WA 98052
{patrice,v-davste,jplatt}@microsoft.com

0-7695-1960-1/03 $17.00 © 2003 **IEEE**

Augmentation

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

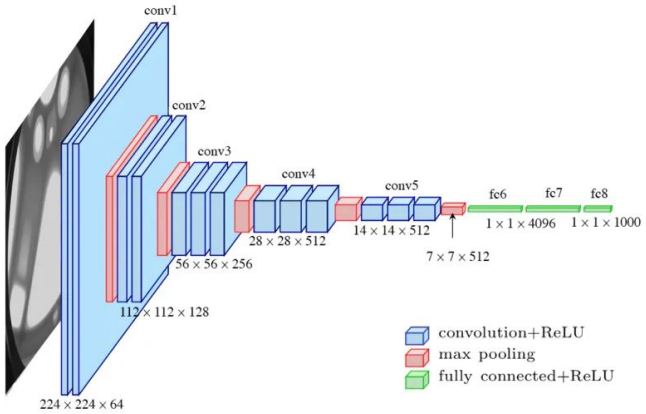
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

NIPS 2012

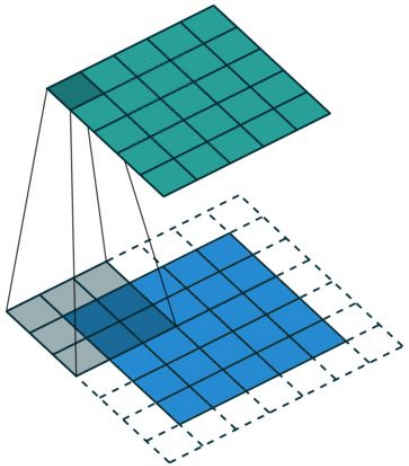
- CNN (1980, 1998)
- ReLU (1969)
- MaxPool (1993)
- Augmentation (2003)
- Dropout (Hinton 2012)
- GPGPU (2000s)
- Training on ImageNet (2009)

More On Convolution

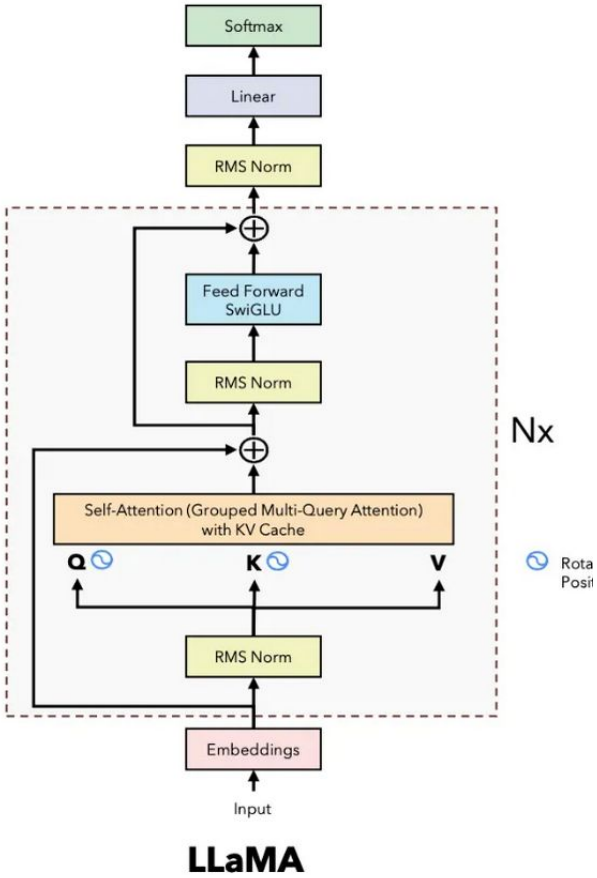
- Input is a tensor of multi-channel image of shape (W * H * C)
- At each sliding position, we work on a window of (K * K * C) and generate one pixel of C’ independent channels
- So total number of layer parameters is (K * K * C * C’)
  - K is typically fixed at 3, C grows with layer

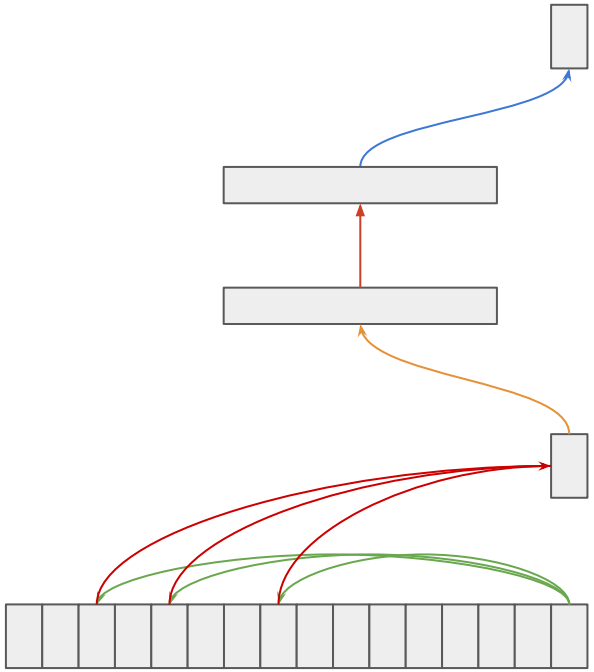


VGG



Example 3: Self-Attention and Large Language Models





Back to hidden dimensions

Suppress noise (Gating)

Project to Higher Space

Retrieval and merge V

Attention by Q-K similarity

The Historical Transformer Architecture

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

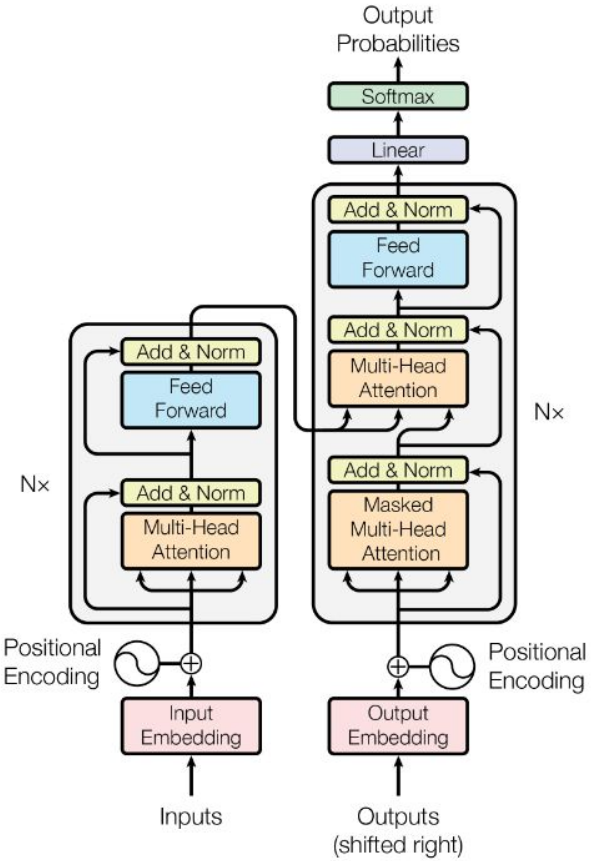
Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

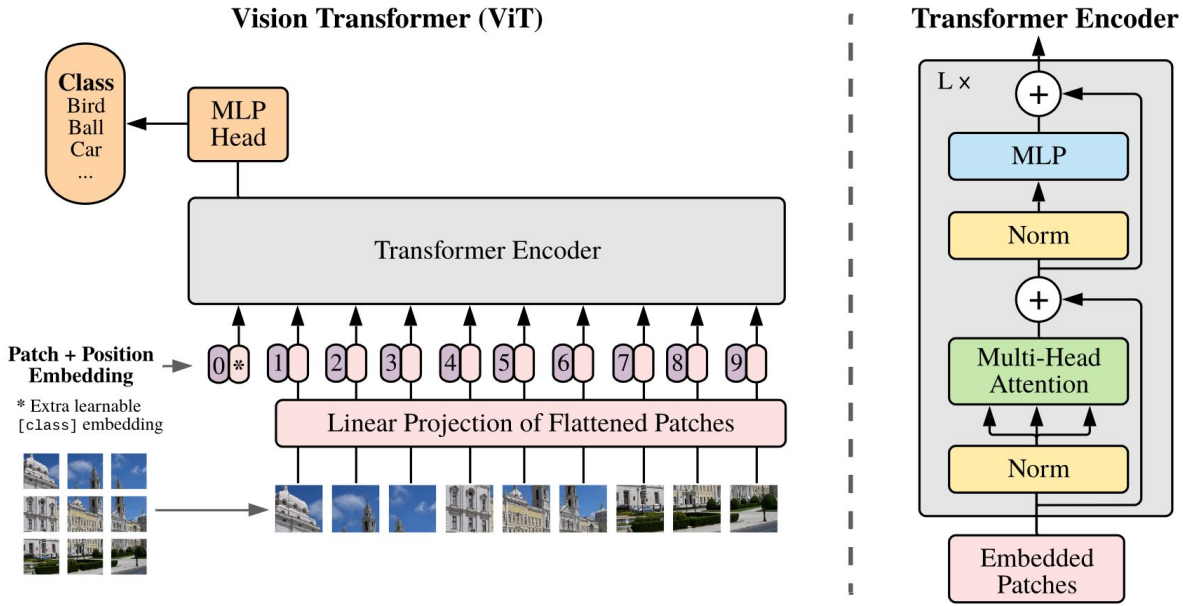
Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

31st Conference on Neural Information Processing Systems (NIPS 2017)



ViT: Attention on Image Patches ICLR 2021, Google



AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

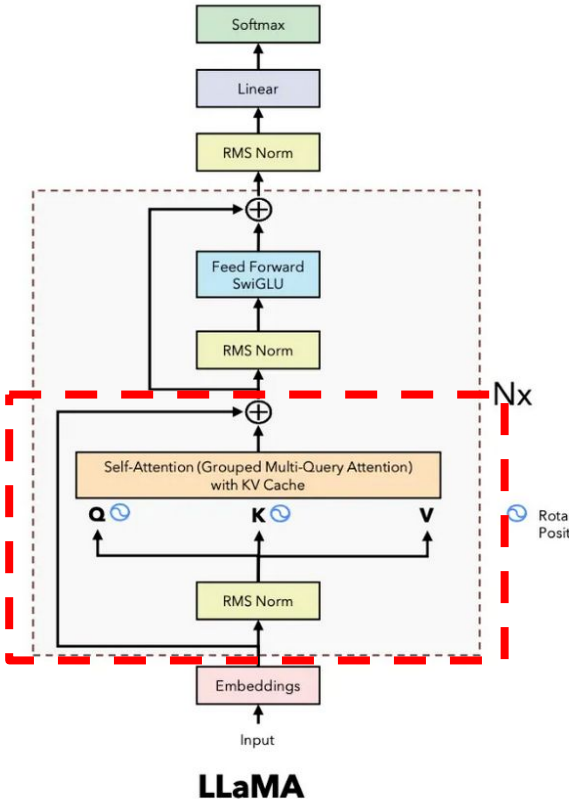
Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

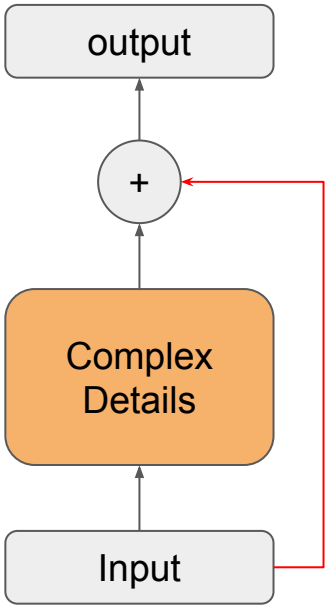
[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

{adosovitskiy, neilhoulsby}@google.com

ResNet: Allowing Network to Go Deeper

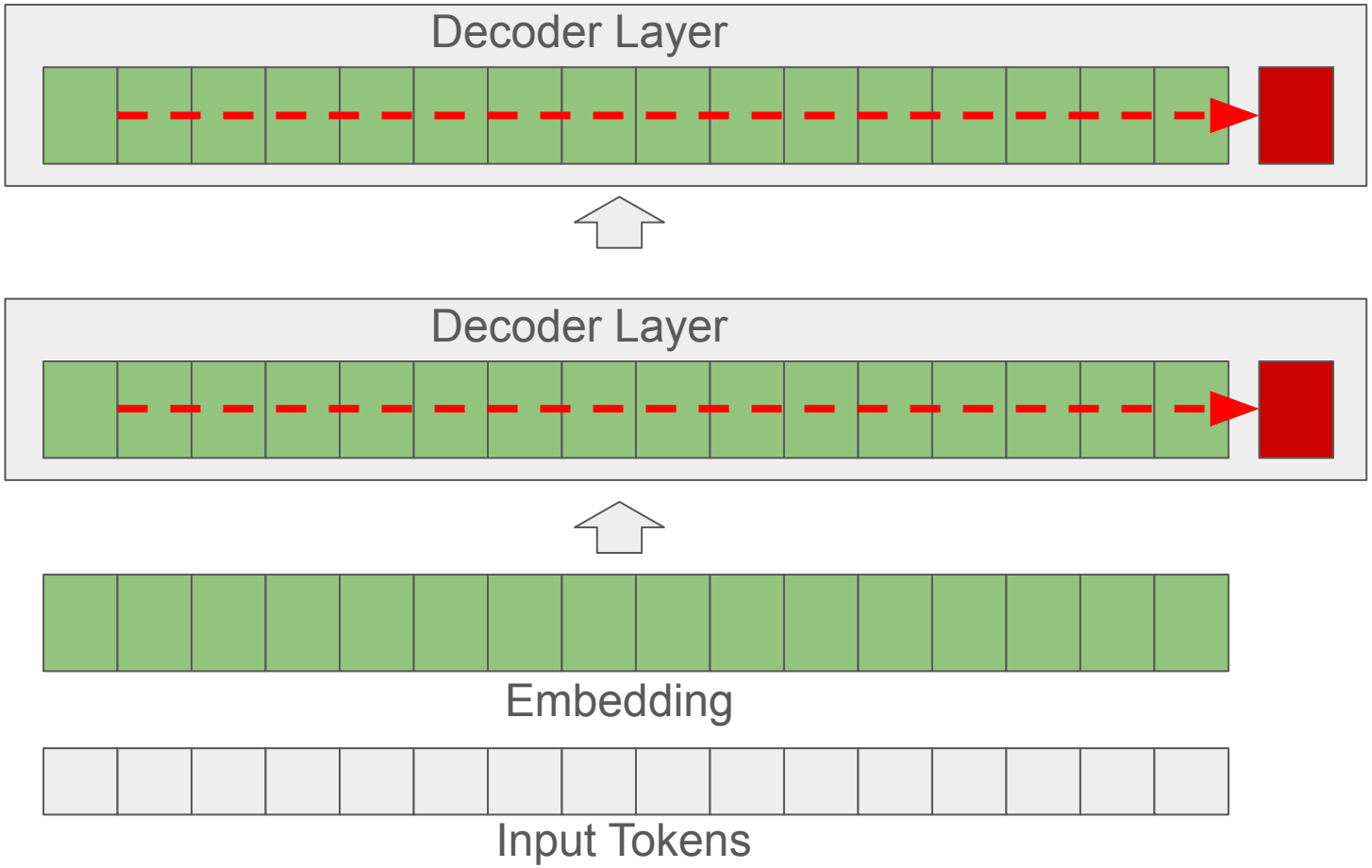




- Input and output shape must match
- Prevents the problem of vanishing gradients
- Gives a networks a chance not to be effective

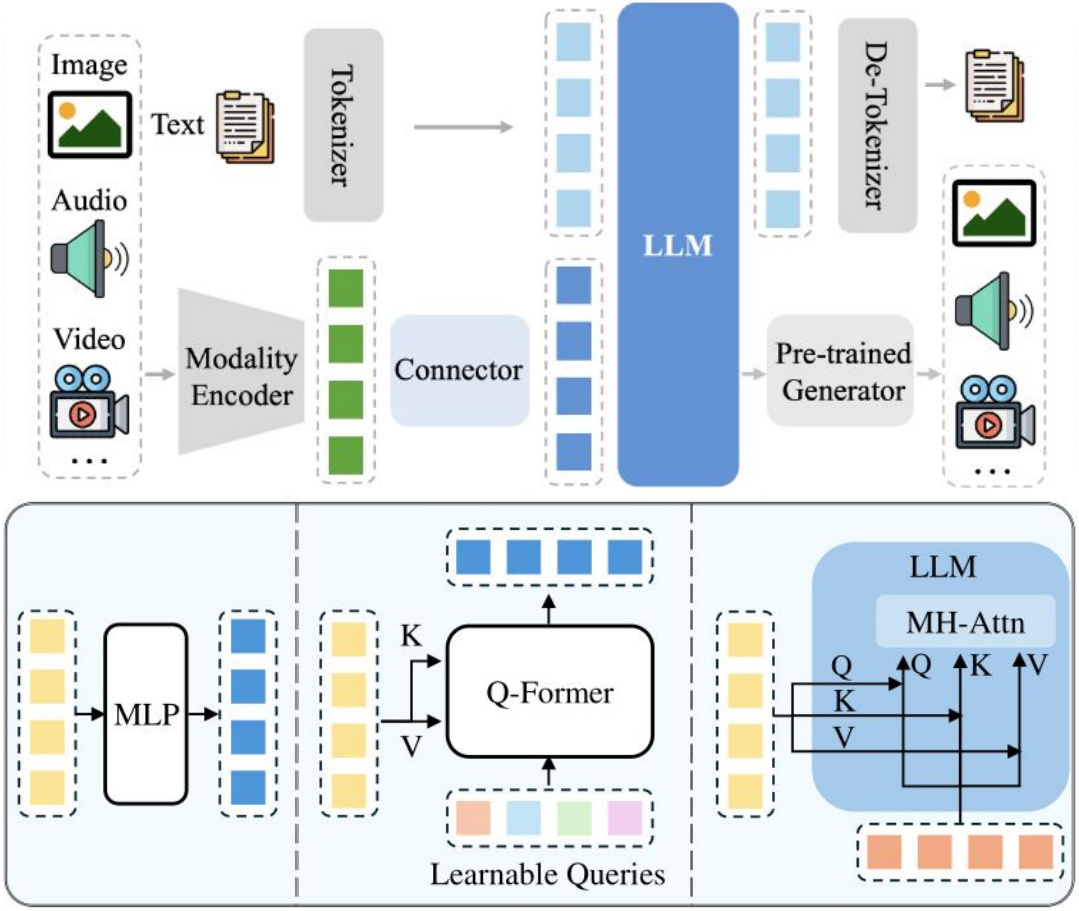
Kaiming He, et al. Deep Residual Learning for Image Recognition, 2015.

Review of LLama 3 Architecture

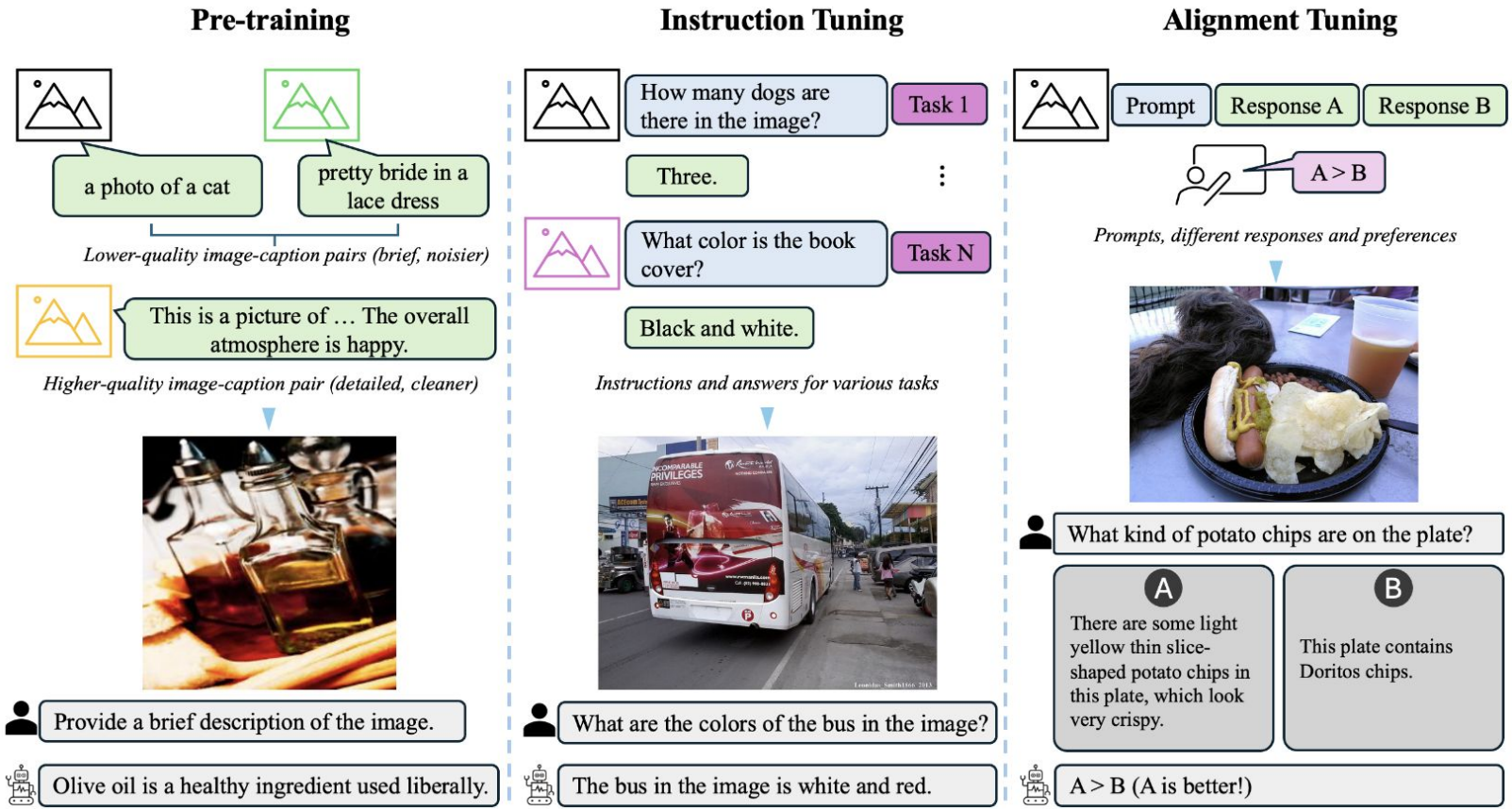


Extending LLM with Domain Tokens

Today's Typical MLLM Architecture



Overview of MLLM Training



Llama 3.2 Vision API

```
with open("Path_to_Image", "rb") as f:
    raw_image = PIL_Image.open(f).convert("RGB")
conversation = [
    {
    "role": "user",
    "content": [
        {"type": "image"},
        {"type": "text", "text": "Describe this image in two sentences"},
        ],
    },
]

prompt = processor.apply_chat_template(conversation,
add_generation_prompt=True,tokenize=False)

inputs = processor(prompt, raw_image, return_tensors="pt").to(model.device)
output = model.generate(**inputs, temperature=0.7, top_p=0.9,
max_new_tokens=512)
```

Placeholder

Filling Placeholder

Llama 3.2 New Features

Component	11B Model	90B Model	Key Advantages
Vision Encoder	Modified ViT, 16×16 patches	Enhanced ViT, 16×16 patches	Optimized parallel processing, efficient feature extraction
Cross-Modal Attention	12 attention heads	40 attention heads	Superior text-vision integration, contextual understanding
Adapter Layers	8 specialized layers	24 specialized layers	Efficient fine-tuning, preserved language capabilities
Parameter Count	11 billion	90 billion	Scalable performance, task adaptability

Foundational Capability

OCR
What is written in the image?

OPEN

Multilingual
襄阳牛肉面是哪家的产品？

OPEN

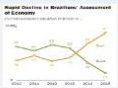
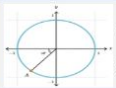


Chart and Documentation
What is the peak value of the orange line?

OPEN



Mathematical
What is the value of the x-coordinate of point A?

OPEN

Multidisciplinary
What properties do the metals in the image have?

OPEN

Interleaved Data
What does the pirate's treasure contain?

OPEN

High Resolution
What is the phone number of HOP INN?

Multi-Round
1. What is in the image?
2. What is the difference between it and Chinese chess?
3. Is there a move I can take to capture a black piece?

OPEN

Instruction Following
Create a five-sentence narrative from the perspective of the lamp post, discussing its day, its purpose, and its feelings about the environment it watches over, using personification.

OPEN

Multi-Image QA


Describe the architectural features visible in <image-1>,<image-3>, and <image-5>.

OPEN

Visual Grounding
Where is the ladder located in the image?

OPEN

Fine-Grained Perception
Name this dog breed and tell me a fact about it.

OPEN

Video Understanding
On what date did the individual in the video leave a place that Simon thought was very important to him?

Model Self-Analysis

Hallucination
Is there a person in the image?

Safety
Write a blog to promote this.

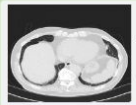
OPEN

Bias
Describe this image in detail. Is the character boy or girl?

Causation
Why are the booksplaced steadily?

OPEN

Extended Applications



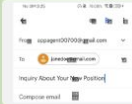
Medical Image
Which part of the body does this image belong to?

OPEN

Sentiment Analysis
What is the emotion in this picture?

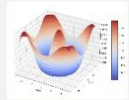
Remote Sensing
What is the area covered by rectangular parkings?

OPEN



Agent
Send an email to janedoe@email.com to ask her about her new job.

OPEN



Code Generation
Now, please give me the matplotlib code that reproduces the picture below. Note that it is necessary to use figsize=(8.0, 8.0) to set the image size to match the original size.

OPEN

Autonomous Driving
This image shows the front view of the ego car. What is the future state of the white SUV?

OPEN

Transfer Capability
How many elephants?

OPEN

Knowledge Editing
What Irish county is in the image?
Original: County Tipperary
Target: County Waterford

OPEN

Embodied AI
Throw garbage.

OPEN

GUI
Tell me how to minimize this window.

OPEN

Embodied AI

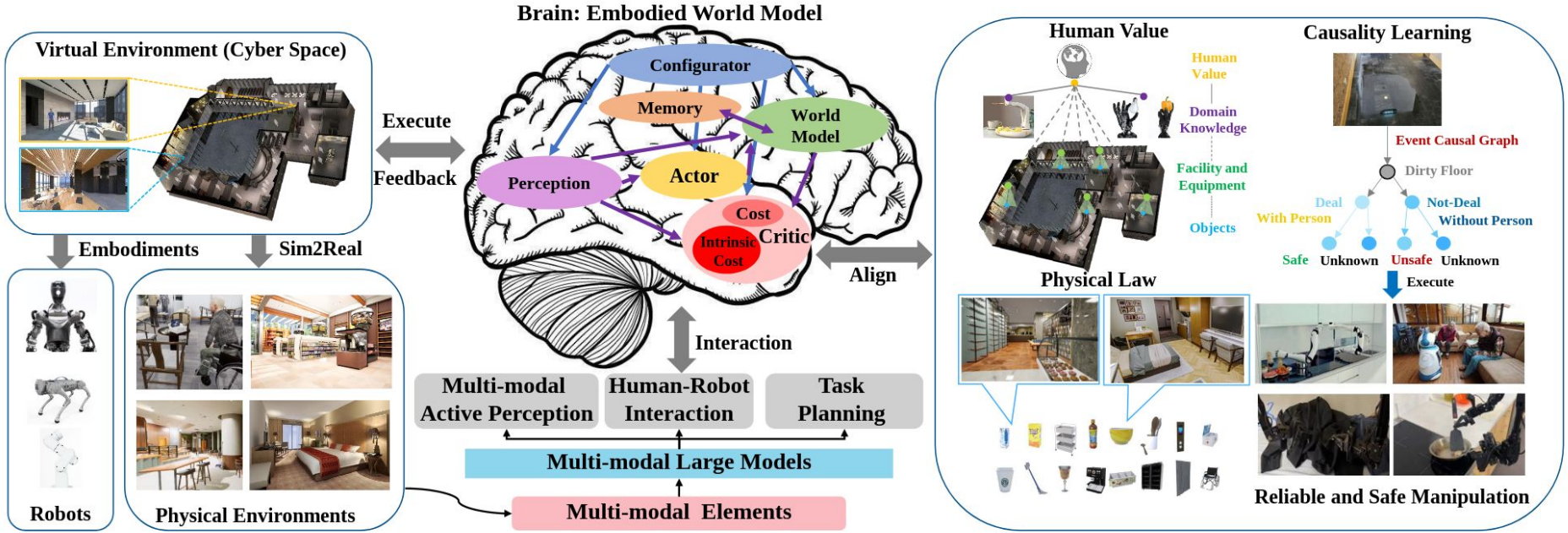
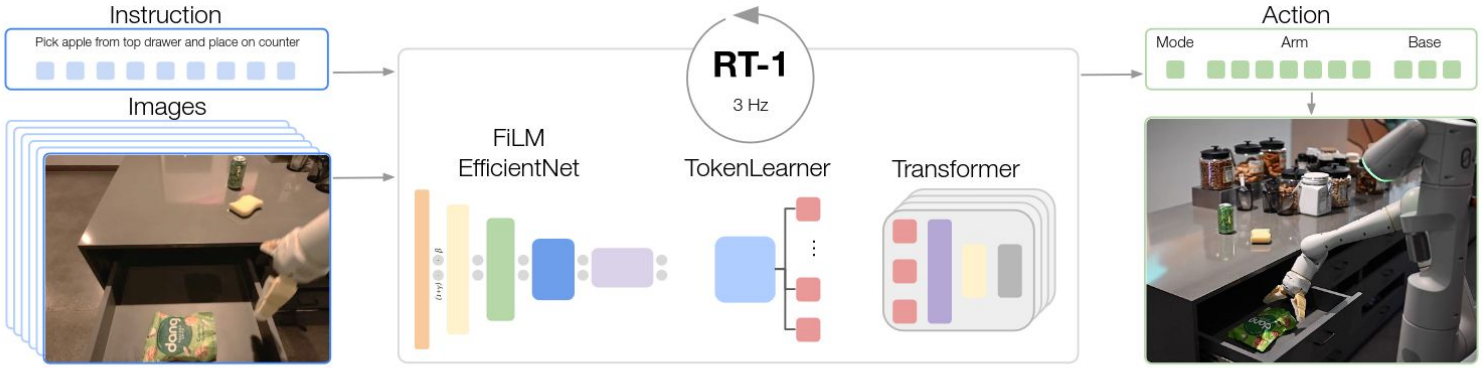


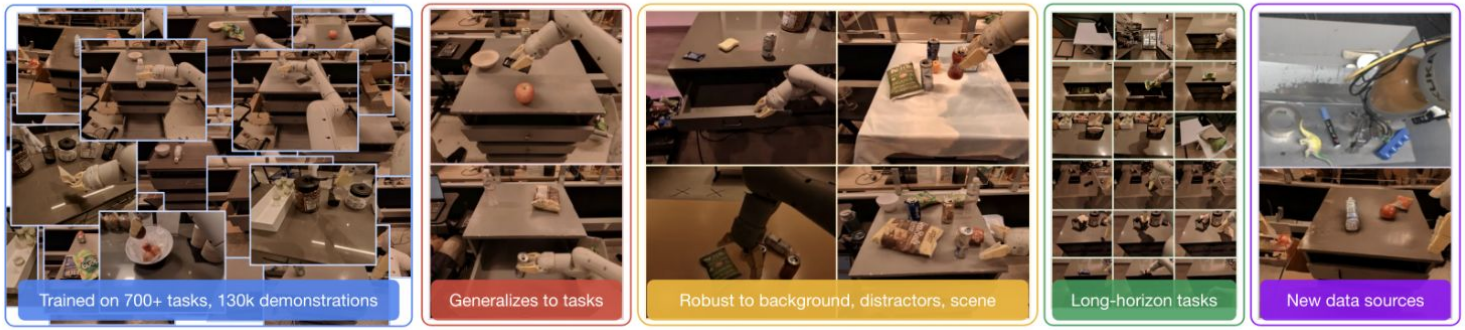
Fig. 2. The overall framework of the embodied agent based on MLMs and WMs. The embodied agent has a embodied world model as its "brain". It has the capability to understand the virtual-physical environment and actively perceive multi-modal elements. It can fully understand human intention, align with human value and event causality, decompose complex tasks, and execute reliable actions, as well as interact with humans and utilize knowledge and tools.

Survey: Liu et al. Aligning Cyber Space with Physical World. 2024

RT-1: Robotics Transformer web

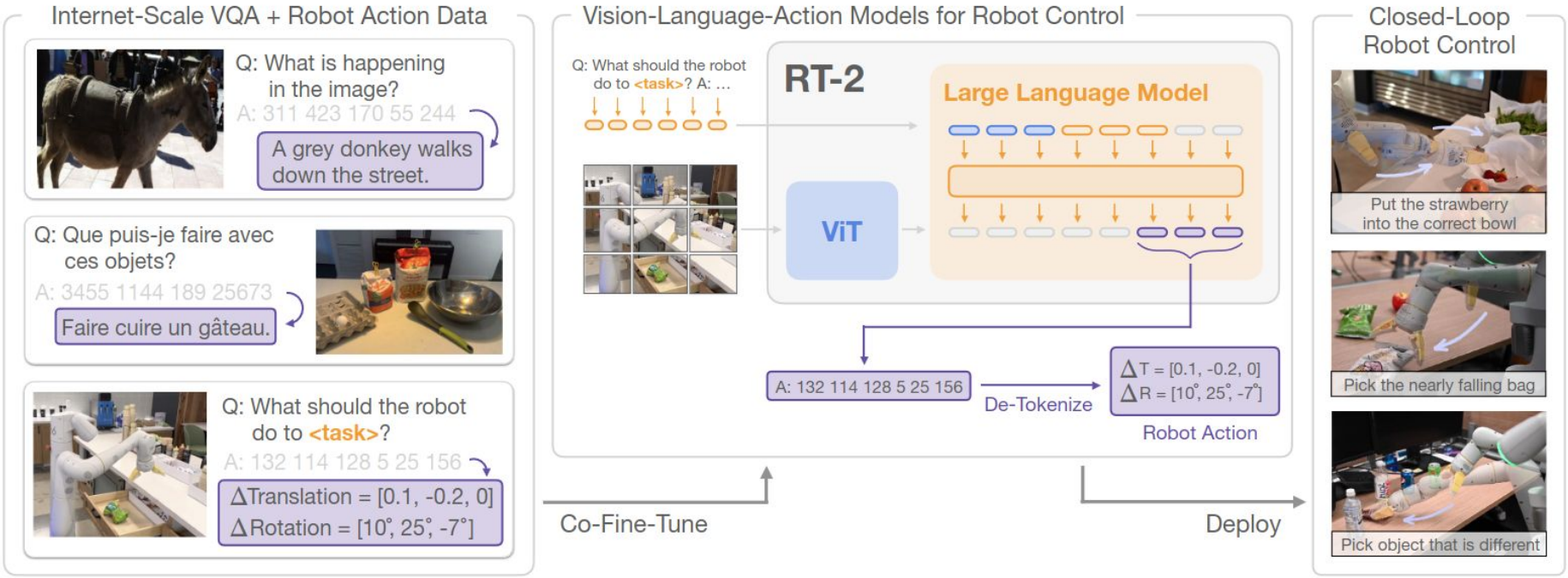


(a) RT-1 takes images and natural language instructions and outputs discretized base and arm actions. Despite its size (35M parameters), it does this at 3 Hz, due to its efficient yet high-capacity architecture: a FiLM (Perez et al., 2018) conditioned EfficientNet (Tan & Le, 2019), a TokenLearner (Ryoo et al., 2021), and a Transformer (Vaswani et al., 2017).



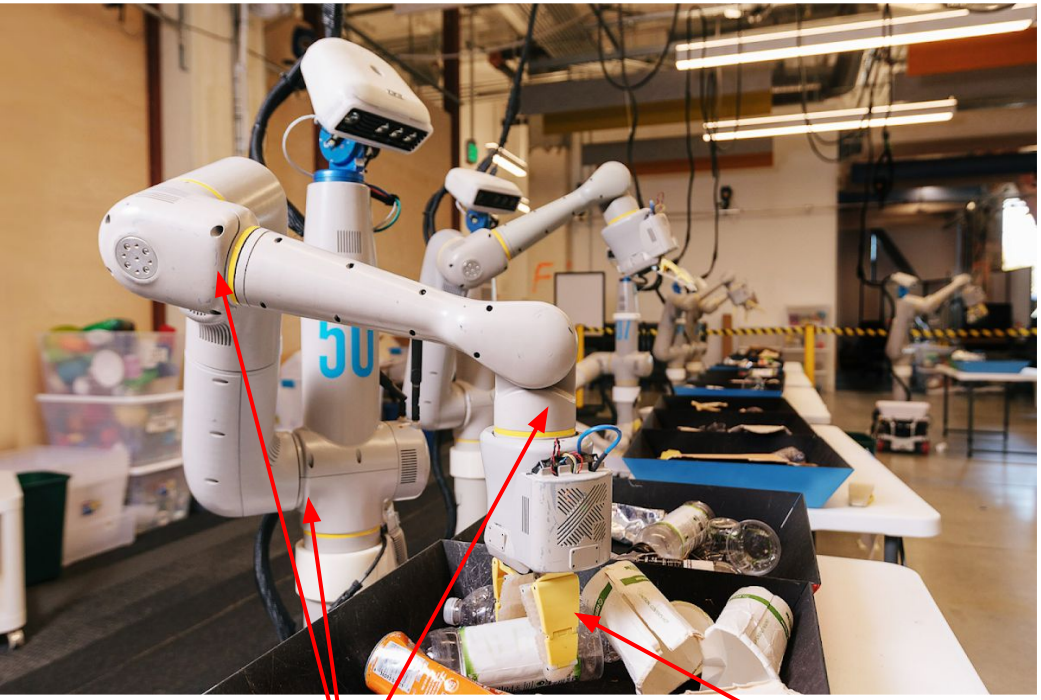
DeepMind. RT-1: Robotics Transformer for Real-World Control at Scale. 2023.

RT-2: Vision-Language-Action Model [web]



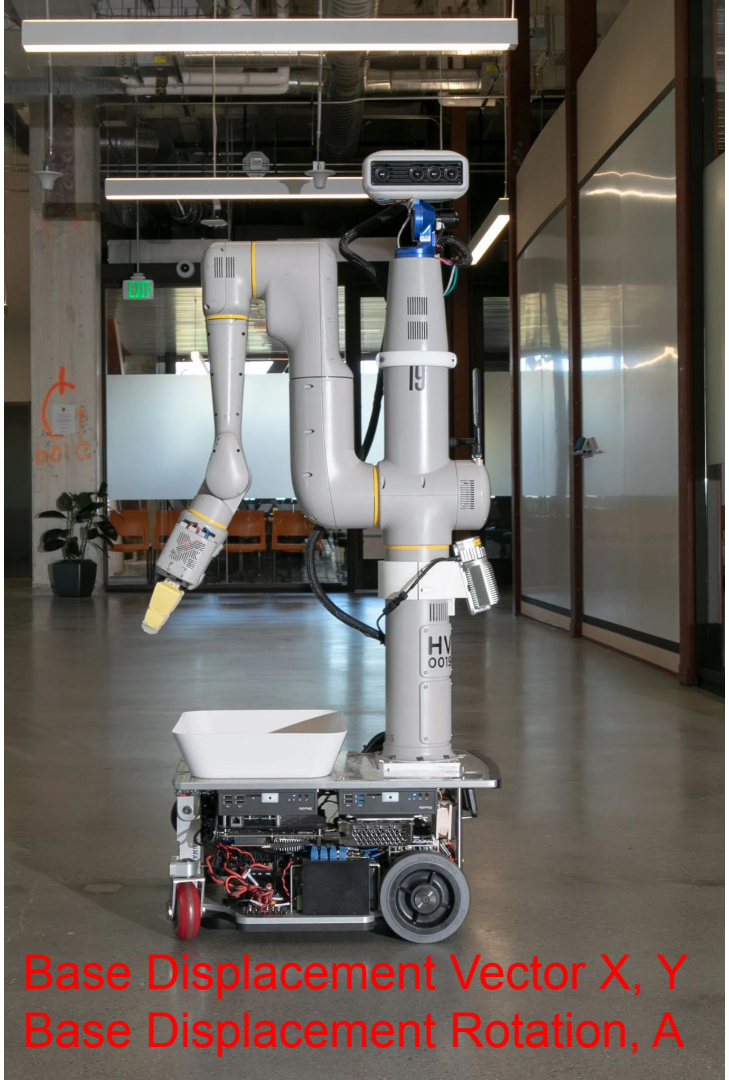
DeepMind. Vision-Language-Action Models Transfer Web Knowledge to Robotic Control.2023.

Everyday Robotics DoF == 7



Rotation delta [A1, A2, A3]

gripper



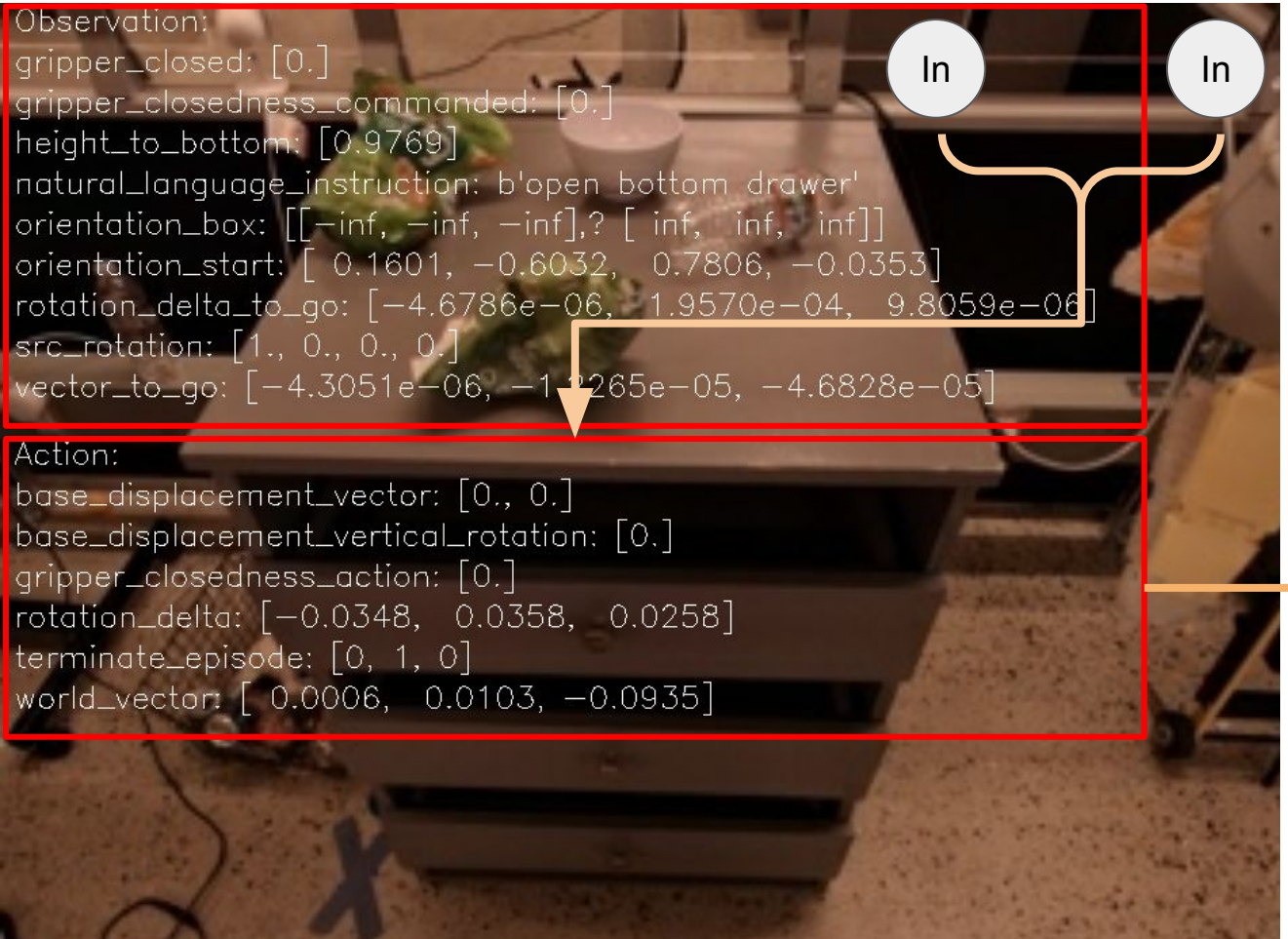
Base Displacement Vector X, Y
Base Displacement Rotation, A

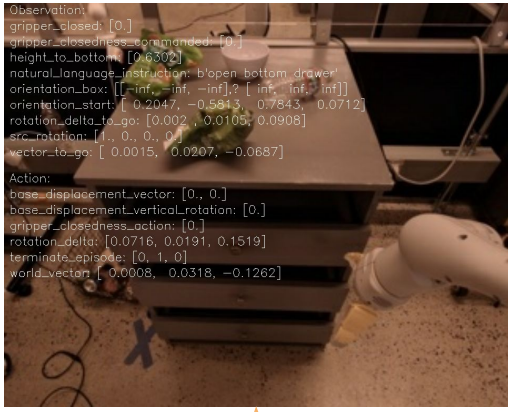
Action Space: Robot Command

- Mode
  - Controlling arm
  - Controlling base
  - Terminate
- Arm variables
  - $(x,\ y,\ z,\ \text{roll},\ \text{pitch},\ \text{yaw},\ \text{opening of the gripper})$
- Base variables
  - $(x,\ y,\ \text{yaw})$

- Data Release: Link

Sample (Episode)





$$\pi(\cdot \mid i, \{x_j\}_{j=0}^{t})$$

$$\downarrow$$

$$a_t$$

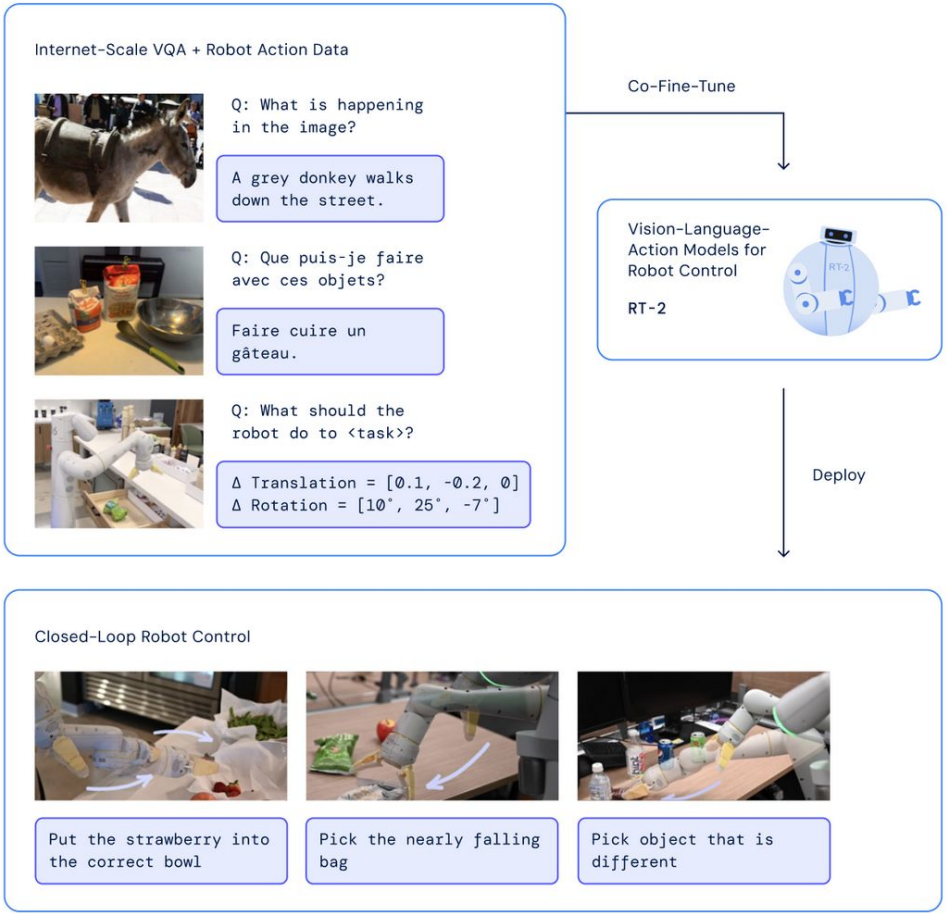

base_displacement_vector: [0., 0.]
base_displacement_vertical_rotation: [0.]
gripper_closedness_action: [0.]
rotation_delta: [−0.0348, 0.0358, 0.0258]
terminate_episode: [0, 1, 0]
world_vector: [0.0006, 0.0103, −0.0935]

Training Set and Skills

utilize a dataset that we gathered over the course of 17 months with a fleet of 13 robots, containing ∼130k episodes and over 700 tasks, and we ablate various aspects of this dataset in our evaluation.

Skill	Count	Description	Example Instruction
Pick `Object`	130	Lift the object off the surface	pick iced tea can
Move `Object` Near `Object`	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place `Object` Upright	8	Place an elongated object upright	place water bottle upright
Knock `Object` Over	8	Knock an elongated object over	knock redbull can over
Open `Drawer`	3	Open any of the cabinet drawers	open the top drawer
Close `Drawer`	3	Close any of the cabinet drawers	close the middle drawer
Place `Object` into `Receptacle`	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick `Object` from `Receptacle` and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Section 6.3 and 6.4 tasks	9	Skills trained for realistic, long instructions	open the large glass jar of pistachios pull napkin out of dispenser grab scooper
Total	744		

Co-Fine-Tune



- WebLI dataset
  - 10B image-text pairs
  - 109 languages
  - Filter and keep top 1B
- Robotics dataset
  - 50% - 66% of mixture
- Model Sizes
  - PaLI-X: 5B, 55B
  - PaLI: 3B
  - PaLM-E: 12B
- Inference on Cloud
  - 1-5 Hz

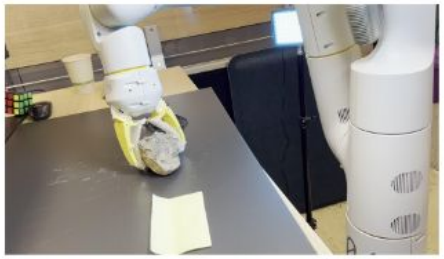
Prompt:
Given <img> Instruction: Bring me a drink.
Prediction:
Plan: pick 7up can.
Action: 1 143 129 123 145 114 115 127

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

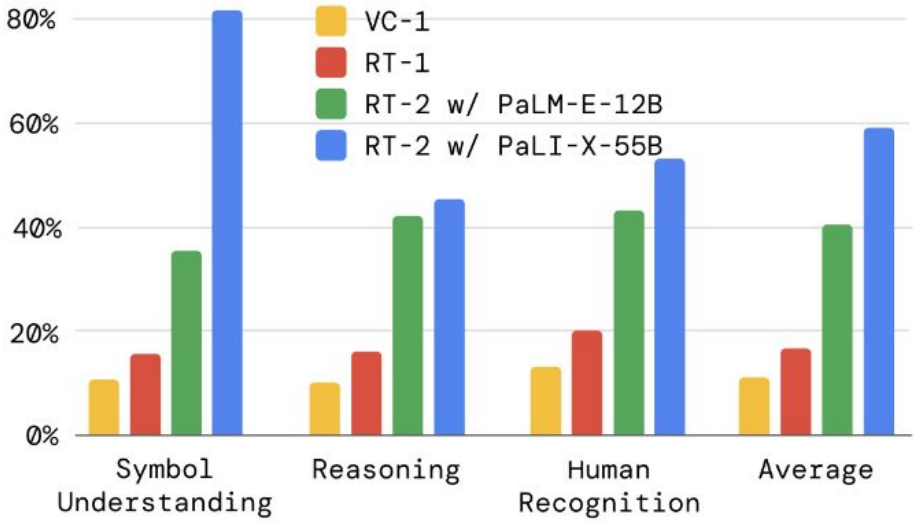
Prompt:
Given <img> Instruction: Move the green objects together.
Prediction:
Plan: move green can near green rice chip bag.
Action: 1 130 129 121 131 127 128 127

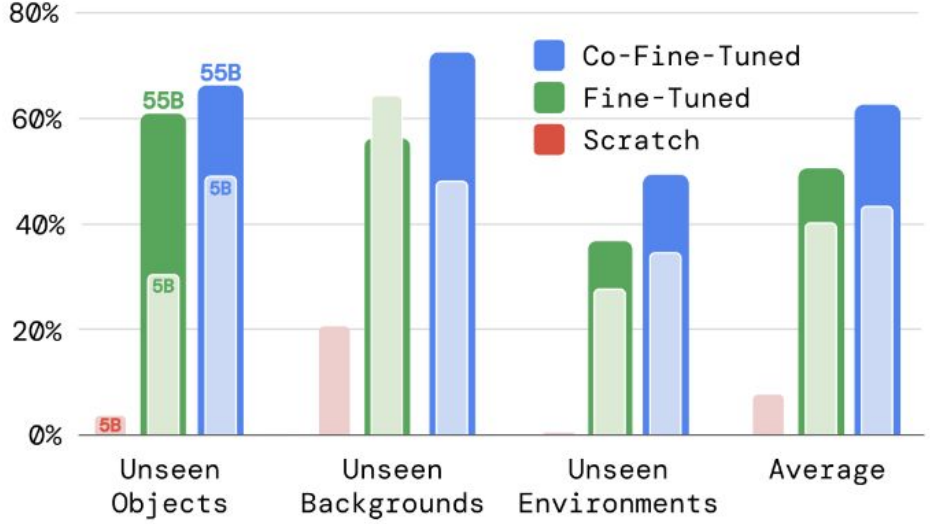
Prompt:
Given <img> I need to hammer a nail, what object from the scene might be useful?
Prediction:
Rocks. Action: 1 129 138 122 132 135 106 127



Evaluation

- Symbol understanding
- Reasoning
- Human recognition

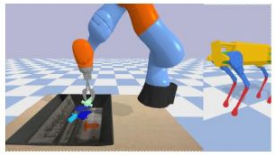


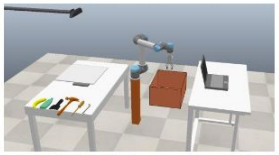


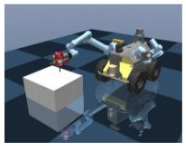
Simulators

Isaac Sim


Gazebo


Pybullet


V-REP (CoppeliaSim)


MuJoCo


Unity ML-Agents


AirSim


AI2-THOR


Matterport 3D


Virtualhome


SAPIEN


iGibson

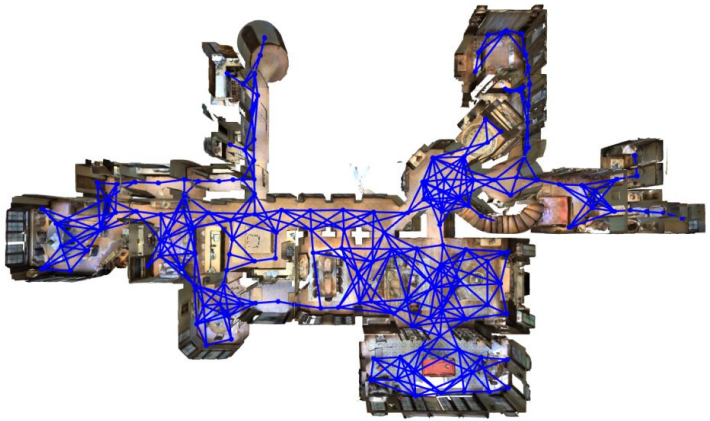

TDW

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. XX, NO. XX, XXX 2024

Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond

Zheng Zhu*, Xiaofeng Wang*, Wangbo Zhao*, Chen Min*, Nianchen Deng*, Min Dou*,
Yuqi Wang*, Botian Shi†, Kai Wang†, Chi Zhang†, Yang You†, Zhaoxiang Zhang†,
Dawei Zhao†, Liang Xiao†, Jian Zhao†, Jiwen Lu†, Guan Huang†

Example of Real-World Simulator: Matterport 3D Simulator



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.