Towards Multimodal AI Agent That Can See, Think and Act

Meta AGI
Jianwei Yang

Why Embodied AI Agents?

- **Autonomous Driving**

- **Industry Robot**

- **Domestic Robot**

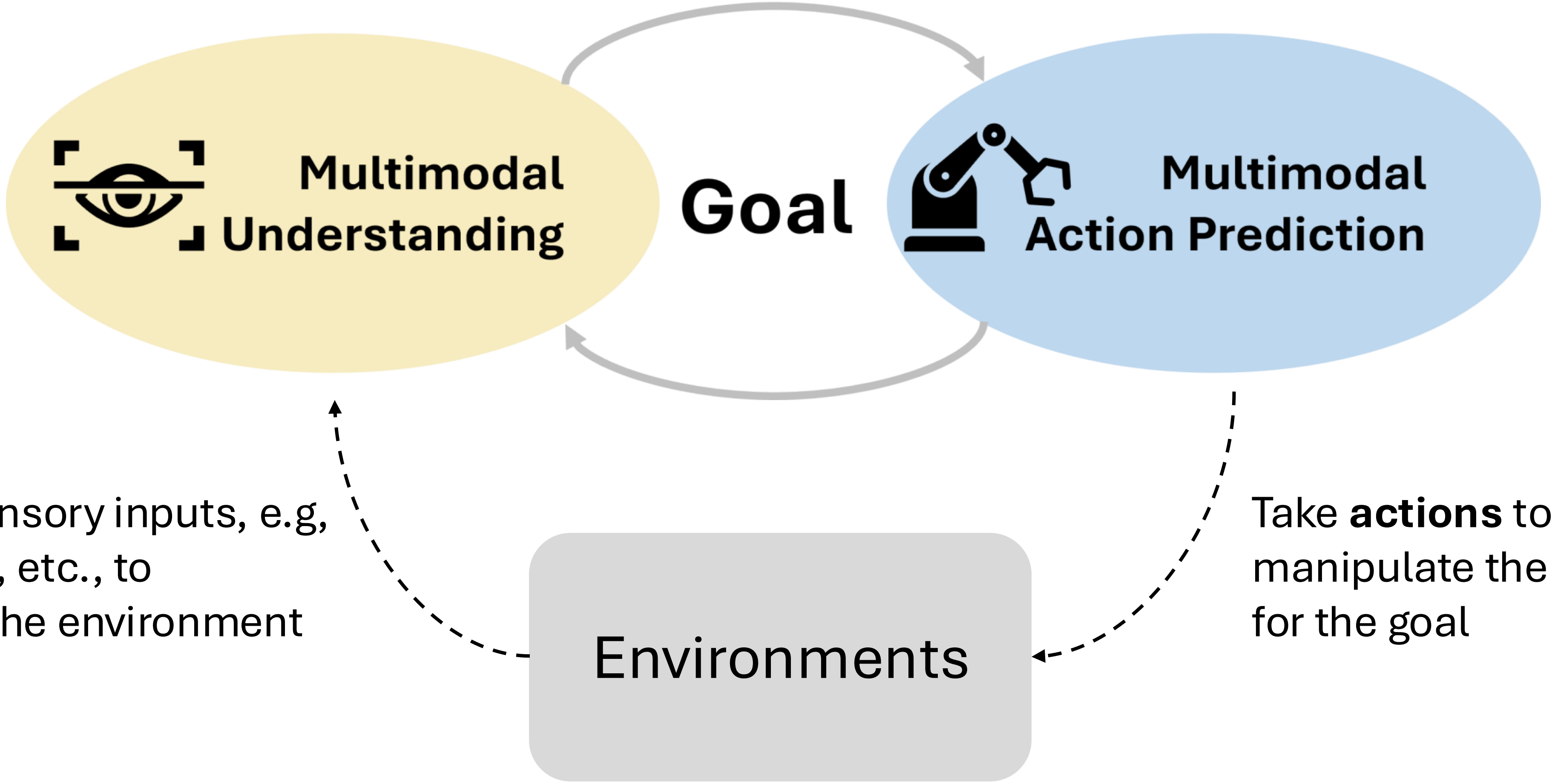
People have been dreaming of having autonomous AI agents help to handle tedious tasks in the daily life

What is an Embodied AI Agent?

What is an Embodied AI Agent?

Wikipedia: An intelligent agent is an entity that perceive its environment, takes actions autonomously to achieve **goals**, and may improve its performance through machine learning or by acquiring knowledge.

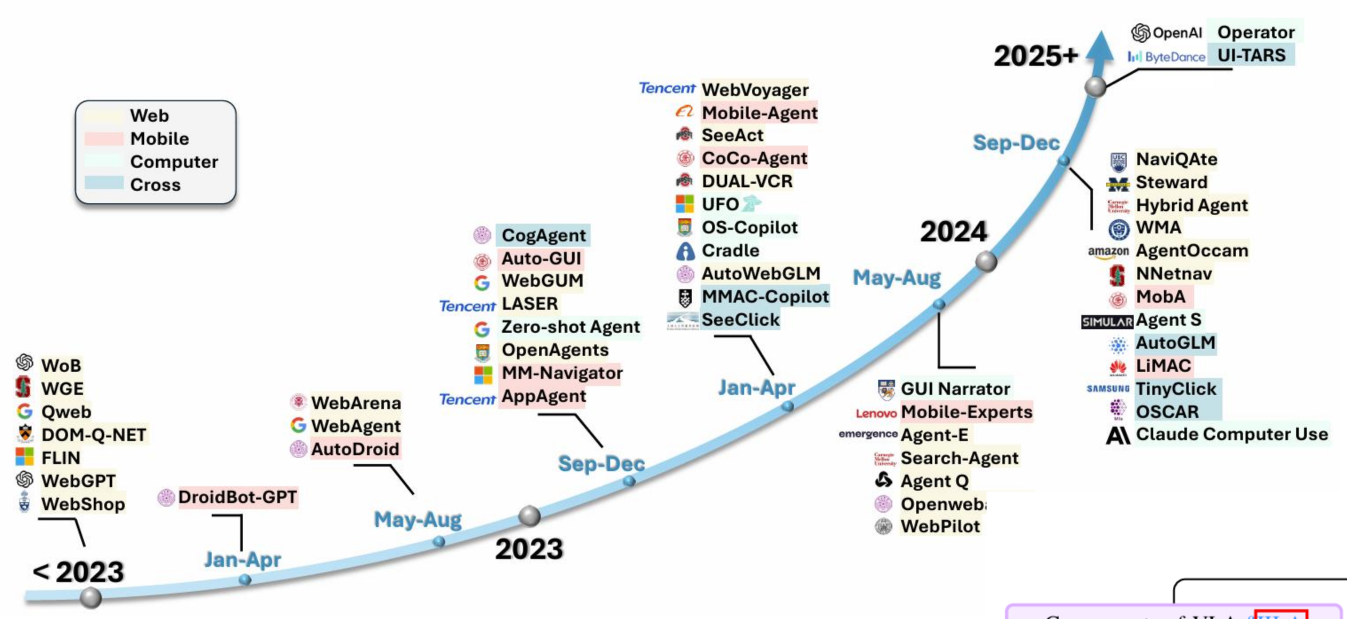
What is an Embodied AI Agent?



Take multi-sensory inputs, e.g, vision, tactile, etc., to **understand** the environment

Take **actions** to interact and manipulate the environment for the goal

The Dawn of Multimodal AI Agents



WebArena
WebAgent
CogAgent
Auto-GUI
MM-Navigator
UFO
SeeAct

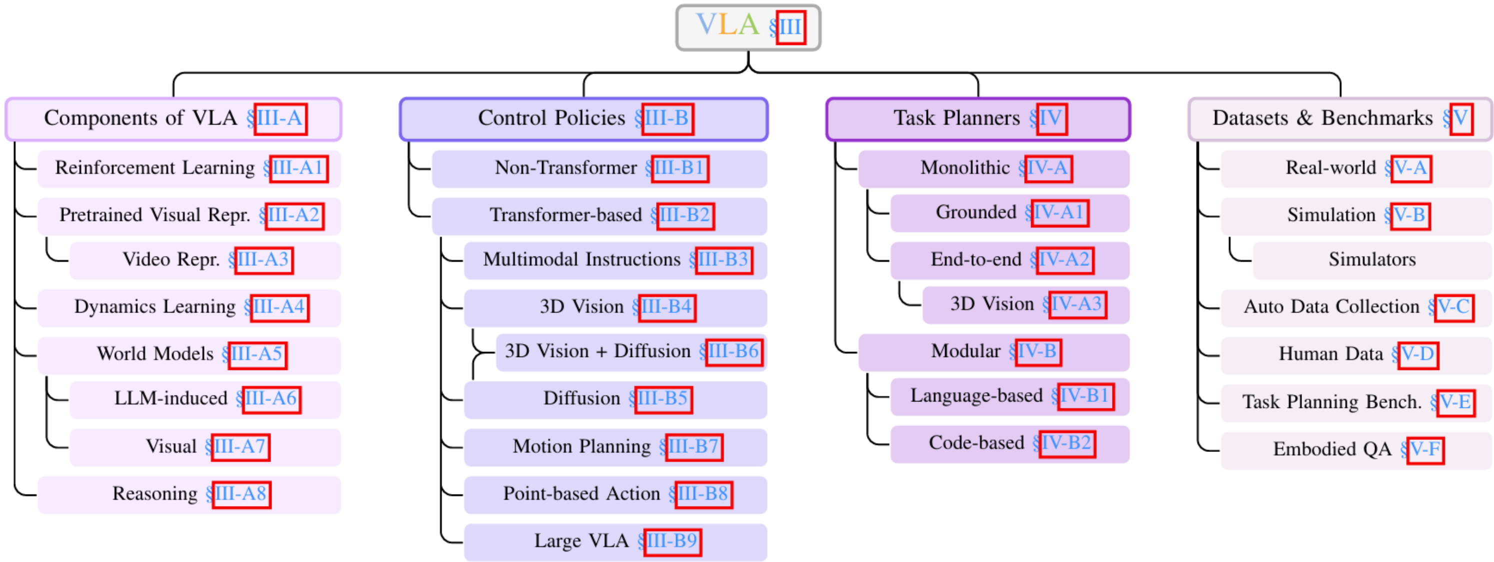
SeeClick
GUI Narrator
Agent Q
Webpilot
AutoGLM
OSCAR
Operator

And many more…

OpenVLA
TraceVLA
LAPA.
Embodied-CoT
Pi0/0.5
GROOT-0
Magma
LLaRA
ChatVLA
CoVLA

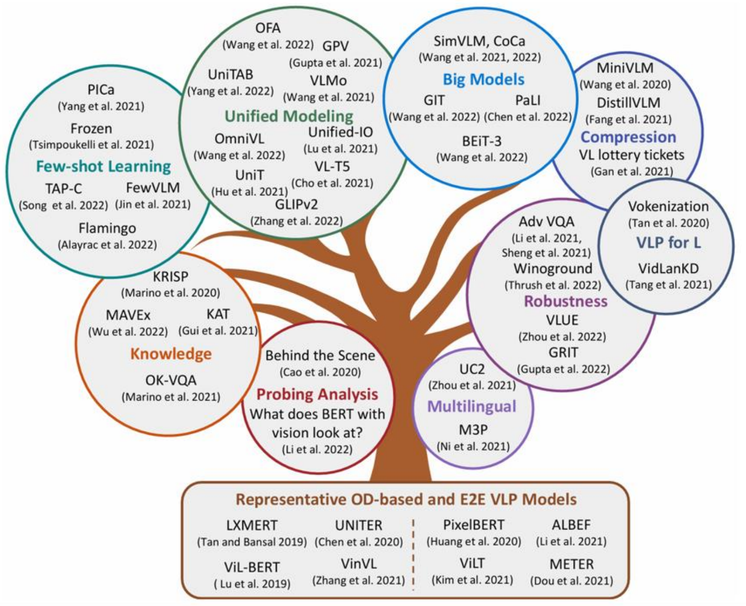
RT-1
RT-2
RT-2X
SmoVLA
TinyVLA
SpatialVLA
AGIBot
RoboBrain
Magma
Helix

And many more…



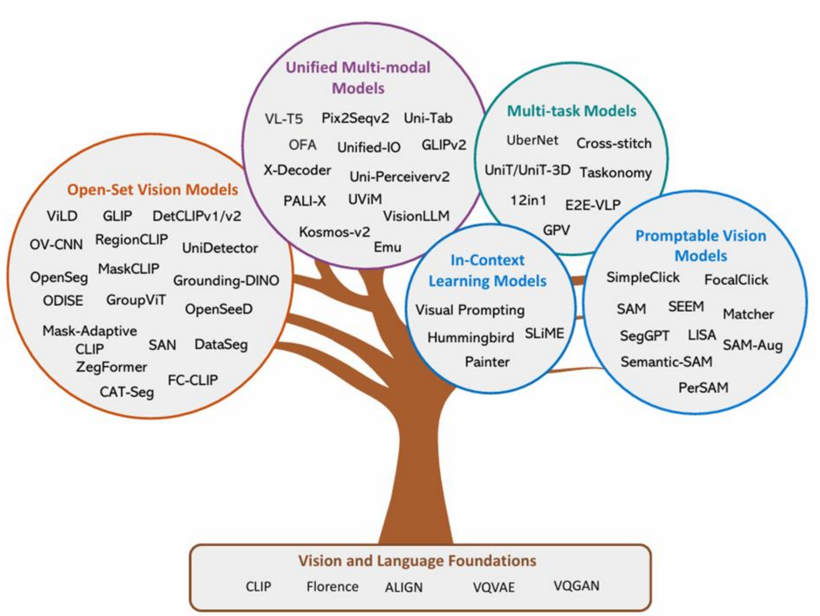
How far we have gone?

2019-2022
Vision-Language Models



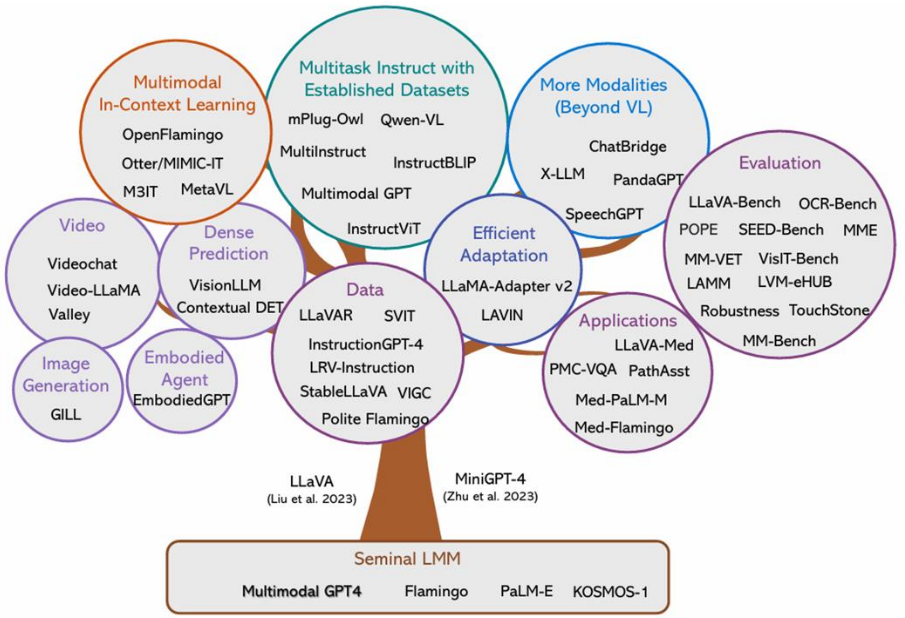
Model: < 1B parameters
Data: < 10M images

2021-2023
Multimodal Vision Models



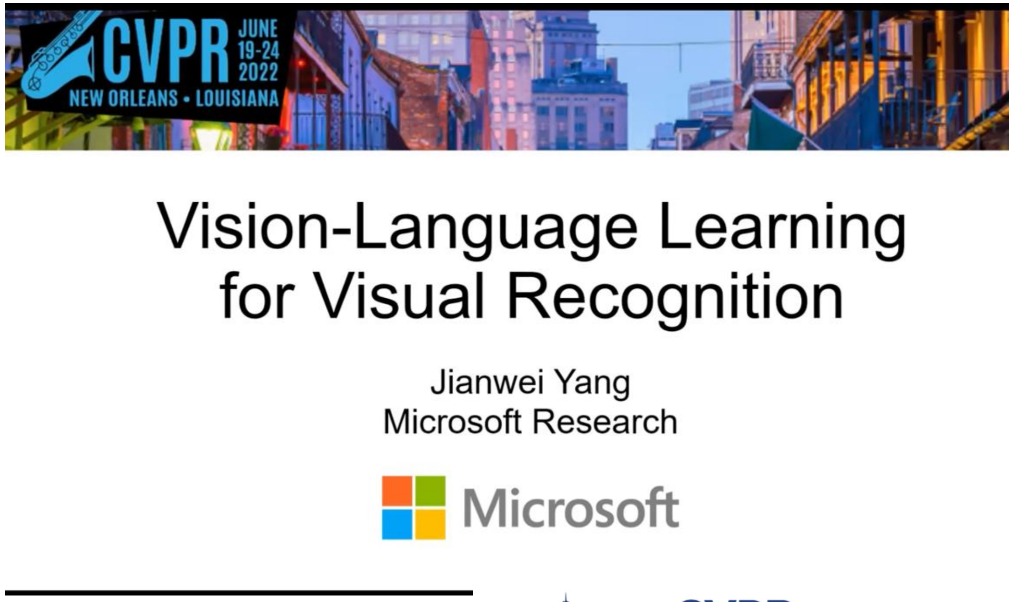
Model: < 1B parameters
Data : up to 1B images

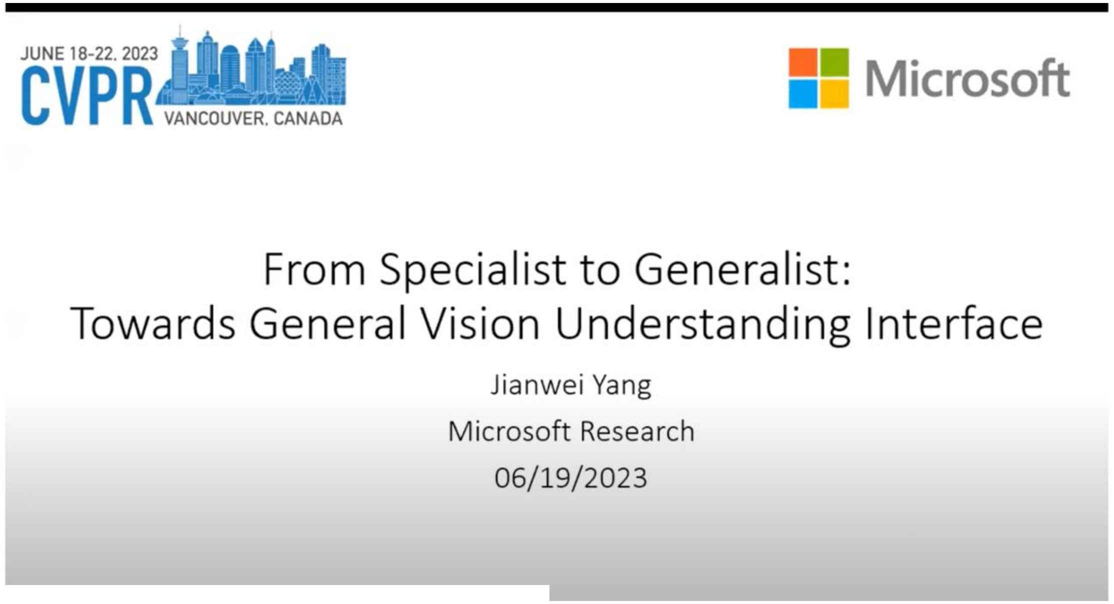
2023-Present
Large Multimodal Models



Model: Up to 1000B parameters!?
Data: 10B images and 100T tokens

Vision-Language Pre-training: Basics, Recent Advances, and Future Trends. Gan et al.
Multimodal Foundation Models: From Specialists to General-Purpose Assistants. Li et al.





A Close Look at Vision in Large Multimodal Models

Jianwei Yang
Microsoft Research
06/17/2024

Opportunities: More Capable of Reasoning and Plannings in Real World

face

- An intelligent AI should be able to understand and interact with human and physical world

- Robotics
- Automatous Driving

(My) Path Towards Multimodal Embodied AI Agents

Multimodal Vision Foundation

Embodied Agent

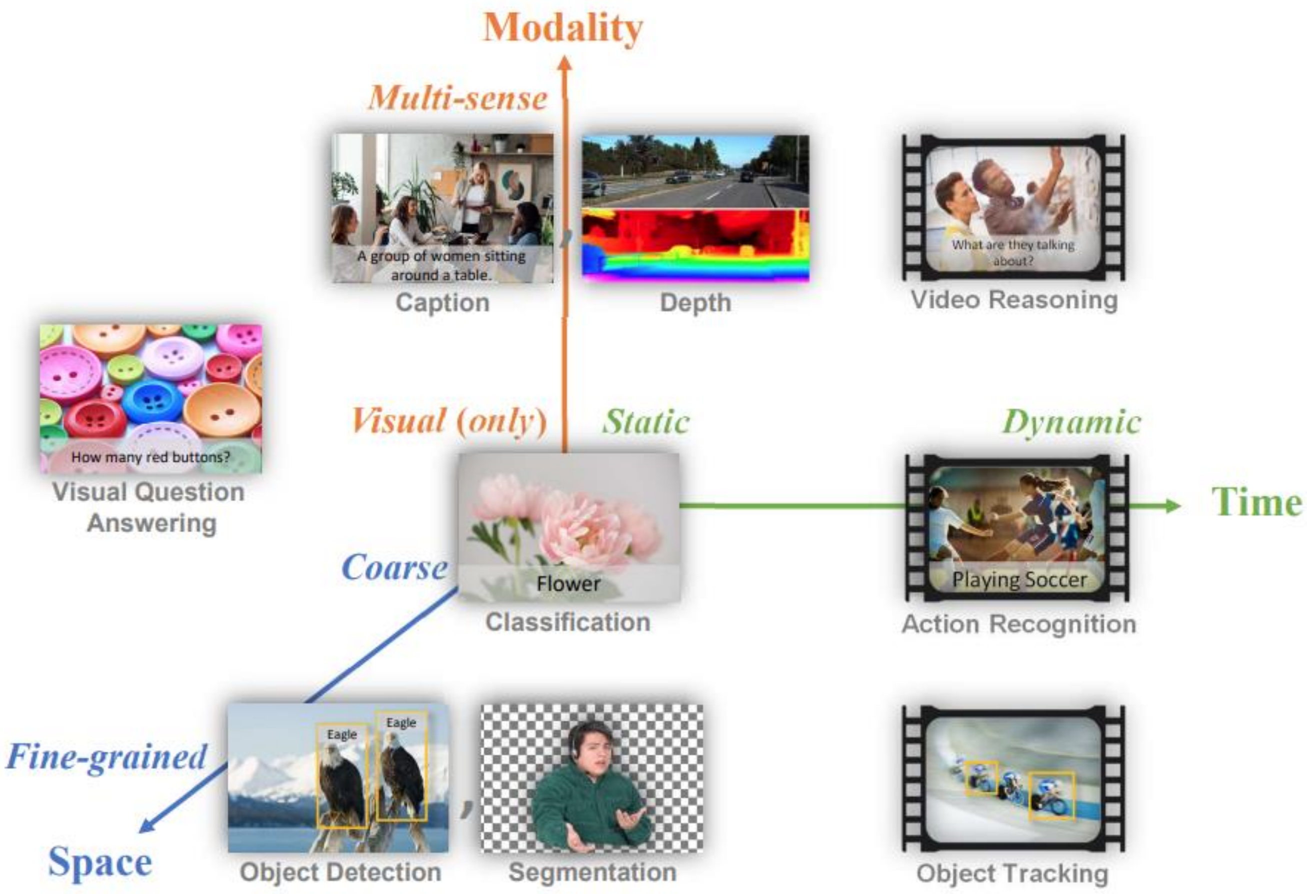
See

Think

Act

Multimodal LLM

Vision is Complicated



a) **Inputs Types:**
- Temporality: static image, video sequence
- Extra modality: text, audio, etc.

b) **Diverse Tasks:**
- Image-level: classification, captioning, etc.
- Region-level: object detection, grounding, etc.
- Pixel-level: segmentation, depth est.. etc.

(c) **Output types:**
- Semantic: class labels, descriptions, etc.
- Spatial: edges, boxes, masks, etc.
- Temporal: traces, optical flows, etc.

Recognize (all) object categories in the wild

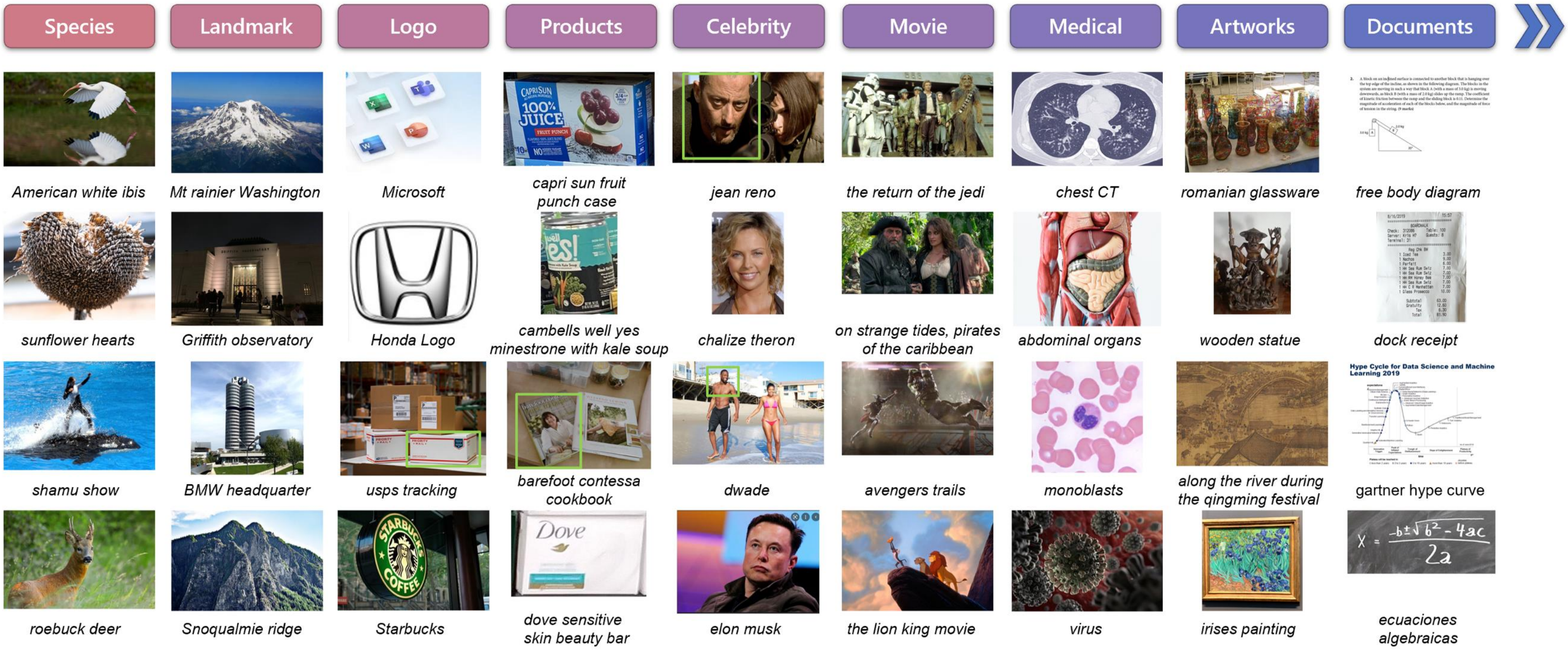
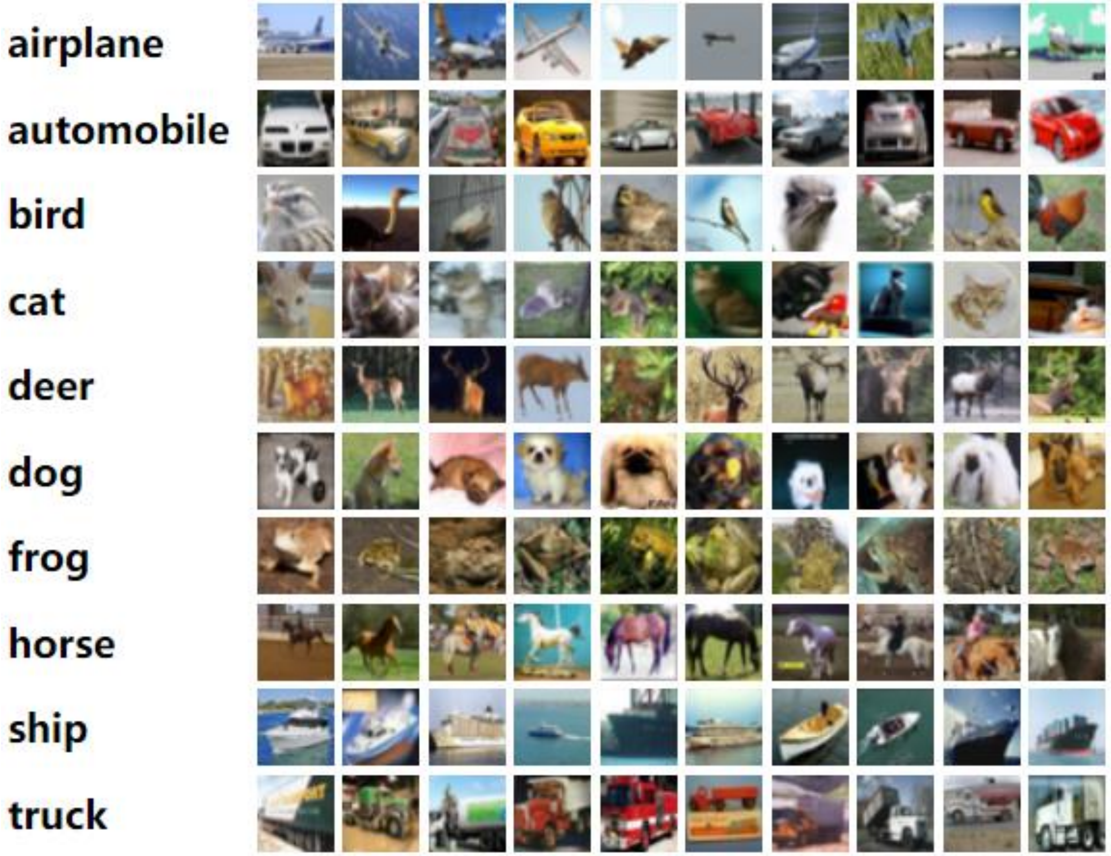


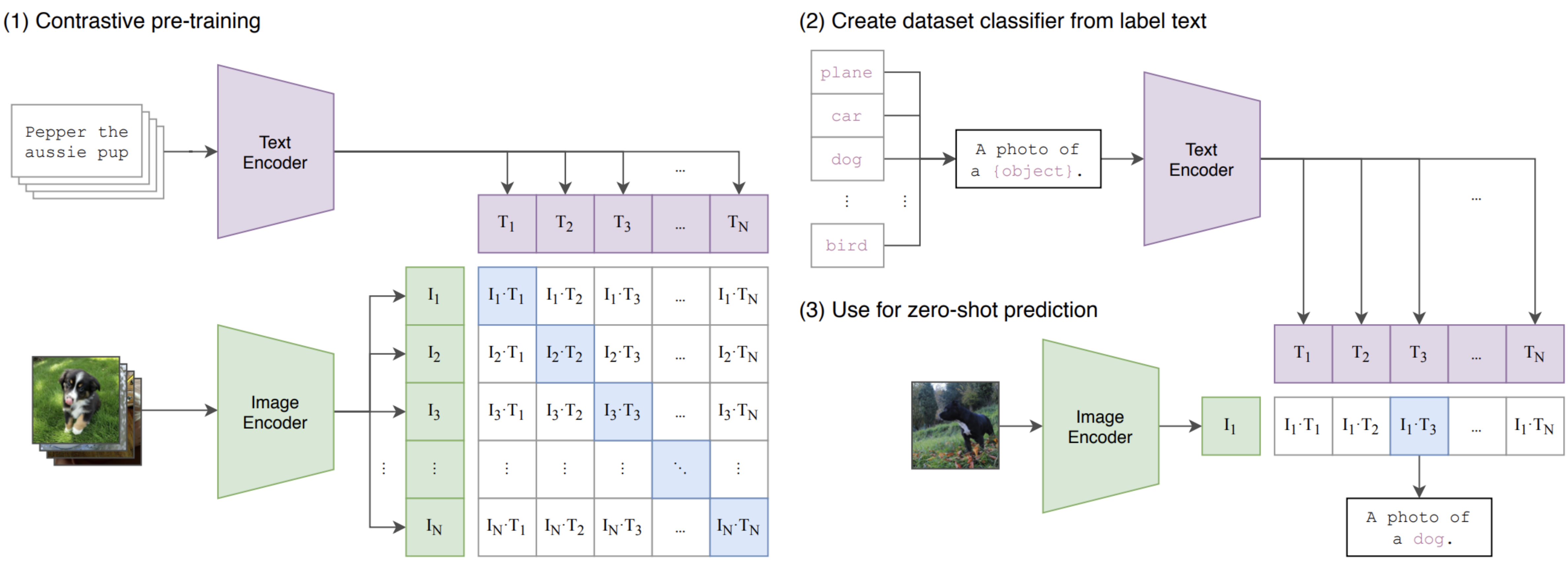
Image is a label



CIFAR-10 (10 labels)

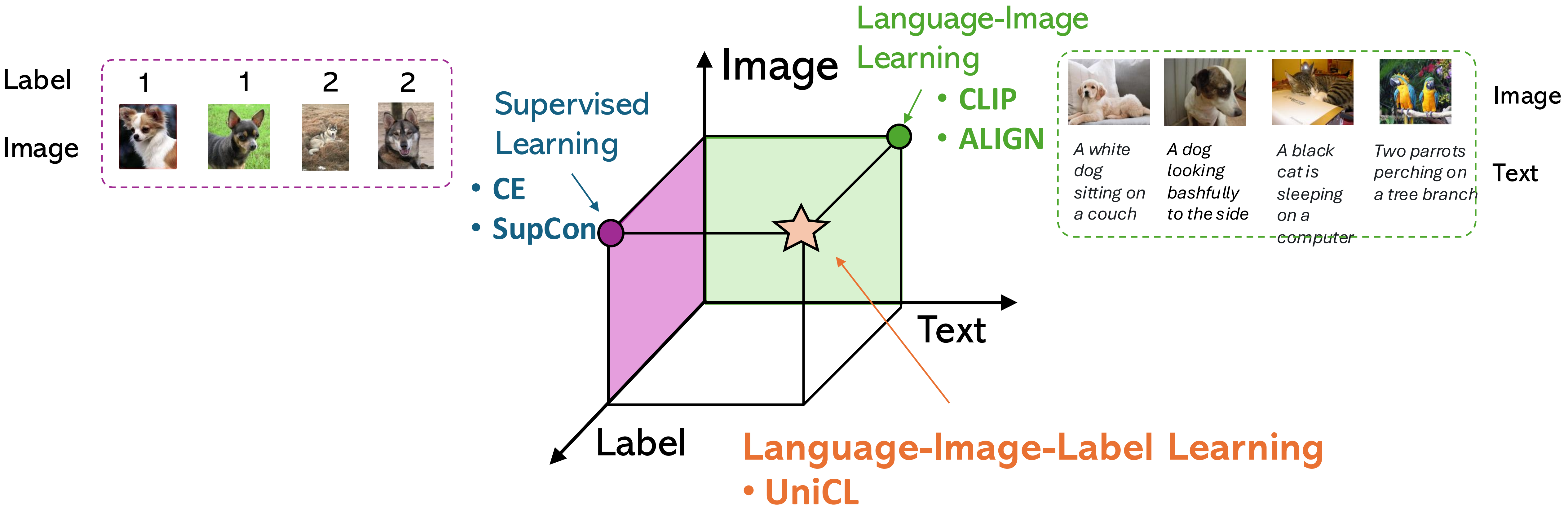
Autonomous Robot Indoor Dataset (ARID)

Image is a text



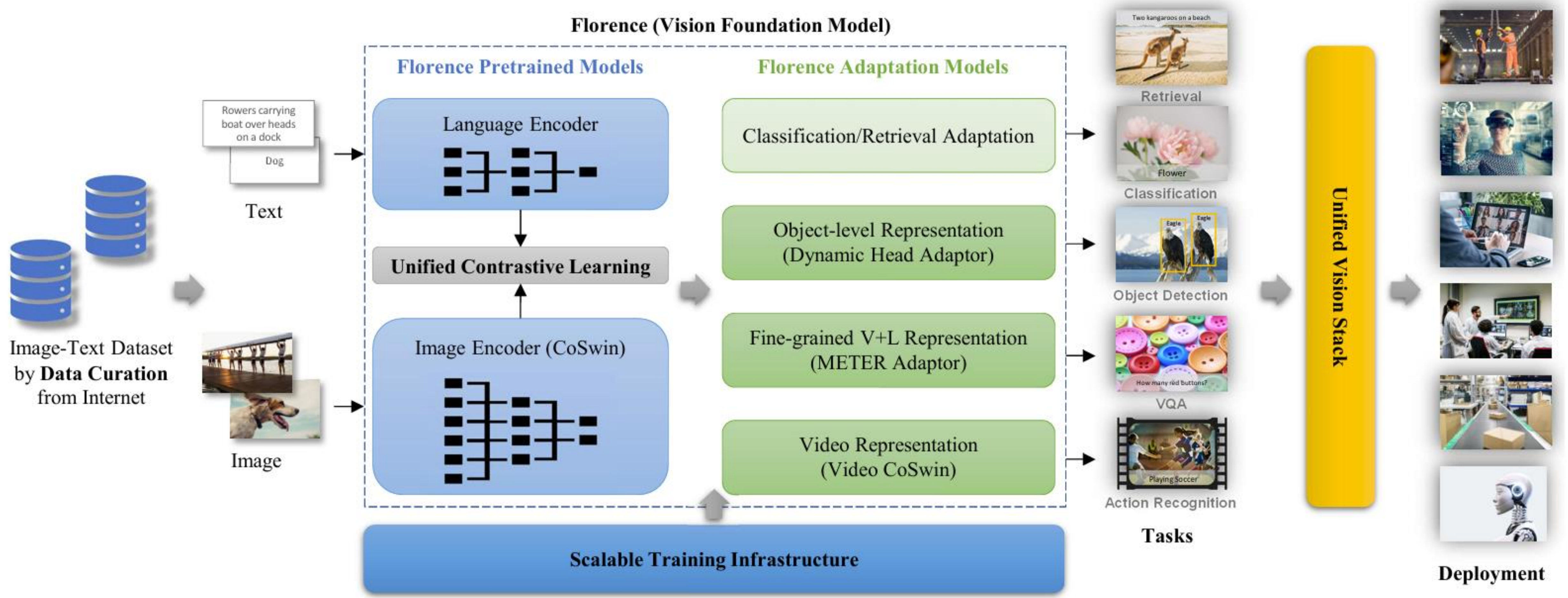
OpenAI CLIP

Learning in the image-label-text space



Leverage both types of data
Discriminative features for image classification tasks
Broad semantic coverage for open-world scenarios

Project Florence



State-of-the-art on 44 vision benchmarks

From Image, to Region, to Pixel

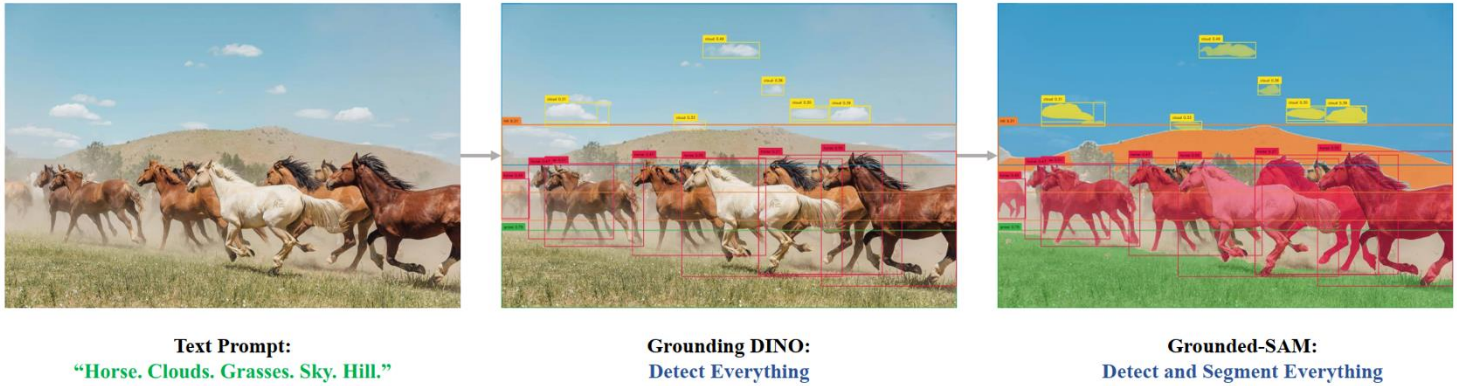
hard times teach us valuable lessons. handwriting on a napkin with a cup of coffee stock photos

person battles with person in the production sedans

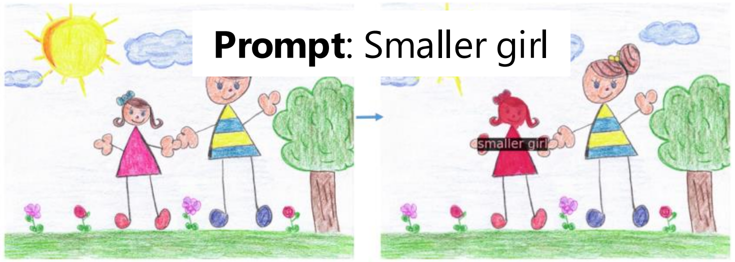
save the straws classic t-shirt

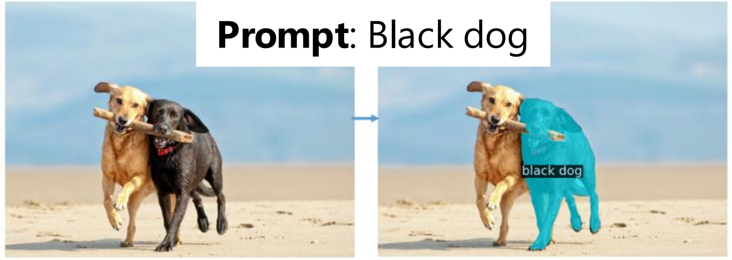
this week i'm going to share 20 ideas with you. 20 different lunchbox ideas. packing school lunch is about nourishment.

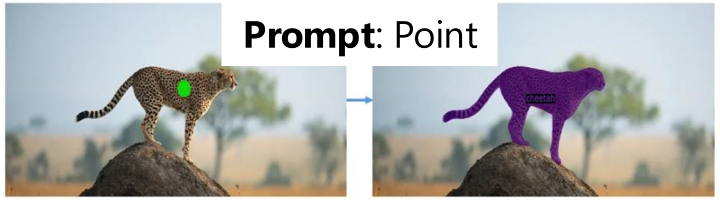
GLIP. 2022



Grounding-DINO. 2022







SEEM. 2023

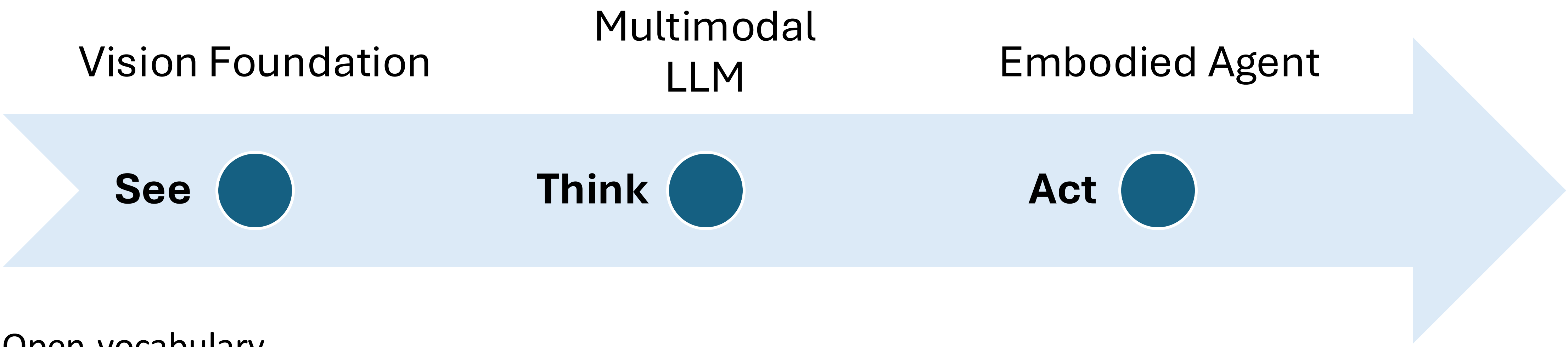
What Vision Foundation Can Give Us?

Vision Foundation — **See**

Multimodal LLM — **Think**

Embodied Agent — **Act**

What Vision Foundation Can Give Us?



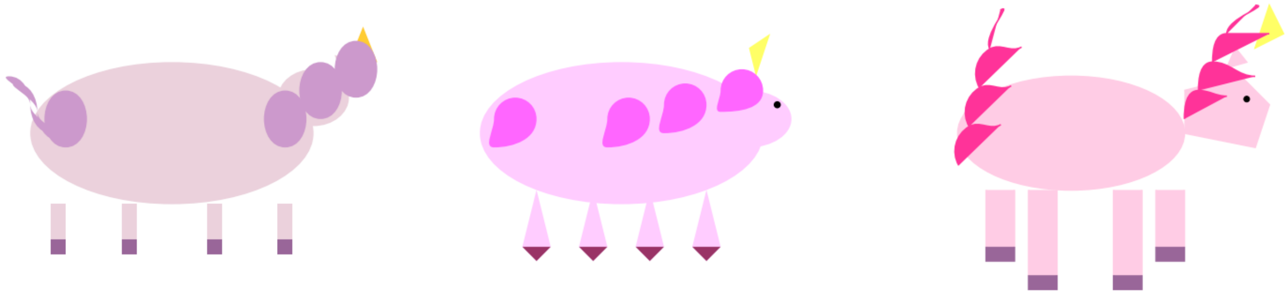
- Open-vocabulary
- Spatially grounded
- Versatile inputs & outputs

A multimodal vision model can effortlessly apply in the wild

Multimodal Large Language Models

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, Yi Zhang



GPT-4 visual input example, Extreme Ironing:

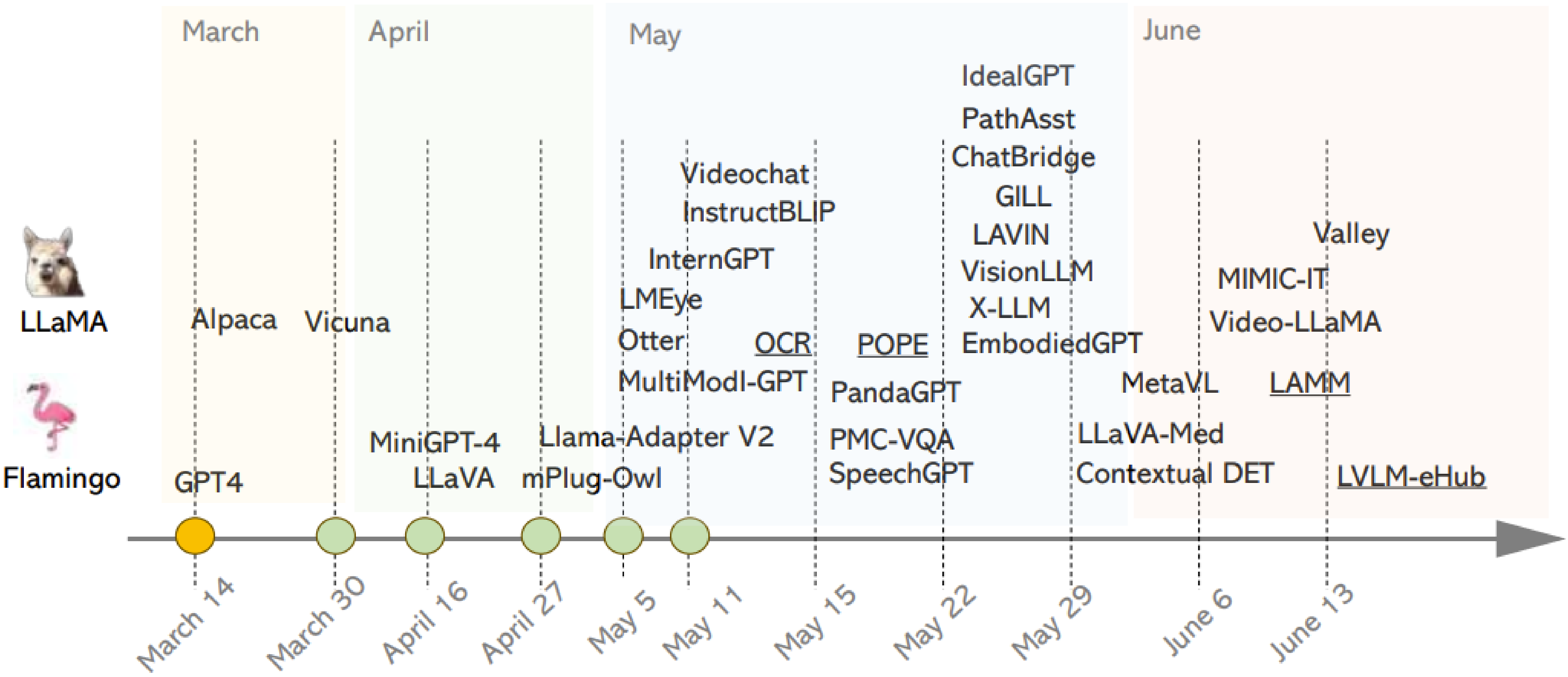
User: What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

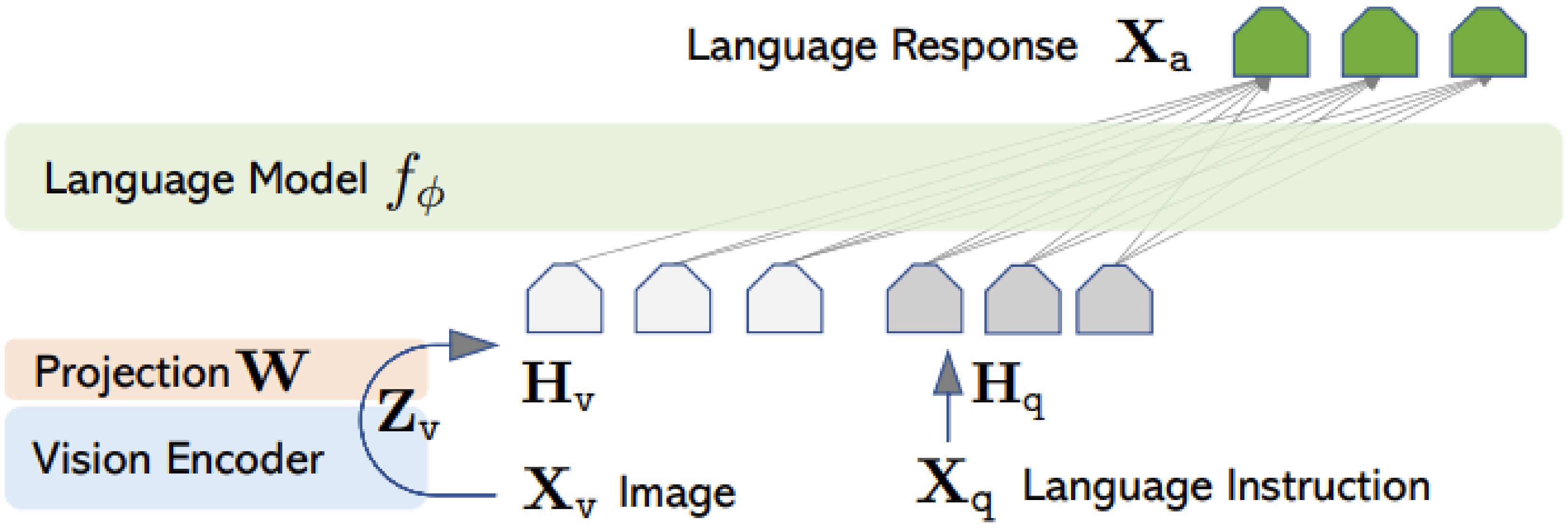
GPT-4: The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

GPT-4V. OpenAI, 2023

The Progression of LMMs



Visual Instruction Tuning: LLaVA-1.0



- **Stage 1 (Alignment)**

  Train only the projection weight W using **image-text pairs**.

  Training cost is as low as taking **a few hours** on 8xA100 GPUs

- **Stage 2 (Instruction Following)**

  Tune the projection and LLMs using **instruction-following data**.

  Training cost is as low as taking **less than 1 day** on 8xA100 GPUs

A Cheap Reproduction of GPT-4V

Visual input example, Extreme Ironing:

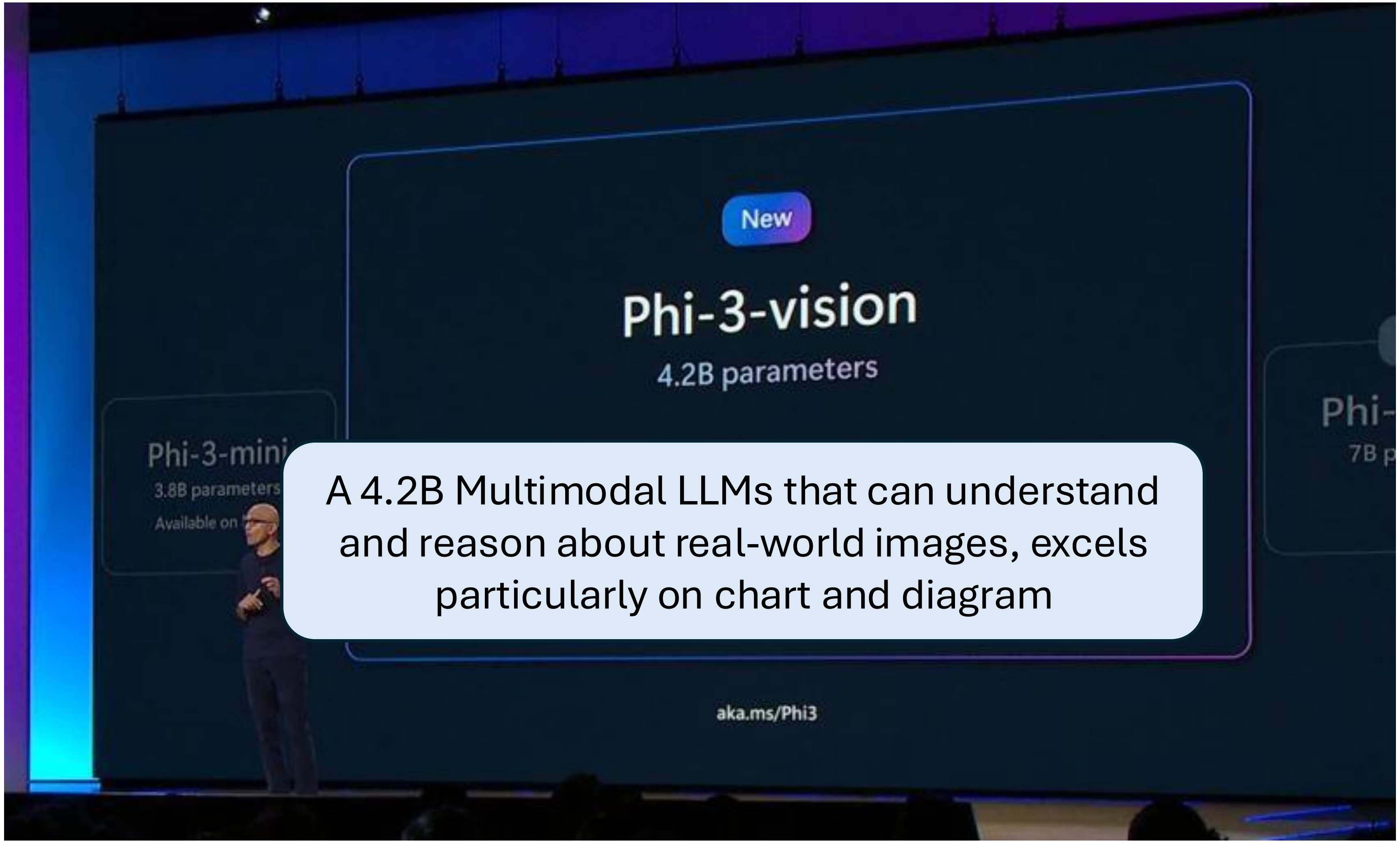
Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.
User	What is unusual about this image?
GPT-4 [32]	The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.
User	Can you explain this meme in detail?
BLIP-2	a man is sitting on the back of a yellow cab
User	Can you explain this meme in detail?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Visual input example, Chicken Nugget Map:

User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all is..." The main image features a tray of cooked chicken nuggets, with the chicken pieces positioned to resemble the continents and islands on a world map. The meme playfully suggests that the chicken nuggets represent the Earth, and the various locations depicted in the photo are actually chicken nugget versions of different places. The punchline of the meme appears in smaller text at the bottom, reading "I mean, it's not the real Earth, but how beautiful it is all is." This lighthearted and entertaining meme combines food and humor, using the image of chicken nuggets to create a funny and imaginative representation of the world.
User	Can you explain this meme?
GPT-4 [32]	This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets. The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world. The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.
User	Can you explain this meme in detail?
BLIP-2	sometimes i just look at pictures of the earth from space and marvel how beautiful it is
User	Can you explain this meme in detail?
OpenFlamingo	It's a picture of a chicken nugget on the International Space Station.

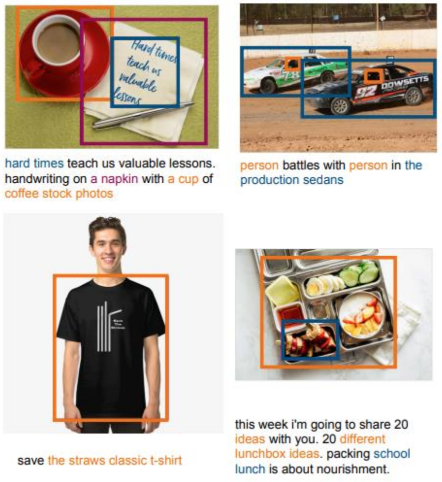
Phi-3 Technical Report:
A Highly Capable Language Model Locally on Your Phone



A 4.2B Multimodal LLMs that can understand and reason about real-world images, excels particularly on chart and diagram

Multimodal Vision Model is spatially grounded but not thinking

Multimodal Large Language Model is thinking but not spatially grounded



Can we reconcile them?

Visual Prompting with Set-of-Marks

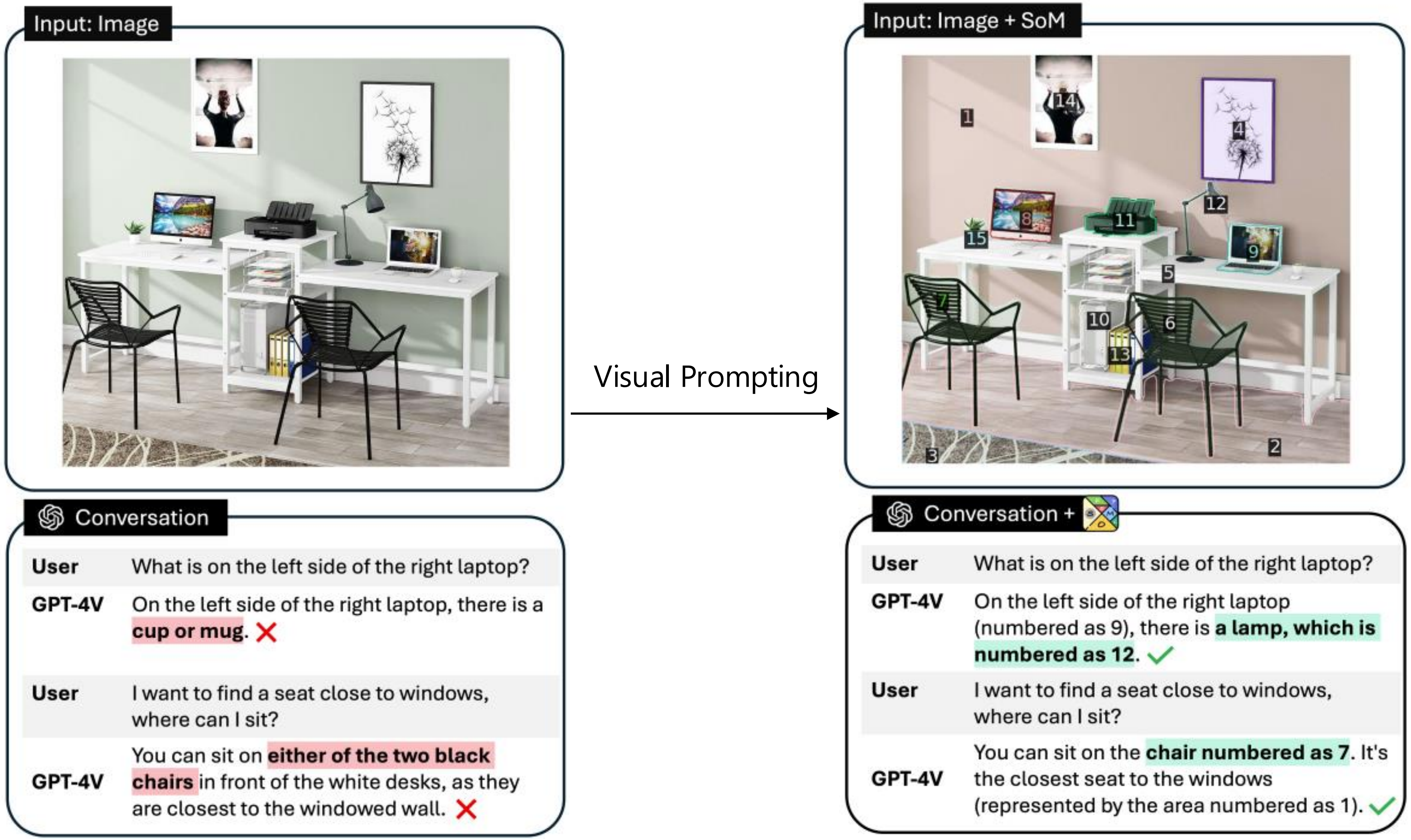
Input: Image

Conversation

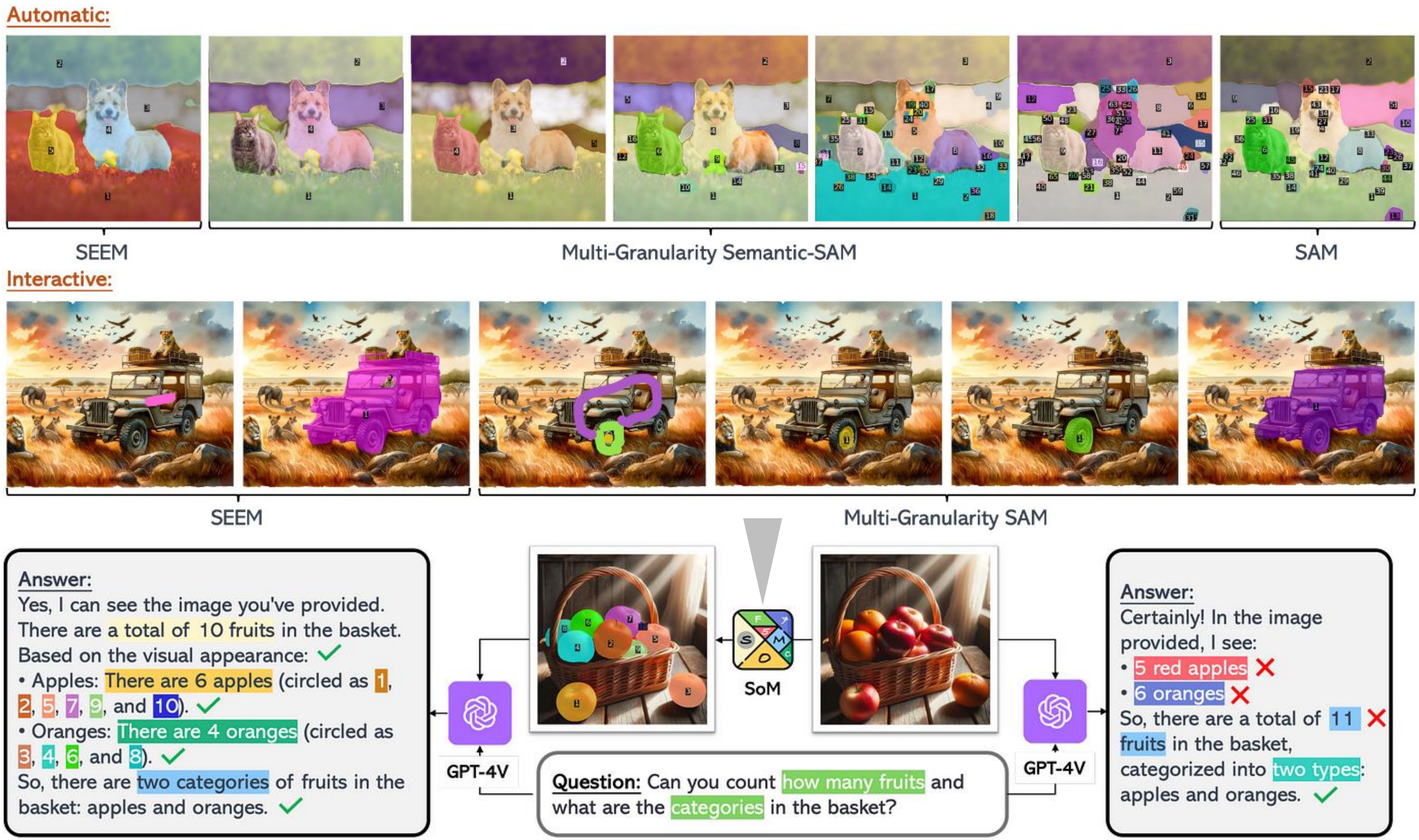
User	What is on the left side of the right laptop?
GPT-4V	On the left side of the right laptop, there is a **cup or mug**. ✗
User	I want to find a seat close to windows, where can I sit?
GPT-4V	You can sit on **either of the two black chairs** in front of the white desks, as they are closest to the windowed wall. ✗

Yang et al. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v

Visual Prompting with Set-of-Marks



Yang et al. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v


Automatic:
SEEM
Multi-Granularity Semantic-SAM
SAM
Interactive:
SEEM
Multi-Granularity SAM
Answer:
Yes, I can see the image you've provided.
There are a total of 10 fruits in the basket.
Based on the visual appearance:
• Apples: There are 6 apples (circled as 1, 2, 5, 7, 9, and 10).
• Oranges: There are 4 oranges (circled as 3, 4, 6, and 8).
So, there are two categories of fruits in the basket: apples and oranges.
GPT-4V
SoM
Question: Can you count how many fruits and what are the categories in the basket?
GPT-4V
Answer:
Certainly! In the image provided, I see:
• 5 red apples
• 6 oranges
So, there are a total of 11 fruits in the basket, categorized into two types: apples and oranges.

Tool Usage Instruction

User: How to use this controller?

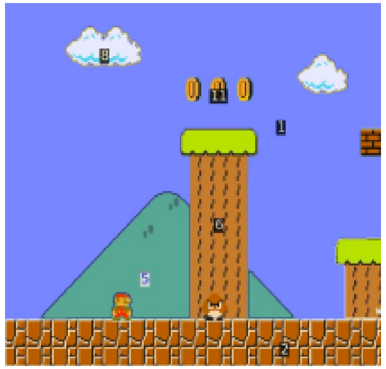
GPT-4V (SoM):

This appears to be an Xbox One or Xbox Series X|S controller. Here' how to use it:

- **Thumbsticks (2 & 3)**: These are primarily used for movement and aiming.
  - **Left Thumbstick (2)**: Controls the movement of the character.
  - **Right Thumbstick (3)**: Controls the camera or aiming reticle.
- **Directional Pad (D-Pad) (1)**: Used for various functions such as navig weapons, or other in-game actions.
- **Face Buttons (4, 5, 6, & 7)**:
  - **A (4)**: Typically the main action or "accept" button.
  - **B (7)**: Often the "back" or "cancel" button.
  - **X (5)**: Varies by game, frequently used for interactions.
  - **Y (6)**: Like X, its function varies but is commonly used for secondary
- **Bumpers and Triggers (8 & 9)**:
  - **Left and Right Bumpers (8)**: Used for quick actions or toggling betv
  - **Left and Right Triggers**: Used for primary actions like shooting o games.
- Menu and View Buttons: Located in the center, used to bring up in-game m
- Xbox Button: Press to go to the Xbox home or turn the controller on/off. menu to turn off the controller or console.
- Pairing Button: Located on the top edge, used for pairing the controller.
- 3.5mm Audio Jack: On the bottom, plug in headphones or a headset.
- USB Port: On the top edge, used for charging or wired play.

Refer to the game's manual or in-game control settings for exact functions for

2D Game Planning



User: How can I get the coins on top?

GPT-4V (SoM):

It looks like you're playing a platformer game reminiscent of Super Maric general strategy based on the layout:

- Elevated Platforms: **The platforms labeled as '4' and '7'** are the key. You ne these platforms to reach greater heights. Depending on the game mechanics, you run-up or utilize a jump-boosting power-up if available.
- Bricks and Blocks: **The bricks labeled as '9'** might contain some power-up breakable. In classic platformers like Super Mario, hitting such blocks from belo ups or even create platforms. **The same goes for the block labeled as '1'**.
- Vertical Jump: **Once on the elevated platform '4'**, try to perform a vertical jum **labeled '11'**. Ensure that you time your jump correctly and use any speed or jump
- Momentum: Some games also allow for gaining extra height in jumps by building game works similarly, try running from a distance and jumping at the last mome
- Look for Hidden Power-ups: Some games hide power-ups in the level that ca reaching difficult places. These might be hidden in unmarked bricks or in other abilities like a double jump, use them.
- Game Mechanics: Every game can have unique mechanics. If there are any mechanics you haven't tried, consider experimenting with them. Maybe there' double jump, or use an enemy to boost your jump height.

If you've tried all the general strategies and still can't reach the coins, it might be b specific strategies or walkthroughs for the particular game you're playing. Good lu

Simulated Navigation

User: I am a bit hungry, where can I find some food?

GPT-4V (SoM):
Based on the image provided, it looks like you're viewing a floor plan of an apartment. For food, **you should head to the kitchen area which is labeled with the number "10"**. That's where you're most likely to find food or kitchen appliances to prepare a meal. If you're in this apartment, head to that section! If you're not, consider checking your own kitchen or a nearby restaurant. Enjoy your meal!

User: cool, where should I sit to eat if I find some food from 10?

GPT-4V (SoM):
Based on the floor plan, the most suitable place to sit and eat would be at **the dining area labeled with the number "14". It appears to have a dining table and chairs**, making it a comfortable spot for enjoying a meal.

User: ok, i got a dirty hand, what should I do?

GPT-4V (SoM):
Based on the floor plan, you should head to the bathroom areas to wash your hands. **You can go to the space labeled with the number "13" which appears to be a bathroom**. It's likely to have a sink and soap where you can clean your hands. Always remember to wash your hands thoroughly, especially before eating!

User: now I need to get some sleep.

GPT-4V (SoM):
For sleep, you should **head to the area labeled with the number "19"**. It appears to be a bedroom with a bed, which would be the most suitable place for rest. Have a good sleep!

User: where is the TV, I want to turn it on.

GPT-4V (SoM):
The TV is located in the area labeled with the number "6". You can head over there to turn it on and enjoy some entertainment.

User: Enough to watch a TV, I want to take a shower.

GPT-4V (SoM):
To take a shower, you should **head to the space labeled with the number "17"**. It appears to be a shower area in the bathroom. Make sure you have your toiletries and towel ready, and enjoy your shower!

User: okay, need to wash some clothes, put them where?

GPT-4V (SoM):
For washing clothes, you should **head to the area labeled with the number "8"**. It appears to be a location designated for a washing machine or laundry area. You can place your clothes there and start the washing process.

Simulated Navigation

User: I am a bit hungry, where can I find some food?

GPT-4V (SoM):
Based on the image provided, it looks like you're viewing a floor plan of an apartment. For food, **you should head to the kitchen area which is labeled with the number "10"**. That's where you're most likely to find food or kitchen appliances to prepare a meal. If you're in this apartment, head to that section! If you're not, consider checking your own kitchen or a nearby restaurant. Enjoy your meal!

User: cool, where should I sit to eat if I find some food from 10?

User: now I need to get some sleep.

GPT-4V (SoM):
For sleep, you should **head to the area labeled with the number "19"**. It appears to be a bedroom with a bed, which would be the most suitable place for rest. Have a good sleep!

User: where is the TV, I want to turn it on.

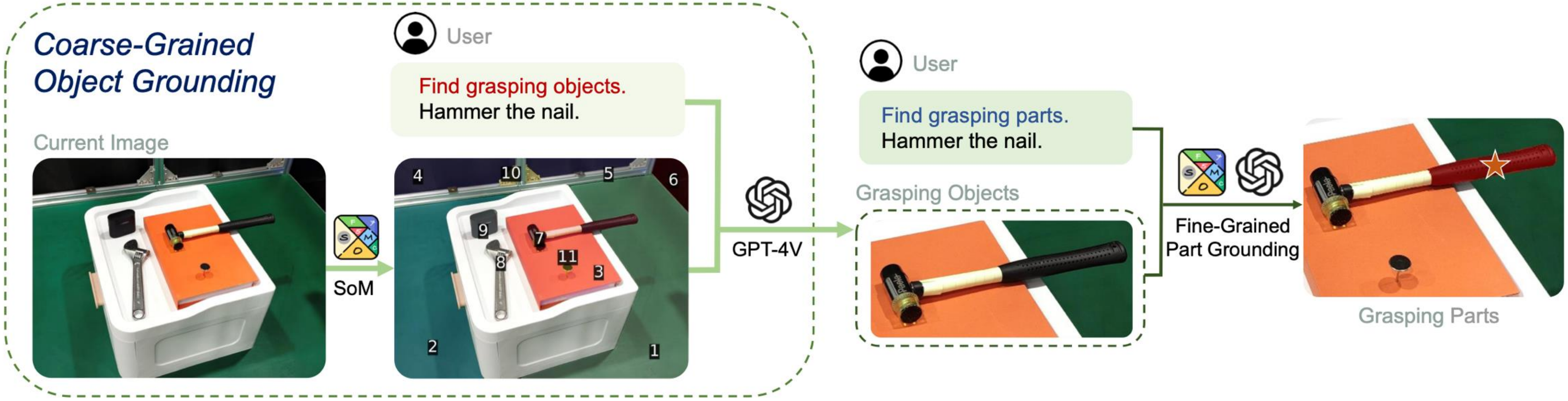
GPT-4V (SoM):
The TV is located in the area labeled with the number "6". You can head over there to turn it on and enjoy some entertainment.

User: Enough to watch a TV, I want to take a shower.

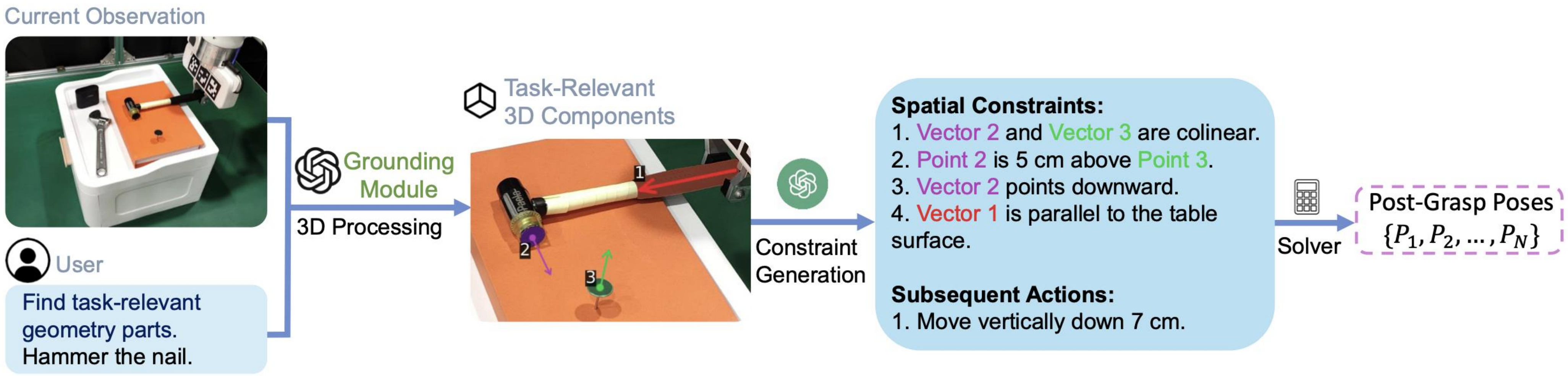
GPT-4V (SoM):
To take a shower, you should **head to the space labeled with the number "17"**. It appears to be a shower area in the bathroom.

Application of SoM for Robotics

Grasping Grounding

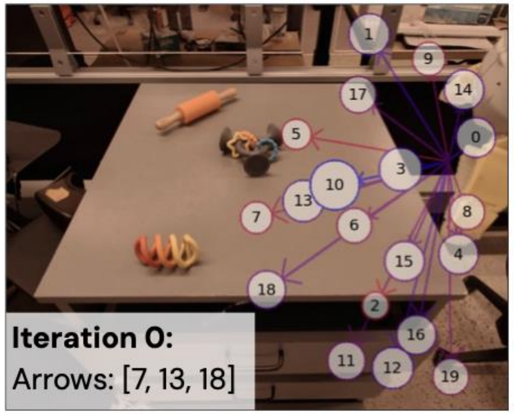


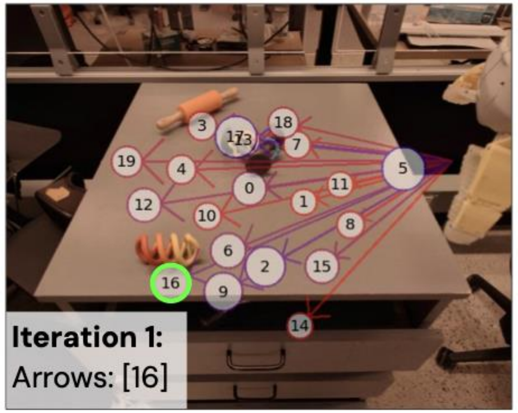
Grasping Planner



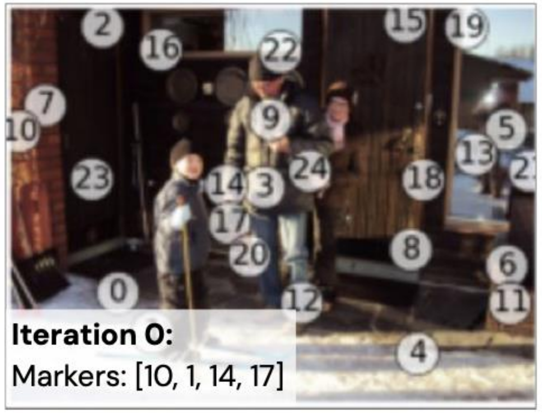
Application of SoM for Robotics

Task: What actions should the robot take to pick up the DNA chew toy?

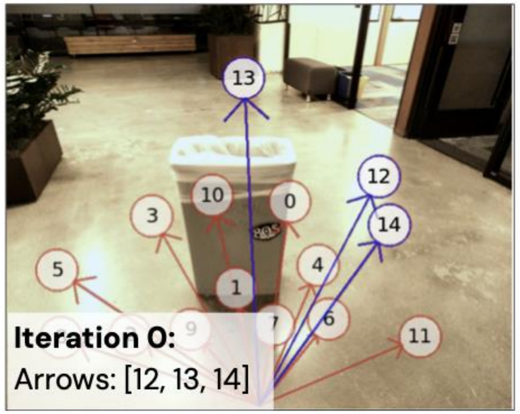


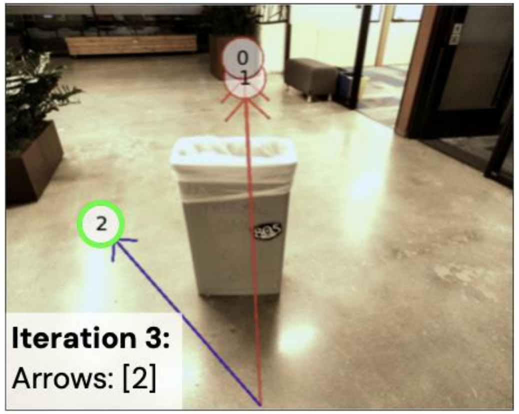


Task: What numbers overlay the "L kid"?

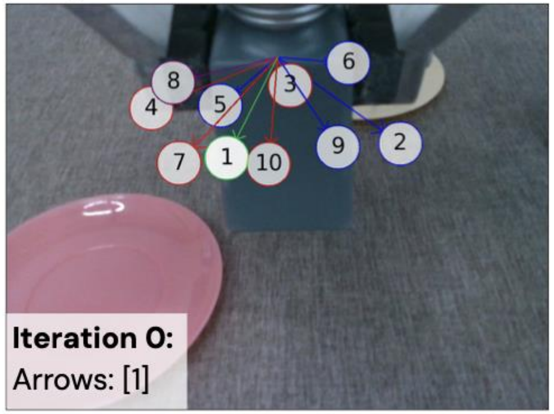


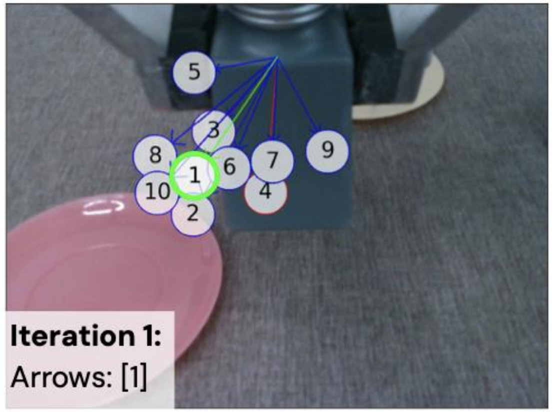
Task: What actions should the robot take to go to wooden bench without hitting the obstacle?





Task: What actions should the robot take to put the pepper shaker on the pink plate?





What MLLM can give us?

Vision Foundation

Multimodal LLM

Embodied Agent

See

Think

Act

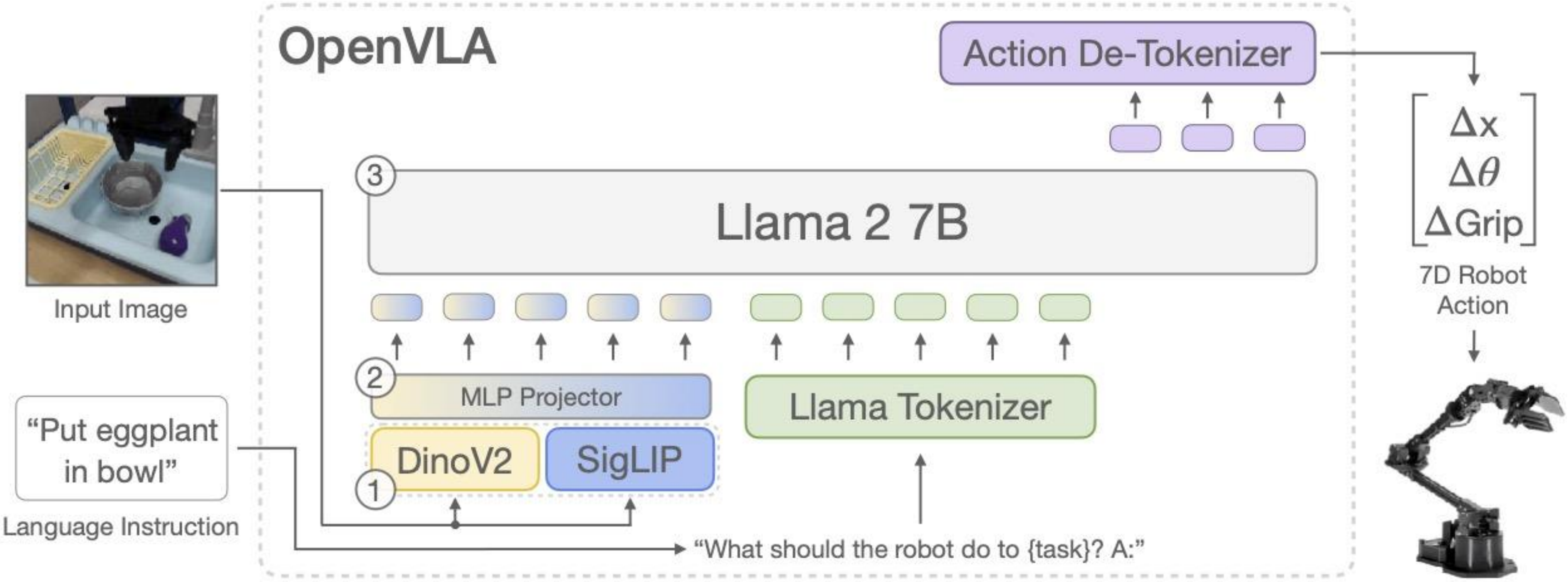
- Open-vocabulary
- Spatially grounded
- Versatile inputs & outputs

A multimodal vision model can effortlessly apply in the wild

- Understand and reasoning
- Free-form instruction
- Think spatially on image

A MLLM can understand the instruction and ground on the observations

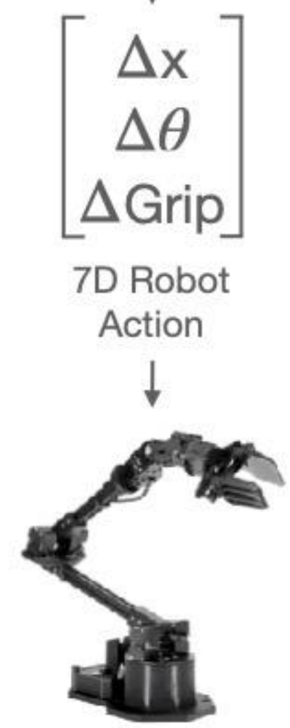
Vision-Language-Action Model

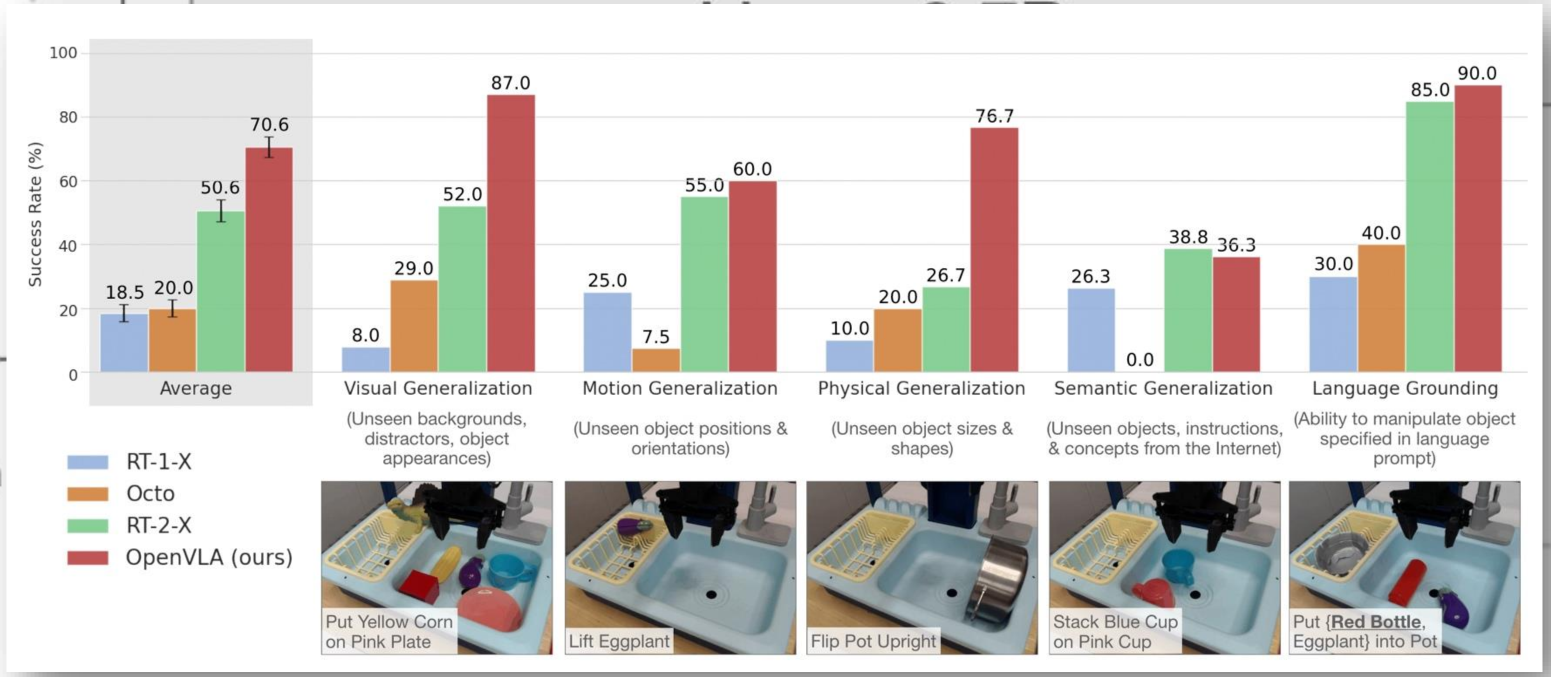


OpenVLA: An Open-Source Vision-Language-Action Model. Kim et al.

Vision-Language-Action Model

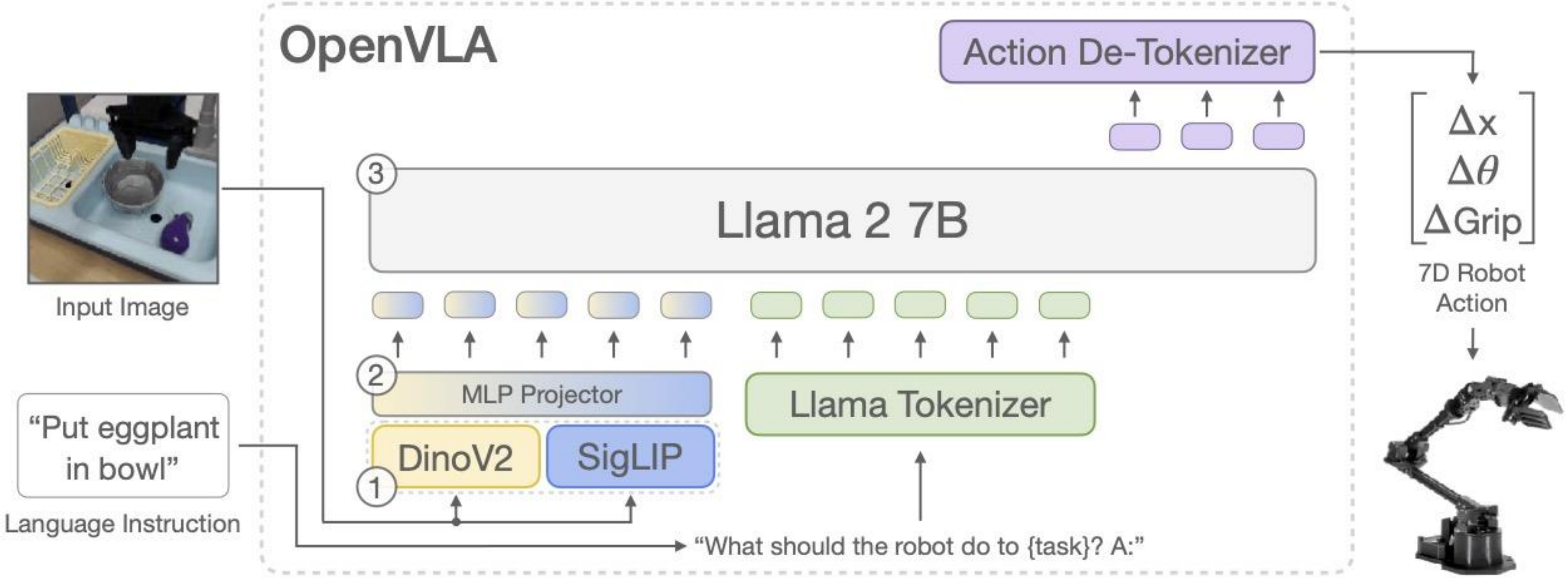
Impressive Generalist Policy, working out of the box





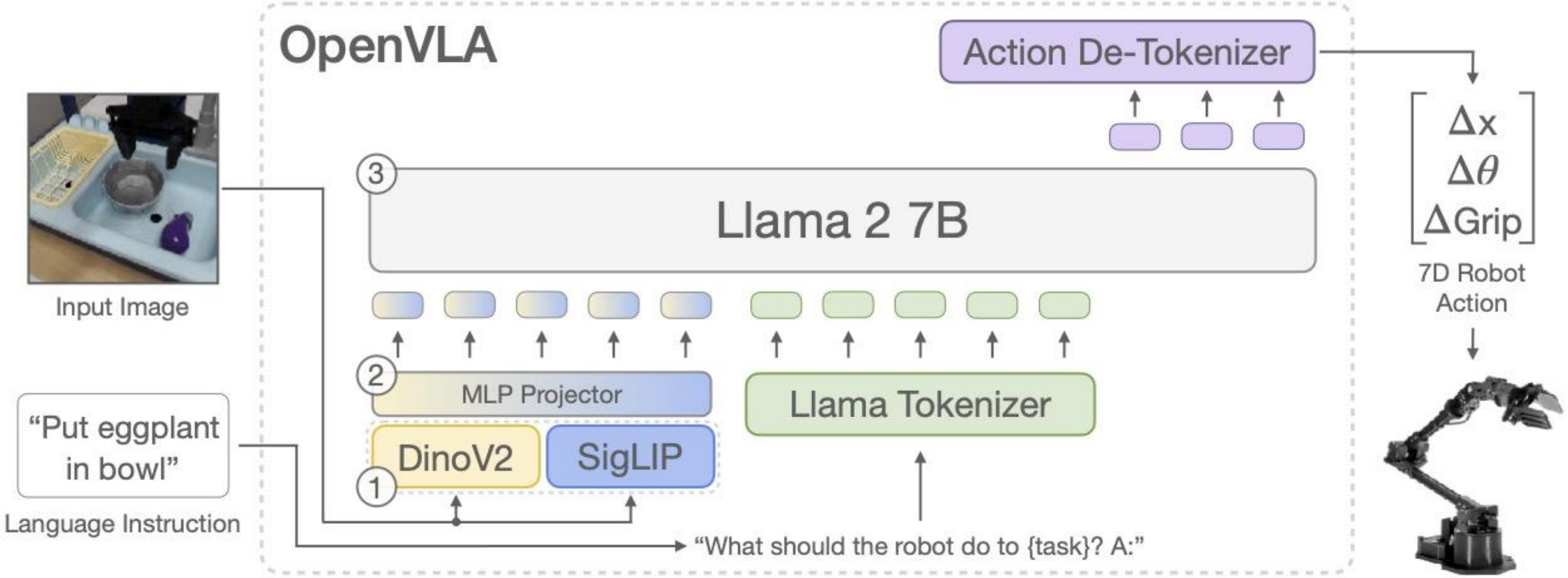
OpenVLA: An Open-Source Vision-Language-Action Model. Kim et al.

Some Obvious Shortcomings



Some Obvious Shortcomings

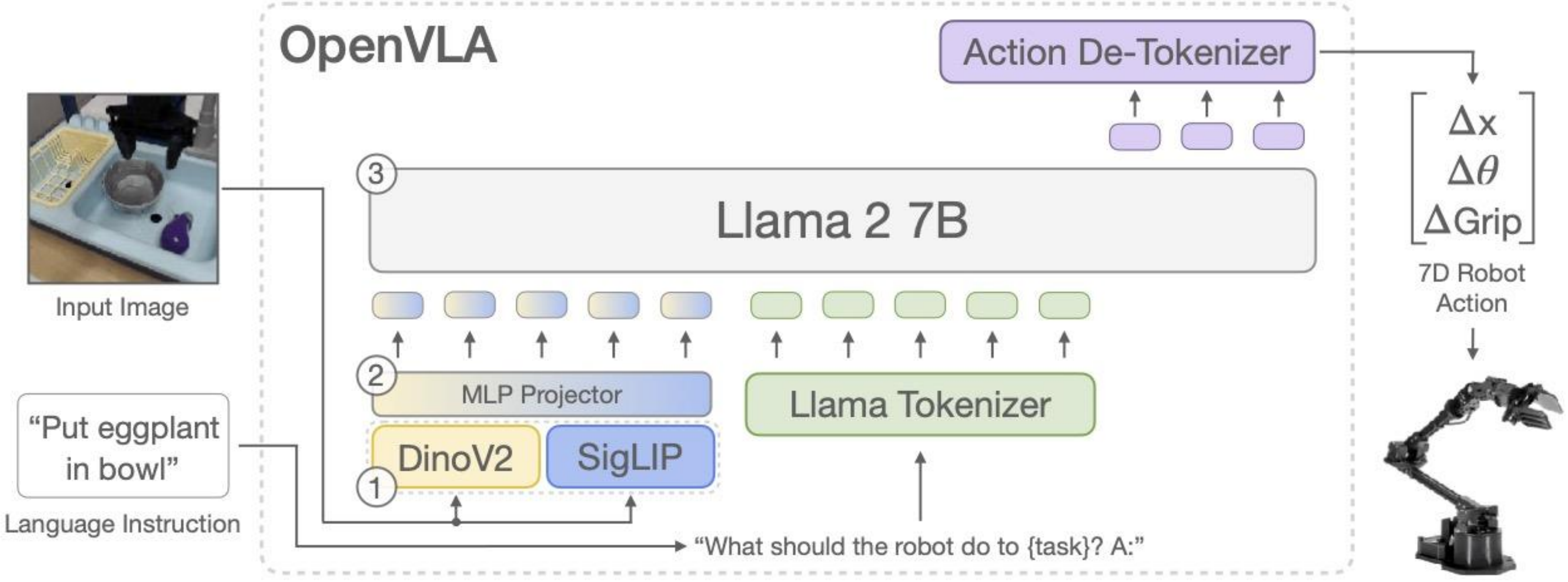
(a) A single image is not sufficient to capture the historical observations



Some Obvious Shortcomings

(b) Action prediction is shortsighted and has limited annotations

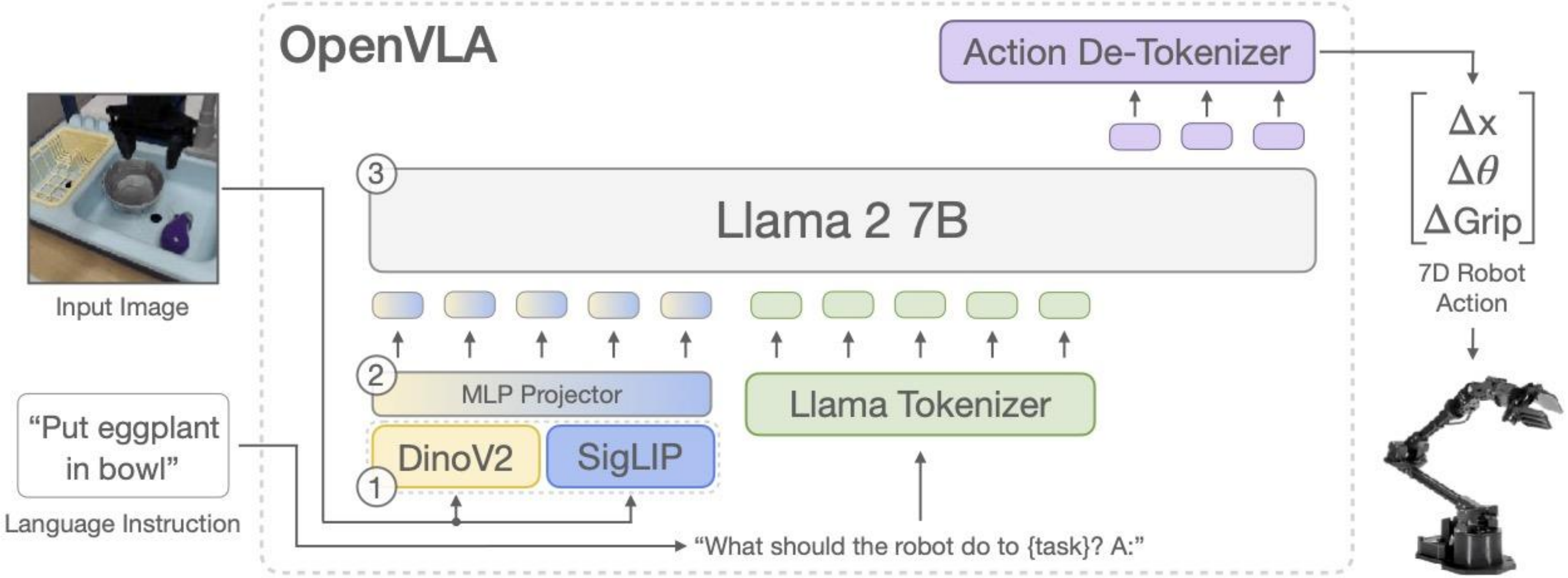
(a) A single image is not sufficient to capture the historical observations



Some Obvious Shortcomings

(b) Action prediction is shortsighted and has limited annotations

(a) A single image is not sufficient to capture the historical observations



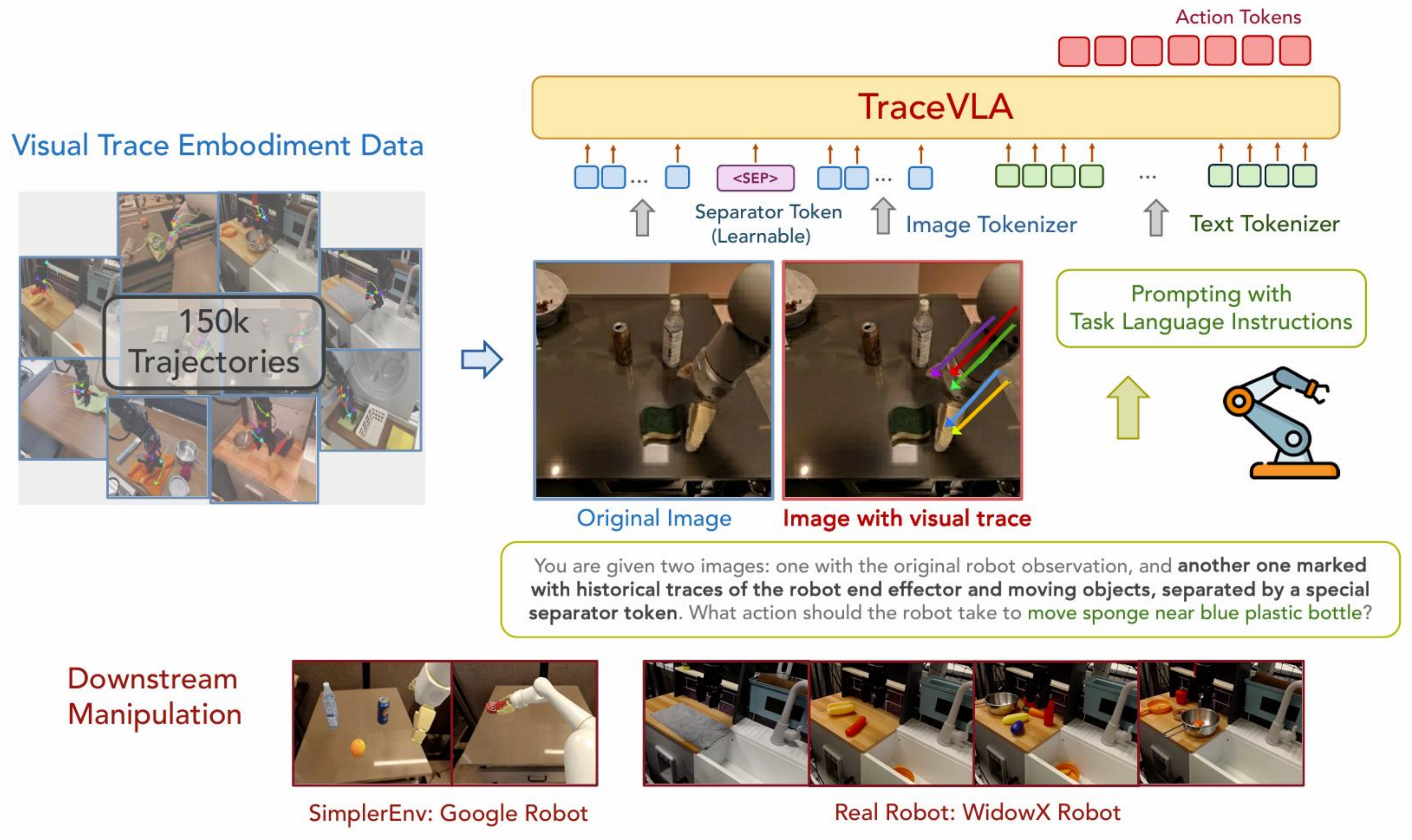
(c) Training only on robotics data (image + goal->action token) may lead to overfit and lose the multimodal understanding and other capabilities

Some Obvious Shortcomings

(a) A single image is not sufficient to capture the historical observations

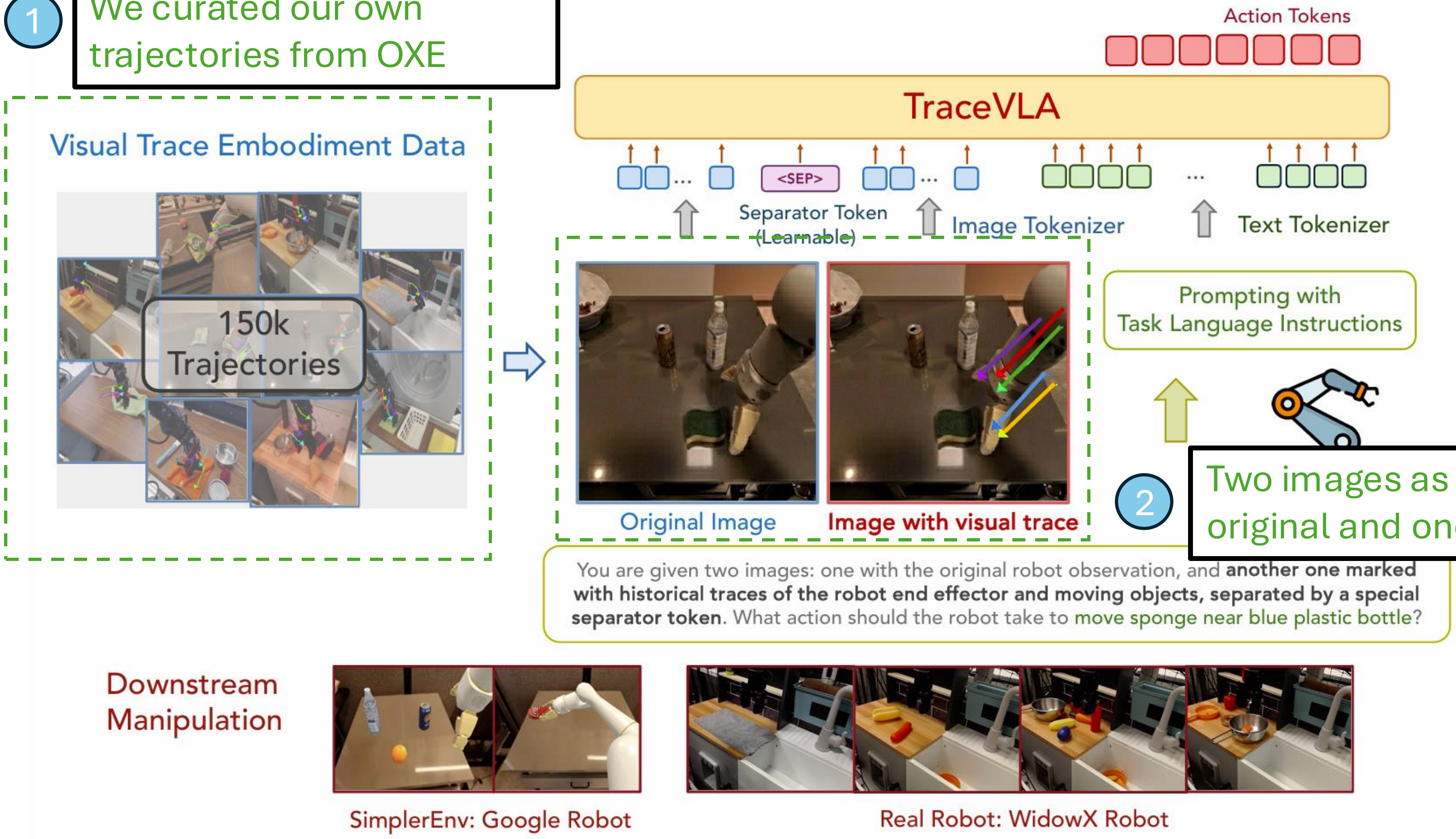
TraceVLA: Visual Trace Prompting for VLA

(a) A single image is not sufficient to capture the historical observations



TraceVLA: Visual Trace Prompting for VLA

1 We curated our own trajectories from OXE



2 Two images as inputs, one original and one prompted

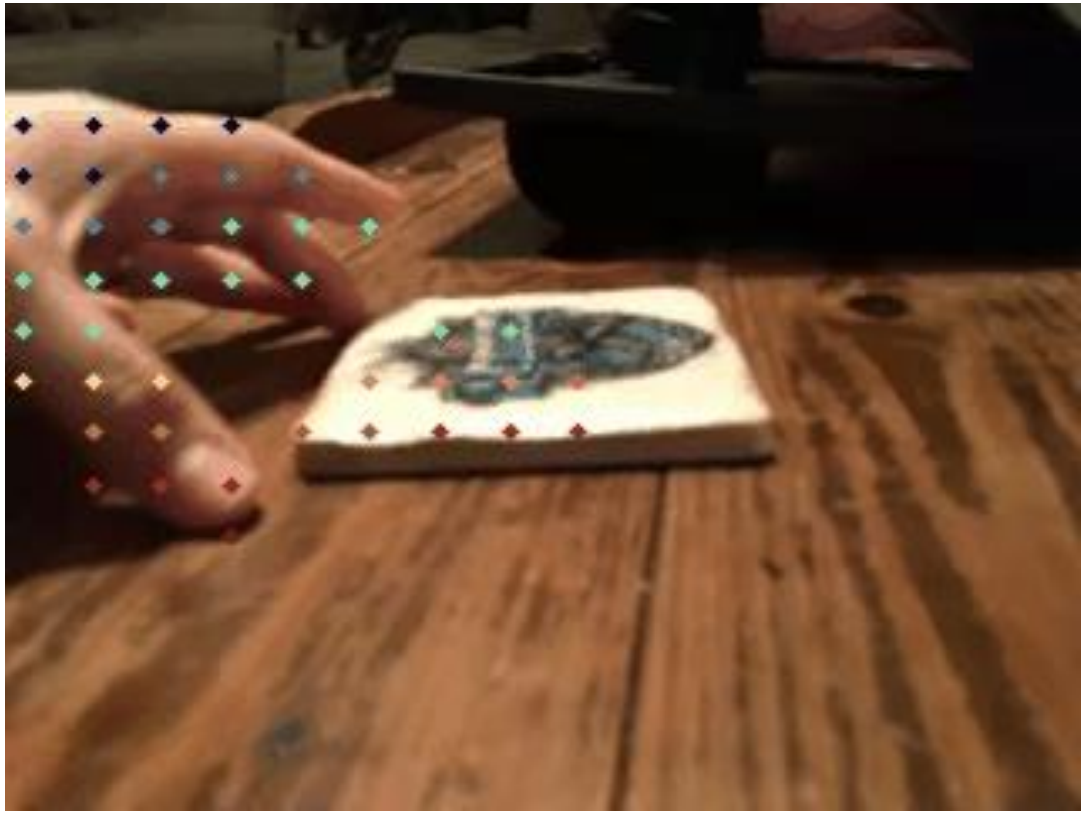
How to encode historical information?

<u>Intuition</u>: A better understanding of the history facilitates better prediction of the future actions

Multi-Images	Text Prompt	Visual Prompt
• Significantly increase the number of visual tokens to VLA. • Vision encoder can hardly capture the subtle changes or motions across adjacent frames.	• Easy to use but can hardly capture the fine-grained movement. • Model is difficult to ground the textual description with spatial-temporal observation.	• Prove to be effective, e.g., Set-of-Mark (SoM) for static image. • Naturally ground the current observation and history spatially and temporally.

TraceVLA. Zheng et al.

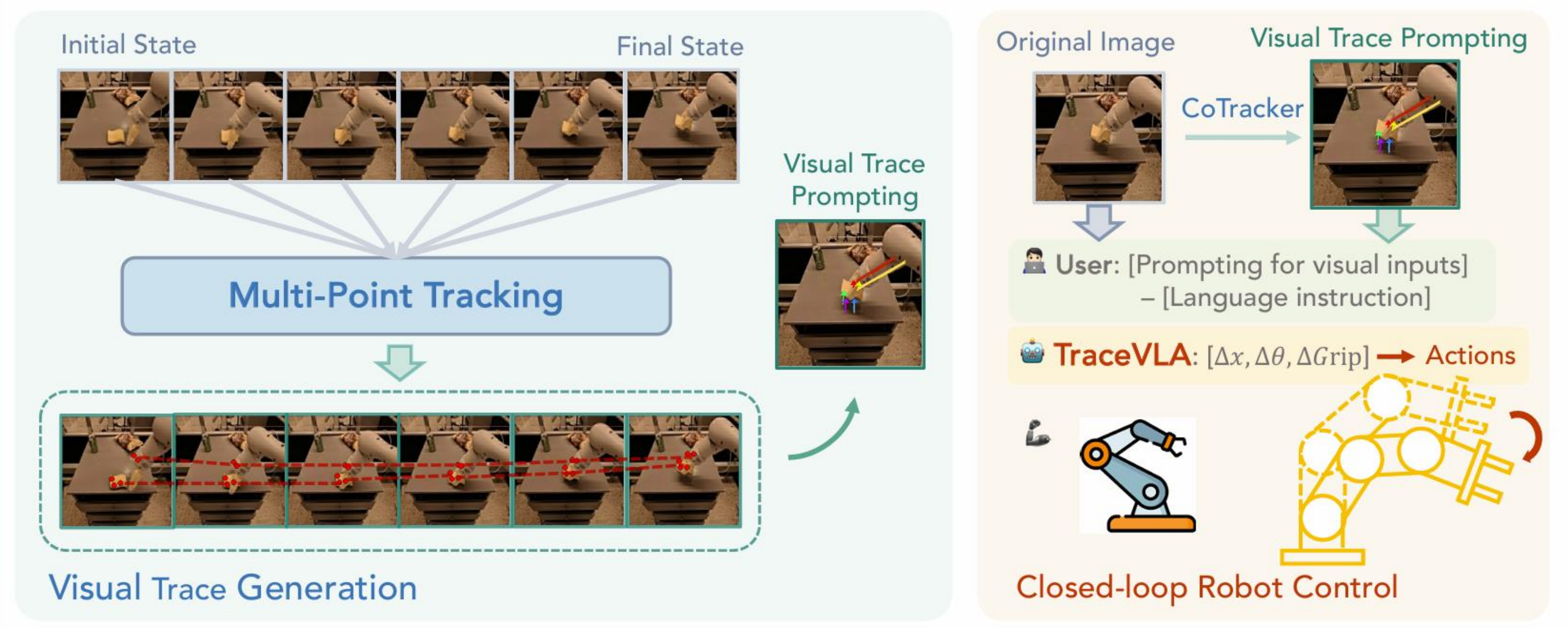
Point trace captures the spatial-temporal dynamics



CoTracker is used to extract the motion traces for keypoints on both objects and robot arms.

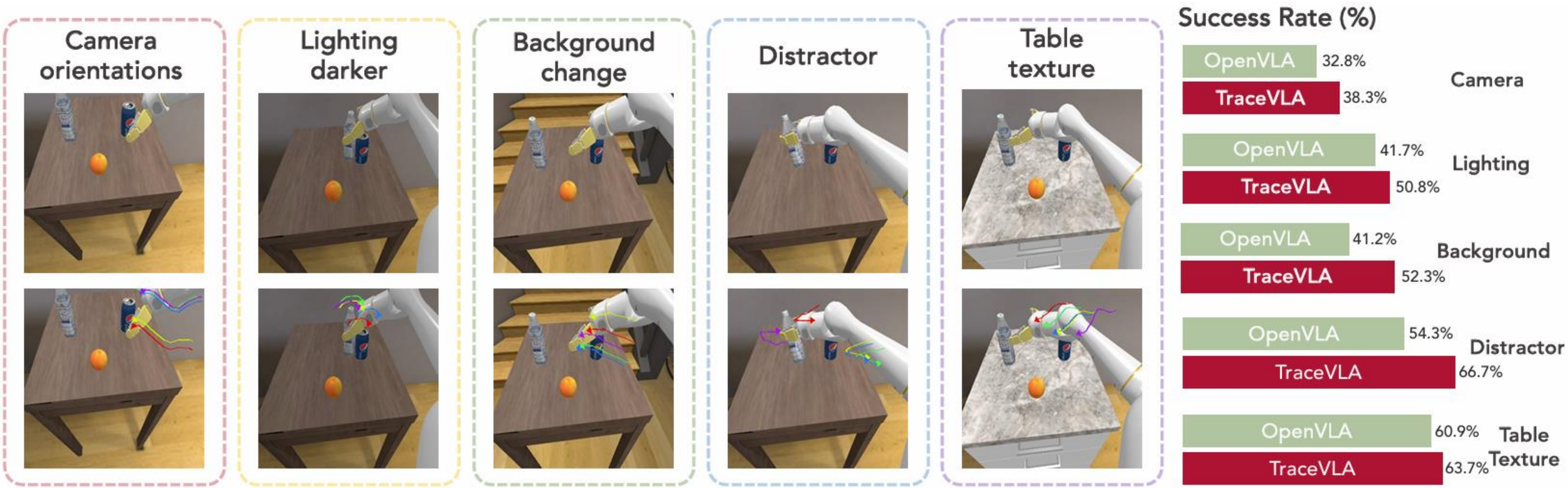
TraceVLA: Visual Trace Prompting

Step 1: Extract visual traces from video sequence **Step 2**: Overlay visual traces on top of the image visually



TraceVLA. Zheng et al.

SimplerEnv Google Robot Tasks

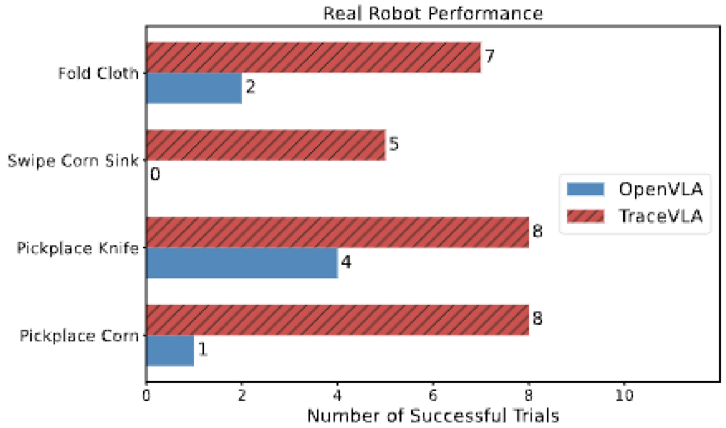


TraceVLA outperforms OpenVLA with clear margin across various domains

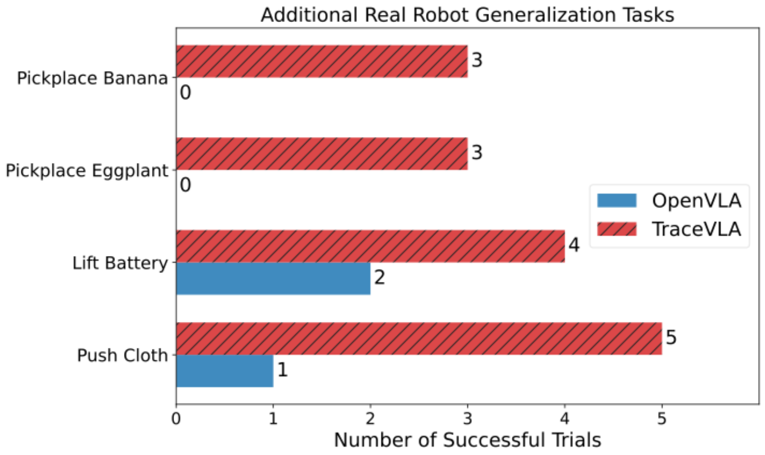
WidowX 250 Robot Arm

Real Robot Manipulation Evaluation

Seen tasks

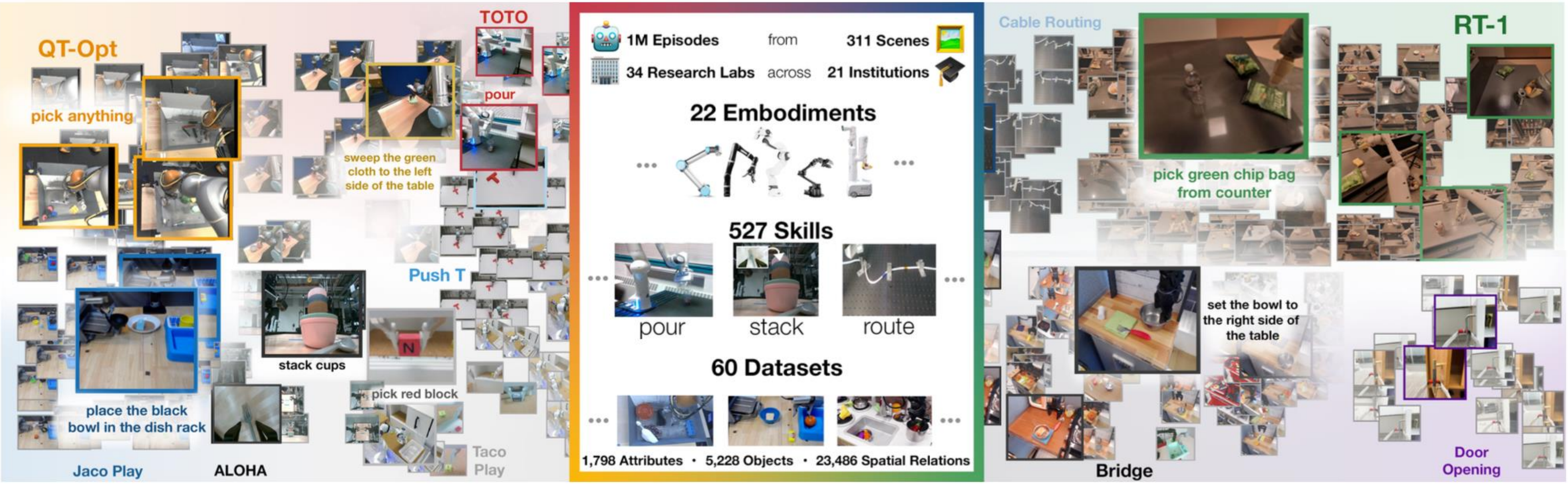


Unseen tasks



TraceVLA. Zheng et al.

Open X-Embodiment

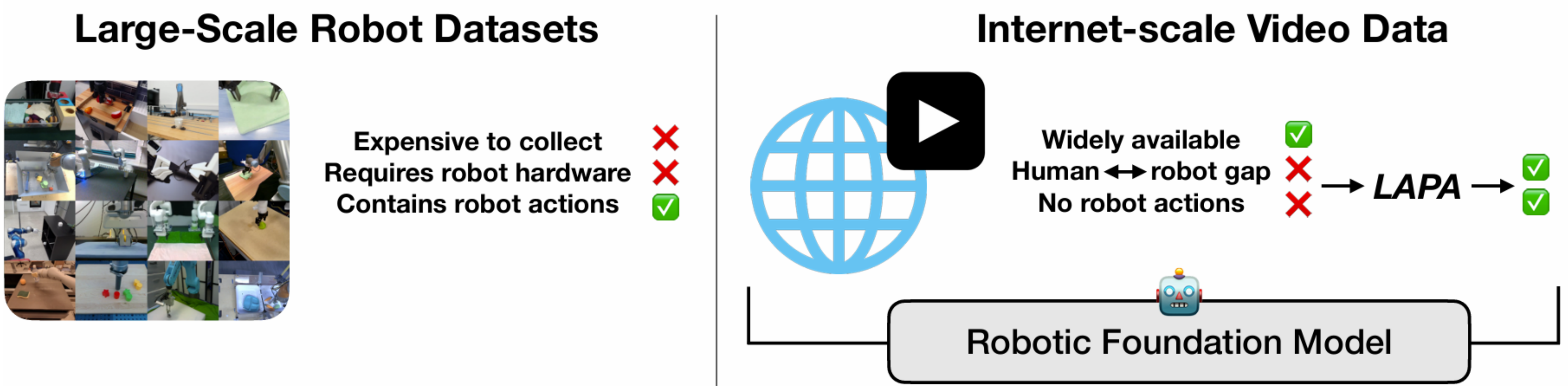


Open X-Embodiment is the largest open-sourced real robotics datasets, but:

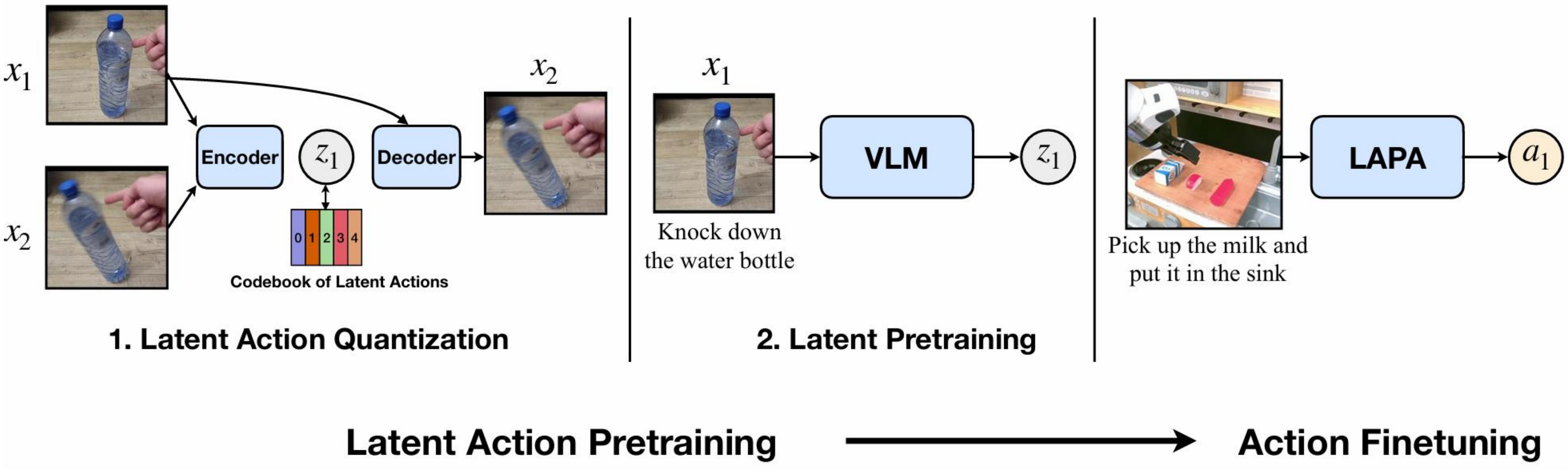
1. Object categories and diversity are limited.
2. Less to no environment variations for each embodiment.
3. Not easy to scale up (collecting real robot data is a headache).

How can we train a robotic foundation model from human videos?

Motivation


Large-Scale Robot Datasets
Expensive to collect
Requires robot hardware
Contains robot actions
Internet-scale Video Data
Widely available
Human ↔ robot gap
No robot actions
LAPA
Robotic Foundation Model

Pretraining Pipeline



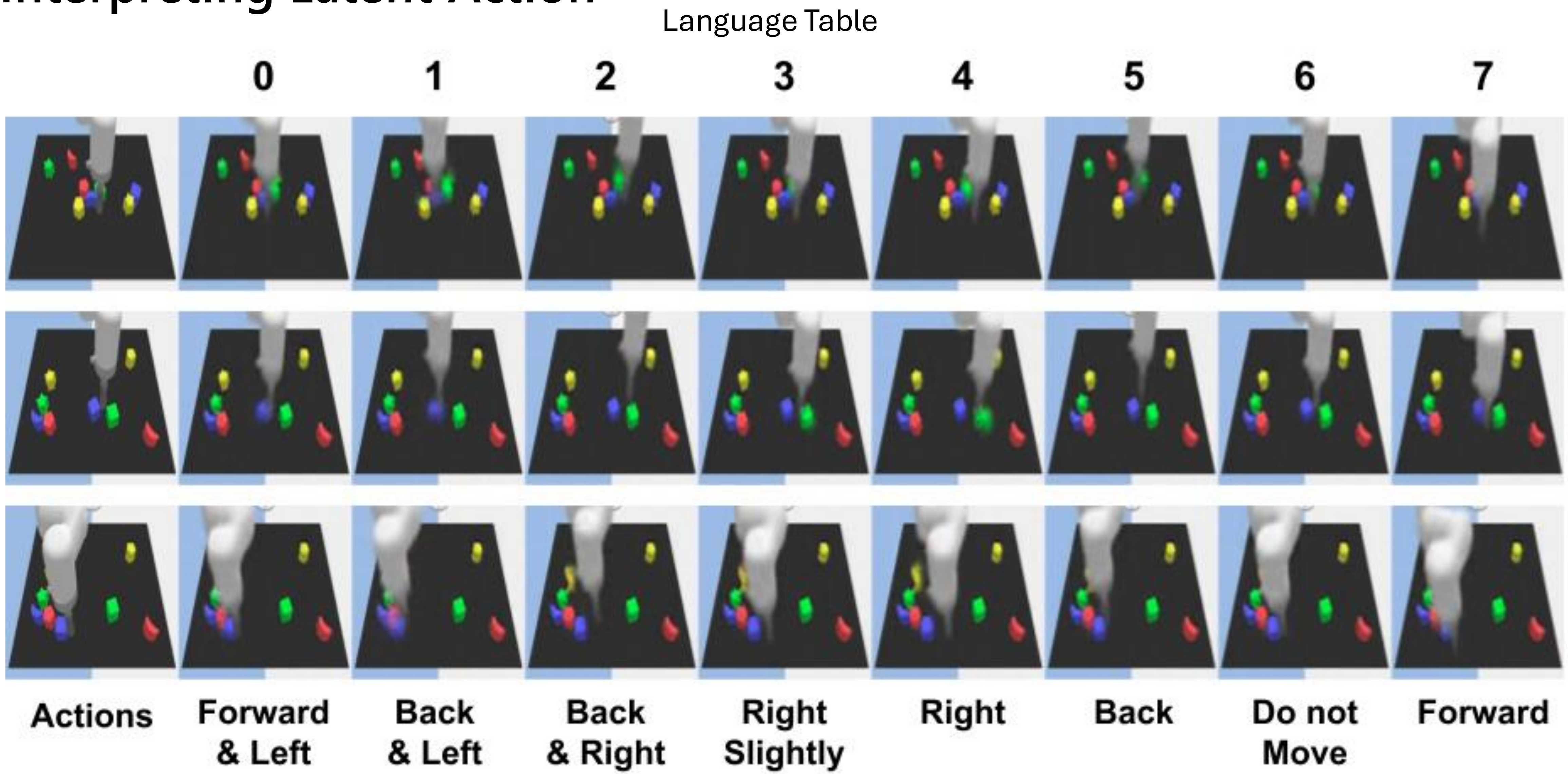
Model arch.: C-ViViT as the video tokenizer

Training method.: VQ-VAE + NSVQ

Model arch.: Large World Model (LWM-7B)

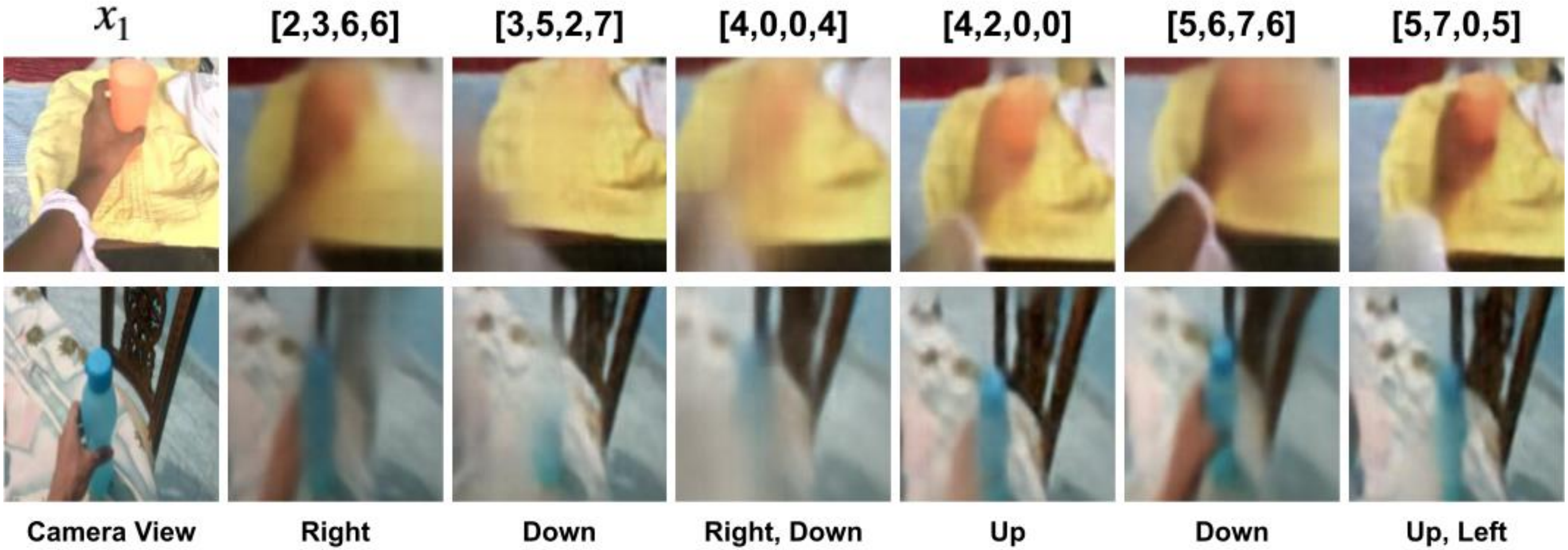
Training method: Next latent/real action token prediction

Interpreting Latent Action



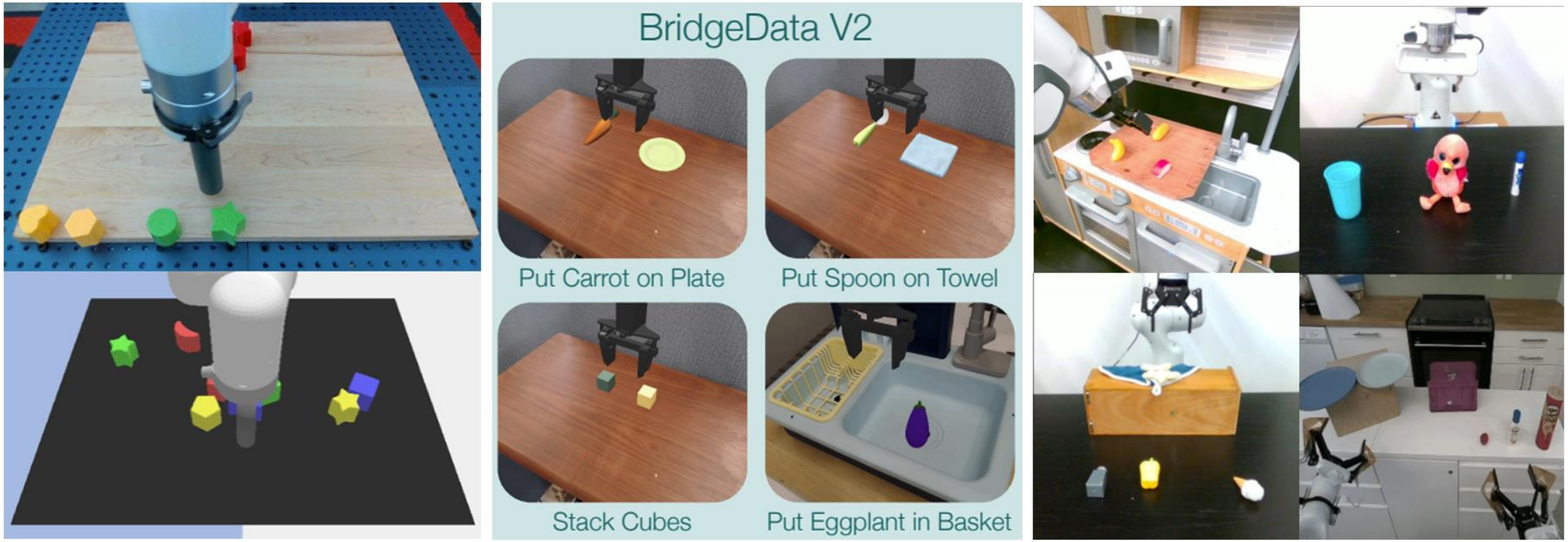
Interpreting Latent Action

Human Video Input



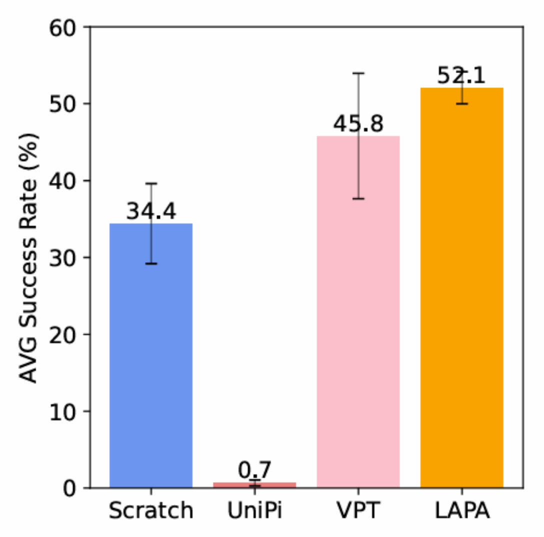
For realistic video input, the action is still interpretable but generated future frames are much blurred

Training and Evaluation Environments

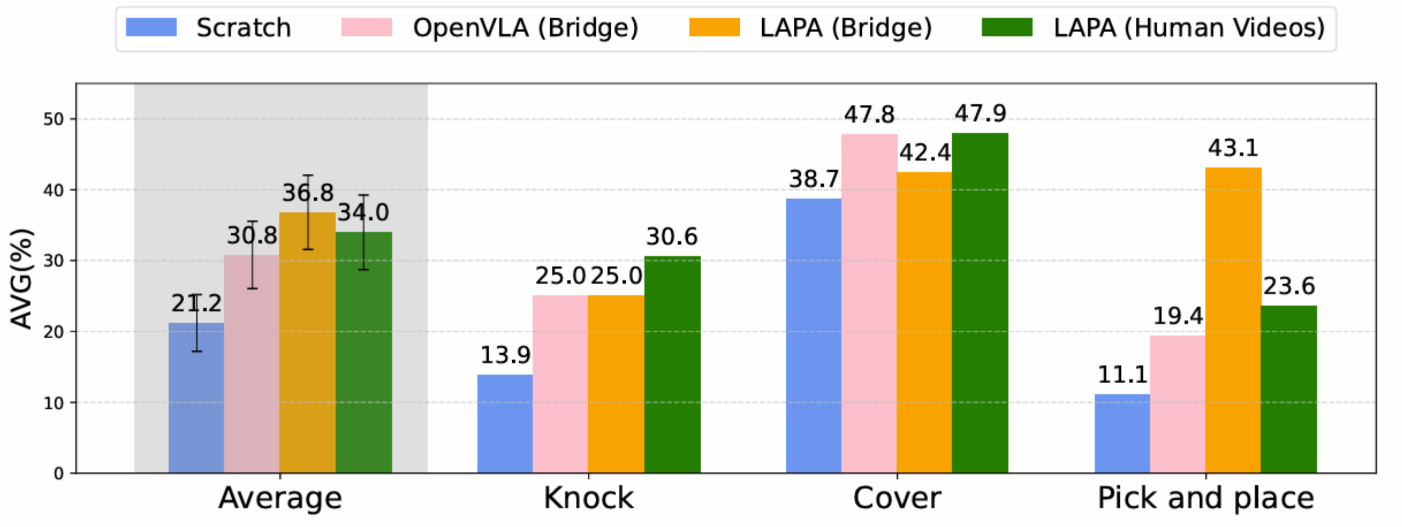


(a) LANGUAGE TABLE (b) SIMPLER (c) REAL

LAPA Learns from Human Videos



(a) SIMPLER Results



(b) Real-world Tabletop Manipulation Robot Results

* We used sthv2 as the human instructional videos

LAPA learned from human videos shows promising results for robotics tasks

It significantly outperforms model trained from scratch

Still a gap with the model pretrained on robotics data w/ or w/o action labels

Seen Objects, Unseen Combinations

Knock mustard down

Scratch

❌

OpenVLA

❌

LAPA

✅

Seen Objects, Unseen Combinations

Knock an object for cleaning

Scratch

❌

OpenVLA

❌

LAPA

✅

Human as an embodied agent do not only produce actions

1. Can we build a single foundation model for **both multimodal understanding and action prediction**?

Not fall back to build a specialized model that only produce action or textual tokens
A good understanding should be beneficial to action-taking, and vice versa

2. How to endow both **verbal and spatial intelligence** for the model? How do we scale up our embodied pretraining?

Video is definitely a good source to learn (as shown our LAPA as well)!
How to transfer the spatially grounded thinking to spatially grounded action?

Magma: A Foundation Model for Multimodal AI Agents

CVPR 2025

Jianwei Yang[*1†] Reuben Tan[1†] Qianhui Wu[1†] Ruijie Zheng[2‡] Baolin Peng[1‡] Yongyuan Liang[2‡]
Yu Gu[1] Mu Cai[3] Seonghyeon Ye[4] Joel Jang[5] Yuquan Deng[5] Lars Liden[1] Jianfeng Gao[1▽]

[1]Microsoft Research [2]University of Maryland [3]University of Wisconsin-Madison [4]KAIST [5]University of Washington
* Project lead. †First authors. ‡Second authors. ▽Leadership

ArXiv | Code | Hugging Face Model | Azure AI Foundry | Twitter / X | Youtube Video

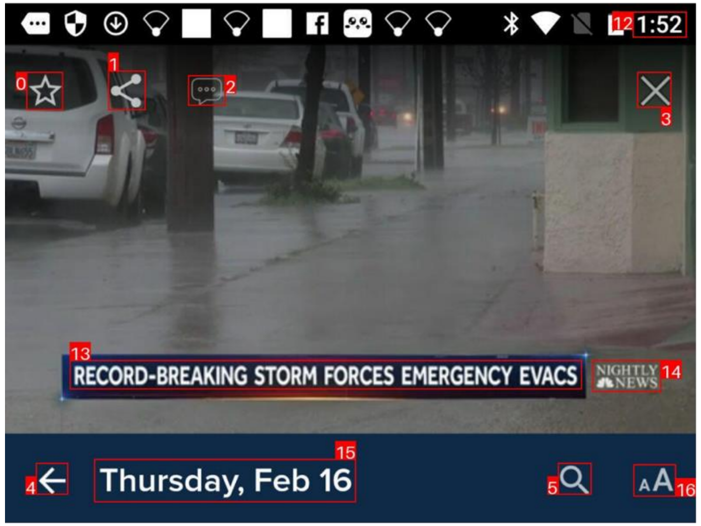
Magma-UI | Magma-Gaming

Poster Location: ExHall D Poster #340, June 14


User: What is they doing?
Magma: They are camping in a forest and hiking on a rocky mountain.
User: On what date did the individual leave a place that Simon thought was very important to him?
Magma: The date is June 9, 2021.
Temporal
User: The figure represents a 5x5 grid containing various animals, where each 1x1 square is considered a block and each block contains an animal from ['cat', 'dog', 'elephant', 'giraffe', 'rabbit']. What is the animal of the block located at the first row, second column of the grid?
Magma: The animal is a cat.
Spatial
User: What is unusual thing about this image?
Magma: The unusual aspect of this image is that a man is ironing clothes on the back of a yellow taxi. ....
User: How to increase the sound?
Magma: Tap on the "Volume" icon. Then, slide the volume up to increase the volume.
Multimodal Understanding
Task: Push the chip bag to left
Task: Put the can to left
Task: Peel the potato
Task: Stir the food in the pot
Planning
Task: Plan a visit to the museum
Task: Lock Screen
Task: View my Account
Multimodal Action Grounding
User: Close the middle drawer
Magma: Trajectory Rollout on SimplerEnv Google Robot
User: Pick up the mushroom to the pot
Magma: Trajectory Rollout on Widow-X Real Robot
User: Find baby shoes under 20 dollars and rated 5 stars.
Magma: Trajectory on Mind2Web
Task: Install app "Instagram"
Magma: Trajectory on AITW
Multimodal Agentic Taks
Physical Environment
Digital Environment

Set-of-Mark for Action Grounding

UI Navigation



Task: Swipe until Thursday, Feb 16.

Set-of-Mark: Mark 15 for bounding box at coordinate: [33, 95, 130, 105]

Robot Manipulation

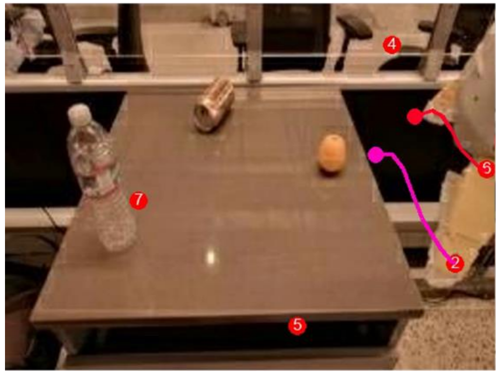
Task: Move the white object above the green cloth.

Set-of-Mark:
Mark 1 at [169,54],
Mark 4 at [152,37],
Mark 8 at [169,86],
Mark 9 at [169,70]

Set-of-Mark: SoM can significantly reduce the search space to take actions, e.g., click a button on the screenshot

Trace-of-Mark for Action Planning

Robot Manipulation

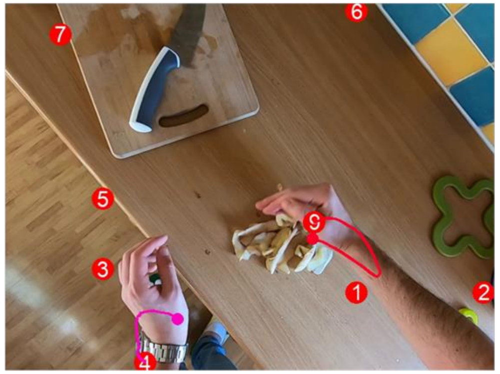


Task: Future 14 steps to move orange near orange can.

Trace-of-Mark:

{
"**Mark 2**":
"[[228,128],[218,112],[212,97],[208,86],[202,79],[193,79],[185,82],[178,85],[172,89],[167,89],[166,86],[166,80],[163,76],[162,78]]“
"**Mark 6**":
"[[248,86],[238,76],[233,65],[229,59],[223,56],[214,59],[203,65],[194,70],[188,74],[184,74],[181,72],[180,69],[178,62]]“
}

Human Action



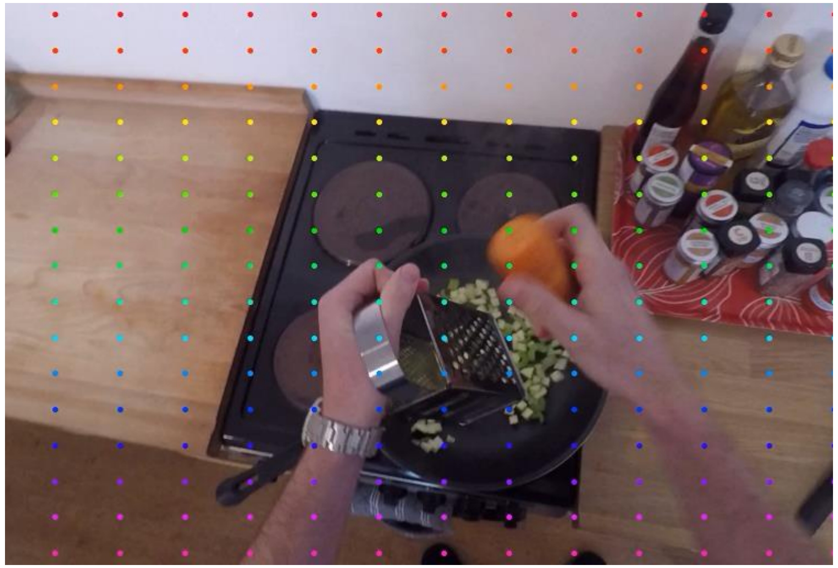
Task: Future 16 steps to move hands towards the pile of potato peels to gather them together.

Trace-of-Mark:

{
"**Mark 4**":
"[[85,248],[84,243],[84,237],[83,230],[83,224],[83,220],[84,217],[86,215],[89,214],[92,215],[96,216],[98,217],[99,219],[99,221],[98,222],[96,225]]",
"**Mark 9**":
"[[157,149],[160,149],[163,151],[166,155],[172,163],[175,173],[177,182],[178,185],[178,190],[176,189],[171,183],[164,174],[156,166],[149,162],[144,160]]“
}

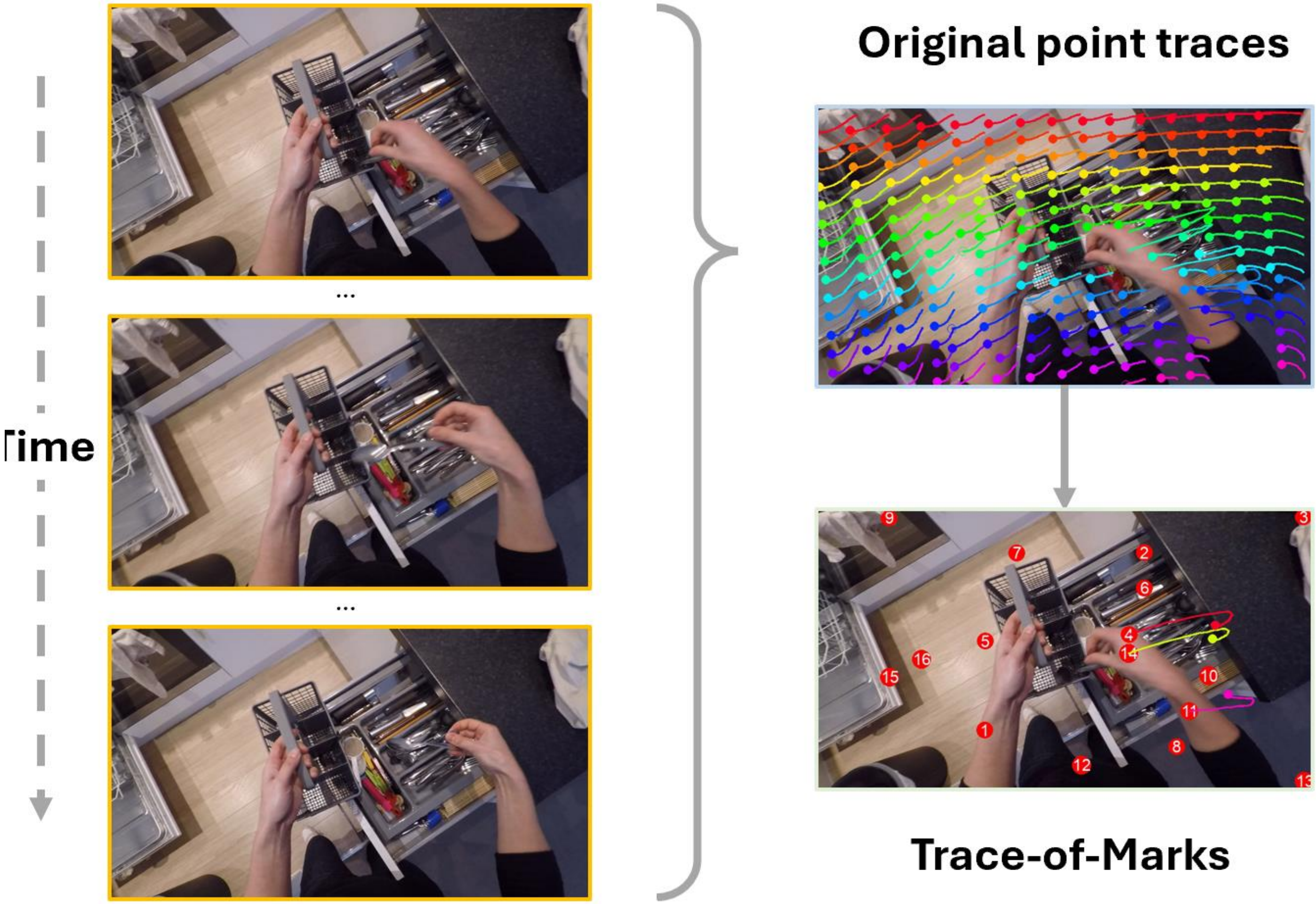
Trace-of-Mark forces the model to learn a longer horizon by predicting distant future “actions”, and more importantly, provides an effective way to leverage unlabeled video data.

Human is a great instructor for embodied agents



Motions are great "action" supervisions, if processed properly

Trace-of-Mark Generation for Videos in the Wild



Algorithm 2 SoM and ToM generation for instructional videos and robotic data

Require: image sequence $\mathcal{I} = \{I_t, ...I_l\}$; grid size s; global motion threshold η; foreground threshold ϵ

$\mathcal{M} = \{M_t, ..., M_l\} \leftarrow \texttt{CoTracker}(\mathcal{I}, s)$

if $\texttt{HasGlobalMotion}(\mathcal{M}, \eta)$ **then**

 $\mathcal{M} \leftarrow \mathcal{H}(\mathcal{M})$ ▷ Apply homography transformation

end if

$\mathcal{M}^f, \mathcal{M}^b = \texttt{ClassifyTraces}(\mathcal{M}, \epsilon)$ ▷ Classify traces into foreground and background ones

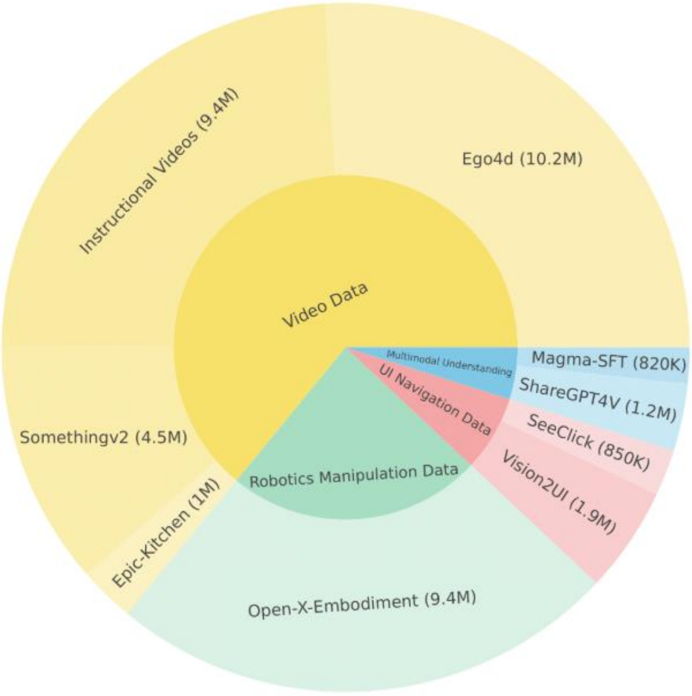
$k \leftarrow \texttt{Random}(1, \min(5, |\mathcal{M}^f|))$

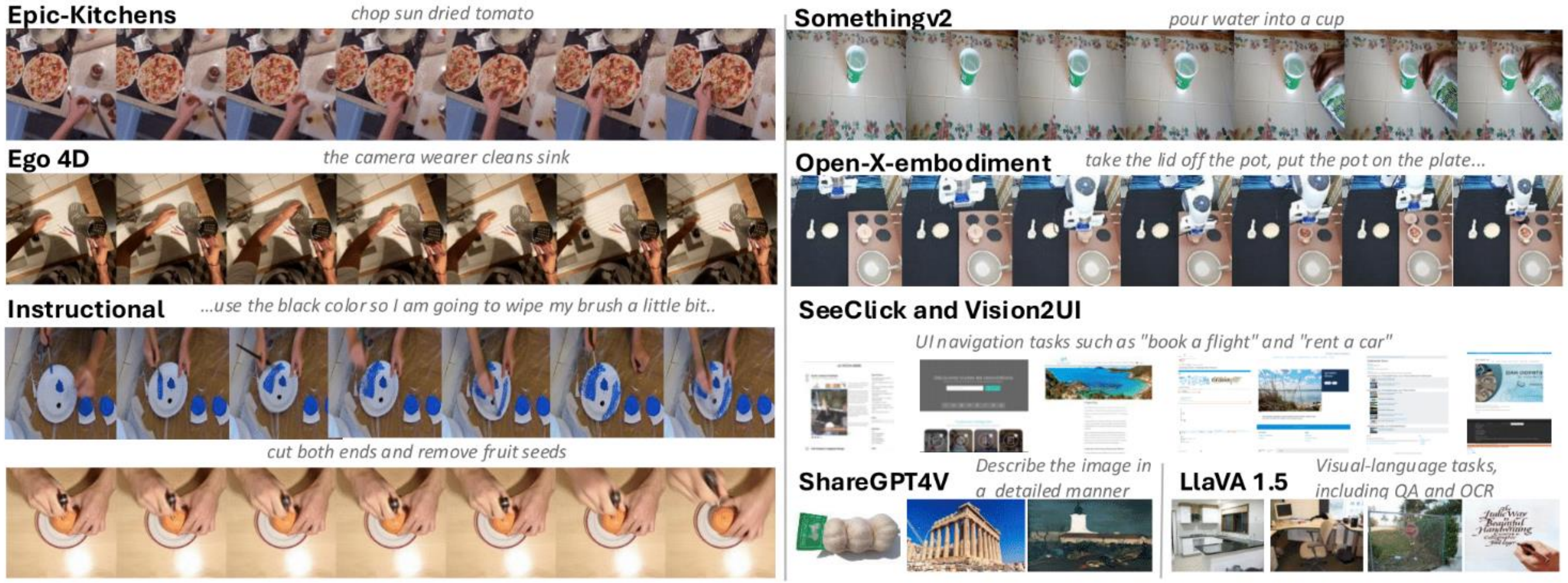
$\mathcal{M}^f, \mathcal{M}^b = \texttt{KMeans}(\mathcal{M}^f, k), \texttt{KMeans}(\mathcal{M}^b, 2k)$ ▷ Cluster foreground and background traces separately

$I_t \leftarrow SoM(I_t, \{M_t^f, M_t^b\})$ ▷ Apply SoM on 1st frame

Return $\mathcal{I}, \mathcal{M}_f^*$

Pretraining on a collection of Images, Videos and Robotics Data

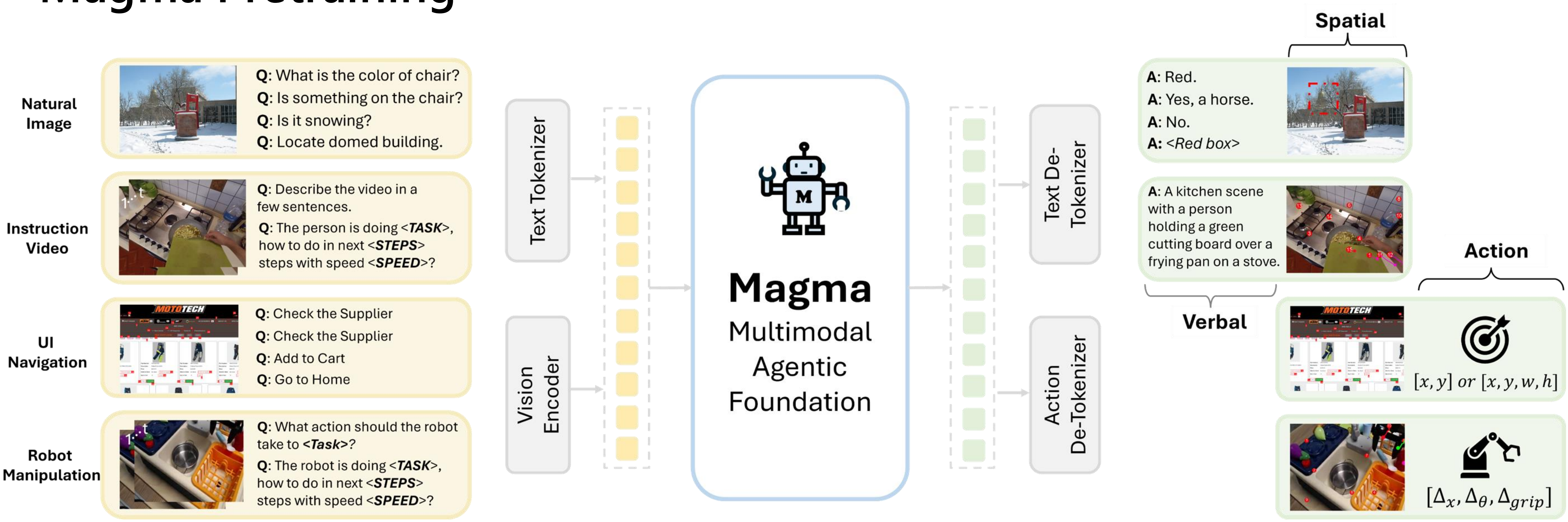




39M Pretraining Data: instructional videos, robotics manipulation, UI navigation, and multimodal understanding.

We apply SoM and ToM for different data types, with SoM enabling unified action grounding across all modalities while ToM is applied to video and robotics data.

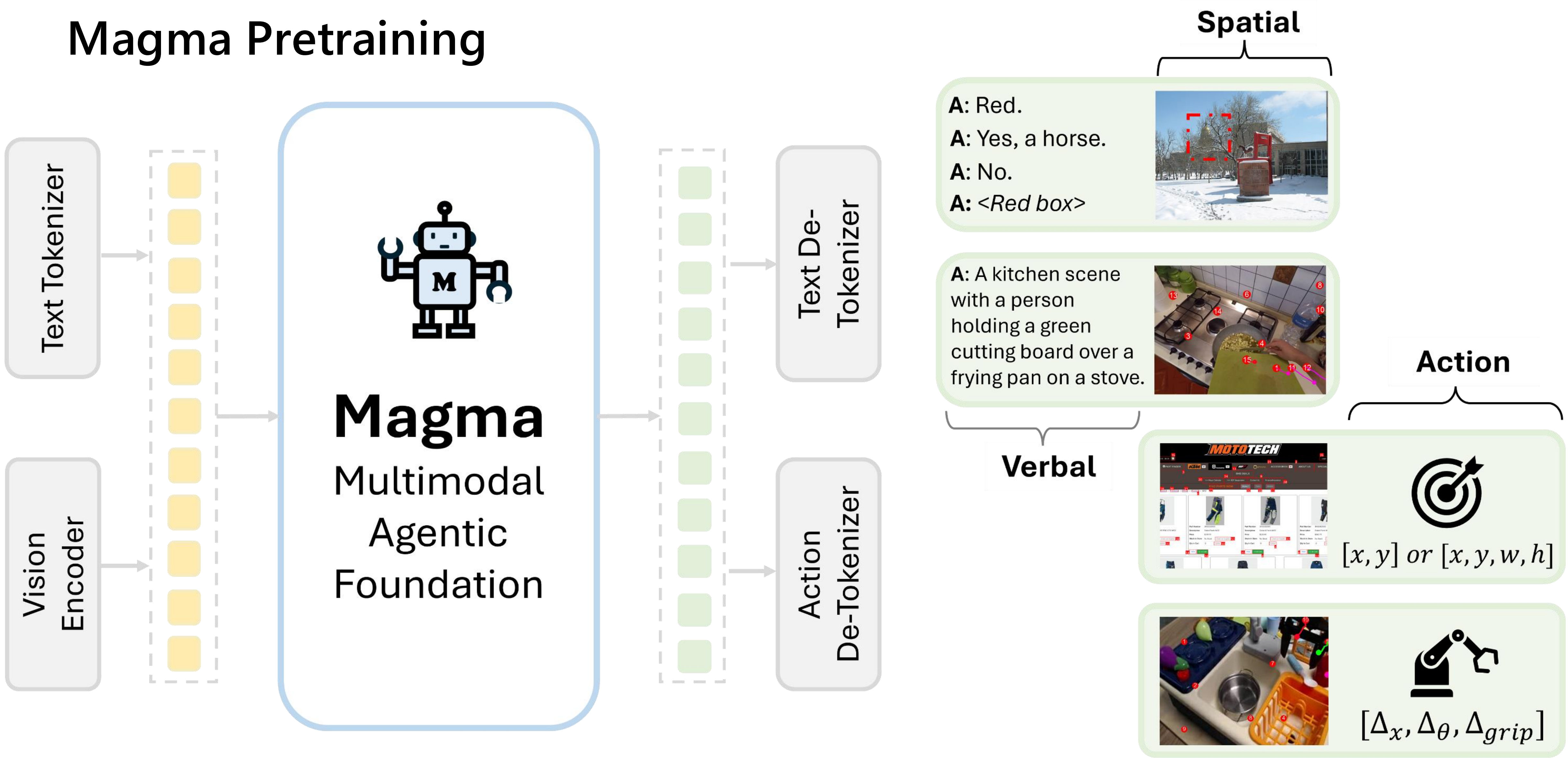
Magma Pretraining



(a) Vision Encoder: ConvNeXt-XXLarge
(b) Language Model: Llama-3-8B

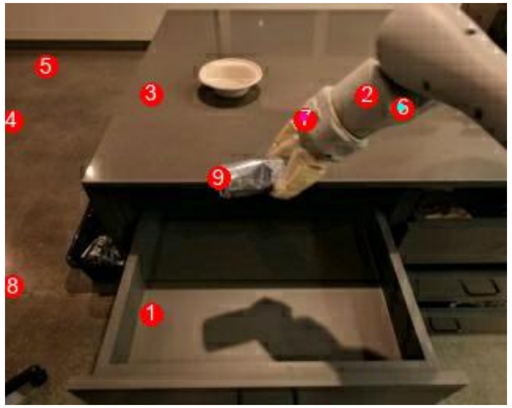
Setup: Pretraining from scratch for 3 epochs on the whole pretraining data

Magma Pretraining

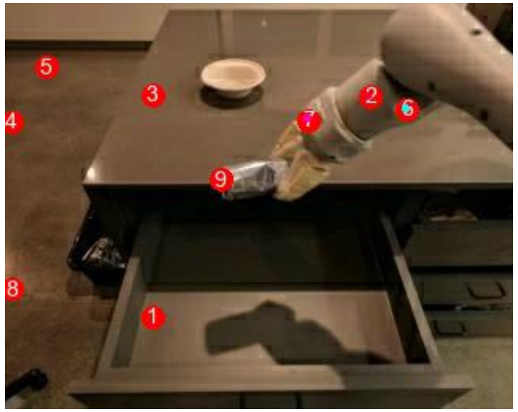


Robot Planning Scaling Property

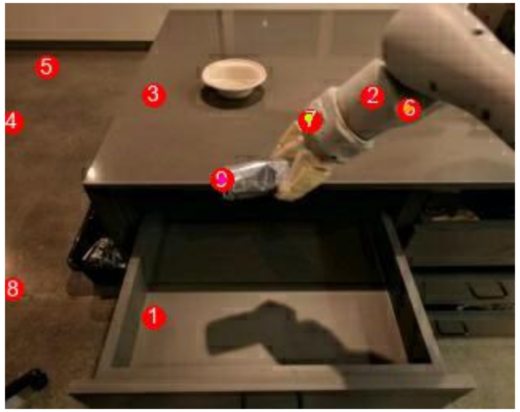
Effect of Data Scaling



Training Data: **1/3**



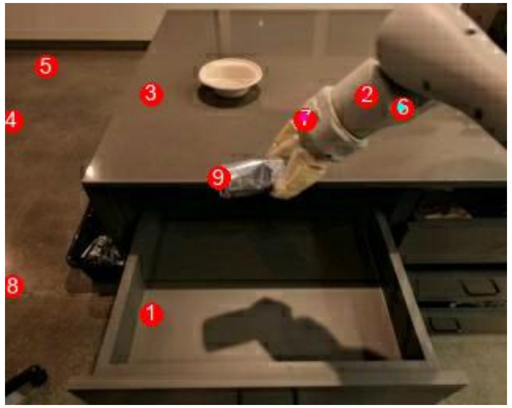
Training Data: **2/3**



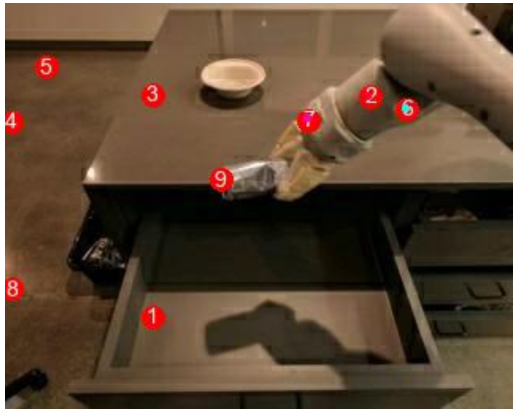
Training Data: **Full**

Robot Planning Scaling Property

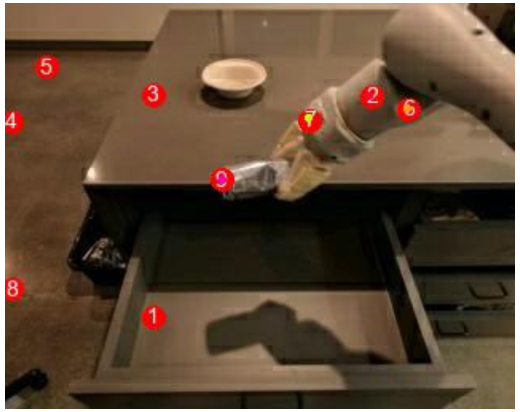
Effect of Data Scaling



Training Data: **1/3**

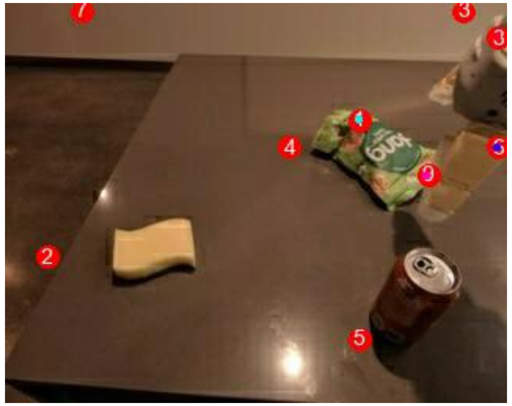


Training Data: **2/3**



Training Data: **Full**

Task Generalization



Pick up the chip bag.

Put the chip bag to coke can.

Push the chip bag to the yellow object.

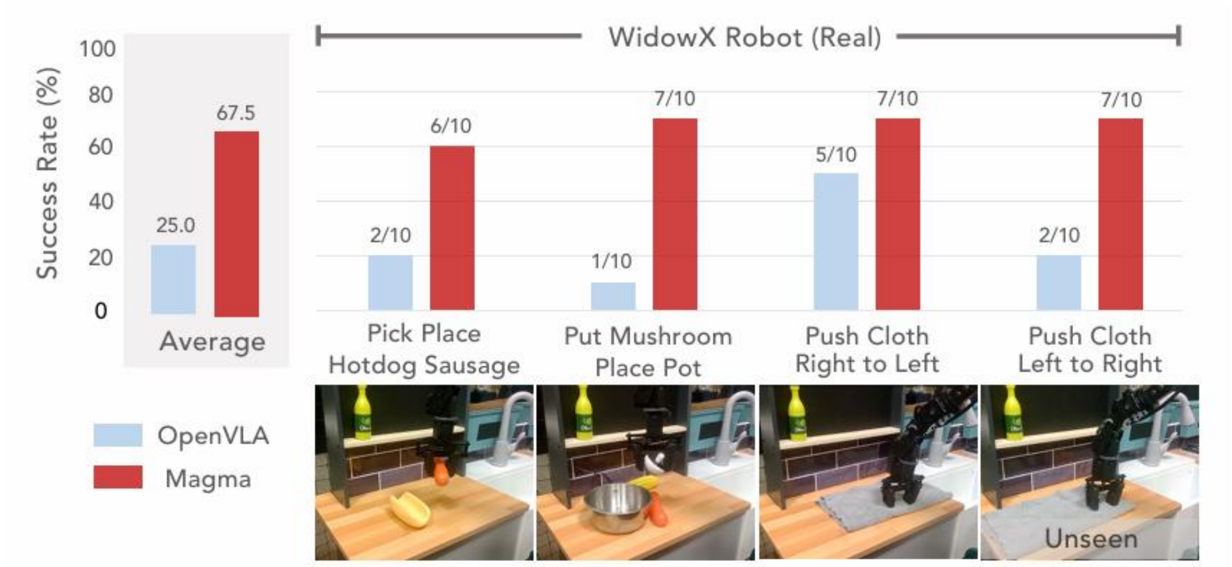
State-of-the-art Performance for Agentic Intelligence

		Multimodal Understanding			UI Action Grounding and Navigation					Robot Manipulation	
Model	Size	VQAv2	TextVQA	POPE	*SS*-Mobile	*SS*-Desktop	SS-Web	VWB-Ele-G	VWB-Act-G	SE-Google Robot	SE-Bridge
GPT-4V [99]	n/a	77.2	**78.0**	n/a	22.6/24.5	20.2/11.8	9.2/8.8	67.5	**75.7**	-	-
GPT-4V-OmniParser [83]	n/a	n/a	n/a	n/a	**92.7**/49.4	64.9/26.3	**77.3**/39.7	-	-	-	-
LLaVA-1.5 [71]	7.4B	78.5	58.2	85.9	-	-	-	12.1	13.6	-	-
LLaVA-Next [75]	7.4B	**81.8**	64.9	86.5	-	-	-	15.0	8.7	-	-
Qwen-VL [3]	9.6B	78.8	63.8	n/a	7.5/4.8	5.7/5.0	3.5/2.4	14.0	10.7	-	-
Qwen-VL-Chat [3]	9.6B	78.2	61.5	n/a	-	-	-	-	-	-	-
Fuyu [4]	8B	74.2	n/a	n/a	41.0/1.3	33.0/3.6	33.9/4.4	19.4	15.5	-	-
SeeClick [19]	9.6B	-	-	-	78.0/52.0	72.2/30.0	55.7/32.5	9.9	1.9	-	-
Octo [113]	93M	-	-	-	-	-	-	-	-	6.0	15.9
RT-1-X [23]	35M	-	-	-	-	-	-	-	-	34.2	1.1
OpenVLA [54]	8B	-	-	-	-	-	-	-	-	31.7	14.5
Magma-8B (Ours)	8.6B	80.0	66.5	**87.4**	60.4/**58.5**	**75.3/52.9**	69.1/**52.0**	**96.3**	71.8	**52.3**	**35.4**

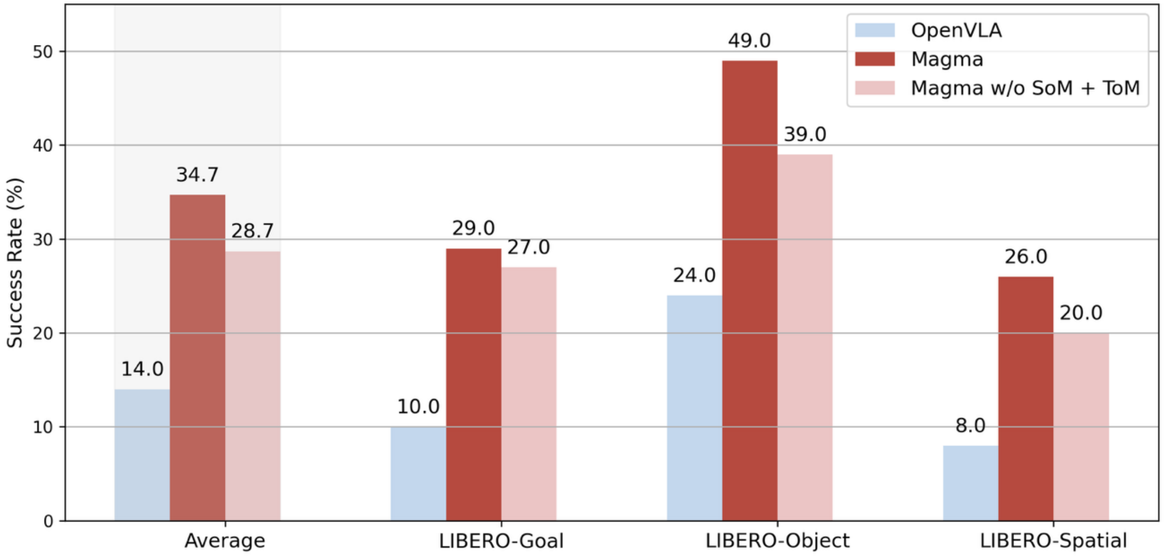
Zero-shot evaluation on agentic intelligence. We report the results for pretrained Magma without any domain-specific finetuning. Magma is the **only model** that can conduct the **full task spectrum**.

Superior Performance for Both Understanding and Action

Evaluated on Real-World Agentic Tasks



Magma outperforms OpenVLA significantly on **real-world robot** tasks (67.5 *v.s.* 25.0)



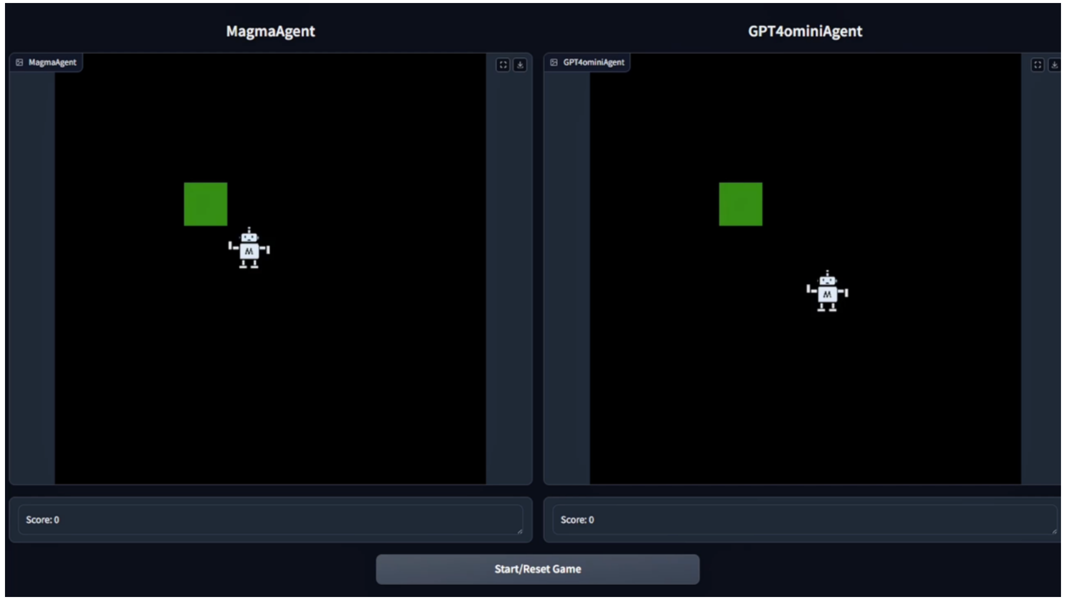
Magma show much better few-shot learning performance than OpenVLA on LIBERO

Superior Performance for Spatial Understanding

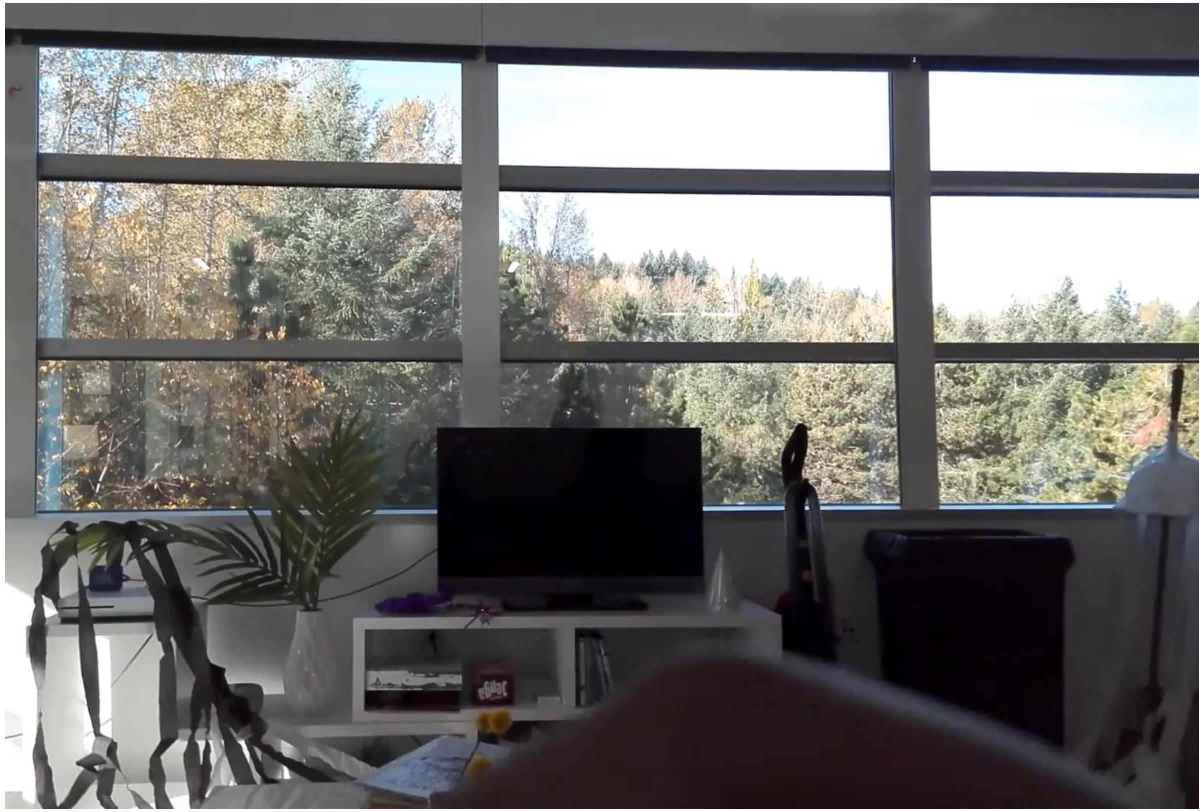
Model	VSR	BLINK-val	SpatialEval[2] Spatial Map	SpatialEval[2] Maze Nav.	SpatialEval[2] Spatial Grid
GPT-4o	74.8	60.0	-	-	-
Gemini	-	61.4	-	-	-
LLaVA-1.5-7B	57.1*	37.1	28.4	28.8	41.6
LLaVA-1.6-7B [75]	52.2*	-	28.0	34.8	32.2
Qwen-VL-9.6B [3]	-	40.3	28.7	31.8	25.7
Magma-8B (Act$^{w/o}$)	62.8	30.1	36.9	**44.8**	37.5
Magma-8B (Full$^{w/o}$)	58.1	38.3	27.5	33.5	47.3
Magma-8B (Full)	**65.1**	**41.0**	**43.4**	36.5	**64.5**

Magma can achieve very god spatial understanding and reasoning capabilities.

Taks: Collect Green Blocks



We ask model to select top, bottom, left or right to move closer to the green block



User: I really want to win this game. Can you suggest the next move that I should make to maximize my chances of winning?
Magma: I suggest **moving the pawn to d4** to maximize your chances of winning.

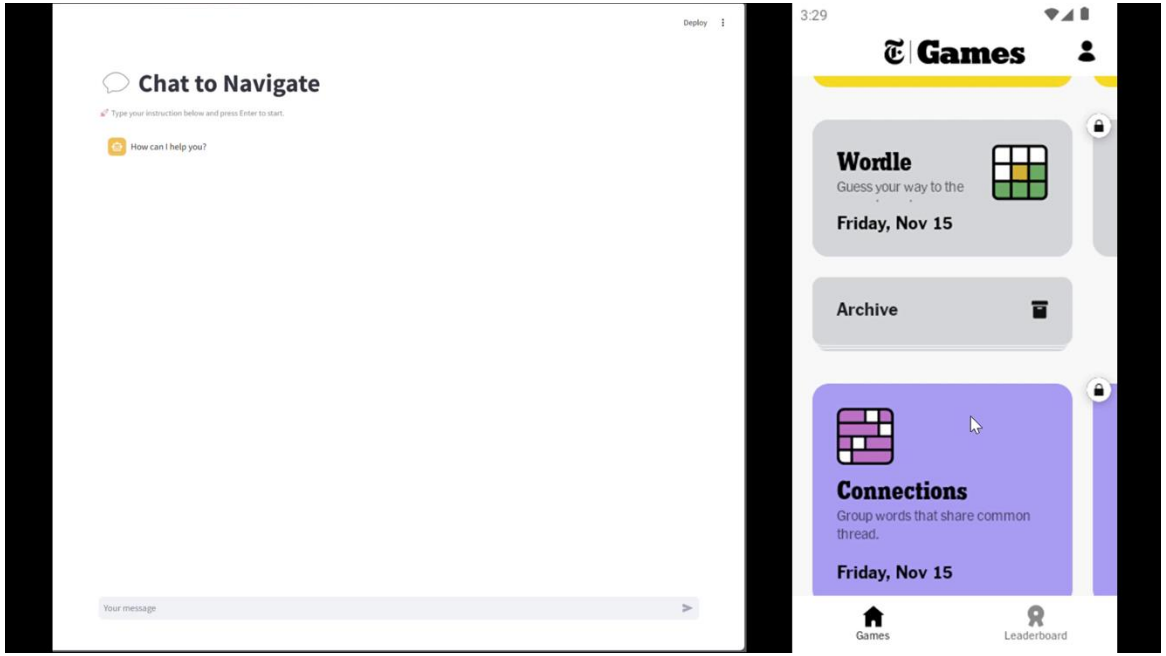
User: I want to relax for a few hours in here. Can you suggest some things for me to do?
Magma: Watch TV, read a book, and enjoy the view of the autumn forest.

Magma

OpenVLA

(Seen task) (Unseen task)

A Bonus of Magma: UI Navigation



Takeaways

- **VLA models should not only look at current frame:**
  - Multi-frames is a straightforward method
  - Visual trace prompting to encode spatial-temporal context efficiently

- **Video is a gold mine for learning generic VLA models:**
  - Implicit: Latent action from human videos
  - Explicit: visual trace dynamics from human motions

- **Embodied AI models should NOT only output action:**
  - Verbal tokens for multimodal understanding
  - Trace tokens for enhanced temporal understanding and long-horizon awareness

An embodied agent that can see, think and act

Vision Foundation	Multimodal LLM	Embodied Agent
See	**Think**	**Act**

- Open-vocabulary
- Spatially grounded
- Versatile inputs & outputs

A multimodal vision model can effortlessly apply in the wild

- Understand and reasoning
- Free-form instruction
- Think spatially on image

A MLLM can understand the instruction and ground on the observations

- Perceive and act
- Steerable to various tasks
- Spatial and temporal intelligence

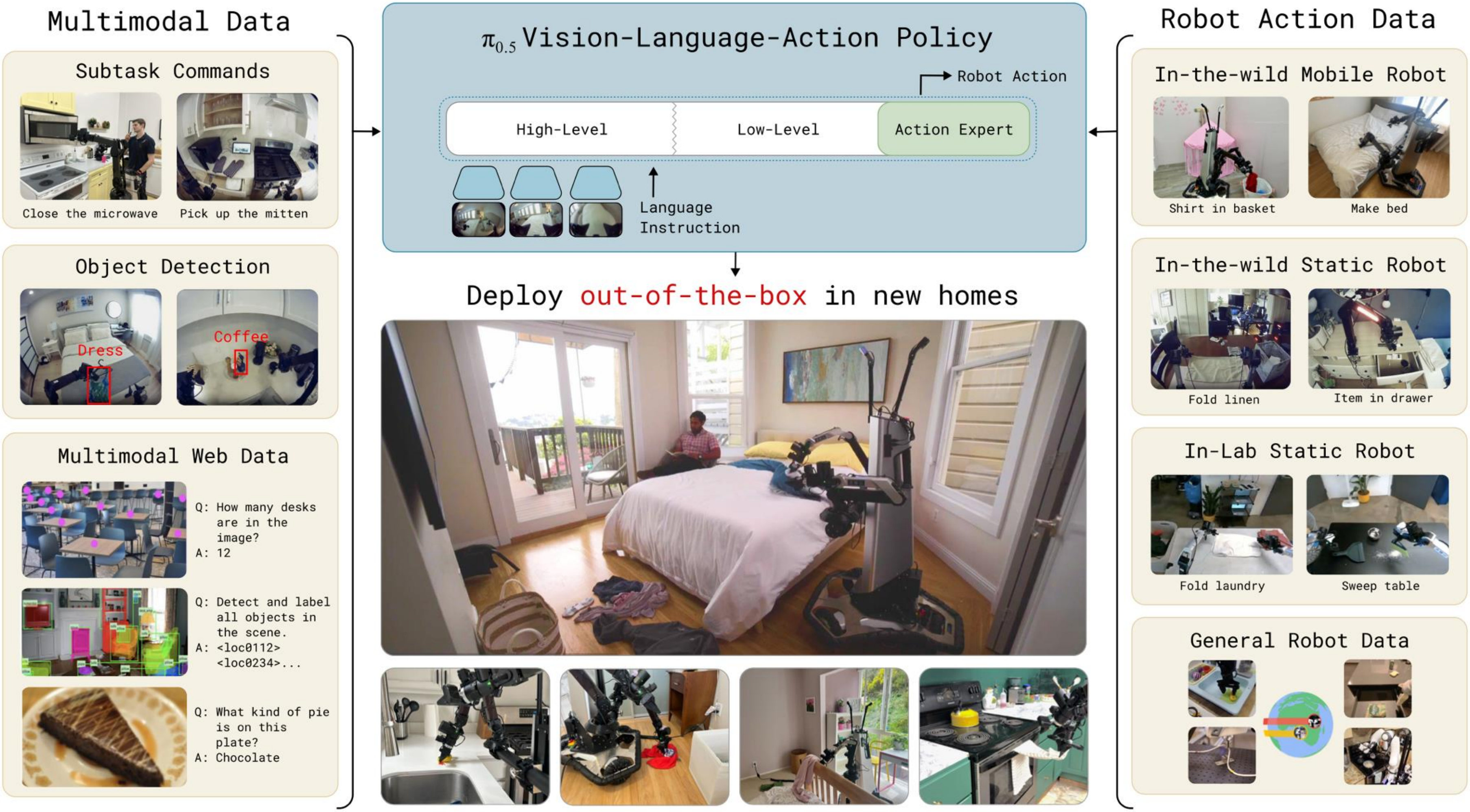
A VLA to drive embodied agent to interact with the environments

However, most models still lack

- **Multimodal Reasoning**
  - **Chain-of-thought reasoning**: as shown in O1 and DeepSeek-R1 for language tasks
  - **Thinking before action**: some tasks does require system-2 thinking to derive the answer
  - Many reasoning works appear recently, e.g., LMM-R1

- **Multimodal Planning**
  - **Task decomposition**: model needs to crack the task into multiple steps and take actions step by step, probably also requires self-correction in the middle
    - *Low-level task*: “pick up the apple”
    - High-level task: “clean up this table”
  - Some recent work: Pi 0.5

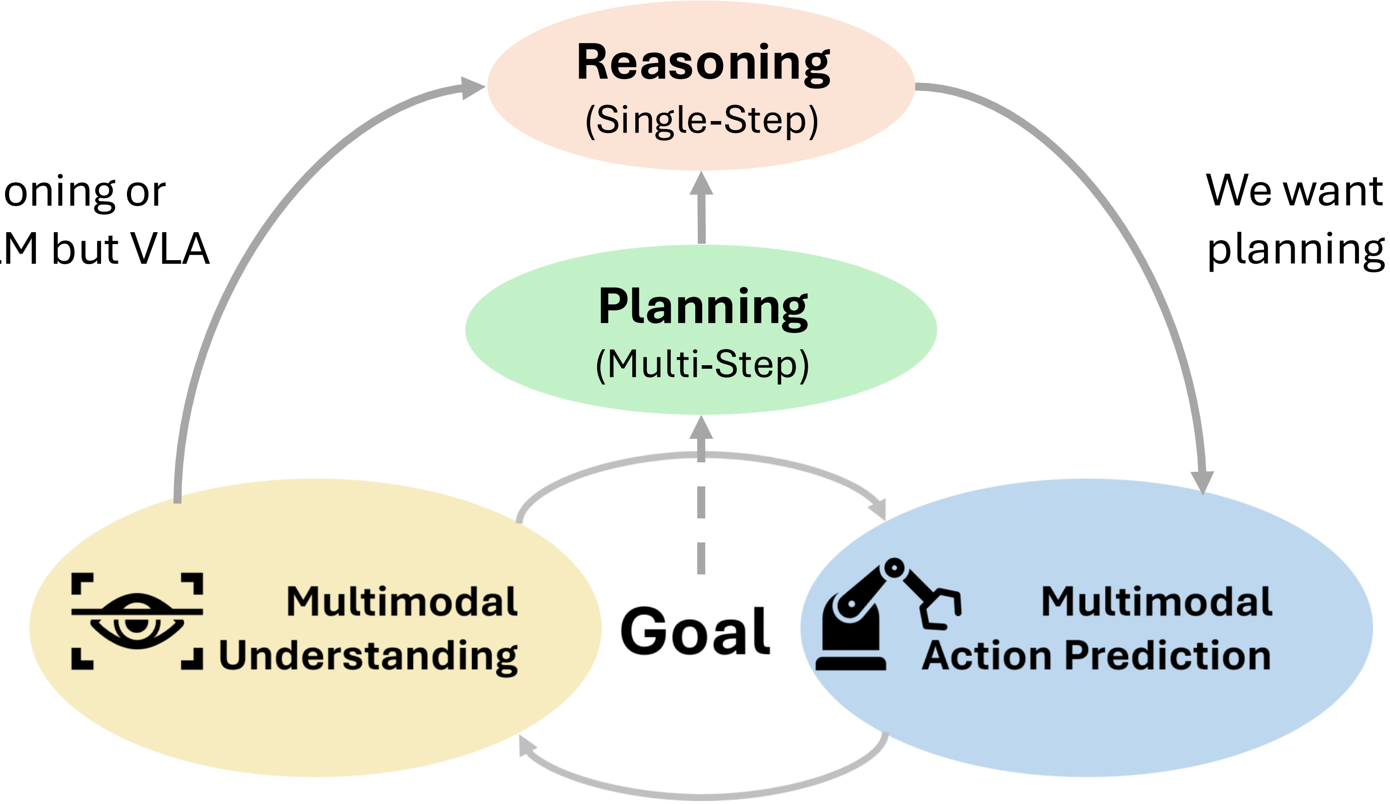
π0.5: a Vision-Language-Action Model with Open-World Generalization



What Embodied Agents Should Look Like

We are NOT building reasoning or planning capability for VLM but VLA

We want to achieve both multi-step planning and single-step reasoning

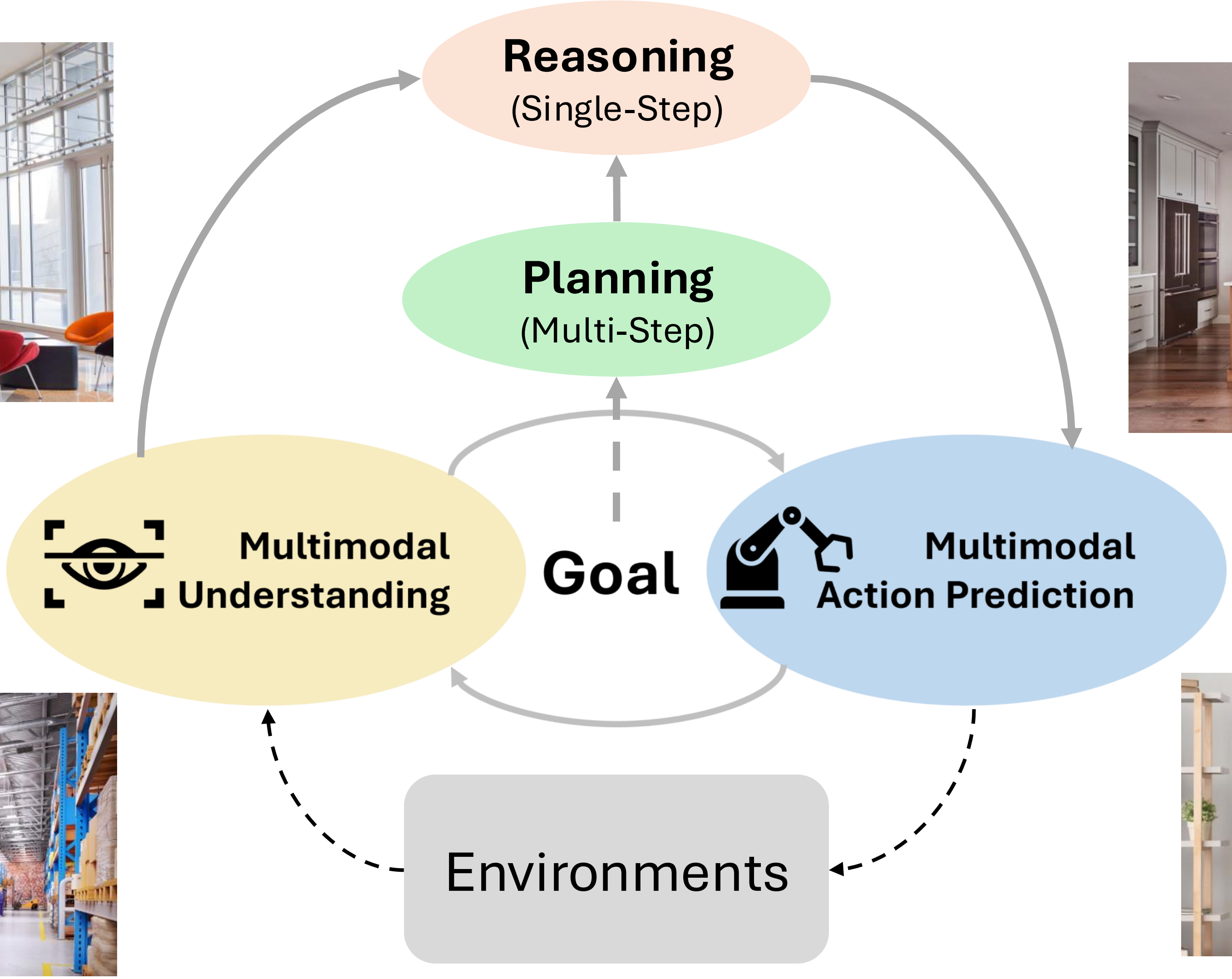


1 **Verbal Intelligence:** the ability to understand and use language, including the ability to think with words.

2 **Visual Intelligence:** the ability to understand the visual observation, and abstract them into semantic meanings.

3 **Spatial + Temporal Intelligence:** ability to capture the position, location, force and motion of objects and interact with the 2D/3D world along time axis

What Embodied Agents Should Look Like



Welcome to the Era of Experience

David Silver, Richard S. Sutton*