Visual Reasoning
will be bigger than language reasoning

Ranjay Krishna
Assistant Professor, University of Washington
Research Director, Allen Institute for AI

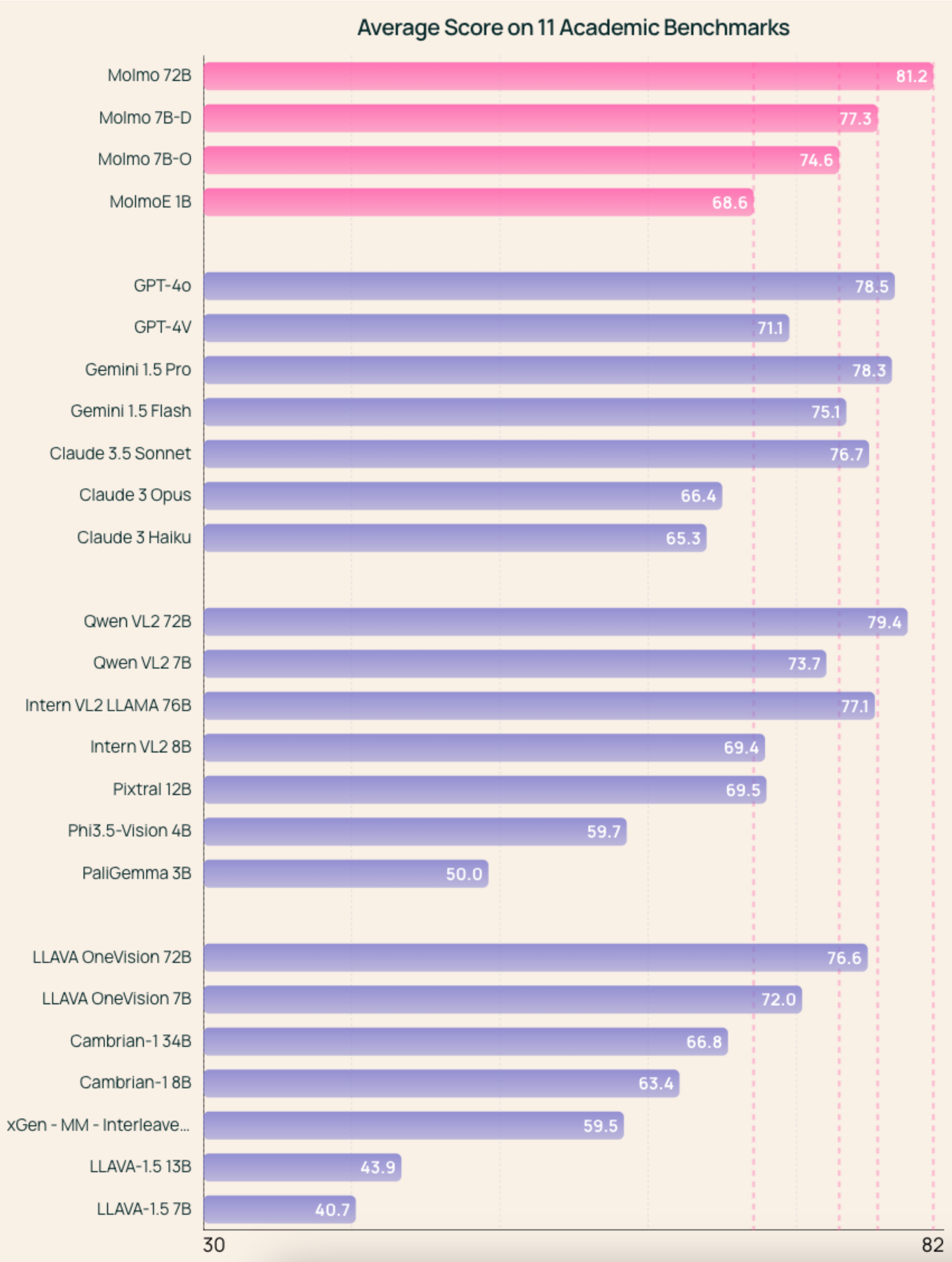
Open
Weights
Data
Code
Evals

API Only

Open
Weights

Distilled

Open



Where I left off at CVPR

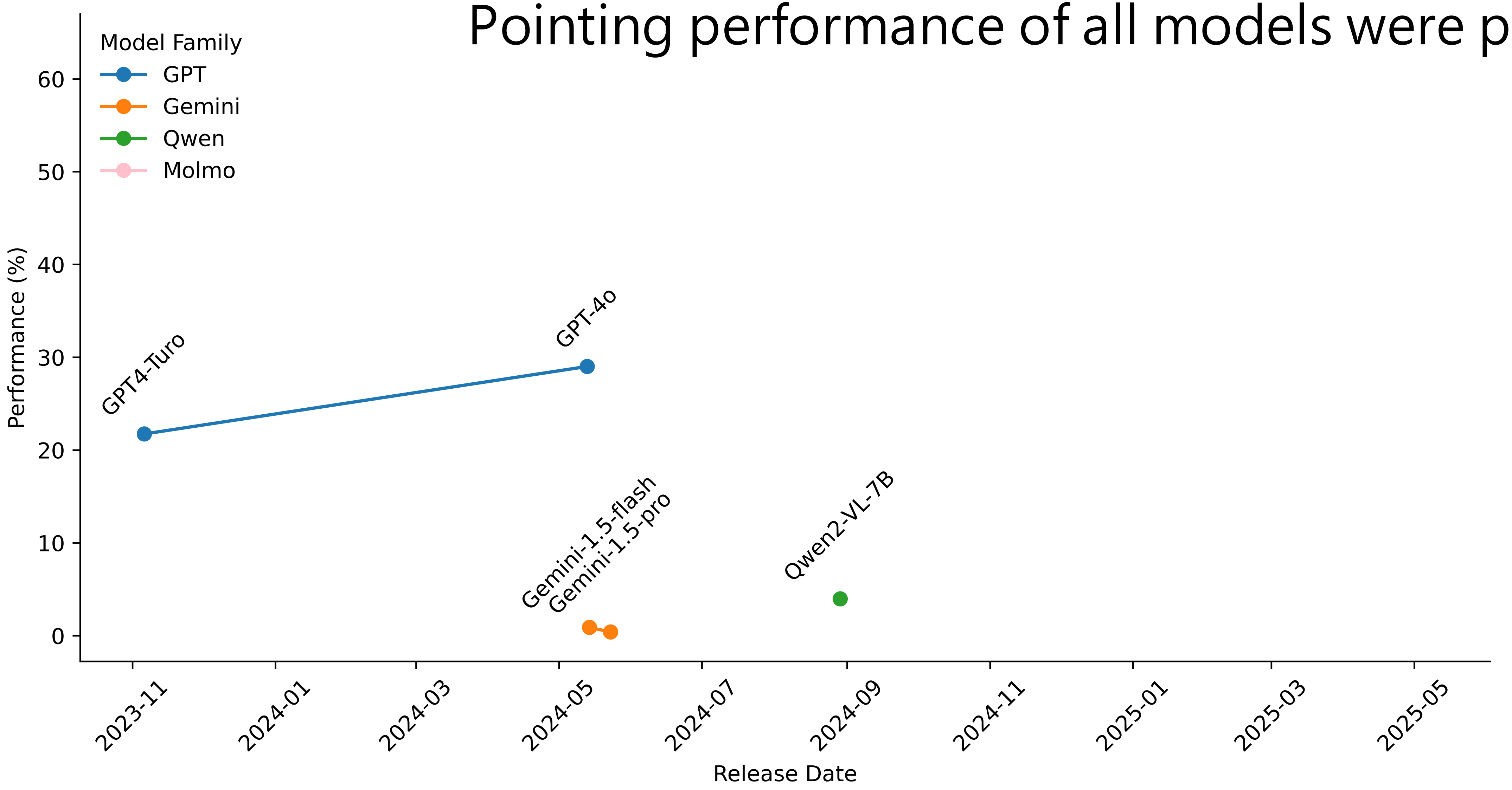
Molmo

Open weights
Open training data
Open training code
Open Evaluations

Molmo reasoning directly in the pixels

When it counts, it points

Pointing performance of all models were poor



Cheng et al. PointArena: Probing Multimodal Grounding Through Language-Guided Pointing. ArXiv 2025

Cheng et al. PointArena: Probing Multimodal Grounding Through Language-Guided Pointing. ArXiv 2025

In today’ s talk:

Visual Reasoning

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

BLINK
Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought

Prioritizing perception

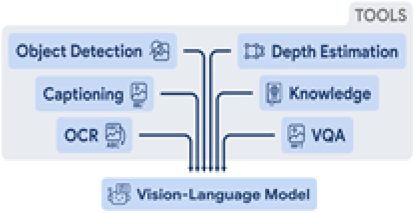
Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought



How can we distill from specialist models into generalist VLMs?

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

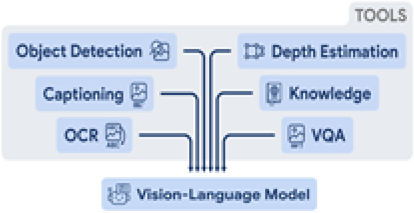
Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought



How can we distill from specialist models into generalist VLMs?

Complete open Action Reasoning model for robotics

Papers we will discuss

BLINK: Multimodal Large Language Models Can See but Not Perceive
Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, Ranjay Krishna
ECCV 2024

Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models
Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, Ranjay Krishna
NeurIPS 2024

Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models
Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, Ariel Fuxman
CVPR 2025

Perception Tokens Enhance Visual Reasoning in Multimodal Language Models
Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, Ranjay Krishna
CVPR 2025

MolmoAct: Action Reasoning Models that can Reason in Space
Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, Ranjay Krishna
ArXiv 2025

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

BLINK
Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought

How can we distill from specialist models into generalist VLMs?

MolmoAct

Complete open Action Reasoning model for robotics

We know that models today can describe images well

What can you see?

In the image, I see a cat with a light brown and white fur coat sitting on a blue chair. The cat appears to be wrapped in or resting on a dark gray fabric, possibly a piece of clothing or a small blanket. The cat's expression seems slightly grumpy or annoyed. The background includes part of the chair and a light-colored cushion.

But can they understand 3D space?

Question: Is the camera moving towards the left or the right?

(A) Left

(B) Right GPT 4V

(C) I don’t know Gemini

Maybe an even easier example

Question: Which image fits the missing part?

(A)

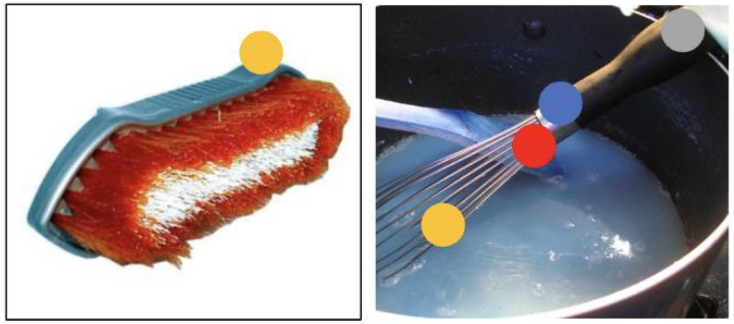
(B)

GPT 4V

Gemini Pro

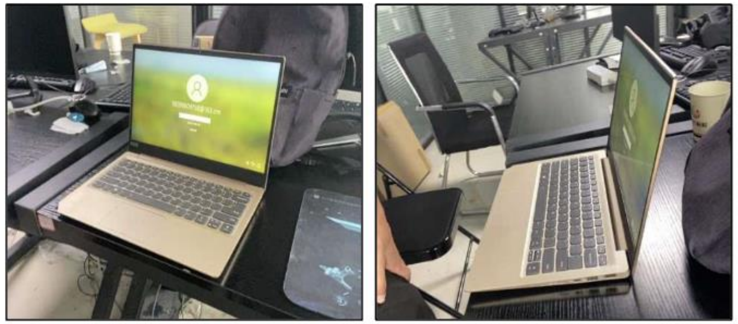
What fundamental perceptual capabilities do we want VLMs to have

Semantic affordance



Q: Which point has similar affordance?

Multiview reasoning

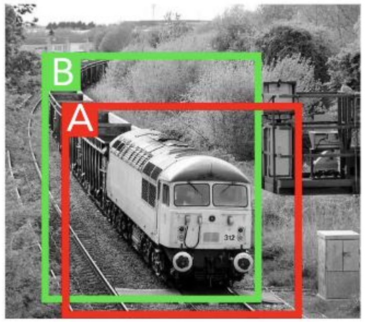


Q: Is the camera moving clockwise around the object?

Visual similarity

Q: Which image is more similar to the left one?

Localization

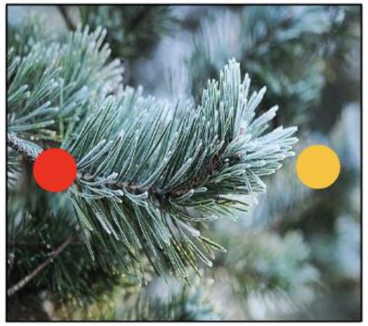


Q: Which box localizes train better?

Inpainting

Q: Which image fits here better?

Depth estimation



Q: Which point is farther?

Color

Q: Which point is darker?

Image forensics

Q: Which image is real?

What are popular VLM benchmarks measuring?

Forcasting

Q: What will happen next?

Social relationships

Q: What is their relationship?

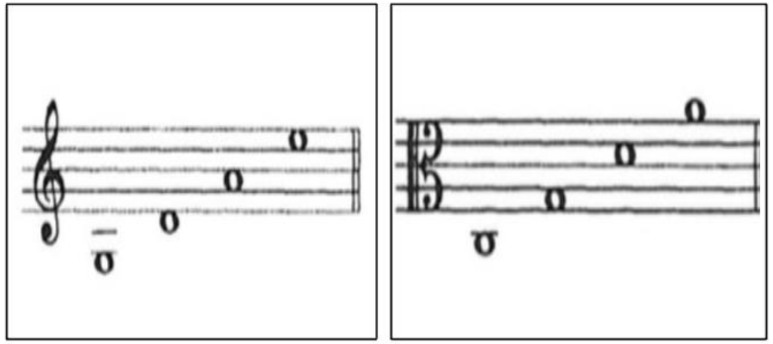
Geolocation

Q: Where is this place?

Recognition

Q: What kind of animal is this?

Music theory?



Q: What is the correct tuning of Violin?

While still valuable, existing benchmarks conflate perception with general knowledge and reasoning

Examples taken from MME – which is a good benchmark

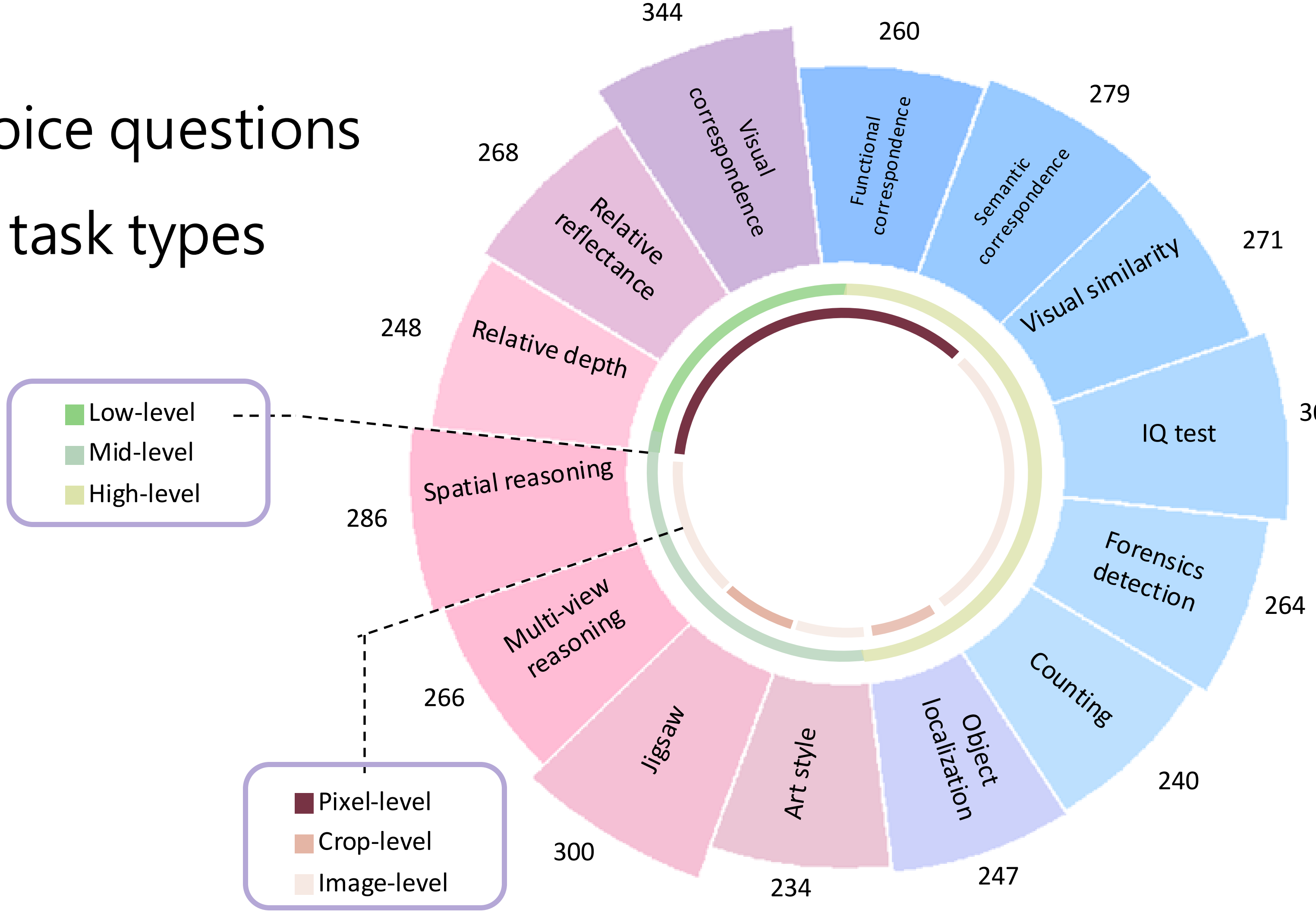
BLINK : A VLM benchmark aimed at measuring classical notions of perception (at ECCV 2024)

Blink has a diverse set of questions

7,358 unique images

3,807 unique multiple-choice questions

Over 14 visual perception task types



Results on BLINK

(1) Humans are very good at these tasks (95%)
(2) Random chance is a reasonable 38%

	Validation (1,901)	Test (1,906)
Random Choice	38.09	38.09
Human	95.67	95.70

Results on BLINK

(1) Humans are very good at these tasks (95%)

(2) Random chance is a reasonable 38%

(3) Open-weight models barely perform better than random chance

	Validation (1,901)	Test (1,906)
Random Choice	38.09	38.09
Human	95.67	95.70
MiniGPT-4-v2 [16]	34.23	34.57
OpenFlamingo-v2 [5]	39.18	38.32
InstructBLIP-7B [24]	39.72	38.65
InstructBLIP-13B [24]	42.24	39.58
LLaVA-internLM2-7B [72]	37.71	36.06
Yi-VL-6B [2]	38.72	41.24
Yi-VL-34B [2]	41.68	42.78
LLaVA-v1.5-7B-xtuner [23]	39.36	40.81
LLaVA-v1.5-13B-xtuner [23]	42.00	41.31
CogVLM [77]	41.54	39.38
LLaVA-v1.5-7B [48]	37.13	38.01
LLaVA-v1.5-13B [48]	42.66	40.55
LLaVA-v1.6-34B [50]	46.80	45.05

Results on BLINK

(1) Humans are very good at these tasks (95%)
(2) Random chance is a reasonable 38%
(3) Open-weight models barely perform better than random chance
(4) Only GPT-4o performs better but not by much (60%).

	Validation (1,901)	Test (1,906)
Random Choice	38.09	38.09
Human	95.67	95.70
MiniGPT-4-v2 [16]	34.23	34.57
OpenFlamingo-v2 [5]	39.18	38.32
InstructBLIP-7B [24]	39.72	38.65
InstructBLIP-13B [24]	42.24	39.58
LLaVA-internLM2-7B [72]	37.71	36.06
Yi-VL-6B [2]	38.72	41.24
Yi-VL-34B [2]	41.68	42.78
LLaVA-v1.5-7B-xtuner [23]	39.36	40.81
LLaVA-v1.5-13B-xtuner [23]	42.00	41.31
CogVLM [77]	41.54	39.38
LLaVA-v1.5-7B [48]	37.13	38.01
LLaVA-v1.5-13B [48]	42.66	40.55
LLaVA-v1.6-34B [50]	46.80	45.05
Qwen-VL-Max [7]	40.28	41.94
Gemini Pro [71]	45.16	45.72
Claude 3 OPUS [1]	44.05	44.11
GPT-4V(ision) [62]	51.14	51.26
GPT-4 Turbo [62]	54.61	53.89
GPT-4o [62]	**60.04**	**59.03**

Are the tasks too difficult to solve? NO!

Task	Vis.Corr.	Depth	Multi-view	Sem.Corr.	Forensic	Reflect.
Random	25.00	50.00	50.00	25.00	25.00	33.33
Human	99.56	99.59	92.10	94.60	100.00	99.63
Gemini Pro	42.44	40.32	44.36	26.62	50.76	45.52
GPT-4V	33.72	59.68	55.64	28.78	34.09	38.81
Specialist	DIFT [70]	DepthAnything [83]	LoFTR [68]	DIFT [70]	DIRE [79]	Ordinal Shading [14]
	96.51	97.58	90.22	71.22	68.94	77.61

Specialist models can do these classifical perception tasks very well

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought

How can we distill from specialist models into generalist VLMs?

Complete open Action Reasoning model for robotics

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

BLINK
Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought

How can we distill from specialist models into generalist VLMs?

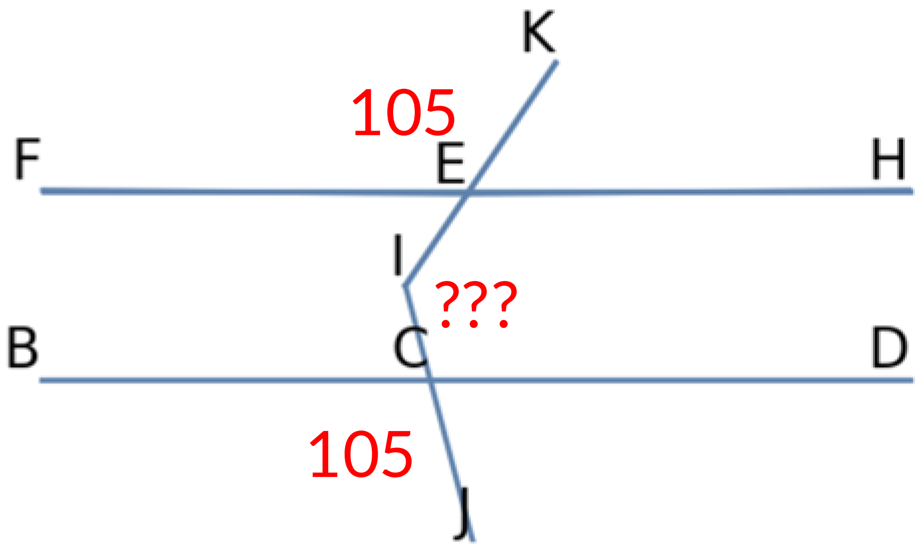
MolmoAct

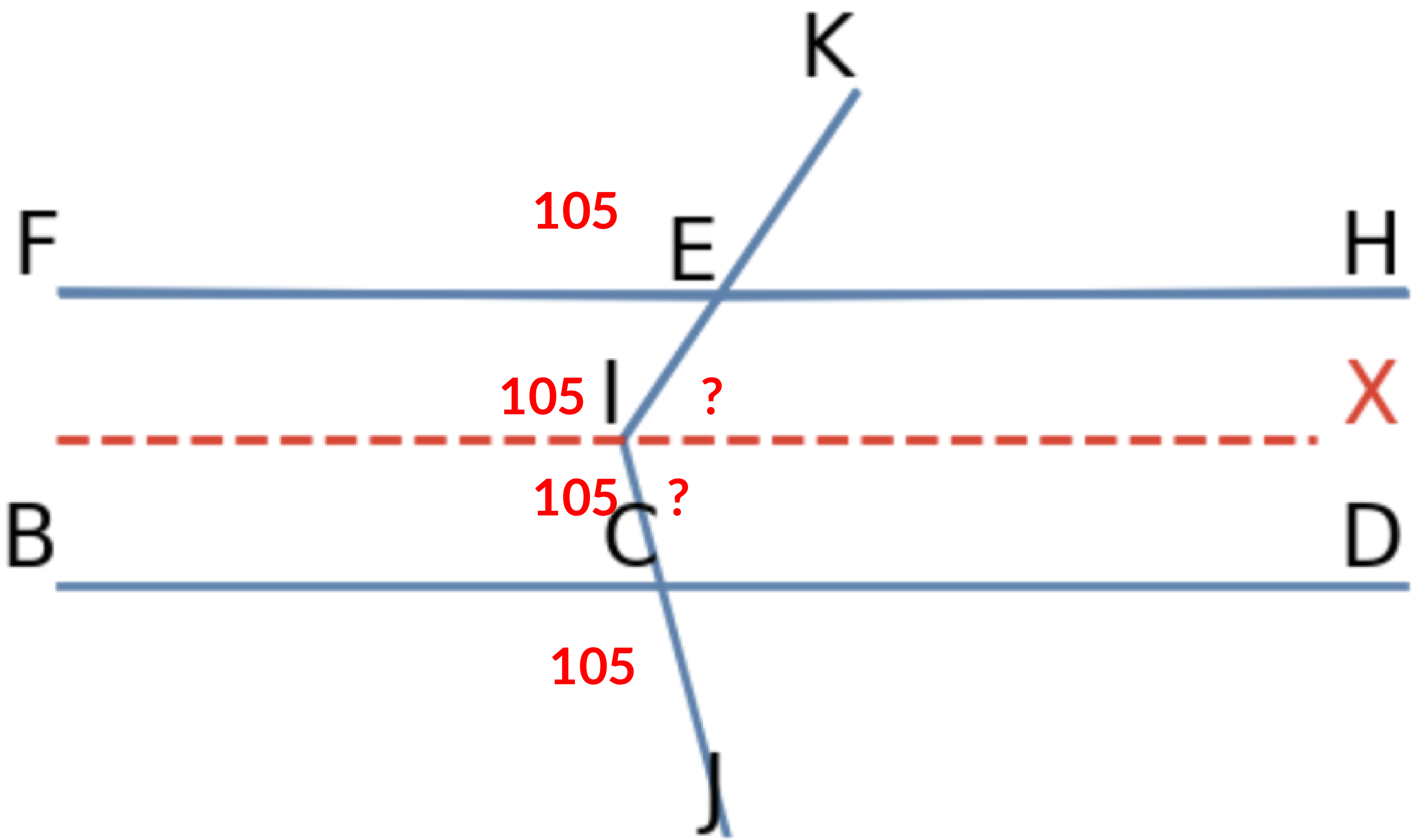
Complete open Action Reasoning model for robotics

Let’ s try to solve some BLINK tasks ourselves

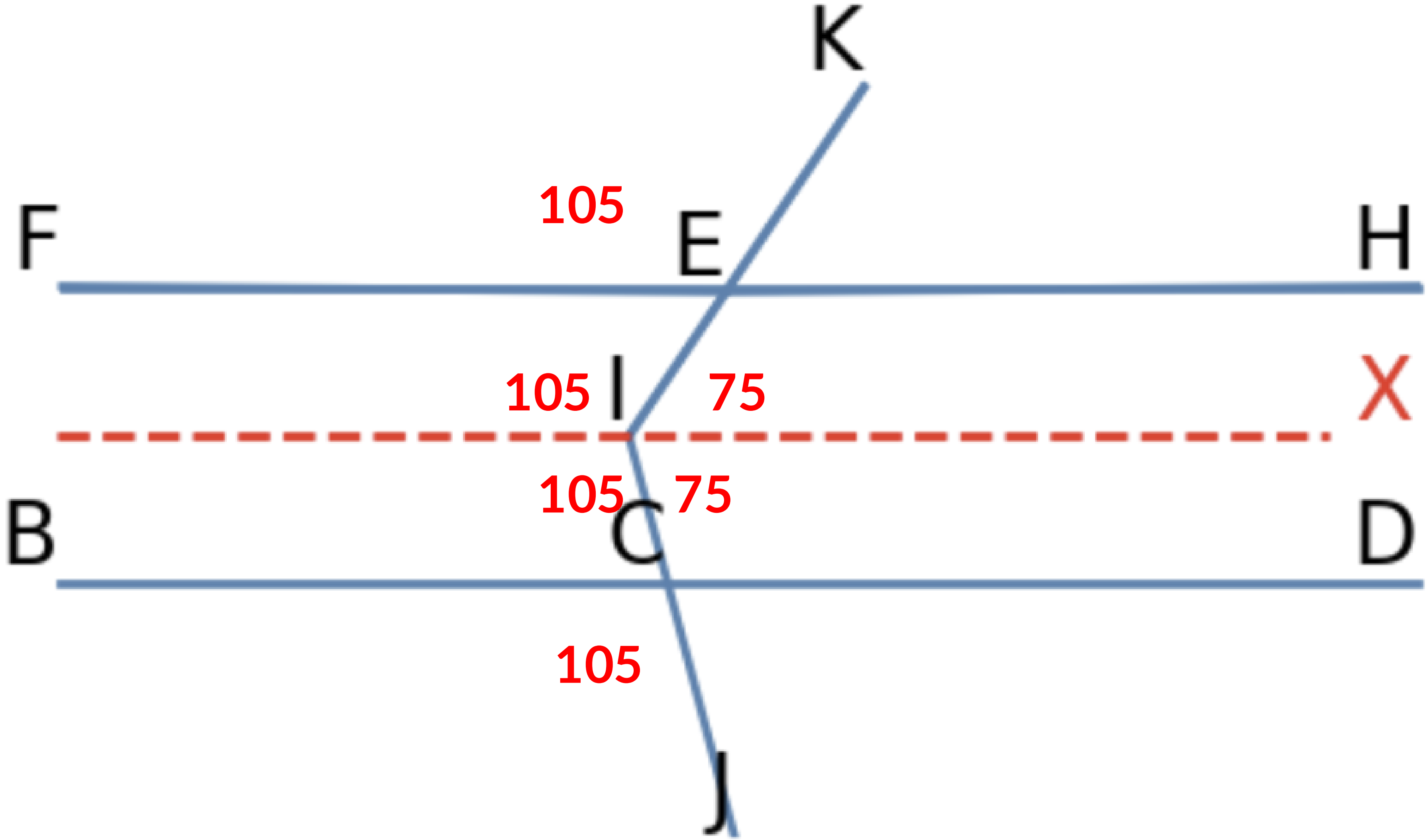
How would you solve this task?

Given ∠BCJ=105°, ∠KEF=105°. Find ∠EIC




K
105
F
E
H
105
I
?
X
105
?
B
C
D
105
J

So, <EIC = ???



Let’ s try another one

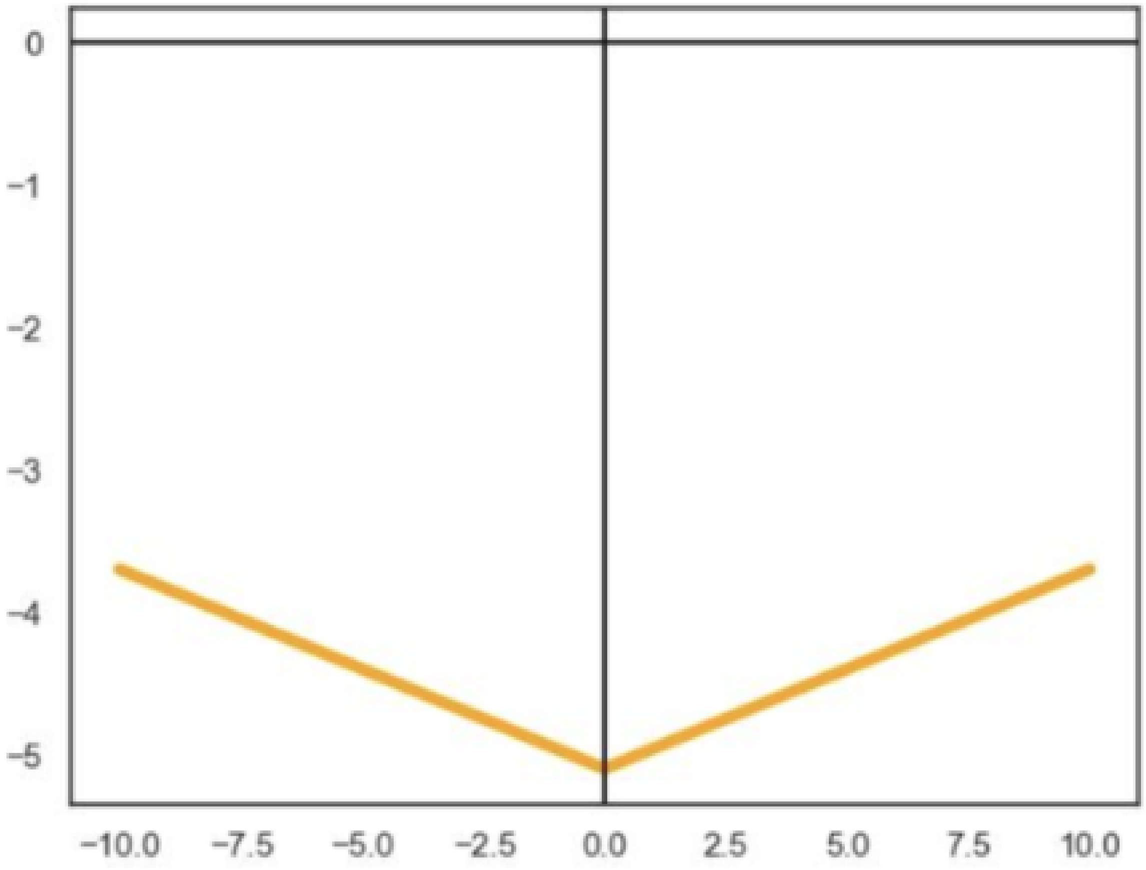
Is f(x) an convex function?

$$f(x) = 0.14\,|x| - 5.09$$

Is this easier?

Is f(x) an convex function?

$$f(x) = 0.14\,|x| - 5.09$$



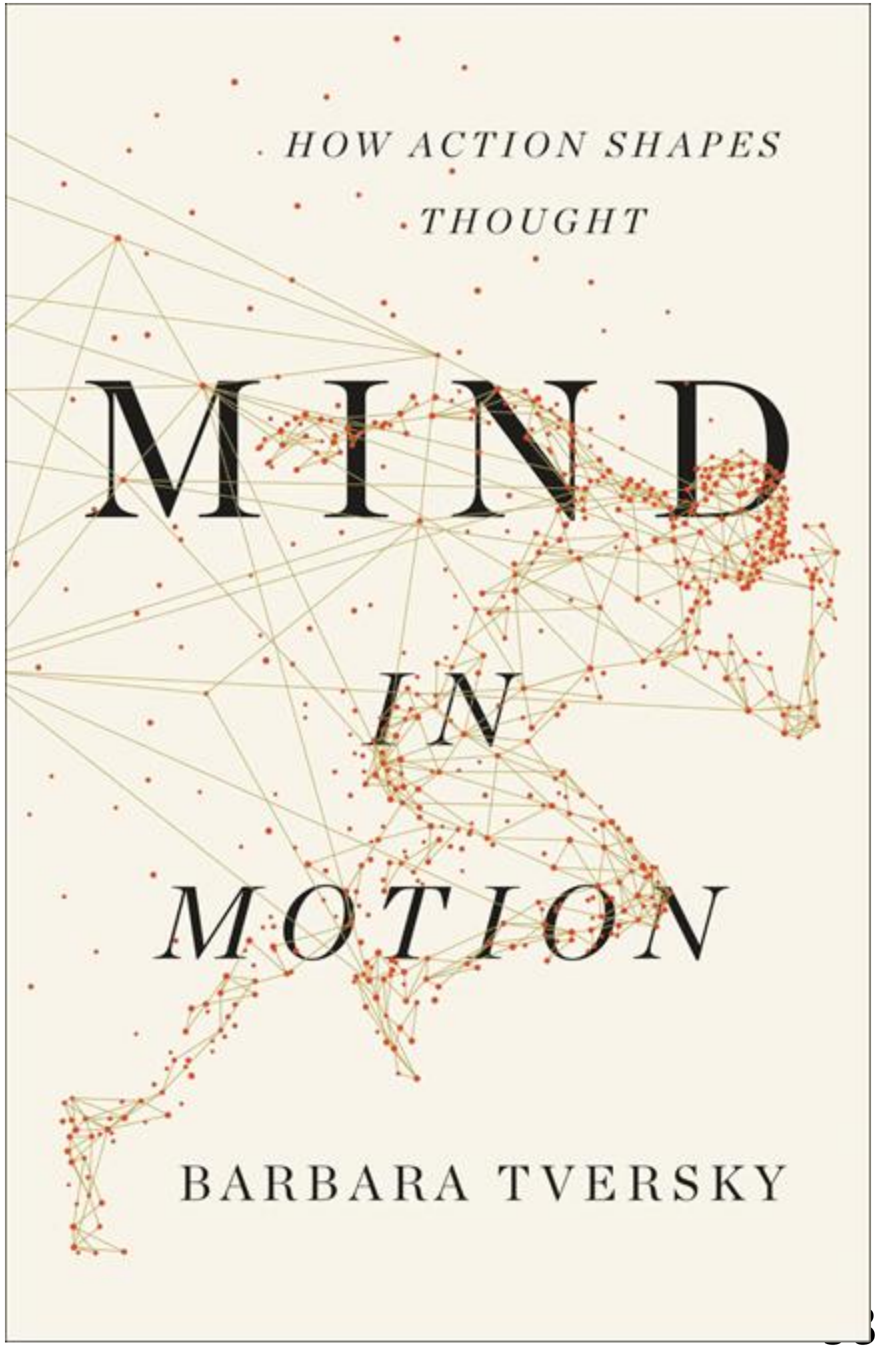
Sketching

This is how we solve problems

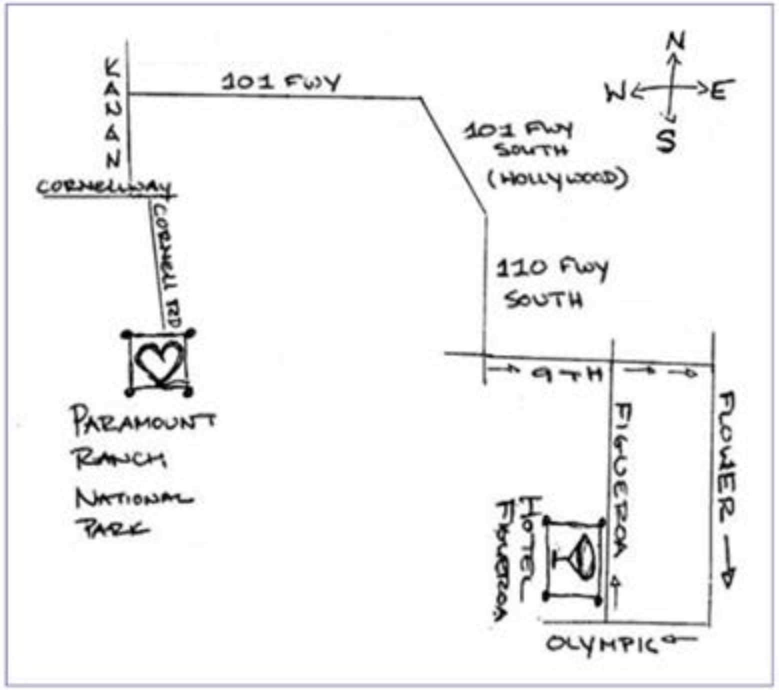
Her work ranges 5 decades and shows culminates to one statement:

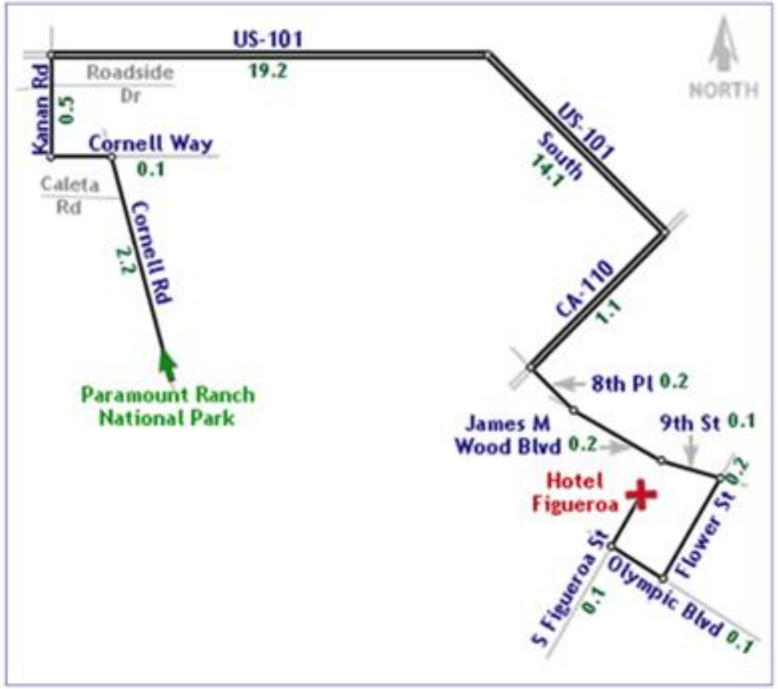
Spatial understanding is the foundation of all intelligence

Barbara Tversky



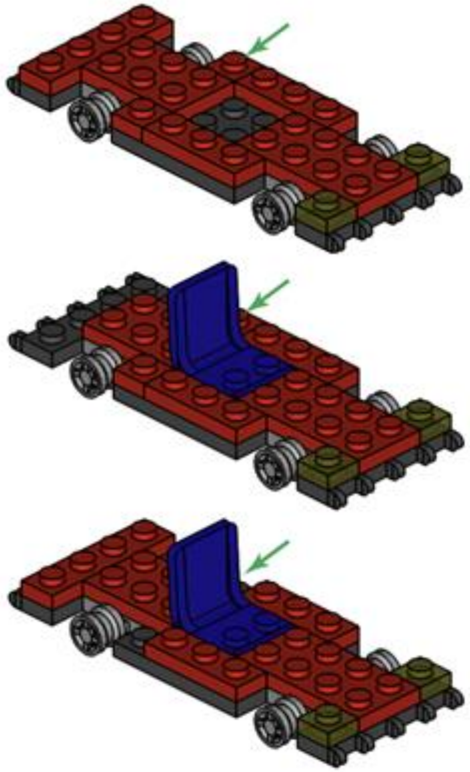
We use sketches to convey directions



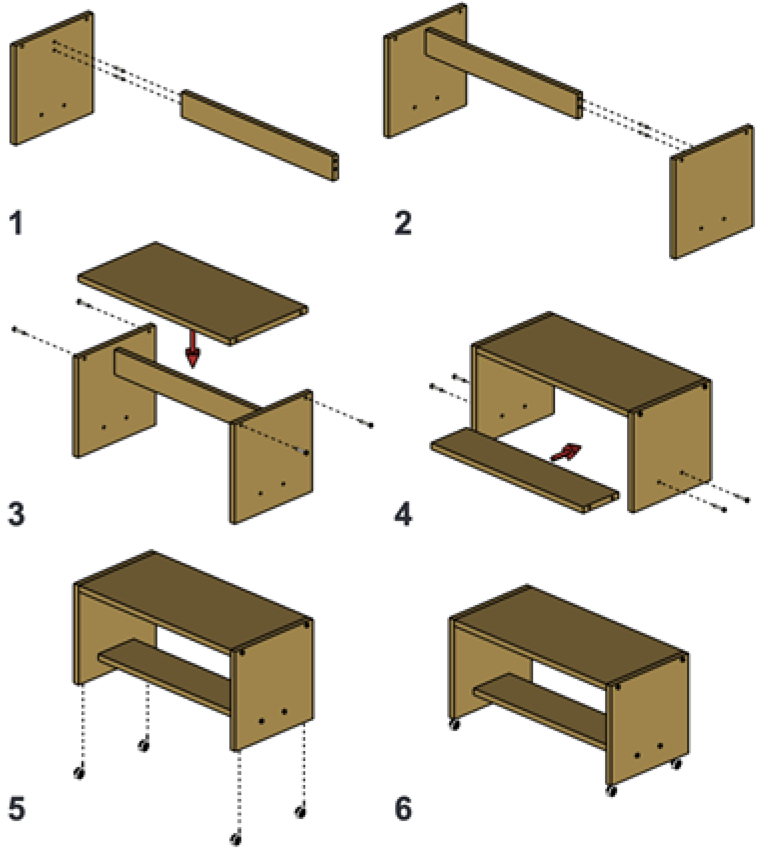


Agrawala and Stolte. Rendering Effective Route Maps: Improving Usability Through Generalization, 2001

Sketches provide instructions

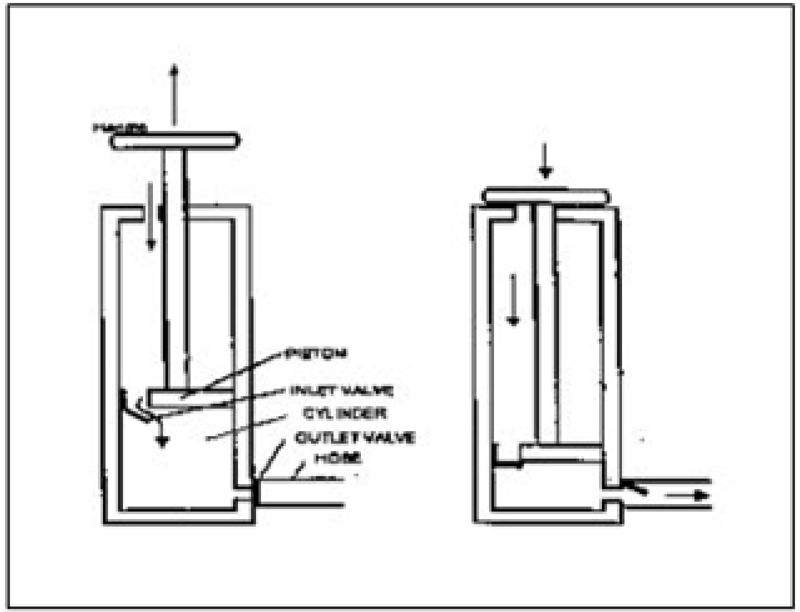


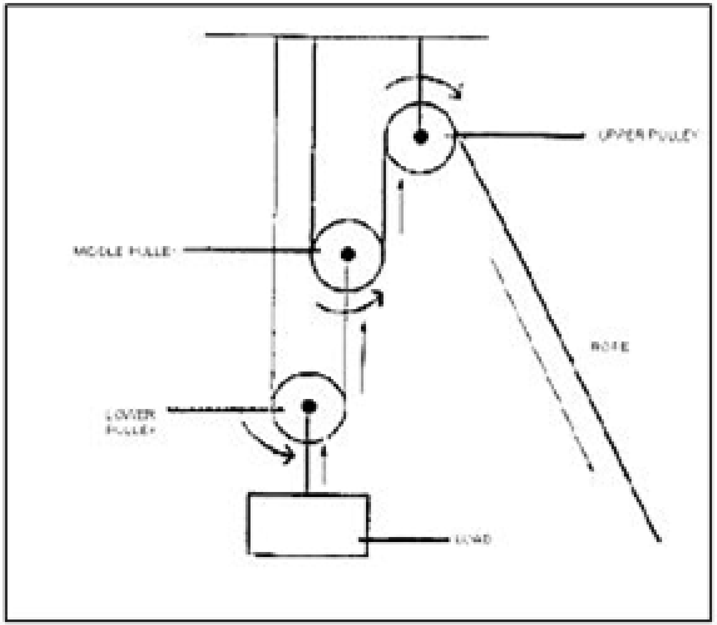
Agrawala et al. Designing Effective Step-By-Step Assembly Instructions, 2004



Sketches explain affordances

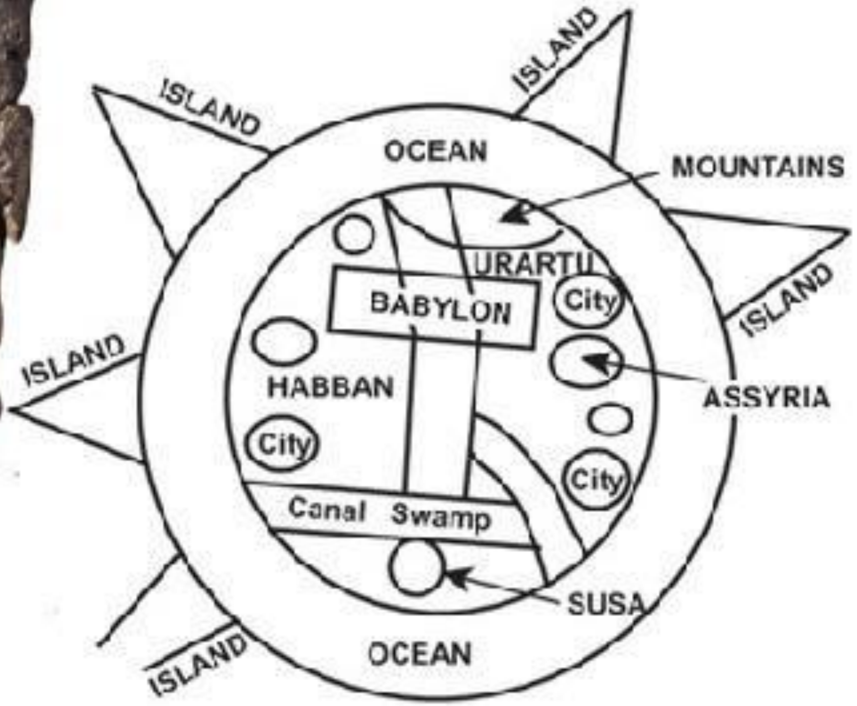
People remember more details when they see sketches than when they read



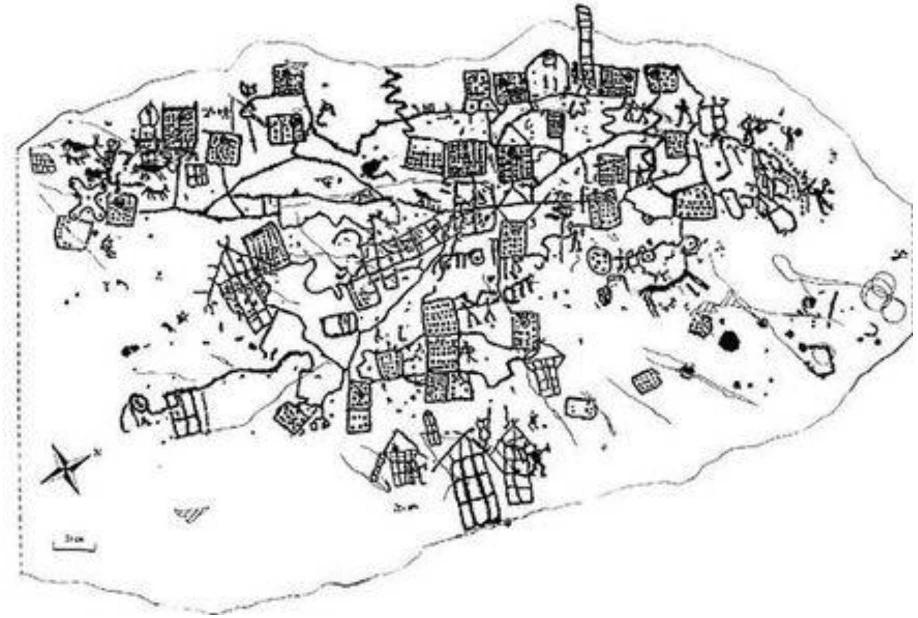


Tversky et al. Animation: can it facilitate?, 2002

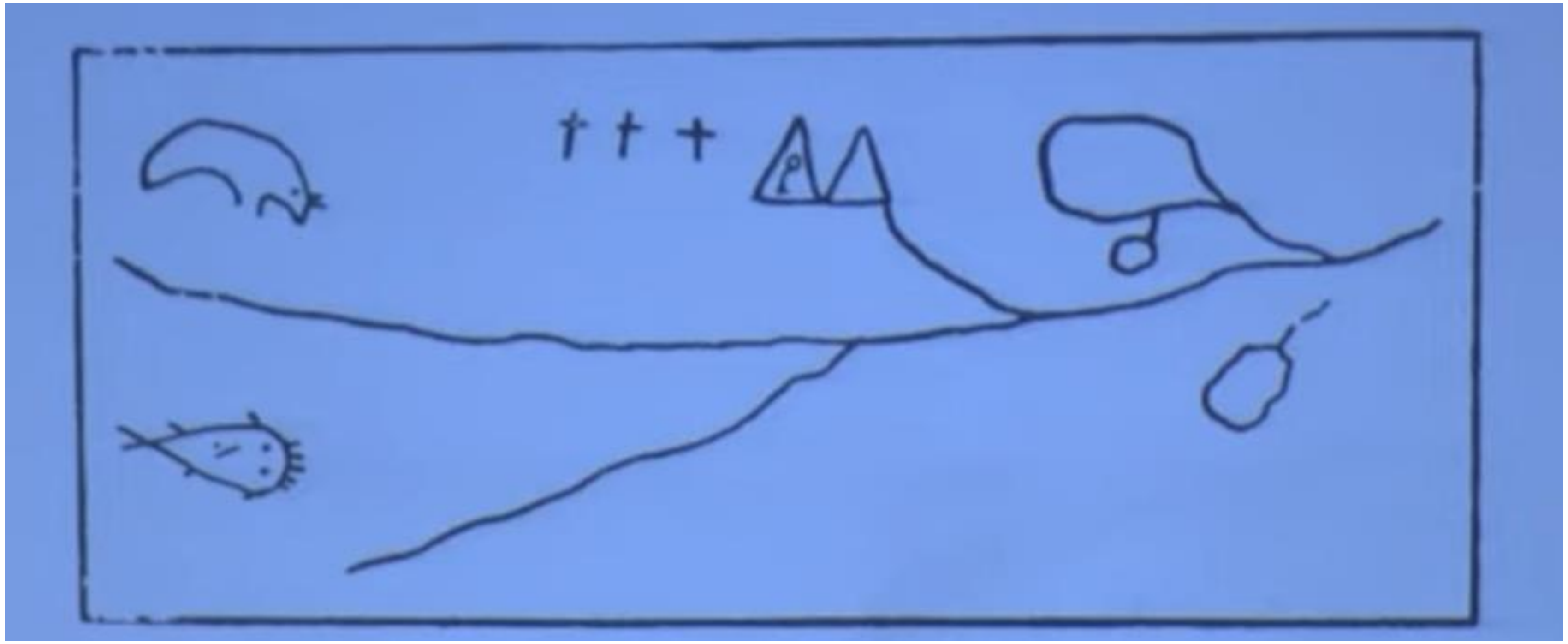
Sketches as early as ancient babylon



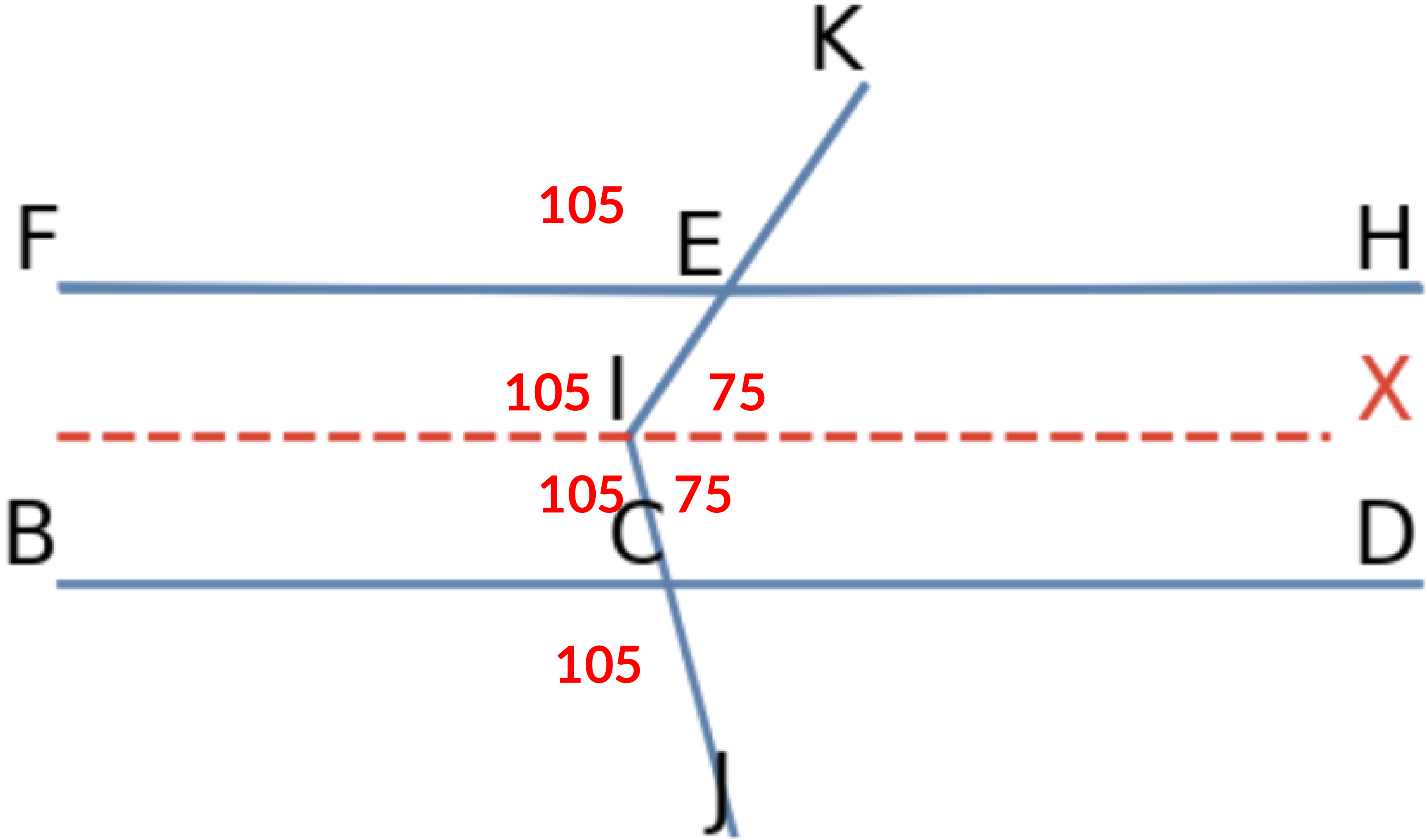
Bedolina map - first known map from 1000-200BC



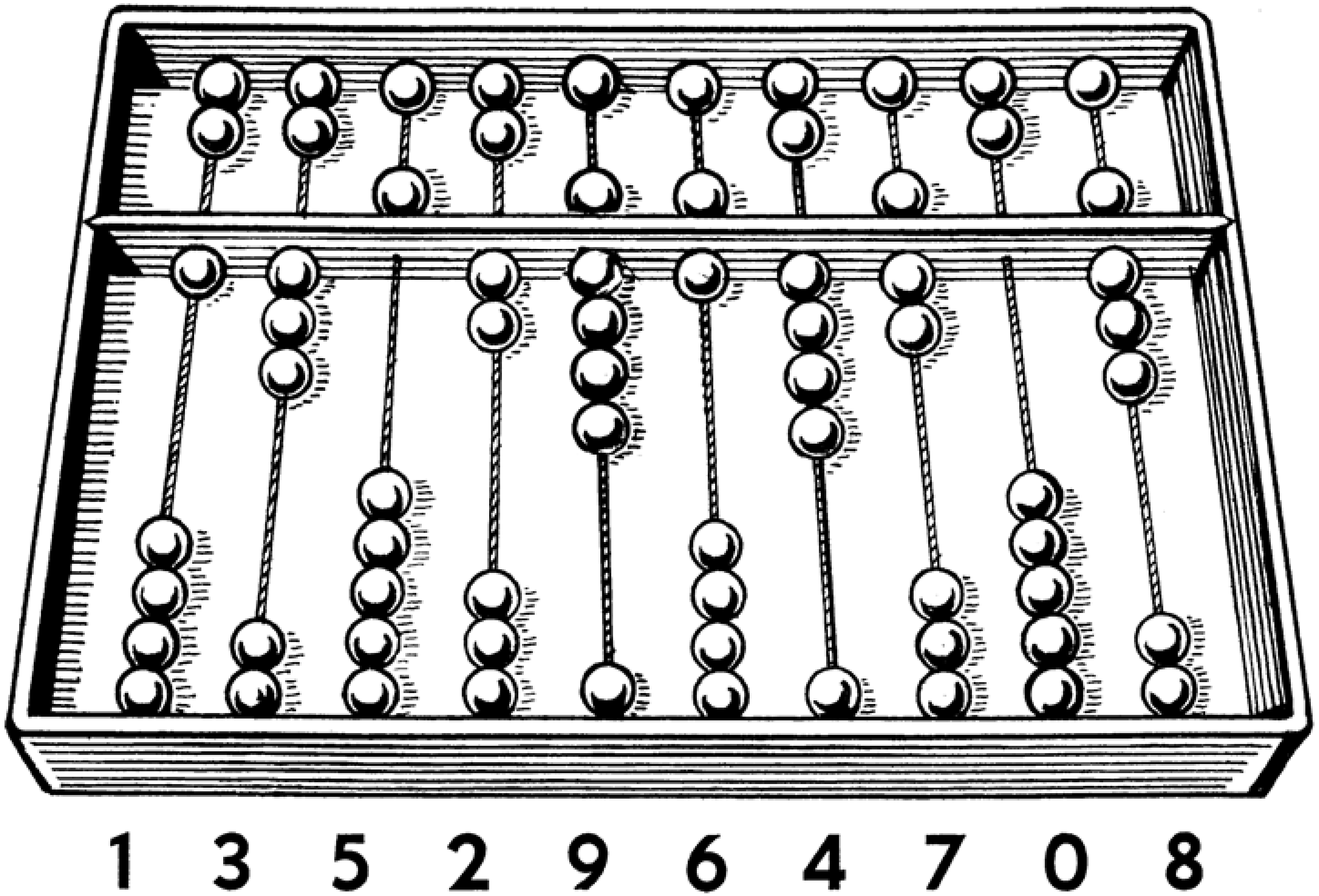
Children learn to sketch very early on



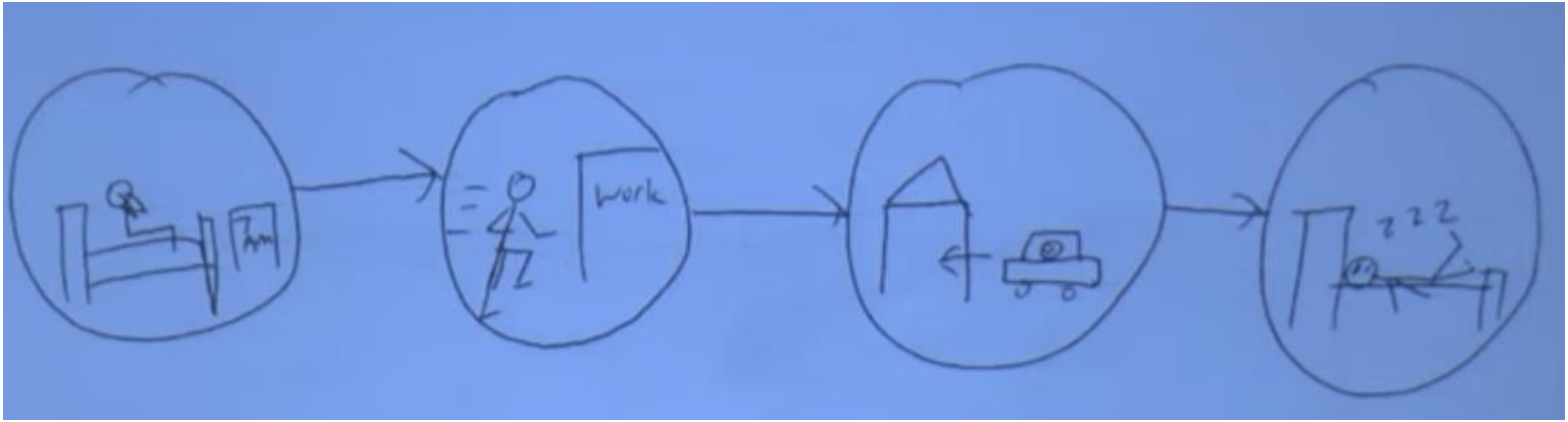
We are taught to sketch to solve math



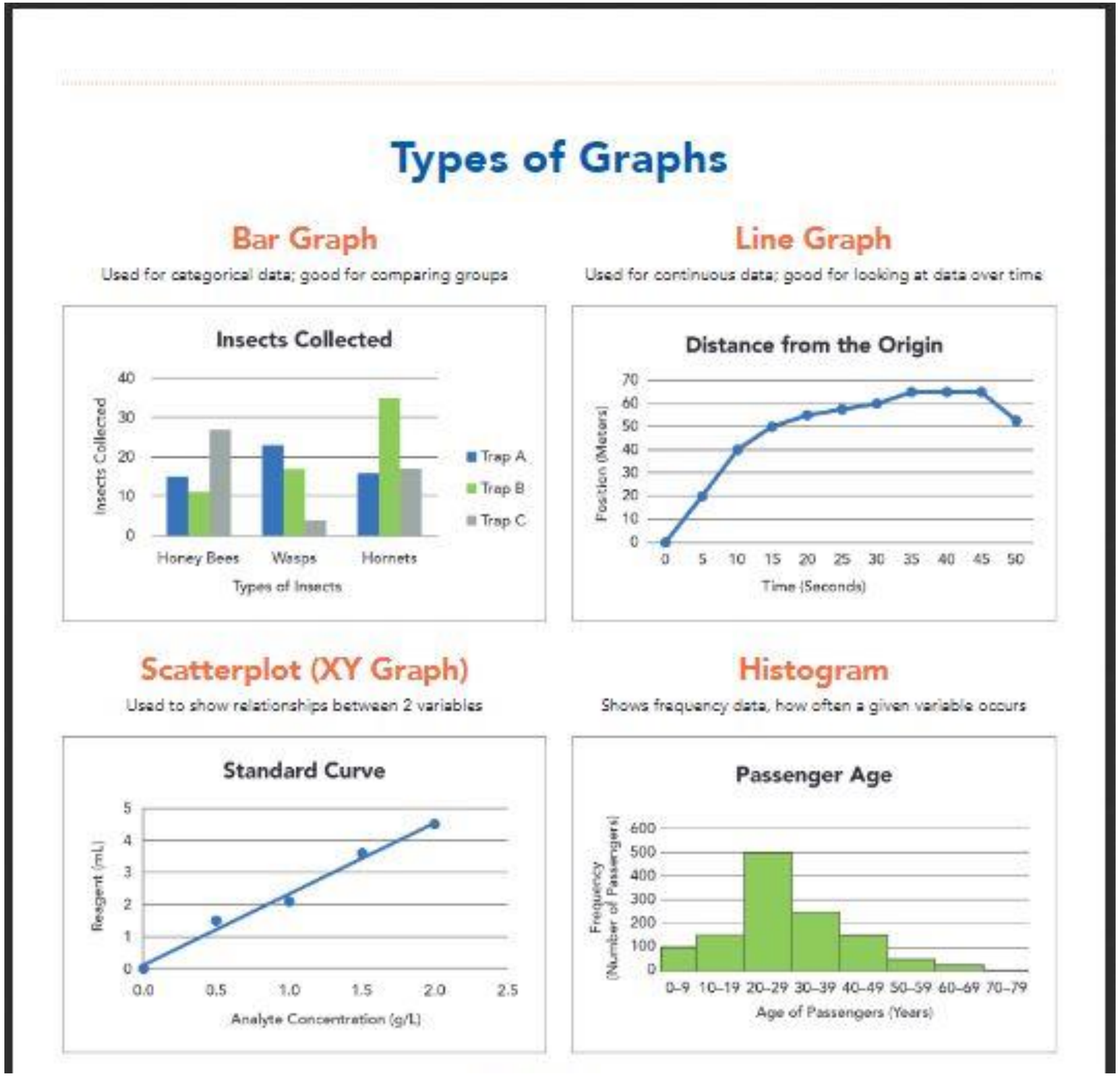
We are built systems that leverage space to solve arithmetic



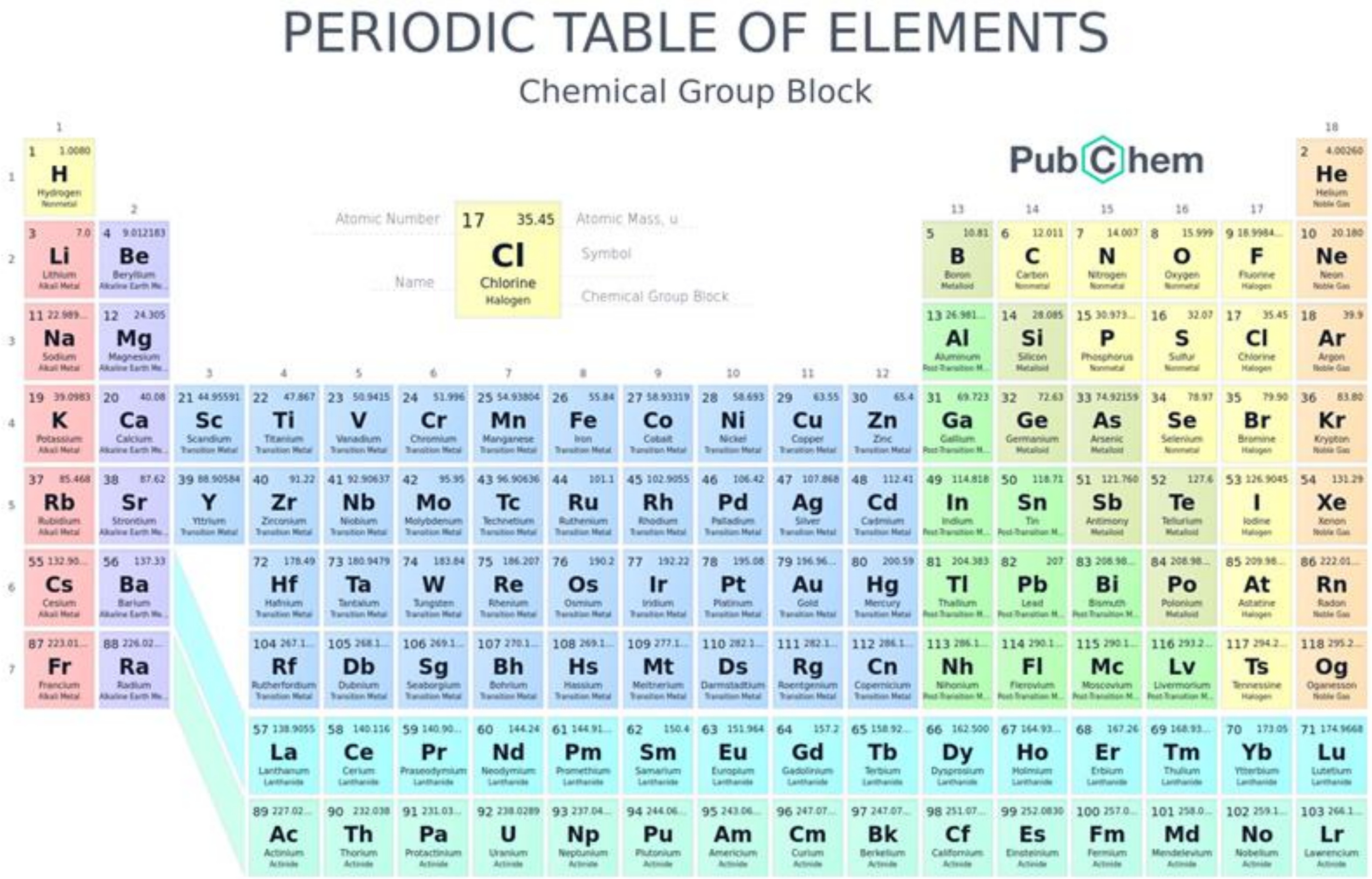
We use sketches to tell stories and convey time



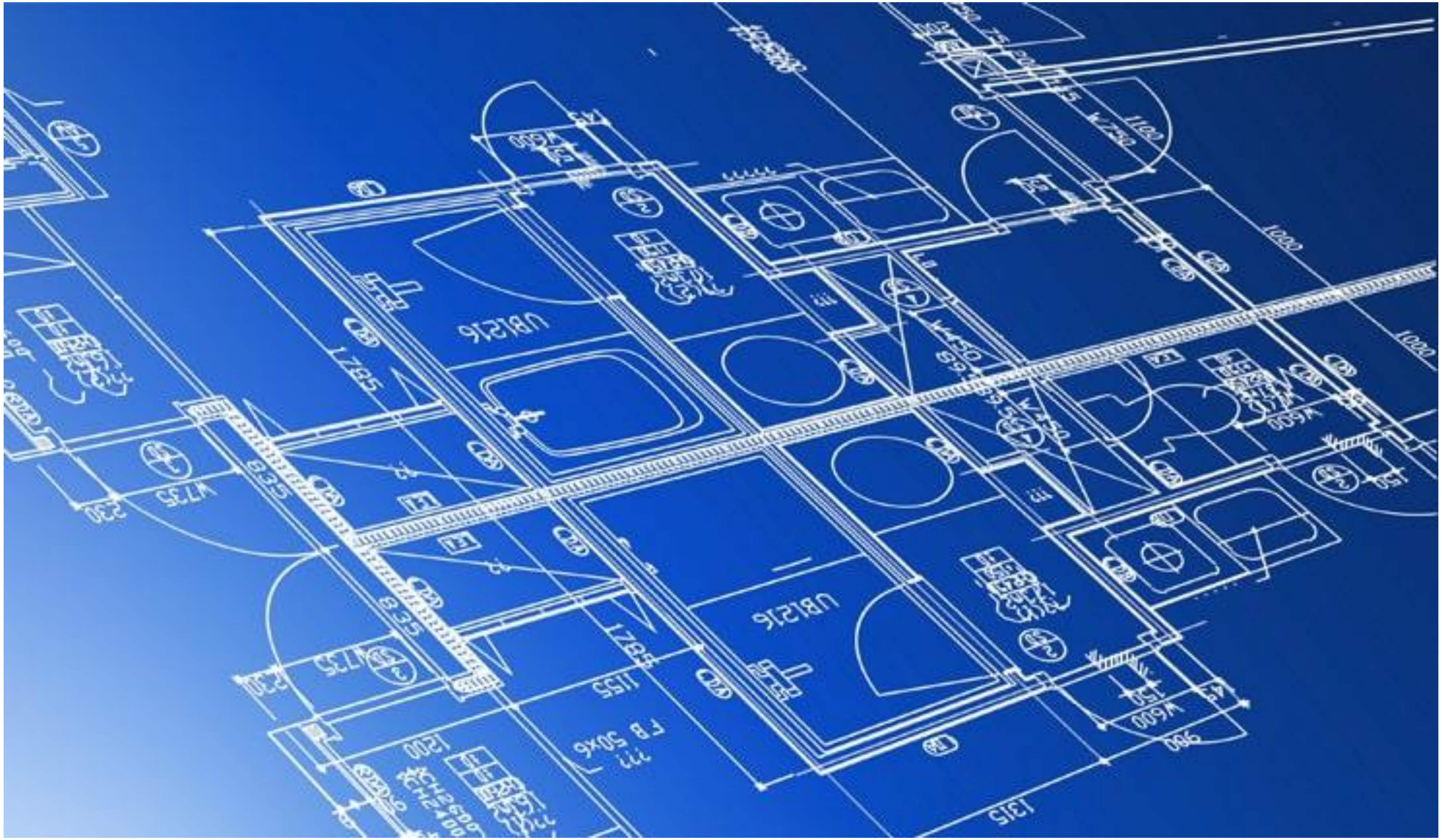
Graphs and visualizations are sketches that are faithful summaries of the underlying data



Sketches in the form of tables help us arrange items



Sketches help us build



Sketches appear in all our research papers

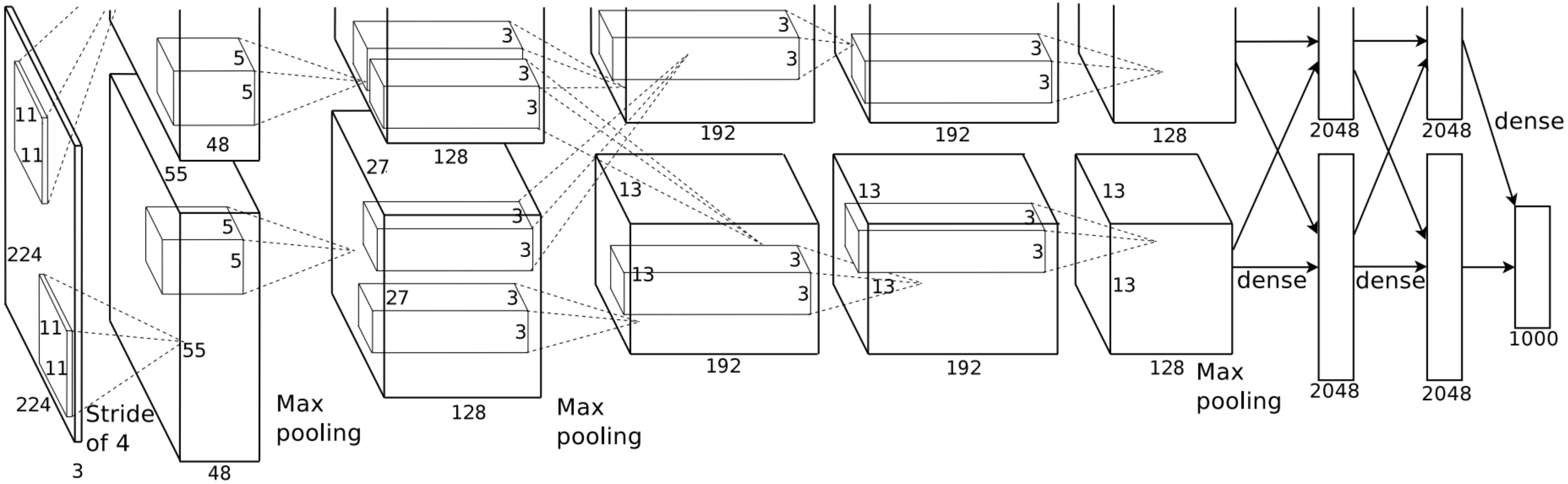
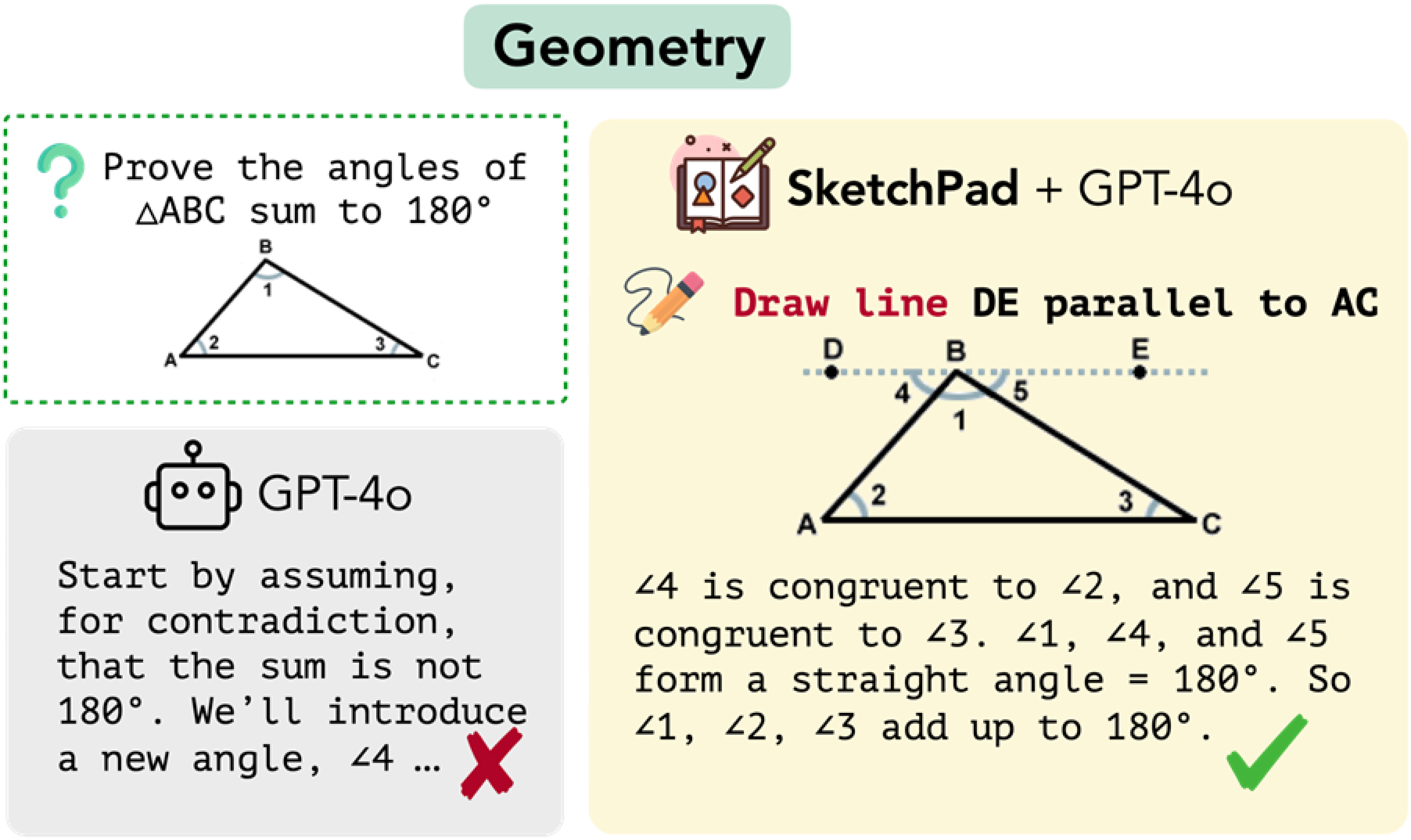


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

Sketchpad for VLMs: **Geometry**



Sketchpad for VLM: **Math**

Math Function

Is f(x) an convex function?

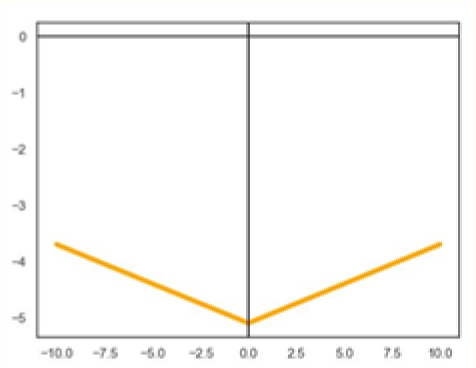
$$f(x) = 0.14\,|x| - 5.09$$

GPT-4o

It is convex for x>0 but concave for x<0

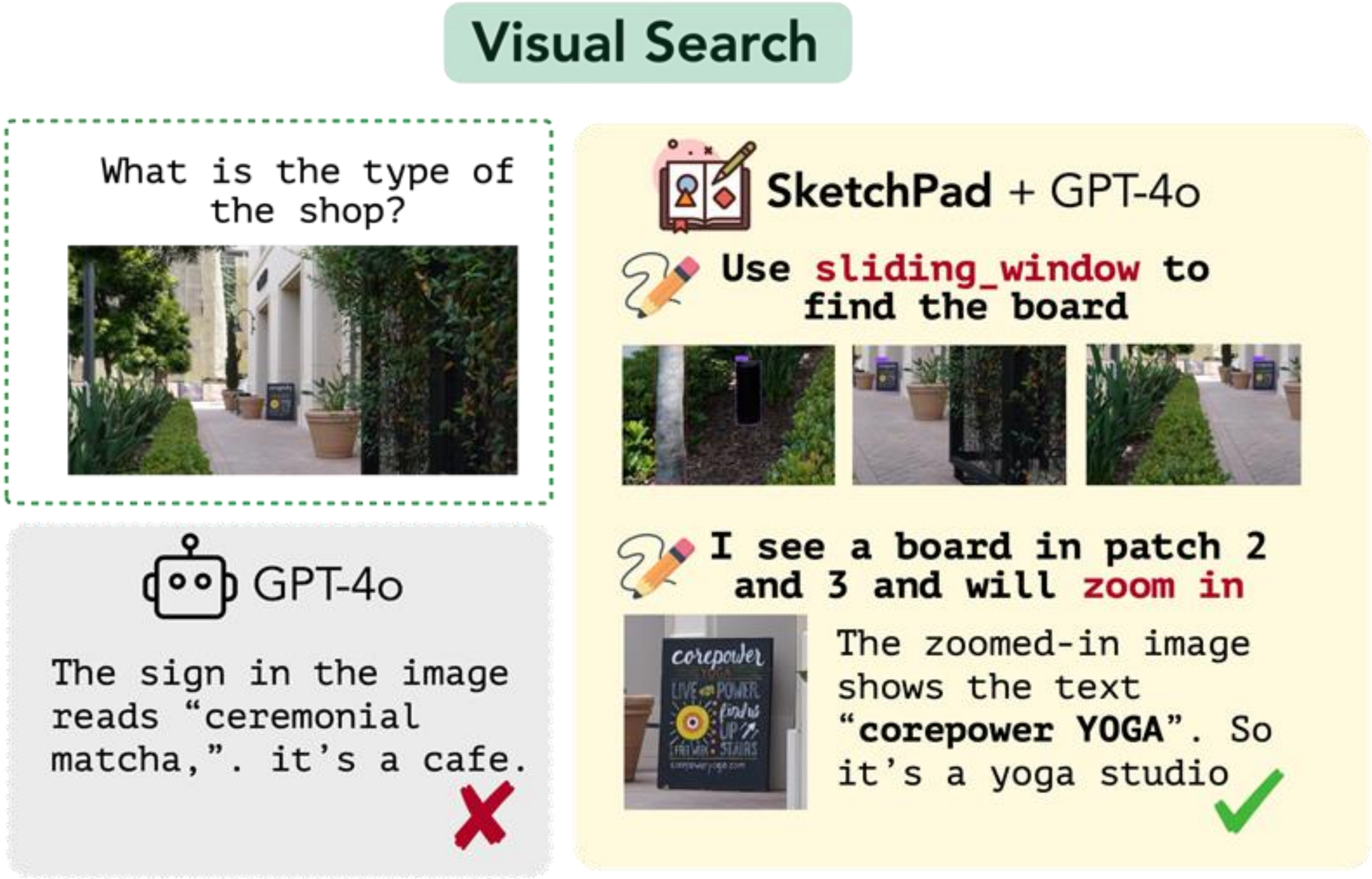
SketchPad + GPT-4o

Plot the function

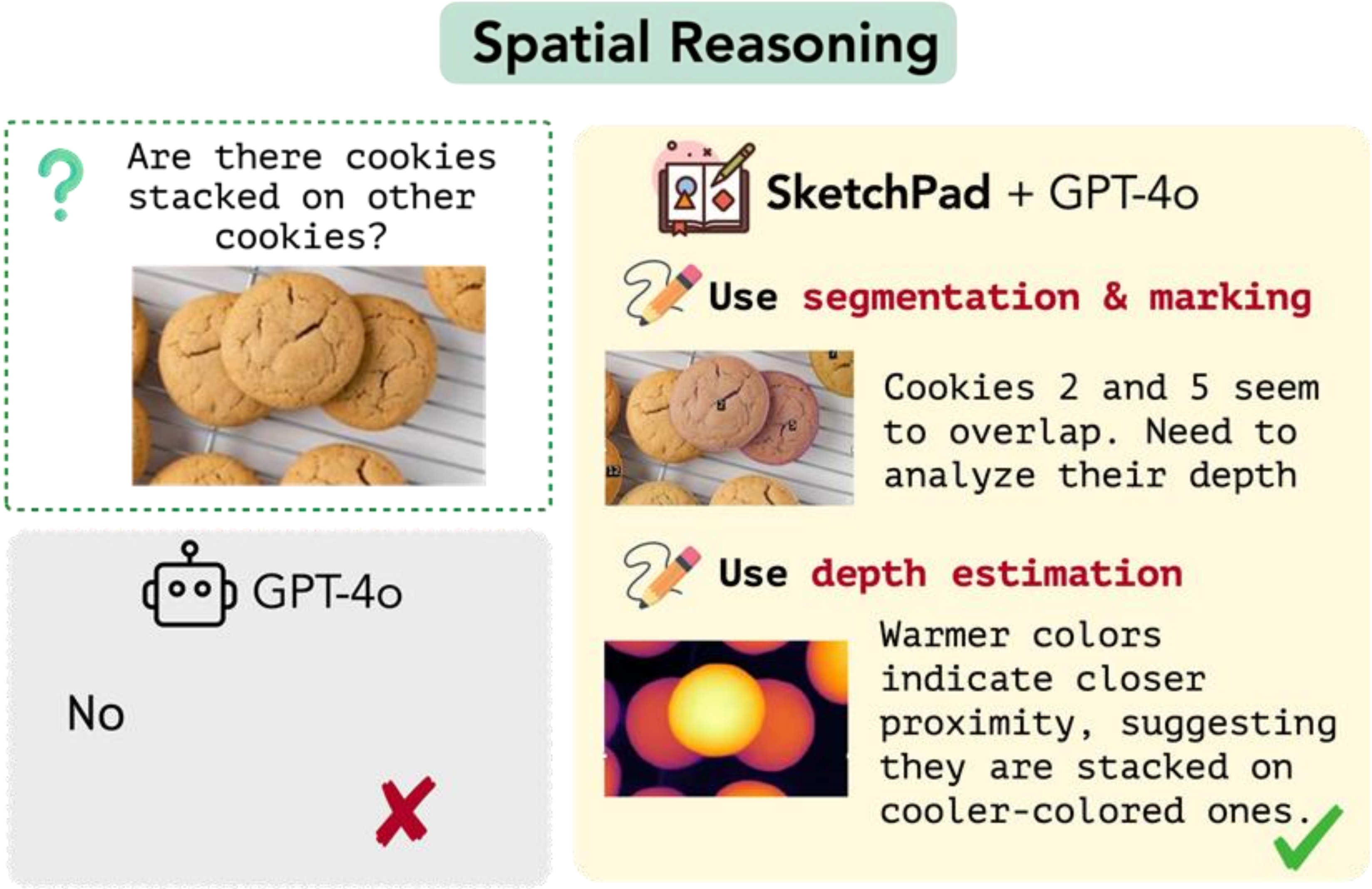


convex because line between any two points lie above the graph.

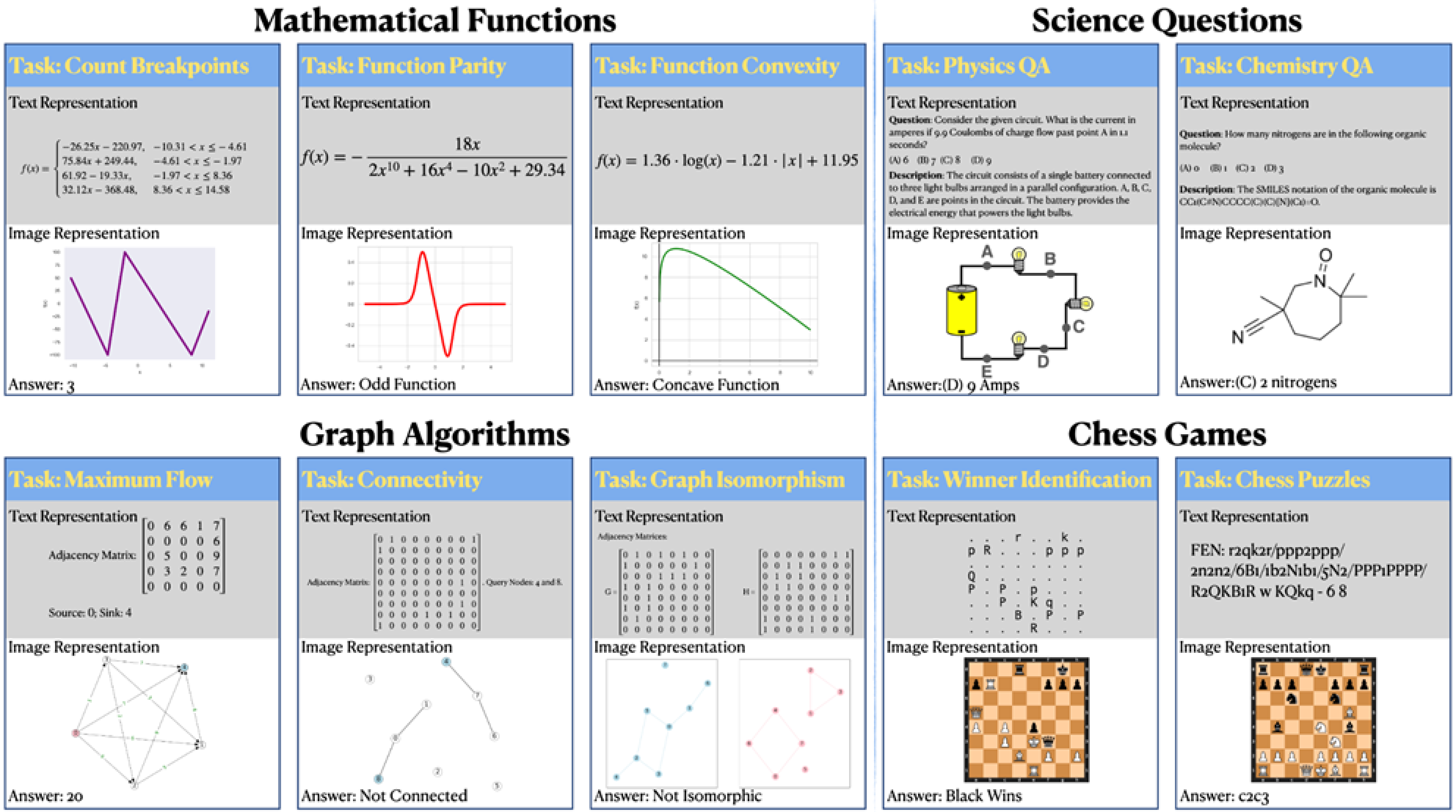
Sketchpad for LLM: **Visual Search**



Sketchpad for LLM: **Spatial Reasoning**



Sketching to Solve Math Problems

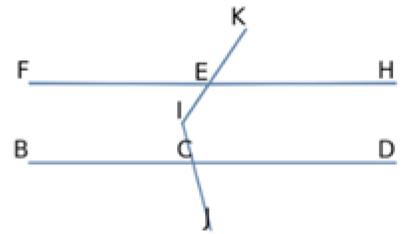


How do we enable VLMs to sketch?

React-style use of tools to sketch

Geometry

Given ∠BCJ=105°, ∠KEF=105°. Find ∠EIC

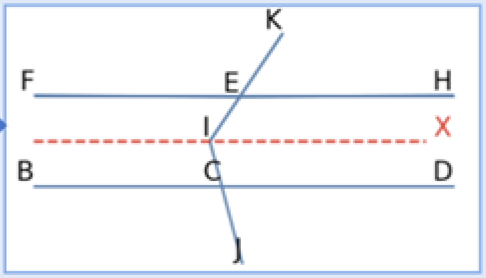


THOUGHT 1: I need to draw an auxiliary ***Line IX*** parallel to ***Line BD***

ACTION 1: Draw the aux. line using Python Code

```
# Define auxiliary point X for the line IX
x_I = points["I"][0]
y_I = points["I"][1]
x_X = x_I + 100
y_X = y_I + slope_bd * (x_X - x_I)
# Plot the auxiliary line IX in red color
ax.plot([x_I, x_X], [y_I, y_X], 'red') …
```

OBSERVATION 1



THOUGHT 2: Given ∠BCJ = ∠KEF = 105°, their supplementary angles, ∠KEH = ∠DCJ = 180° - 105° = 75°. Given BD parallel to IX, it follows that ∠EIX = ∠KEH, and ∠XIC = ∠DCJ, both measuring 75°. Therefore, ∠EIC = ∠EIX + ∠XIC = 150°

How do we enable GPT-4o to sketch?

React-style use of tools to sketch

Graph Connectivity

Is there a path between node 0 and node 4 in the graph?

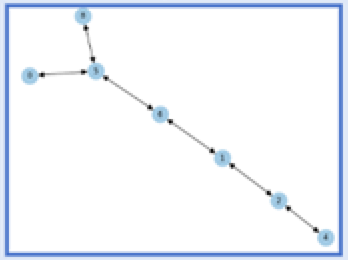
adjacency matrix =
```
[0, 0, 0, 0, 0, 1, 0, 0, 0],
[0, 0, 1, 0, 0, 0, 1, 0, 0],
[0, 1, 0, 0, 1, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 1, 0, 0, 0, 0, 0, 0],
[1, 0, 0, 0, 0, 0, 1, 0, 1],
[0, 1, 0, 0, 0, 1, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 1, 0, 0, 0]
```

THOUGHT 1:We can use the networkx library to plot the graph

ACTION 1: Call Python Ploting tool

```
G = nx.from_numpy_matrix(np.array(adjacency_matrix),
create_using=nx.DiGraph)
pos = nx.spring_layout(G)
nx.draw(G, pos, with_labels=True)
```

OBSERVATION 1



THOUGHT 2: From the graph, I saw there is a path between node 0 and node 4.

Sketching can solve math problems

	Geometry	Graph			Math		Game
Model	Geometry	Maxflow	Isomorphism	Connectivity	Convexity	Parity	Winner ID
Prior LLMs without visual inputs							
Gemini-Pro	\	15.6	47.7	50.0	87.9	48.2	8.1
Claude 3 OPUS	\	56.3	50.0	82.0	93.0	77.6	74.4
Mixtral 8x7B [19]	\	8.6	50.0	62.5	69.1	41.7	7.4
LLaMA-2-70B [45]	\	18.0	50.0	50.0	74.2	33.3	12.4
Latest multimodal LLMs + Visual Sketchpad							
GPT-4 Turbo	37.5	32.8	62.5	66.0	57.0	80.5	50.4
+ Sketchpad	45.8	63.3	64.2	95.1	**93.1**	**93.1**	74.3
	+8.3	+30.5	+1.7	+29.1	+25.4	+12.6	+23.9
GPT-4o	62.5	25.0	50.8	96.1	87.2	84.4	61.1
+ Sketchpad	**66.7**	**66.3**	**65.3**	**98.1**	90.1	88.1	**81.2**
	+4.2	+41.3	+14.5	+2.0	+2.9	+3.7	+20.1

Table 1: Accuracy scores on geometry problems, graph algorithms, mathematical functions, and game. **SKETCHPAD yields large performance gains across all tasks and outperform all baselines.**

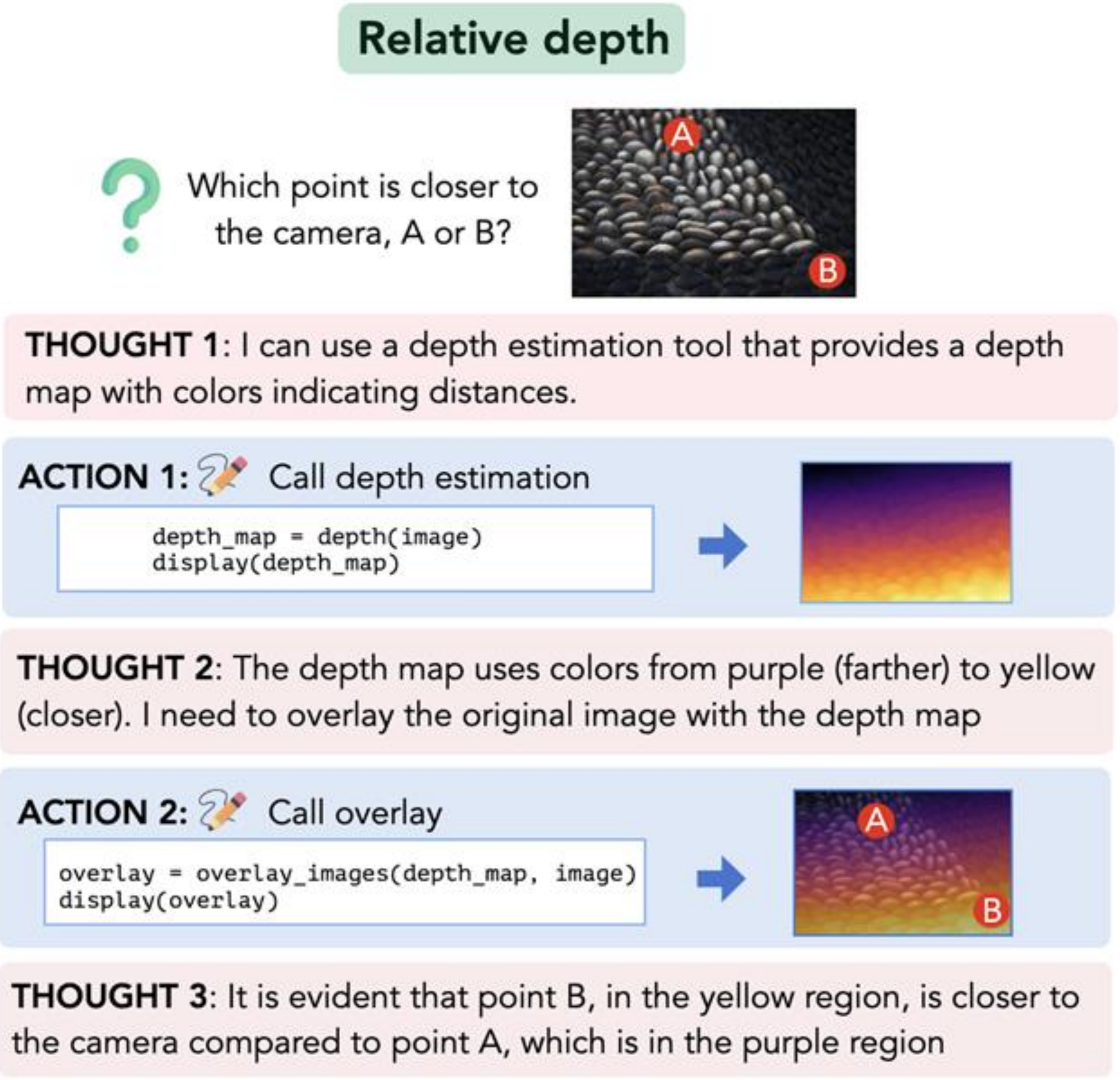
Large improvements across multiple benchmarks

	Geometry	Graph			Math		Game
Model	Geometry	Maxflow	Isomorphism	Connectivity	Convexity	Parity	Winner ID
Prior LLMs without visual inputs							
Gemini-Pro	\	15.6	47.7	50.0	87.9	48.2	8.1
Claude 3 OPUS	\	56.3	50.0	82.0	93.0	77.6	74.4
Mixtral 8x7B [19]	\	8.6	50.0	62.5	69.1	41.7	7.4
LLaMA-2-70B [45]	\	18.0	50.0	50.0	74.2	33.3	12.4
Latest multimodal LLMs + Visual Sketchpad							
GPT-4 Turbo	37.5	32.8	62.5	66.0	57.0	80.5	50.4
+ Sketchpad	45.8	63.3	64.2	95.1	**93.1**	**93.1**	74.3
	+8.3	+30.5	+1.7	+29.1	+25.4	+12.6	+23.9
GPT-4o	62.5	25.0	50.8	96.1	87.2	84.4	61.1
+ Sketchpad	**66.7**	**66.3**	**65.3**	**98.1**	90.1	88.1	**81.2**
	+4.2	+41.3	+14.5	+2.0	+2.9	+3.7	+20.1

Table 1: Accuracy scores on geometry problems, graph algorithms, mathematical functions, and game. **SKETCHPAD yields large performance gains across all tasks and outperform all baselines.**

Sketching to solve BLINK

Example sketching on BLINK



Example sketching on BLINK

Sketching to solve MMVP (Eyes wide shut)

Sketching can solve perception tasks

Model	V^*Bench	MMVP	Depth	Spatial	Jigsaw	Vis. Corr.	Sem. Corr.
Prior multimodal LLMs							
LLaVA-1.5-7B [26]	48.7	-	52.4	61.5	11.3	25.6	23.0
LLaVA-1.5-13B [26]	-	24.7	53.2	67.8	58.0	29.1	32.4
LLaVA-NeXT-34B [27]	-	-	67.7	74.8	54.7	30.8	23.7
Claude 3 OPUS [1]	-	-	47.6	58.0	32.7	36.6	25.2
Gemini-Pro [41]	48.2	40.7	40.3	74.8	57.3	42.4	26.6
GPT-4V-preview [35]	55.0	38.7	59.7	72.7	70.0	33.7	28.8
Previous state of the art	75.4 [50]	49.3 [10]	67.7 [27]	76.2 [42]	70.0 [35]	42.4 [41]	33.1 [48]
Latest multimodal LLMs + Visual Sketchpad							
GPT-4 Turbo	52.5	71.0	66.1	68.5	64.7	48.8	30.9
+ Sketchpad	71.0	73.3	68.5	80.4	68.5	52.3	42.4
	+18.5	+2.3	+2.4	+11.9	+3.8	+3.5	+11.5
GPT-4o	66.0	85.3	71.8	72.0	64.0	73.3	48.6
+ Sketchpad	**80.3**	**86.3**	**83.9**	**81.1**	**70.7**	**80.8**	**58.3**
	+14.3	+1.0	+12.1	+9.1	+6.7	+7.5	+9.7

Table 2: Accuracy on complex visual reasoning tasks. **SKETCHPAD enhances both GPT-4 Turbo and GPT-4o performance, establishing new SOTA performance levels on all the tasks.**

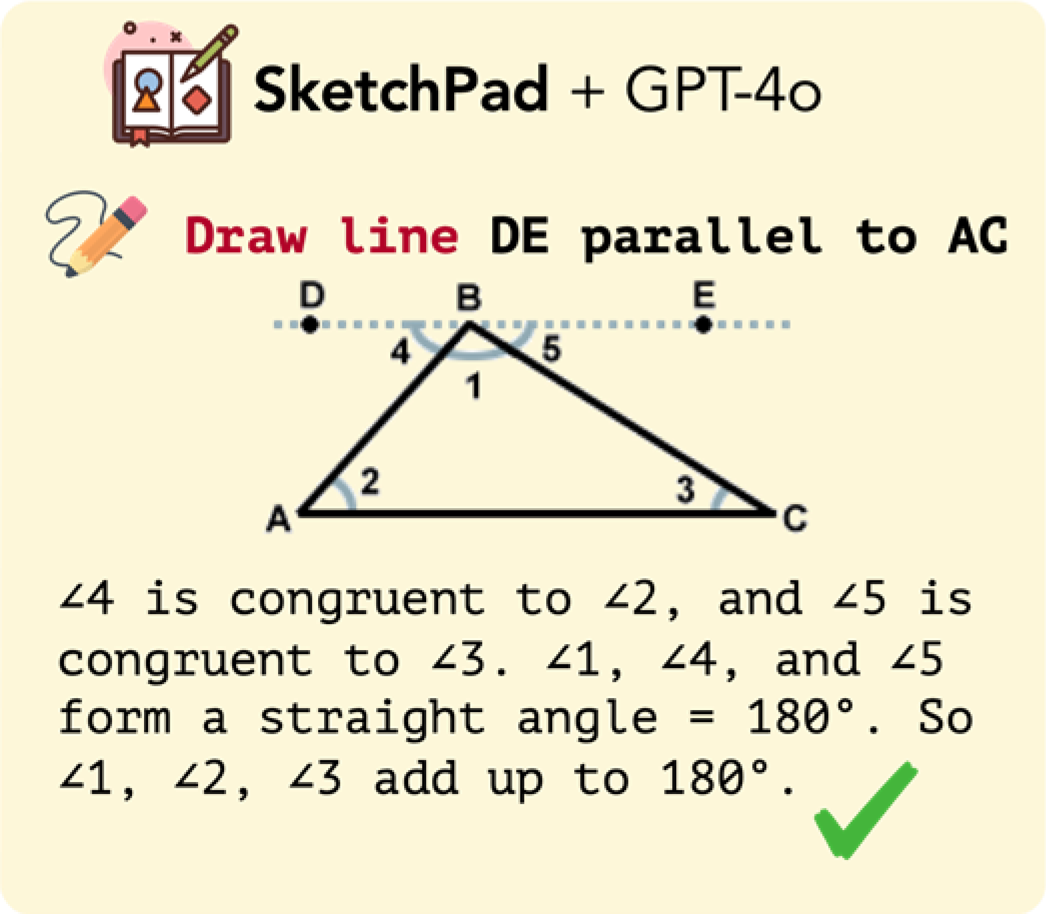
Large improvements across multiple benchmarks

Model	V^*Bench	MMVP	Depth	Spatial	Jigsaw	Vis. Corr.	Sem. Corr.
Prior multimodal LLMs							
LLaVA-1.5-7B [26]	48.7	-	52.4	61.5	11.3	25.6	23.0
LLaVA-1.5-13B [26]	-	24.7	53.2	67.8	58.0	29.1	32.4
LLaVA-NeXT-34B [27]	-	-	67.7	74.8	54.7	30.8	23.7
Claude 3 OPUS [1]	-	-	47.6	58.0	32.7	36.6	25.2
Gemini-Pro [41]	48.2	40.7	40.3	74.8	57.3	42.4	26.6
GPT-4V-preview [35]	55.0	38.7	59.7	72.7	70.0	33.7	28.8
Previous state of the art	75.4 [50]	49.3 [10]	67.7 [27]	76.2 [42]	70.0 [35]	42.4 [41]	33.1 [48]
Latest multimodal LLMs + Visual Sketchpad							
GPT-4 Turbo	52.5	71.0	66.1	68.5	64.7	48.8	30.9
+ Sketchpad	71.0	73.3	68.5	80.4	68.5	52.3	42.4
	+18.5	+2.3	+2.4	+11.9	+3.8	+3.5	+11.5
GPT-4o	66.0	85.3	71.8	72.0	64.0	73.3	48.6
+ Sketchpad	**80.3**	**86.3**	**83.9**	**81.1**	**70.7**	**80.8**	**58.3**
	+14.3	+1.0	+12.1	+9.1	+6.7	+7.5	+9.7

Table 2: Accuracy on complex visual reasoning tasks. **SKETCHPAD enhances both GPT-4 Turbo and GPT-4o performance, establishing new SOTA performance levels on all the tasks.**

Humans draw the same auxiliary lines for Geometry questions

(N=2) human participants, when asked to solve a geometry problem, draw the same auxiliary lines as GPT-4o
80% of the time.



GPT-4o correctly uses vision tools to sketch on images

Human evaluation of GPT-4o plans finds that the tool usage is valid in 92.8% of instances.

Most of the remaining errors on these benchmarks are because of failures of specialized models, not planning

Open-source models also improve if they have GPT-4o generated sketches

Model	Geometry	Maxflow	Convexity	Winner ID
LLaVA-NeXT-13B	11.1	7.8	50.39	5.8
+ oracle Sketchpad	22.2	10.2	50.0	36.7
LLaVA-NeXT-34B	26.1	0.8	81.6	49.0
+ oracle Sketchpad	28.3	14.1	87.1	49.4

But open sourced models aren't able to effectively sketch

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

Most fundamental vision capabilities are still out of reach

Enabling sketching: visual chain of thought

How can we distill from specialist models into generalist VLMs?

Complete open Action Reasoning model for robotics

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

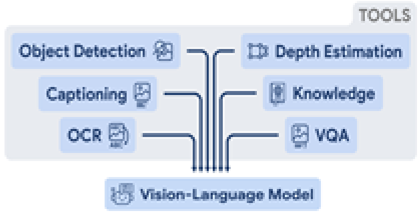
Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

BLINK
Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Enabling sketching: visual chain of thought



How can we distill from specialist models into generalist VLMs?

MolmoAct

Complete open Action Reasoning model for robotics

Strengths of sketching

- LLMs are good at **generating code**.
- We have good **specialized models** for tasks such as:
    - low level vision tasks (e.g., object detectors, segmentation, depth)
    - search & retrieval
    - encyclopedia knowledge
    - domain-specific models

Limitations of sketching

- Error accumulation:
  - Programs can be wrong
  - Tools can be wrong

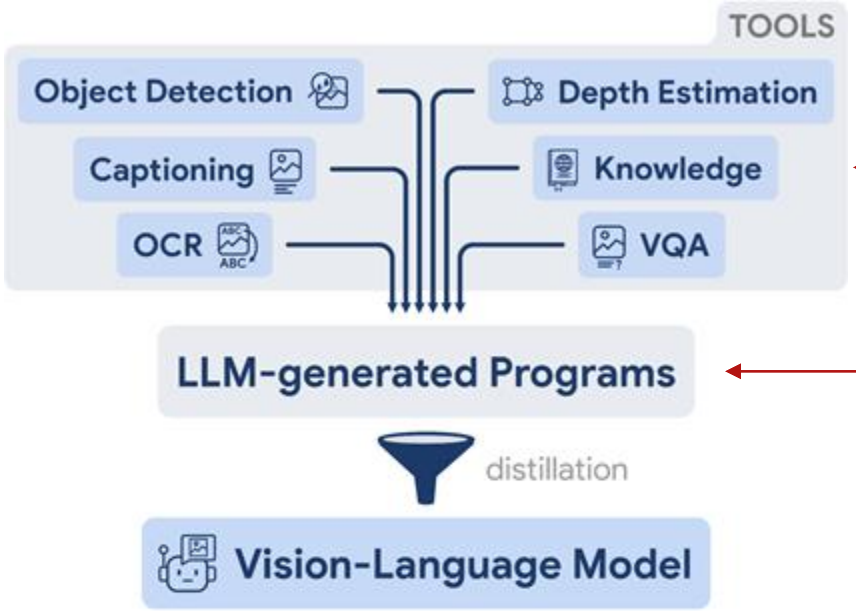
- Latency:
  - Inference for each sample requires generating code and calling multiple tools.

Q: How many red balloons are there?

GPT-4o + sketching: 1

Can we distil perceptual capabilities from specialist models to VLMs?

Our solution: visual program distillation (VPD)



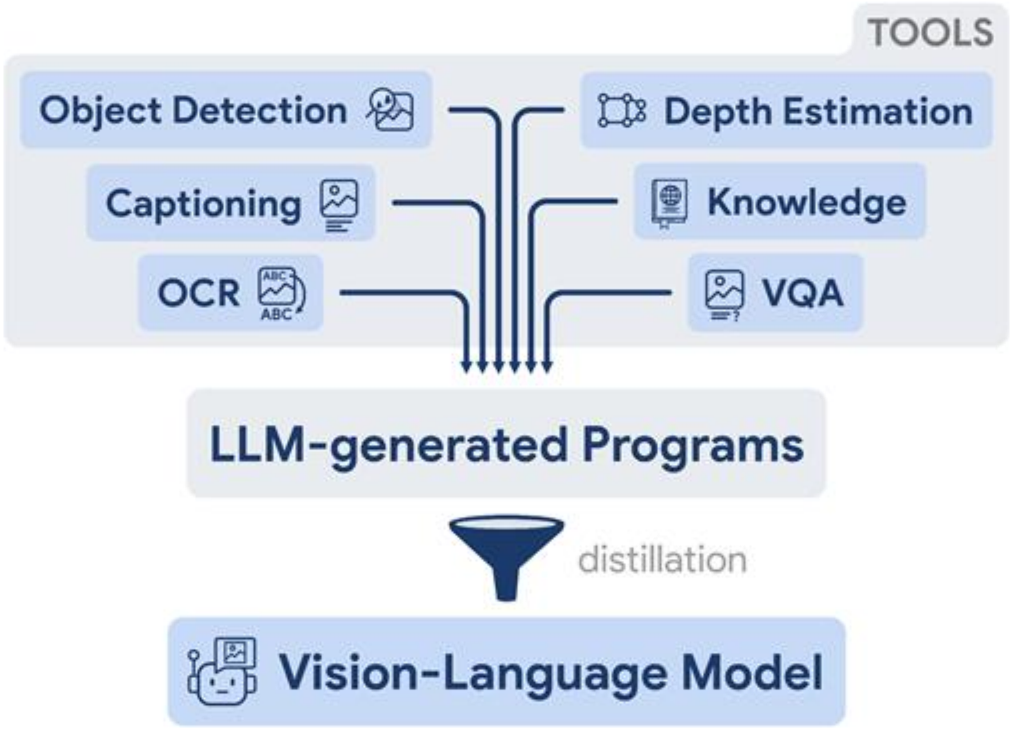
Learn from specialist models

Learn how to compose skills from programs

Visual Program Distillation (VPD): Training phase

Step 1: Generate training data using programs.

Step 2: Fine-tune VLMs on the generated data.



Visual Program Distillation (VPD)

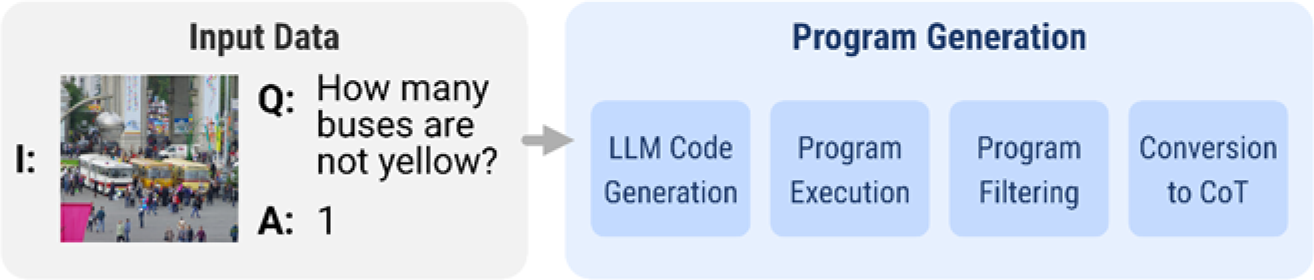
Input Data

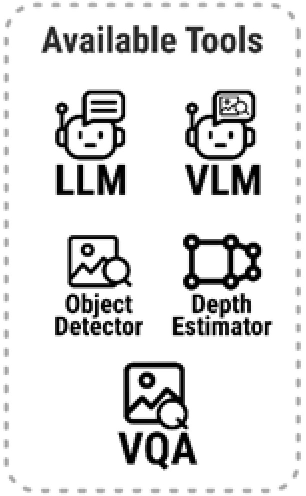
I: 

Q: How many buses are not yellow?

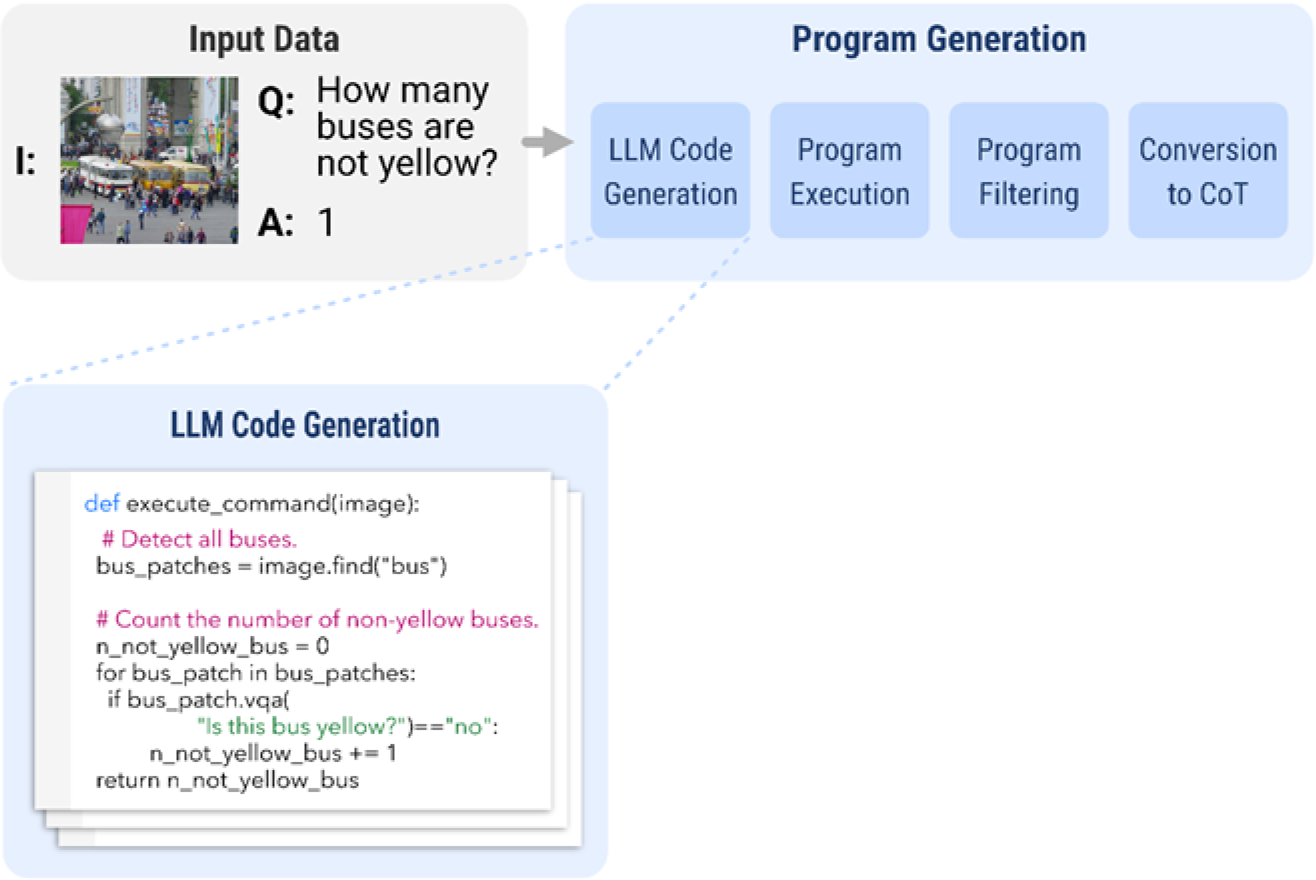
A: 1

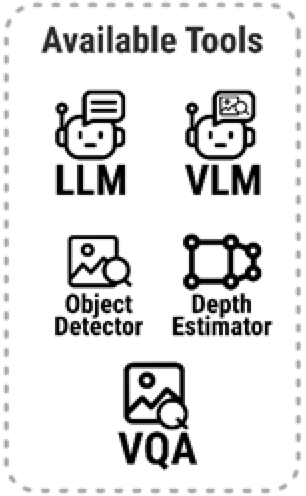
Visual Program Distillation (VPD): Training phase



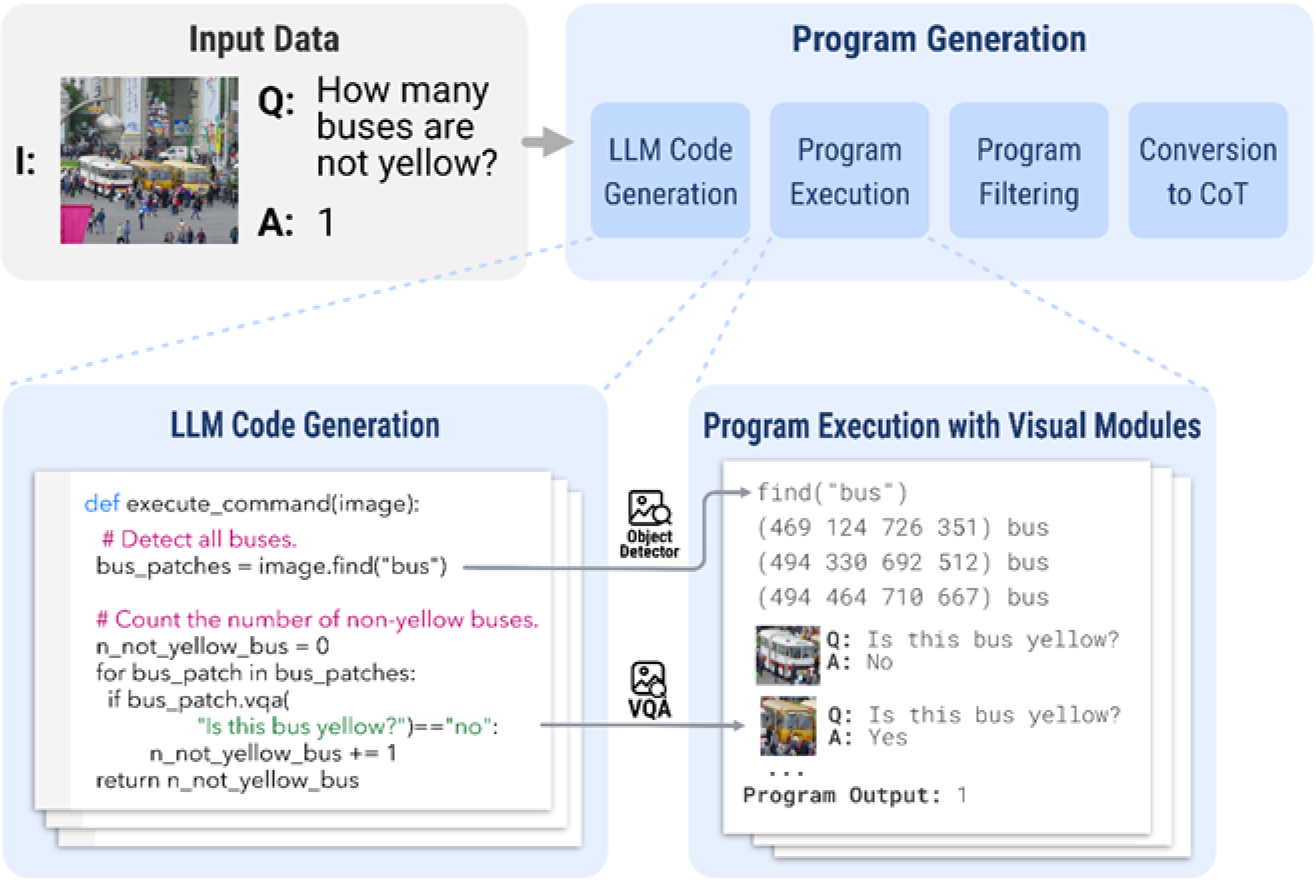


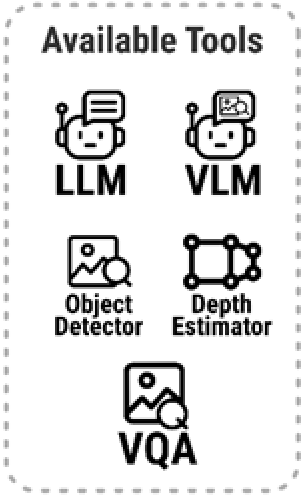
Visual Program Distillation (VPD): Training phase



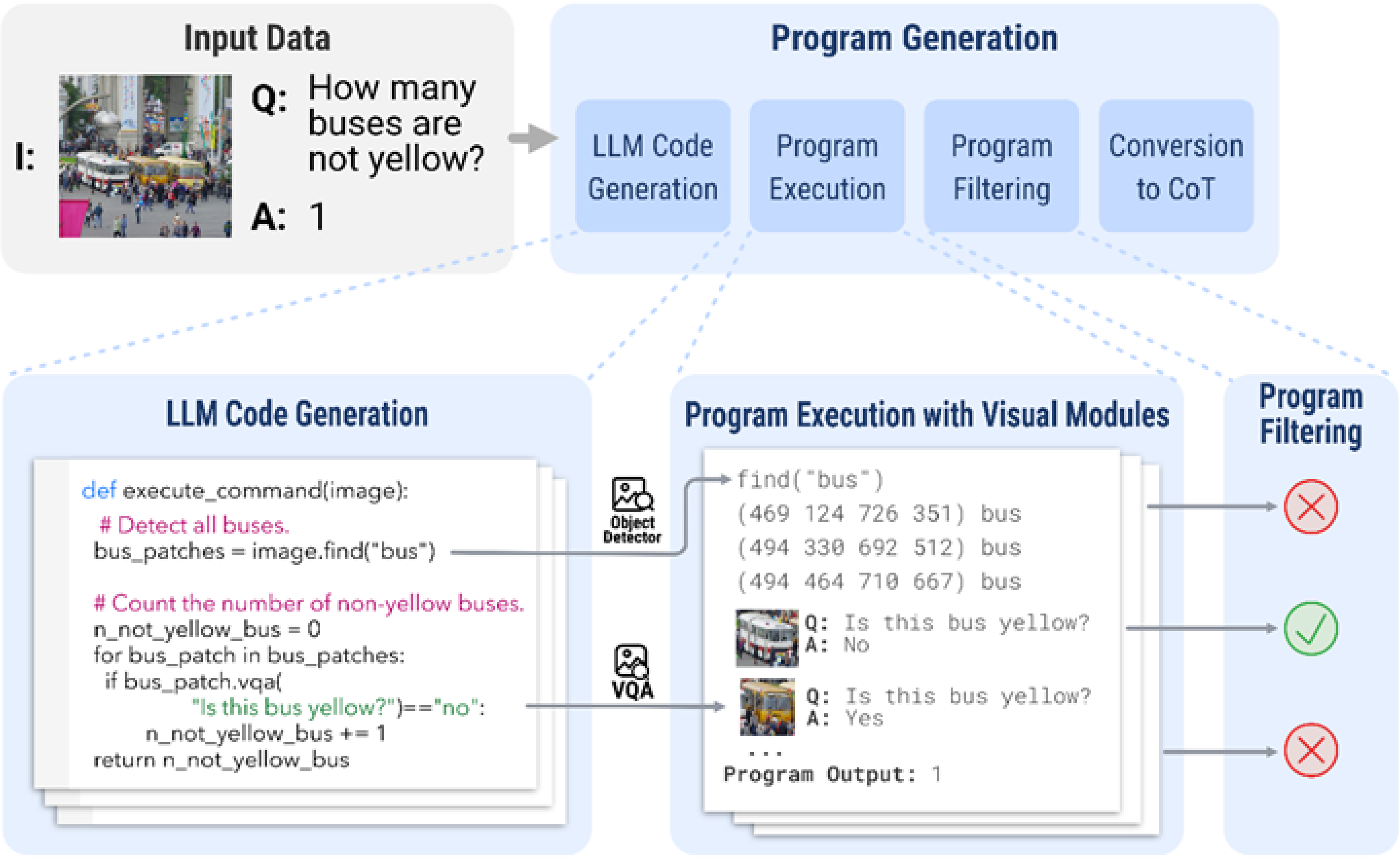


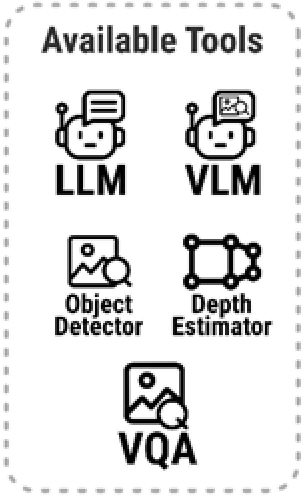
Visual Program Distillation (VPD): Training phase



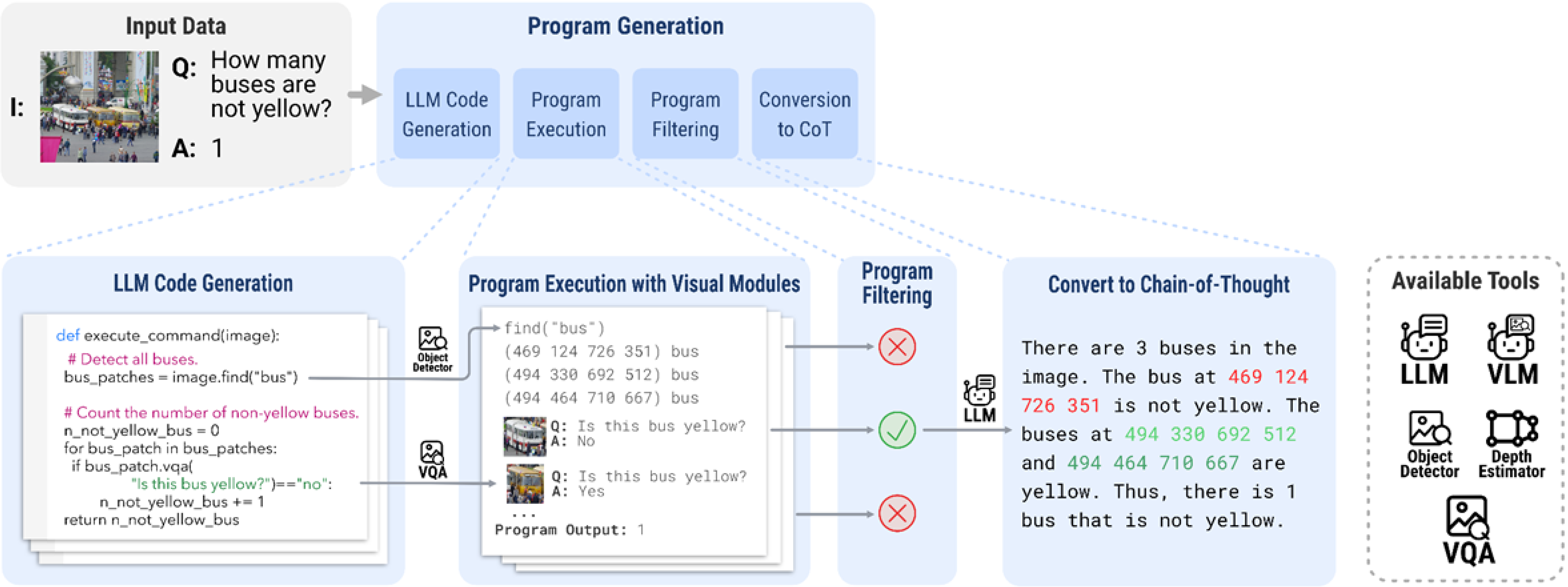


Visual Program Distillation (VPD): Training phase

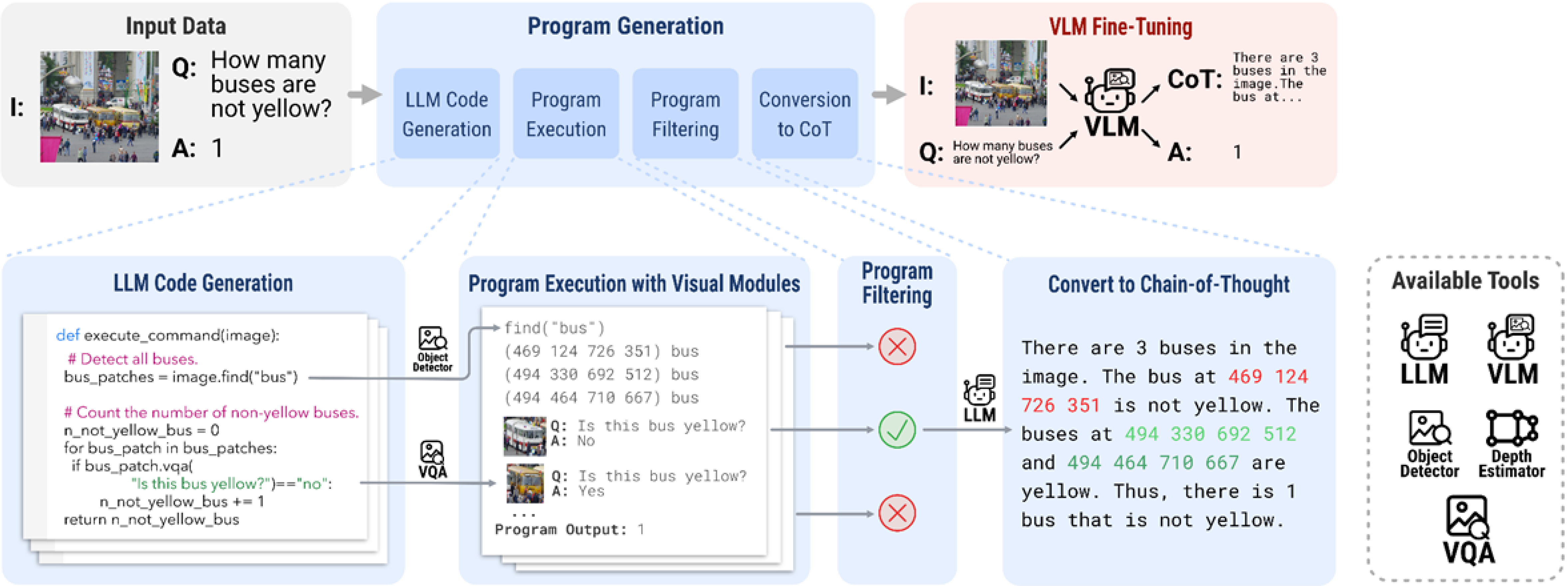




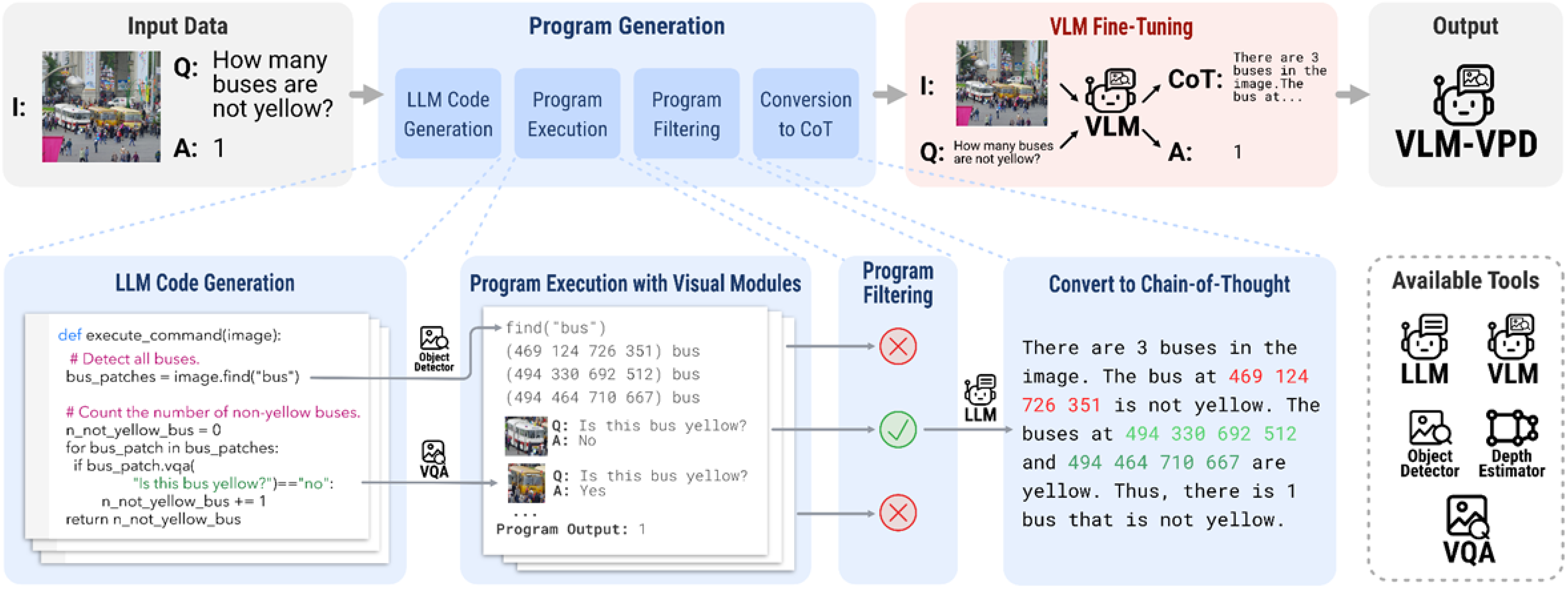
Visual Program Distillation (VPD): Training phase



Visual Program Distillation (VPD): Training phase



Visual Program Distillation (VPD): Training phase



Visual Program Distillation (VPD): Inference phase

- One forward pass!
- No code generation, tool usage, etc.
- VLMs produce interpretable reasoning steps.

How many green vases are there?

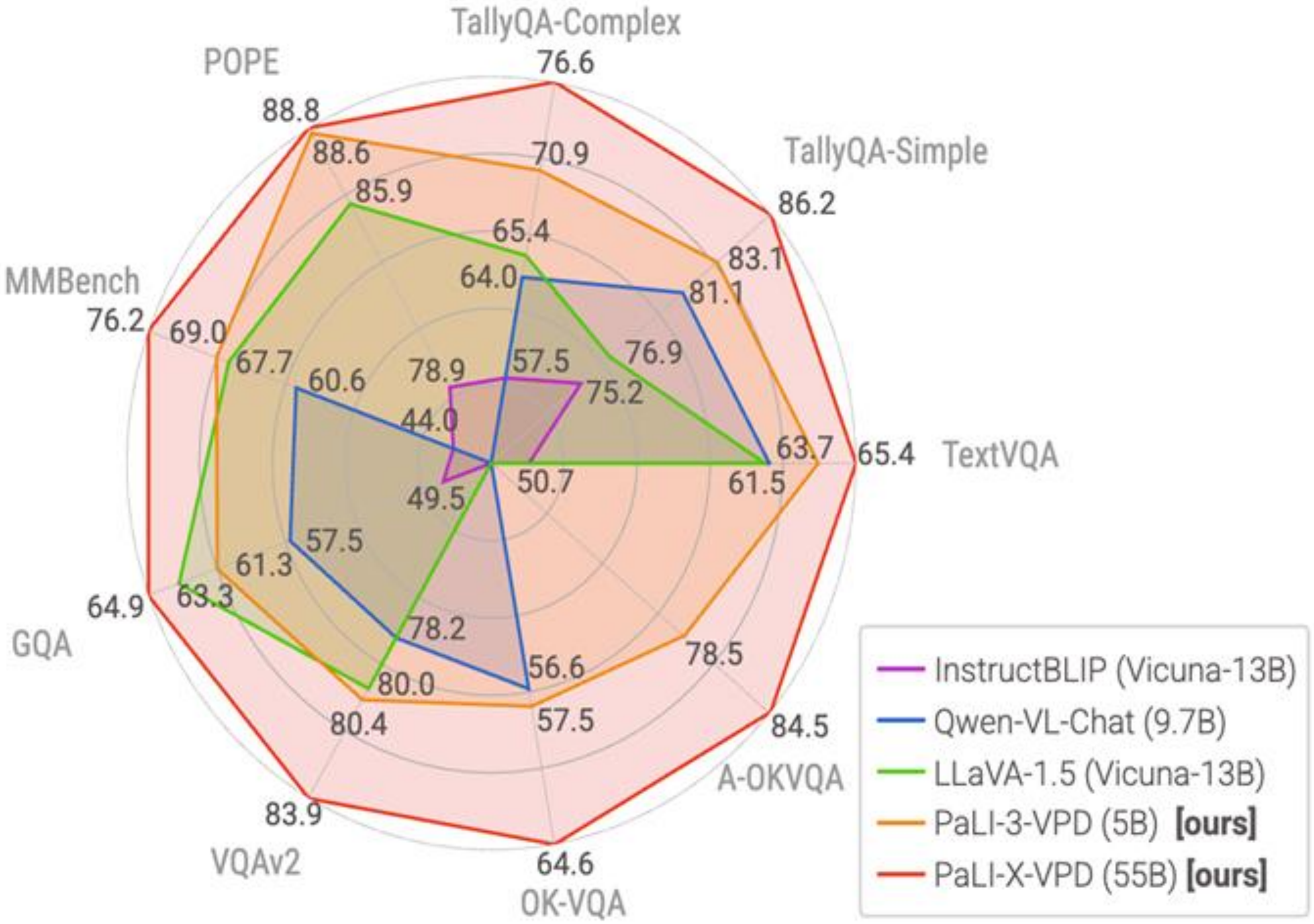
There are 5 green vases.

The vases at 348 139 771 371, 207 409 410 549, and 286 613 659 792 are green. Thus, there are 3 green vases.

Key results

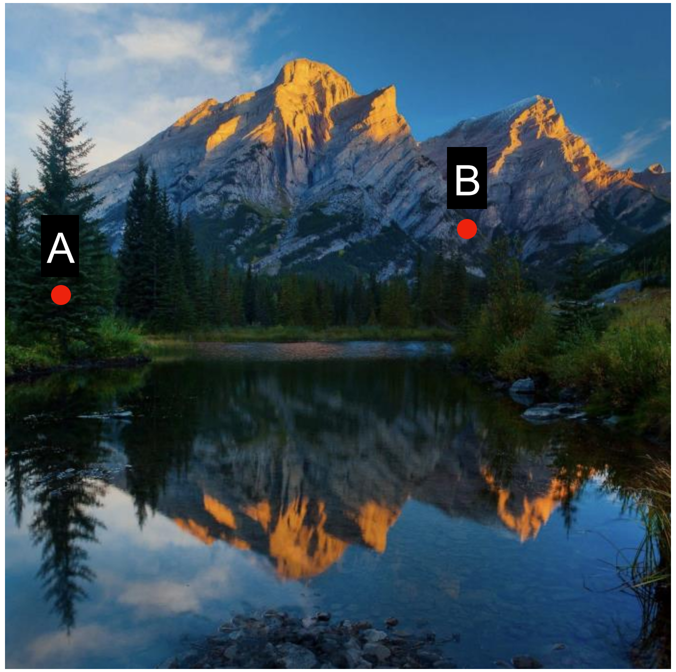
- VPD improves VLM’ s accuracy, consistency, and factuality.
- During inference, VPD model produces interpretable and faithful visual reasoning in one forward pass.
- Sets a new SotA on a wide range of VQA benchmarks.

VPD held SOTA results for most benchmarks until GPT-4o1



Let's go back to our BLINK task for a second

What chain-of-thought would help solve this task?

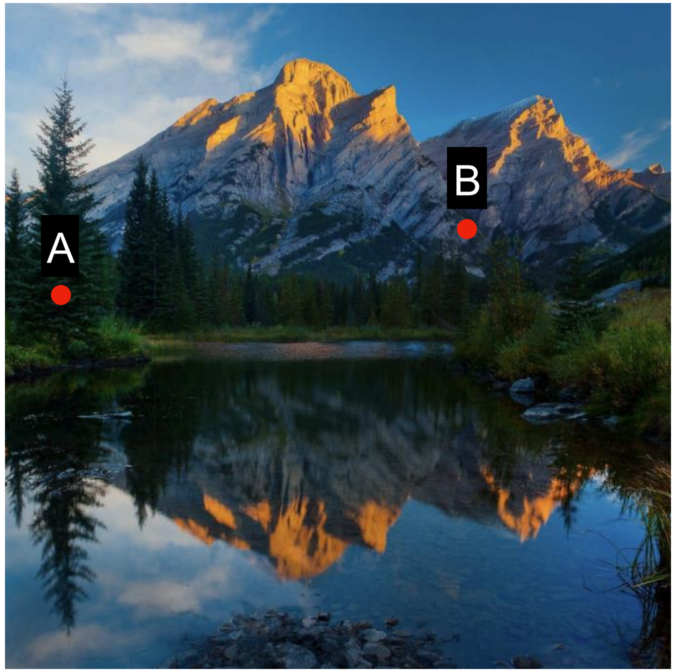


Which point is closer to the camera?

Let' s go back to our BLINK task for a second

What chain-of-thought would help solve this task?

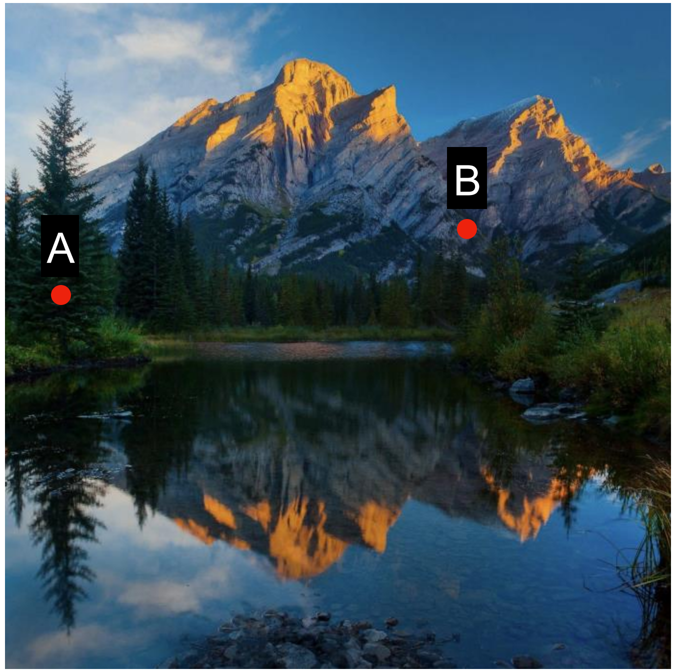
Depth estimation at:
- point A
- point B



Which point is closer to the camera?

Let’ s go back to our BLINK task for a second

What chain-of-thought would help solve this task?

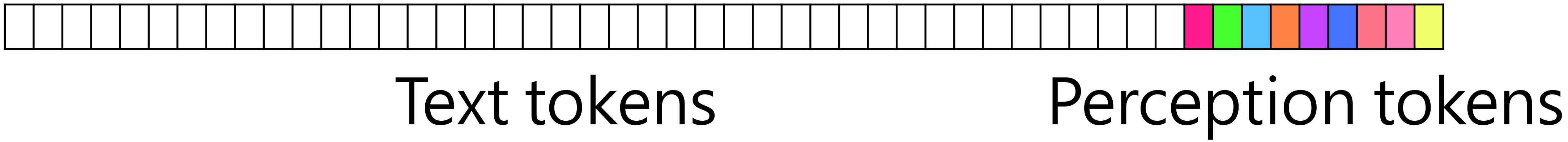


Which point is closer to the camera?

Depth estimation at:
- point A
- point B

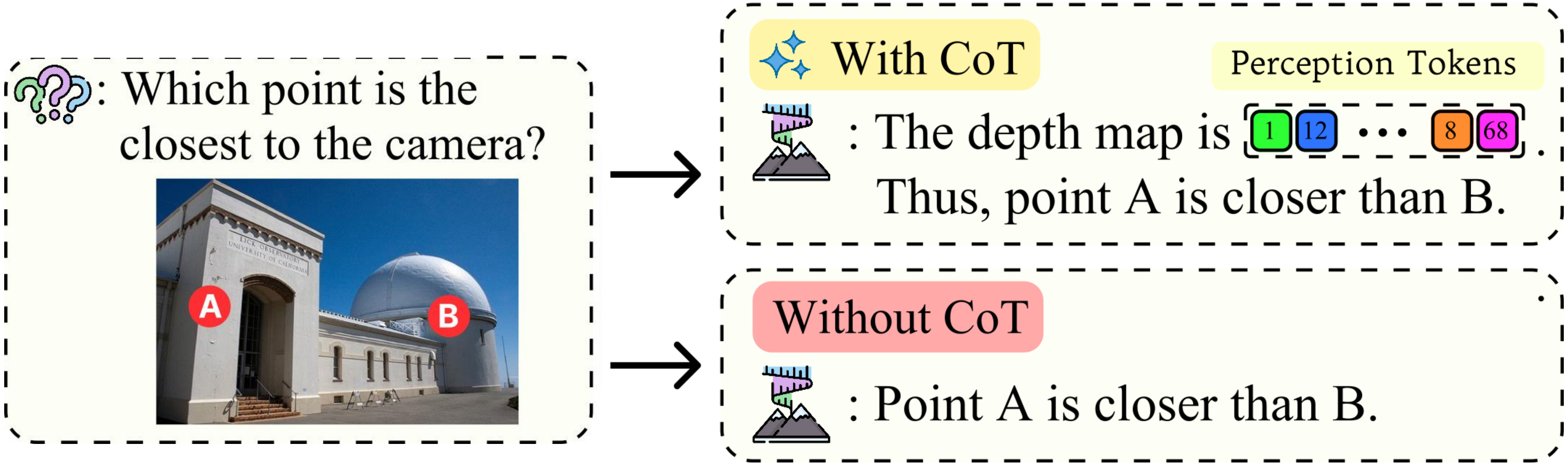
but expressing depth estimations in language leads to hallucinations

We introduce perception tokens



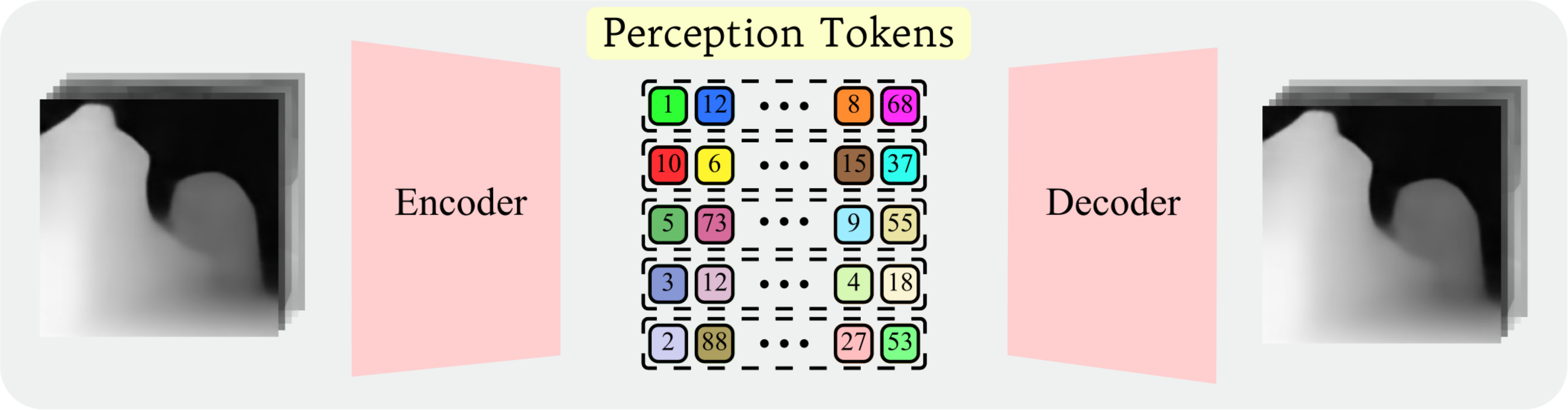
We enable models to generate tokens that produce tokens that can be decoded into implicit images

We enable models to use perceptual tokens to reason

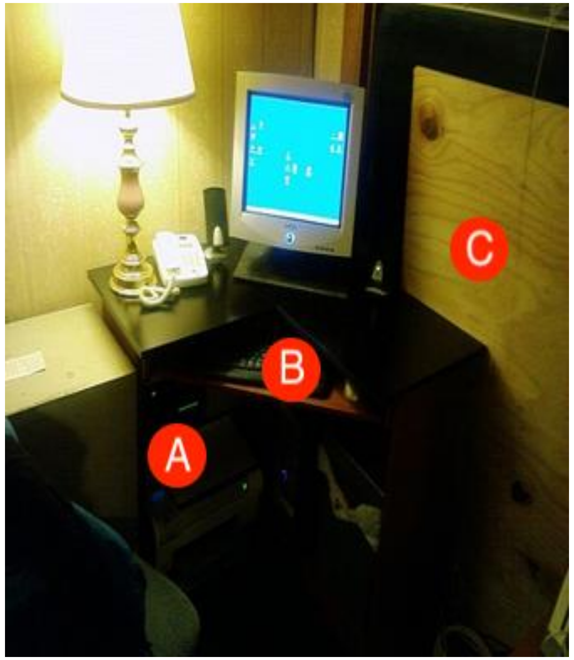


Bigverdi, Luo, Hsieh, Shen, Chen, Shapiro, Krishna. **Perception Tokens Enhance Visual Reasoning in Multimodal Language Models**. ArXiv 2025

The model doesn’ t produce explicitly depth maps
Instead, it produces implicit depth maps



Bigverdi, Luo, Hsieh, Shen, Chen, Shapiro, Krishna. **Perception Tokens Enhance Visual Reasoning in Multimodal Language Models**. ArXiv 2025

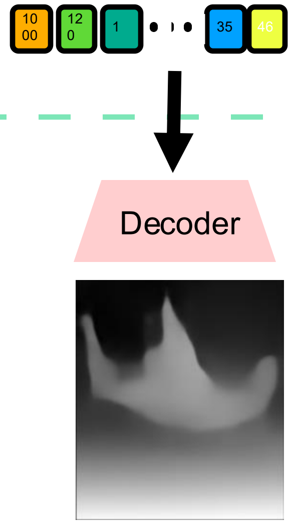


Which point is the closest to the camera?

Baseline Answer: C

Ours: Answer: B

A: (100,230), B:(170,190), C:(280,125) and depth map:
Point B has higher pixel value on the depth map.



Bigverdi, Luo, Hsieh, Shen, Chen, Shapiro, Krishna. **Perception Tokens Enhance Visual Reasoning in Multimodal Language Models**. ArXiv 2025

Visual reasoning outperforms GPT-4o on BLINK

Model	Training: Direct Labeling Data	Training: Depth Generation Data	Training: CoT Data	BLINK [9] 2 Points	HardBLINK 3 Points	HardBLINK 4 Points	HardBLINK 5 Points	Average
LLaVA OneVision	✗	✗	✗	51.6	33.1	22.6	18.5	31.4
LLaVA 1.5 13B	✗	✗	✗	54.0	35.5	37.9	29	39.1
Fine-tunned LLaVA	✓	✗	✗	68.5	58.9	52.4	41.1	55.2
AURORA (Ours)	✓	✓	✓	64.5	**66.9**	**60.5**	**54.8**	**61.6**
GPT-4o	✗	✗	✗	53.2	58.9	50	36.3	49.6
GPT-4 Turbo	✗	✗	✗	58.1	54.8	41.9	32.2	46.7
GPT-4 Turbo + Tool	✗	✗	✗	70.2	57.2	44.3	26.6	49.6

Bigverdi, Luo, Hsieh, Shen, Chen, Shapiro, Krishna. **Perception Tokens Enhance Visual Reasoning in Multimodal Language Models**. ArXiv 2025

How many people are in the iamge?

Ours: Answer: 4

Bounding Box 1: (130 30 225 300 115) Bounding Box 2: (240 115) 300 115 250

Bounding Box 3: (250 130 335 335) Bounding Box 4: (250 150) 335 230

Our training algorithm

- Auto-encode perception tokens
- Curate data such that chain of thought prompting produces perception tokens
- Multi tasking data (with and without CoT)
- Curriculum learning

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

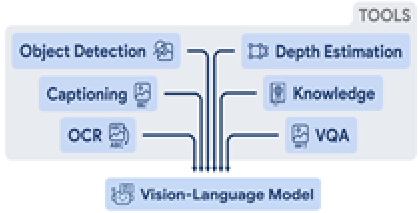
Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

Most fundamental vision capabilities are still out of reach

Enabling sketching: visual chain of thought



How can we distill from specialist models into generalist VLMs?

Complete open Action Reasoning model for robotics

Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

Most fundamental vision capabilities are still out of reach

Enabling sketching: visual chain of thought

How can we distill from specialist models into generalist VLMs?

MolmoAct

Complete open Action Reasoning model for robotics

We want the models to work in our homes

Sampled from Pi & TRI

MolmoAct

A Family of Completely Open
Vision-Language-Action Model

How does Molmo-Act work?

Input Image

"Put the plate in dishwasher"

Language Instruction

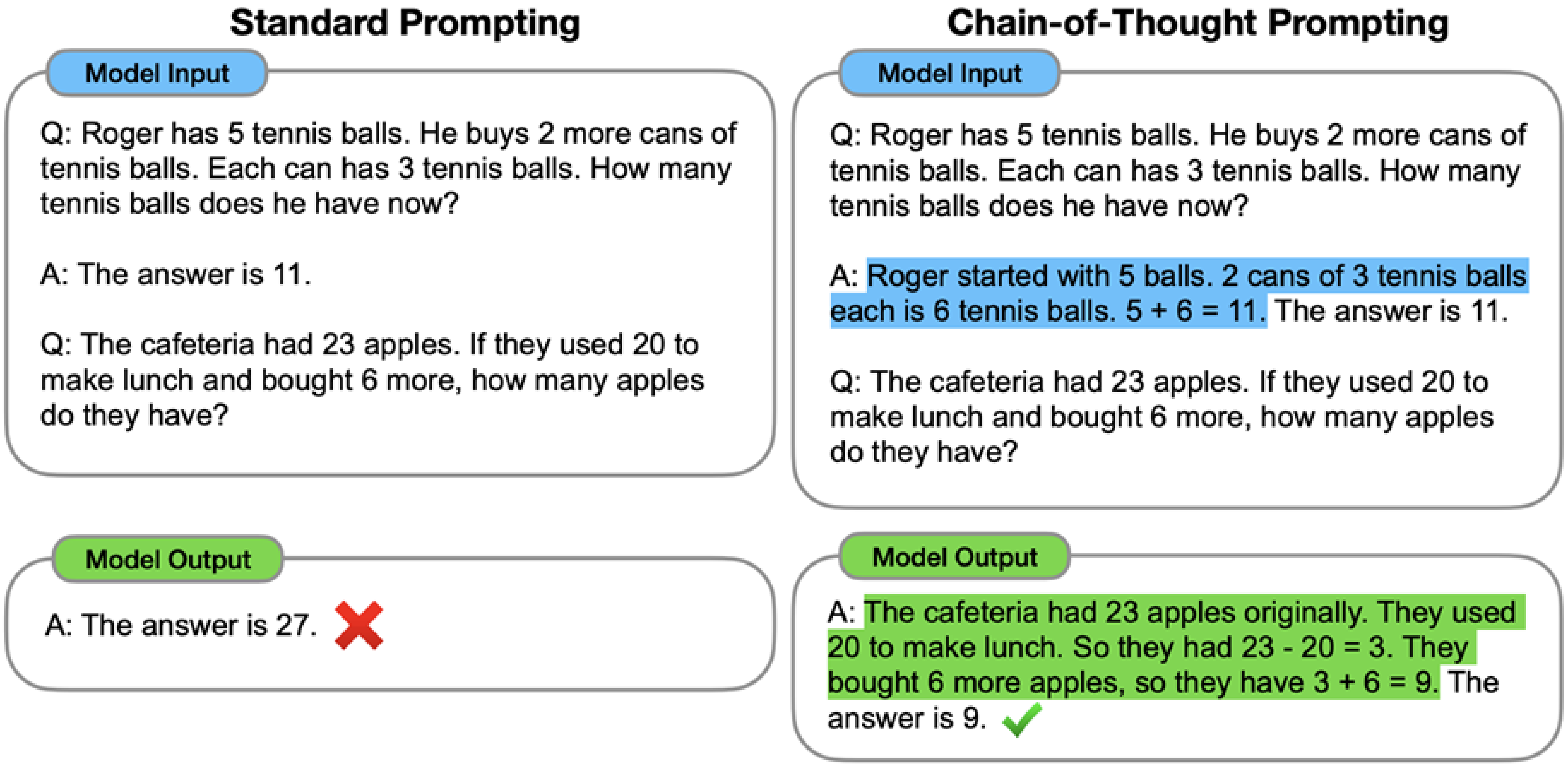
Molmo-Act

$$\begin{bmatrix} \Delta x \\ \Delta\theta \\ \text{Gripper} \end{bmatrix}$$

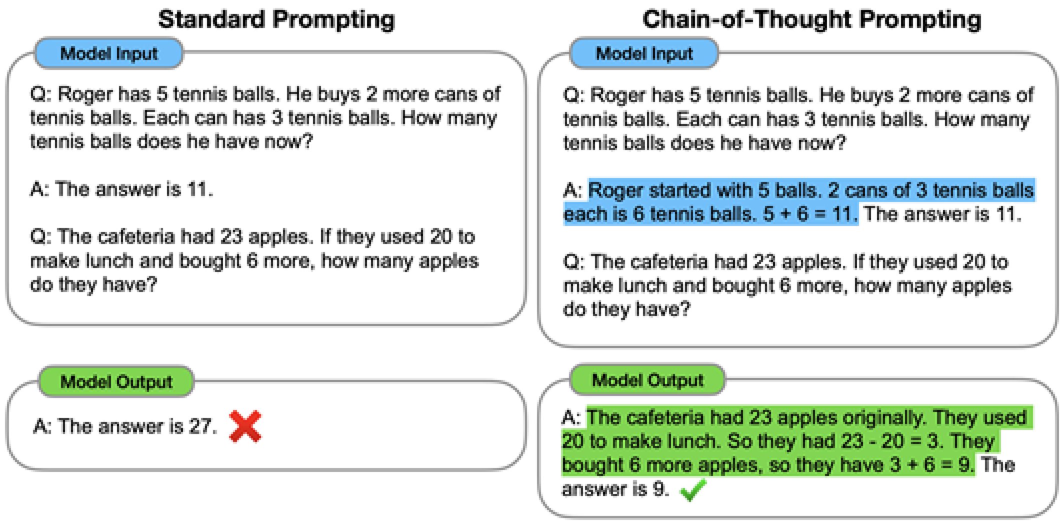
7D Robot Action

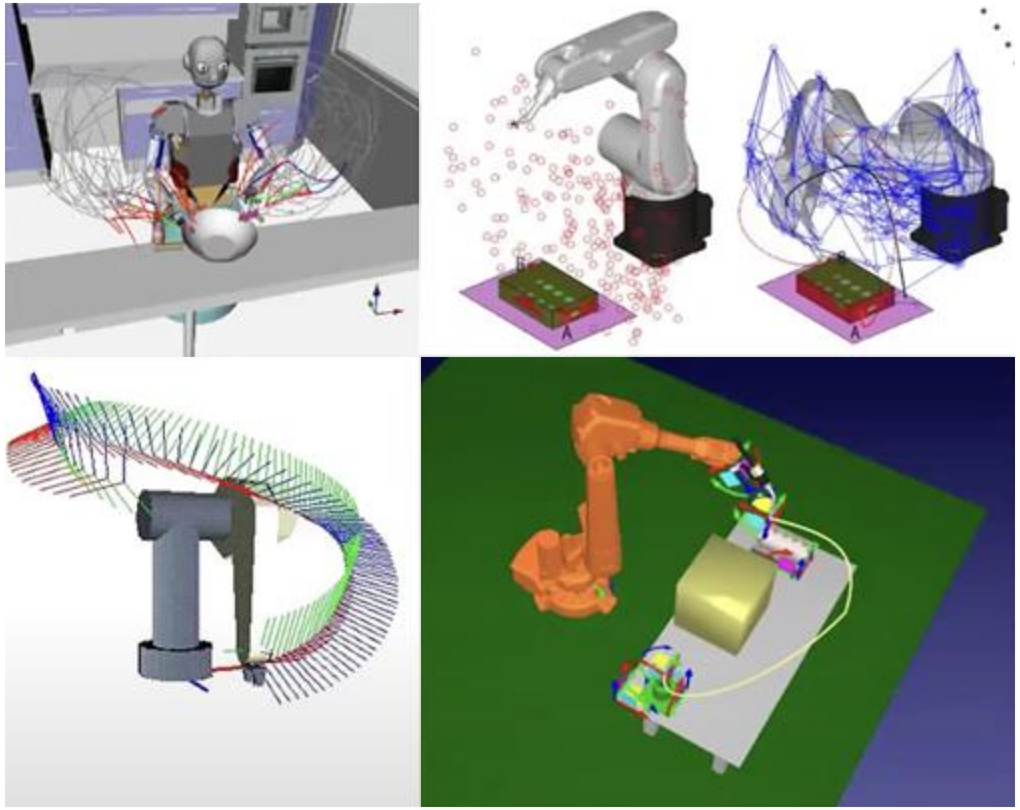
Many manipulation tasks are long horizon

Language models usually use intermediate reasoning to solve long horizon tasks

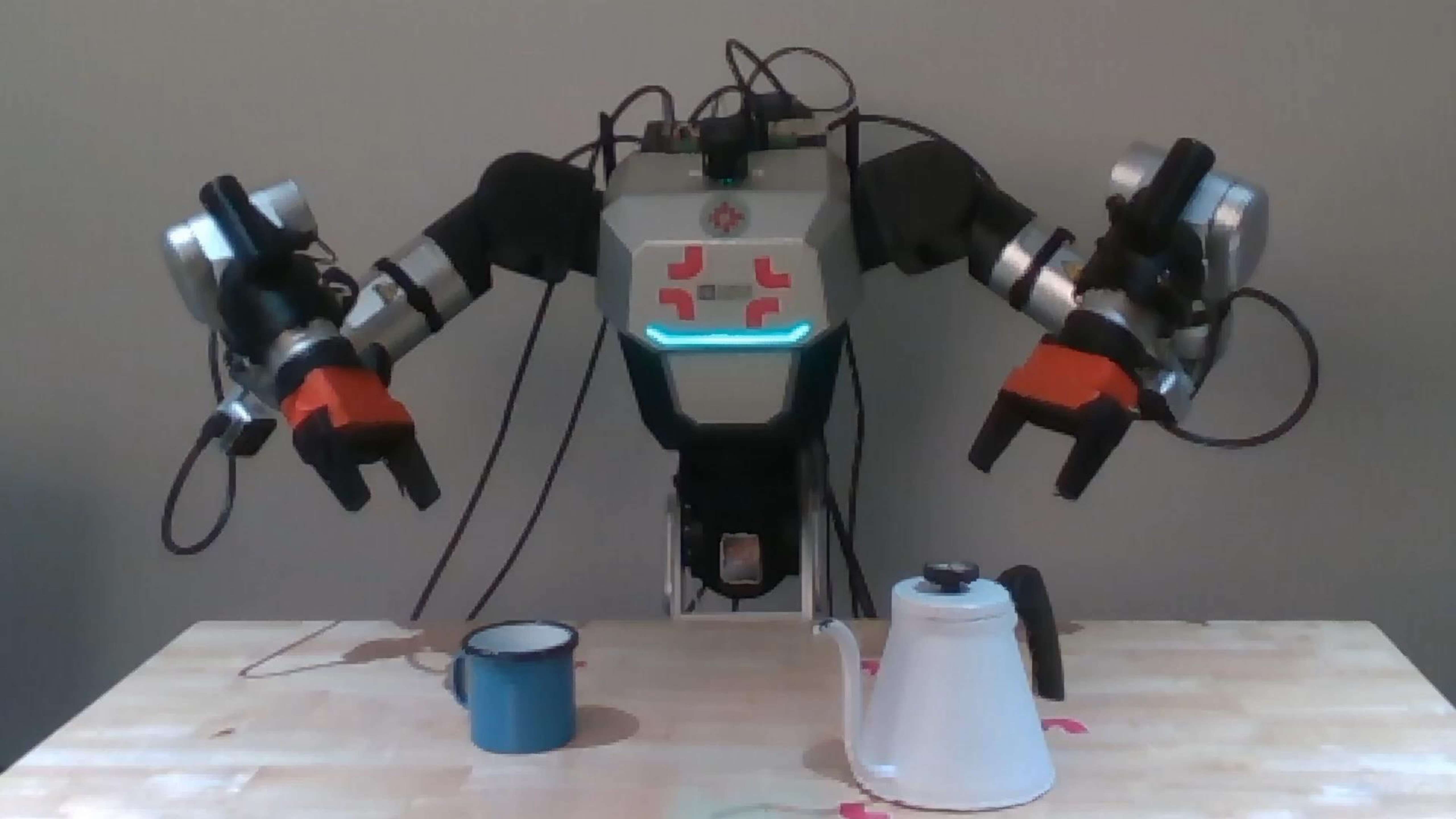


But language reasoning doesn’t allow models to reason in space





Need to reason about space



Molmo-Acts reasons in space – it sketches a plan in 2.5D

“Move pepsi can near Red bull”

MolmoAct Reasoning

Q: To figure out the action that the robot should take to **put plate into dishwasher**, let's think through it **step by step**.

First, what is the **depth map** for this image?

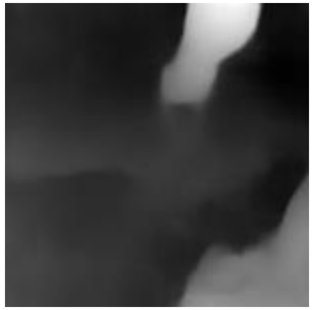
Second, what is the **trajectory** of the end effector?

Based on these, what is the action that the robot should take?

MolmoAct

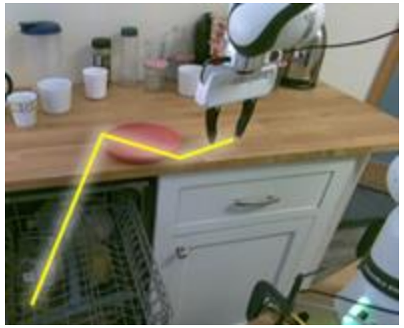
Depth Perception Token

[1 12 • • • 8 68]



Visual Reasoning Trace

[[202,15], [153,178], [130,145], [52,145]]

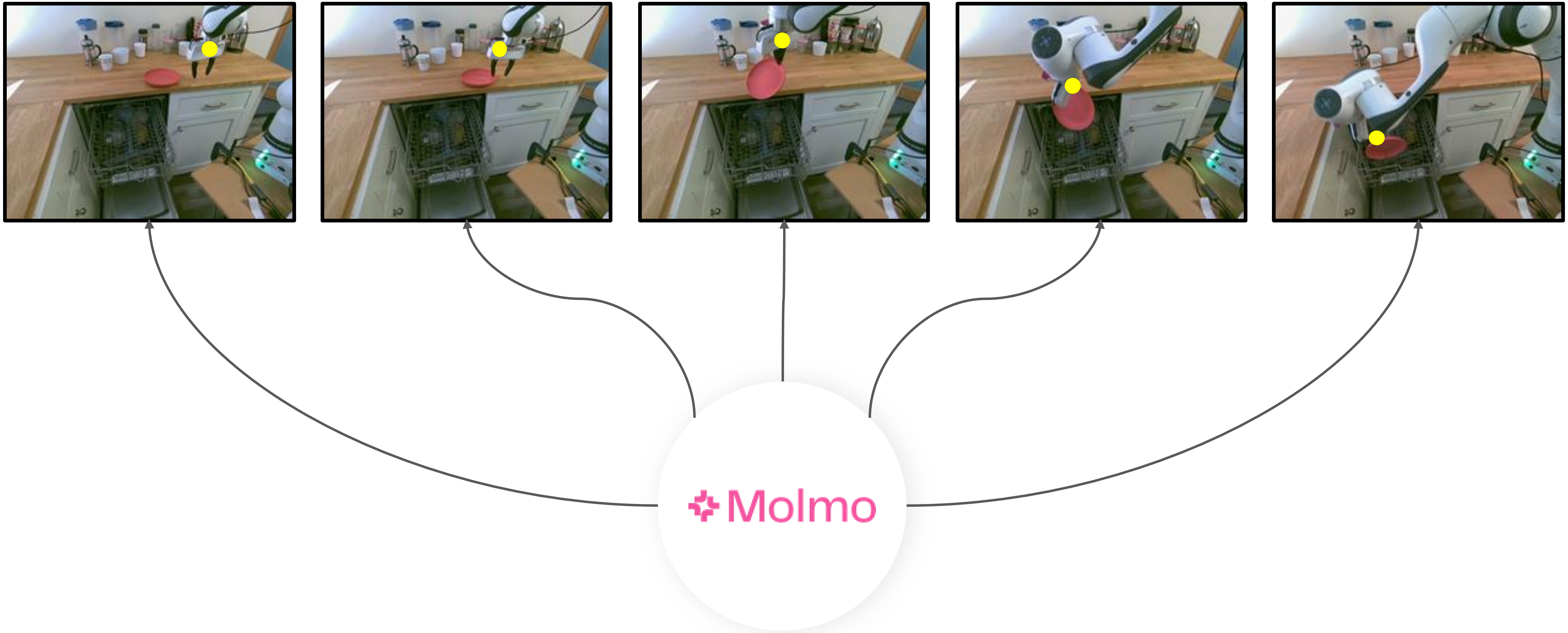


Robot Actions

[Δx, Δθ, Gripper]

Ai2

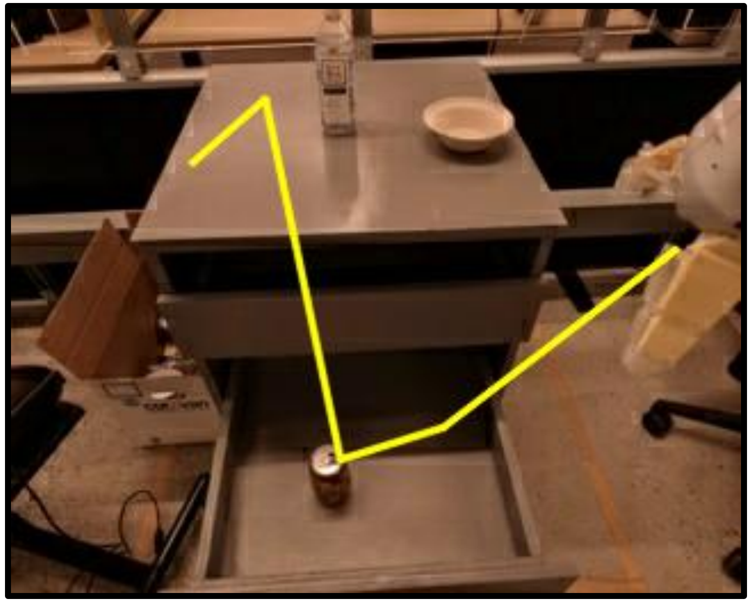
How do we curate ground truth trajectory

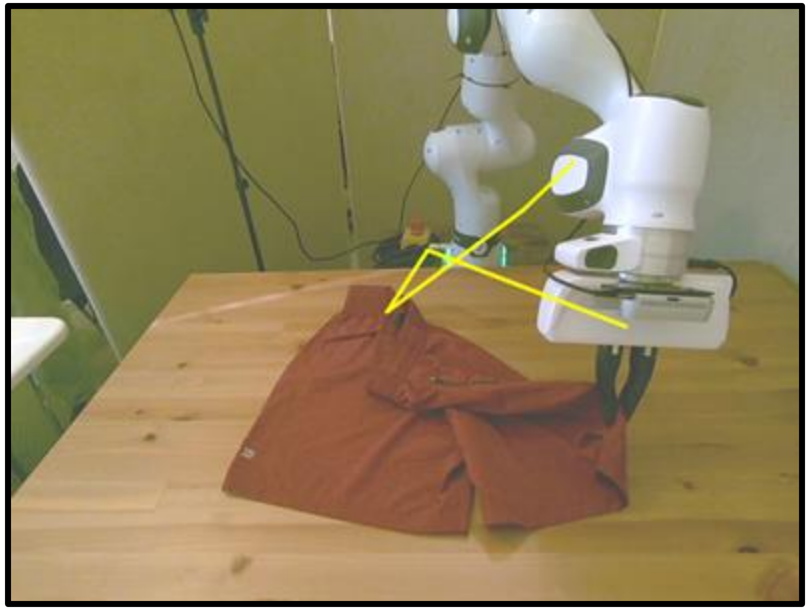


"Point to the gripper"

How do we curate ground truth trajectory

How do we curate ground truth trajectory





8M Image, Trajectory pairs

~170k Robot Trajectories

Reasoning in 2D is not enough!

So, we also reason using depth!

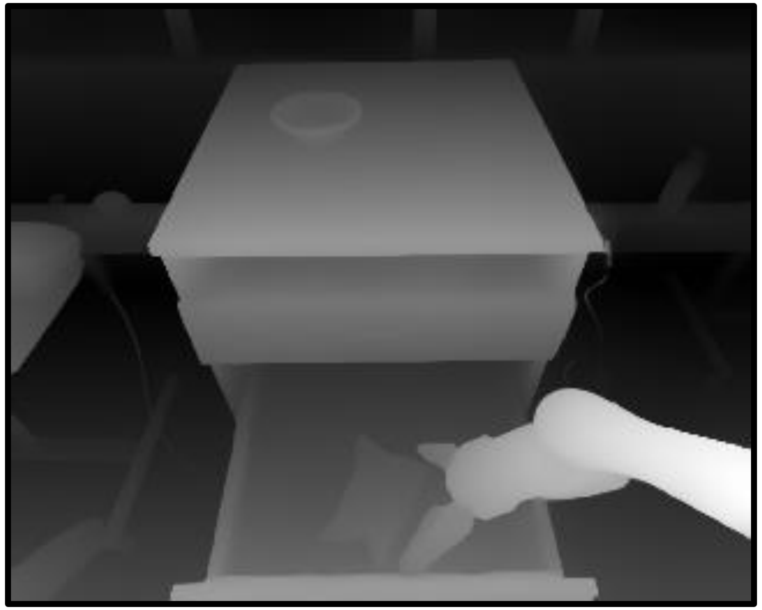
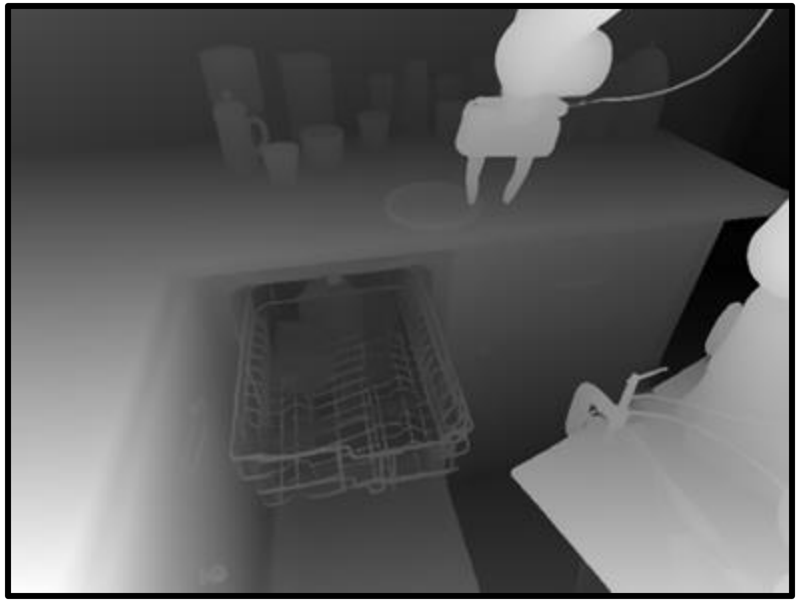
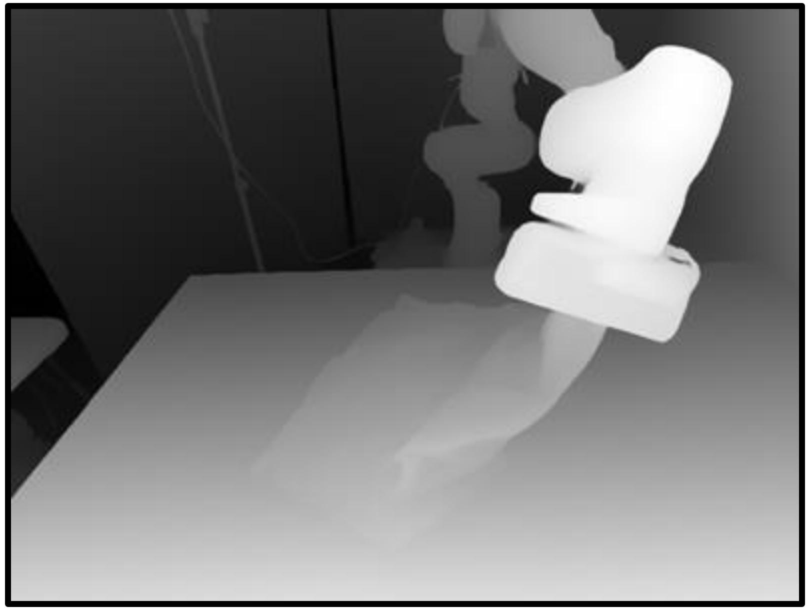
“Put the plate in dishwasher”

Molmo-ACT

The trajectory of the end effector is [234, 68], [245, 71]...

The action that the robot should take is [Δx, Δθ, Gripper].

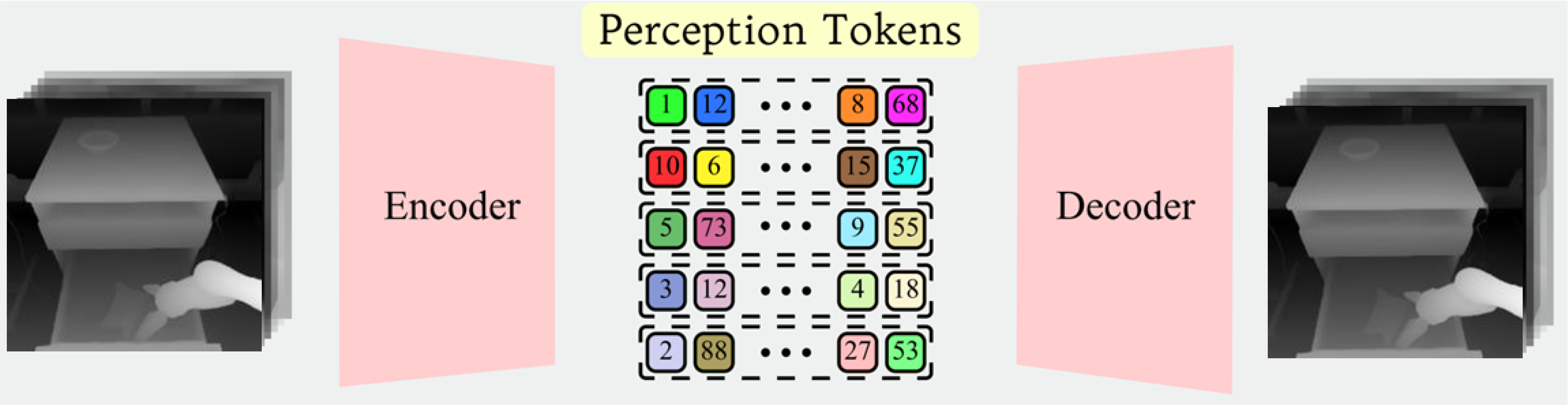
We automatically generate depth data

8M Image, Depth pairs

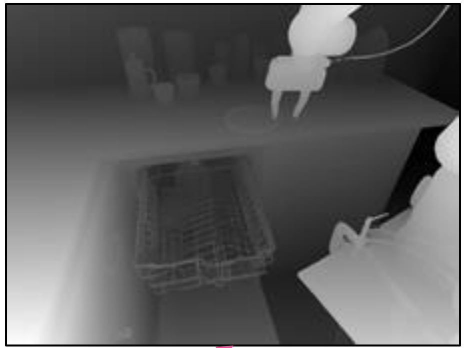
~170k Robot Trajectories

Reasoning in depth requires new innovations
-> Perception Tokens



Bigverdi et al. 2024 "Perception Tokens Enhance Visual Reasoning in Multimodal Language Models"

Grounding Action with Depth



"Put the plate in

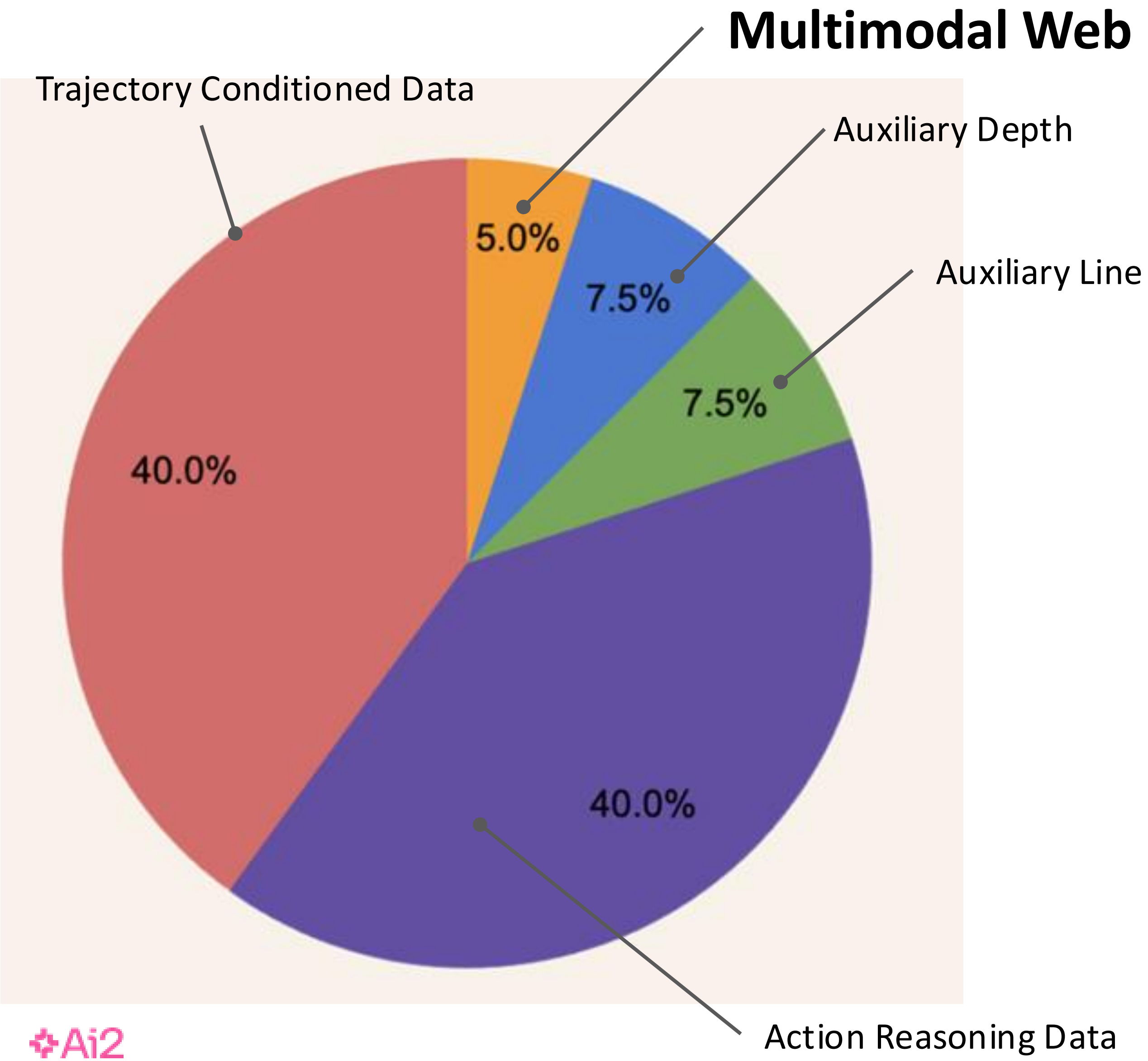
Molmo-ACT

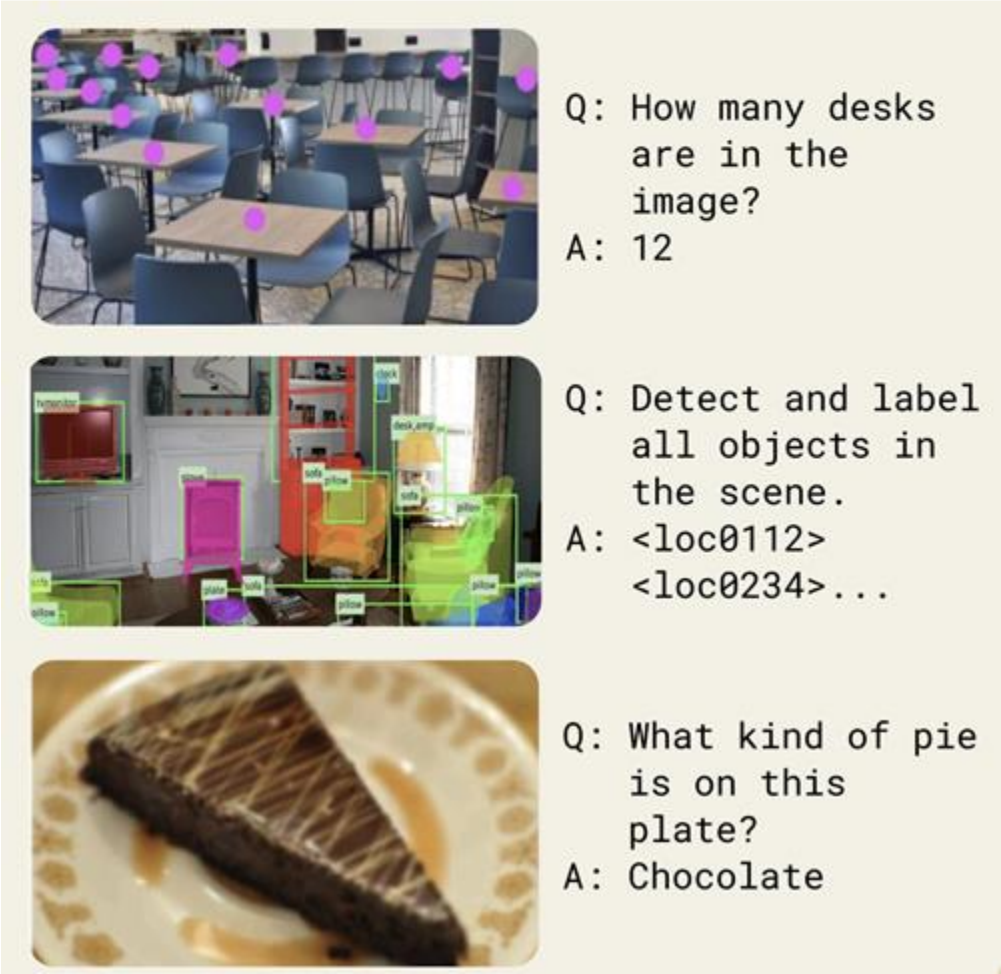
The depth map is

[1 12 • • • 8 68]

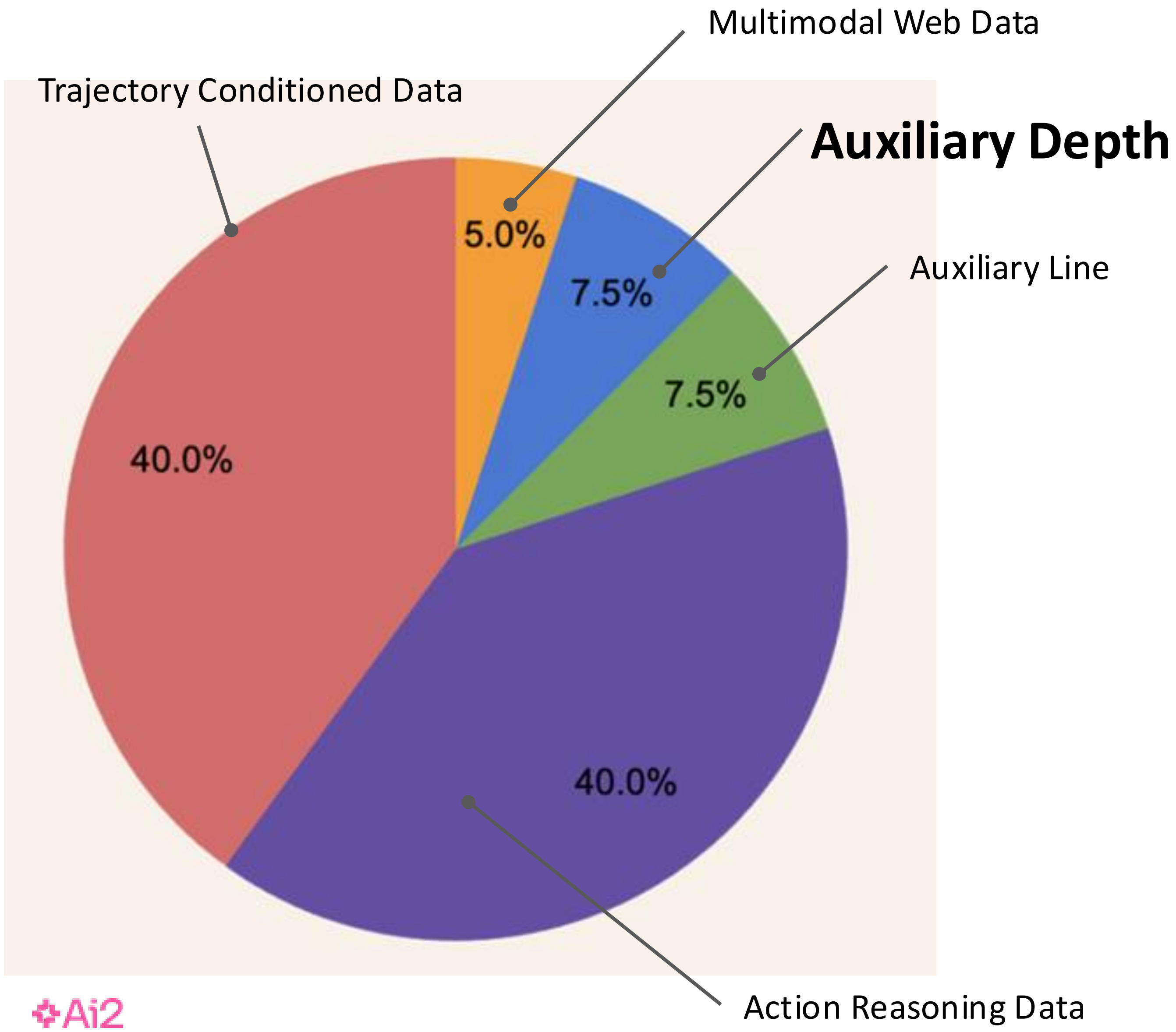
The action that the robot should take is [Δx, Δθ, Gripper].

Our pretraining datamix has general multimodal



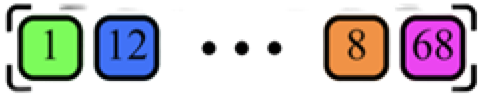


Our pretraining datamix has produce depth estimation



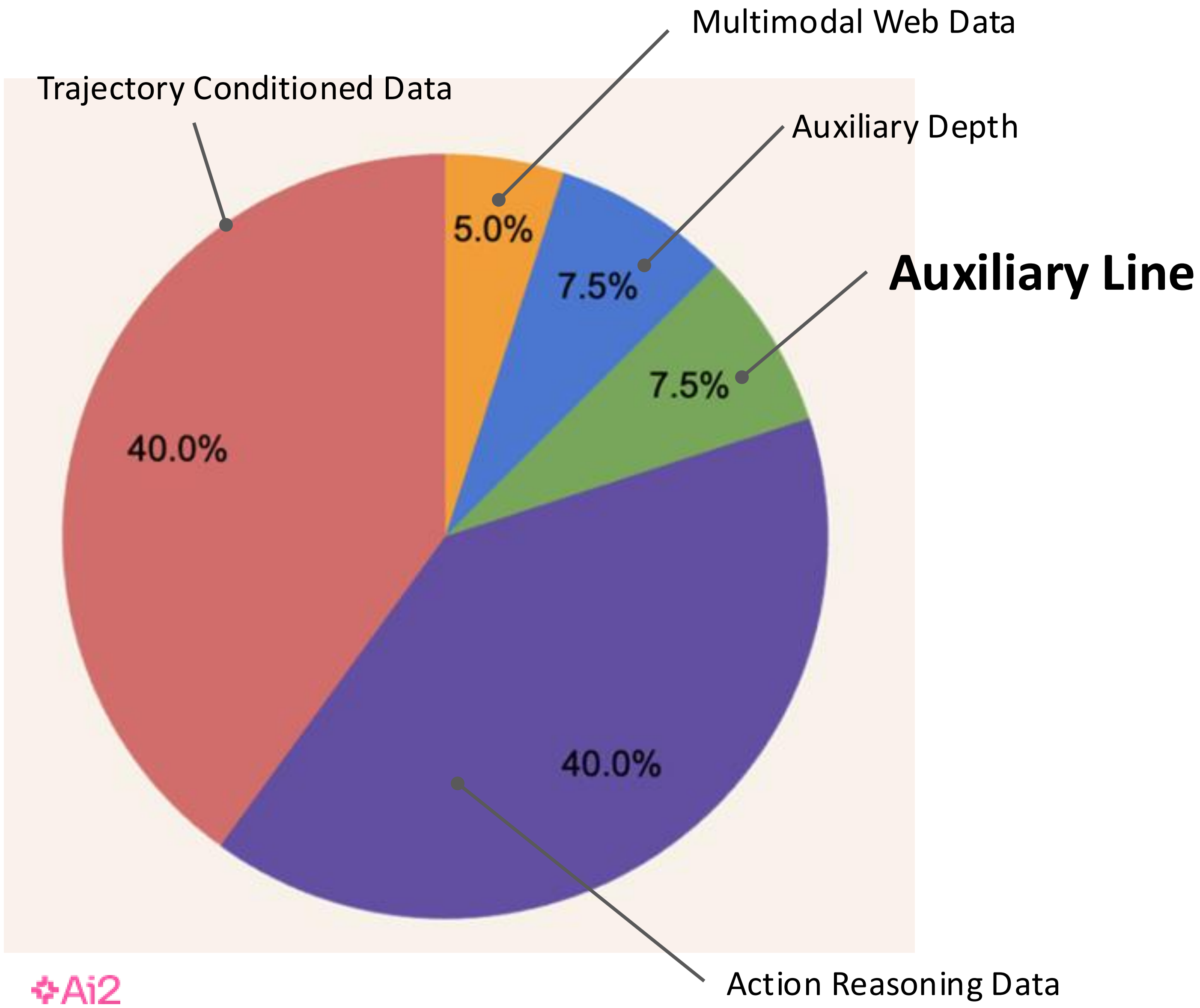
Q: The task is put the plate in the dishwasher, what is the depth map?

A: The depth map is



Ai2

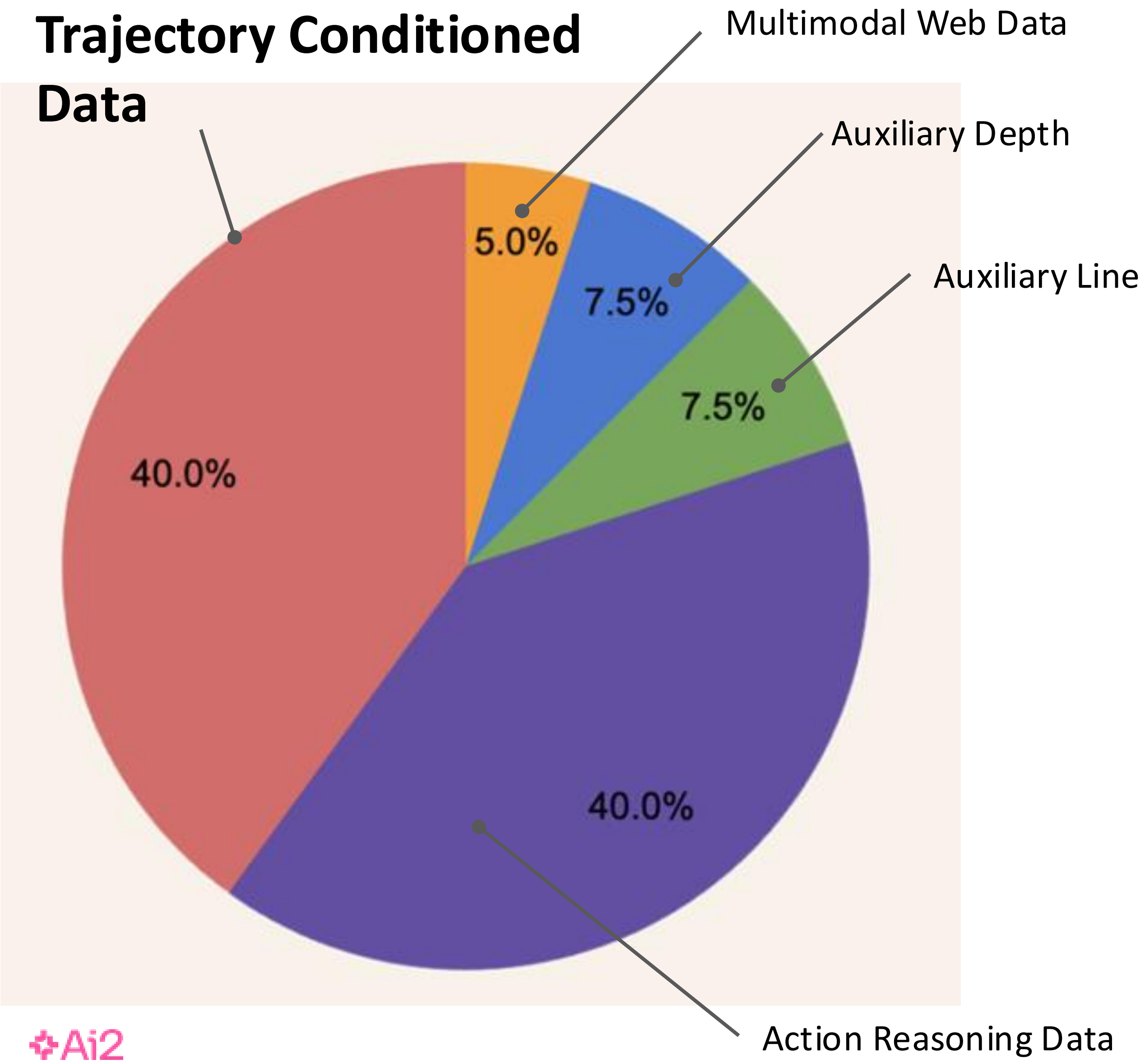
Our pretraining datamix has grounded reasoning



Q: The task is put the plate in the dishwasher, what is the trajectory of the end effector?

A: The trajectory that the robot should take is [194, 24],[203,44]...

Our pretraining datamix has trajectory conditioned

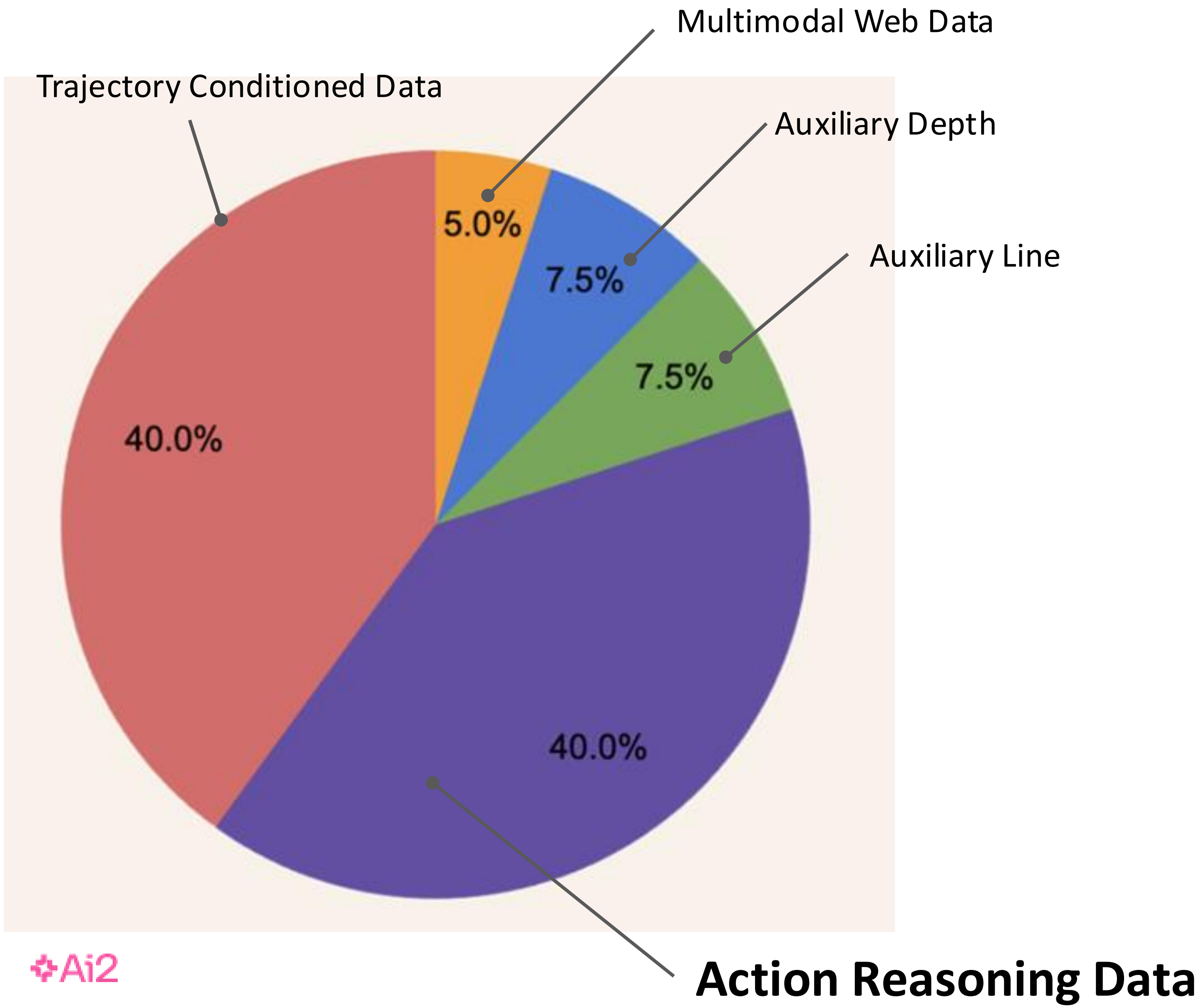


Q: The task is put the plate in the dishwasher, what action should the robot take based on the traj?

A:

The action that the robot should take is [Δx, Δθ, Gripper].

Our pretraining datamix has large robot behavior



Q: The task is put the plate in the dishwasher, what action should the robot take?

A: The depth map is [1 12 • • • 8 68]

The trajectory that the robot should take is [194, 24],[203,44]...

The action that the robot should take is [Δx, Δθ, Gripper].

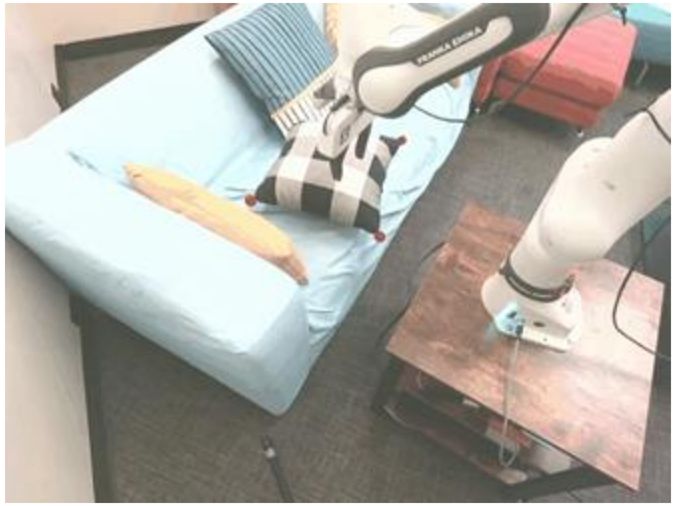
Ai2

MolmoAct dataset contains different household environments

Kitchen

Bedroom

Bathroom



Living Room

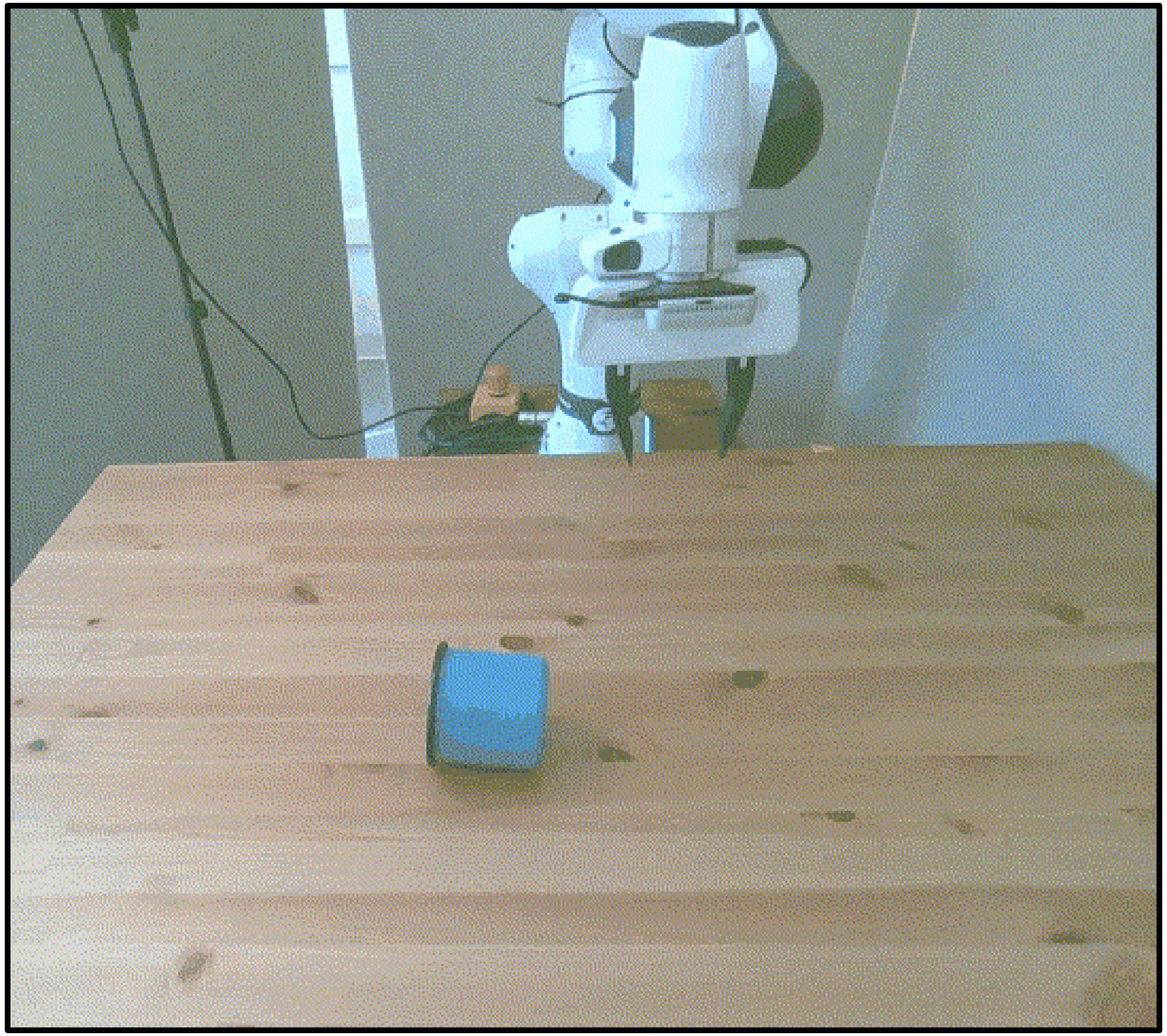
Breaking subtask into atomic motions

Load the dishwasher

- Open the dishwasher
- Pick up the dishes/utensils
- Flip the plates/bowls upright
- Place it in the dishwasher
- Close the tray and the dishwasher

Place the plate on the stand

Flip the mug



Close the drawer

MolmoAct Overview

Molmo Initiation

MolmoAct Training

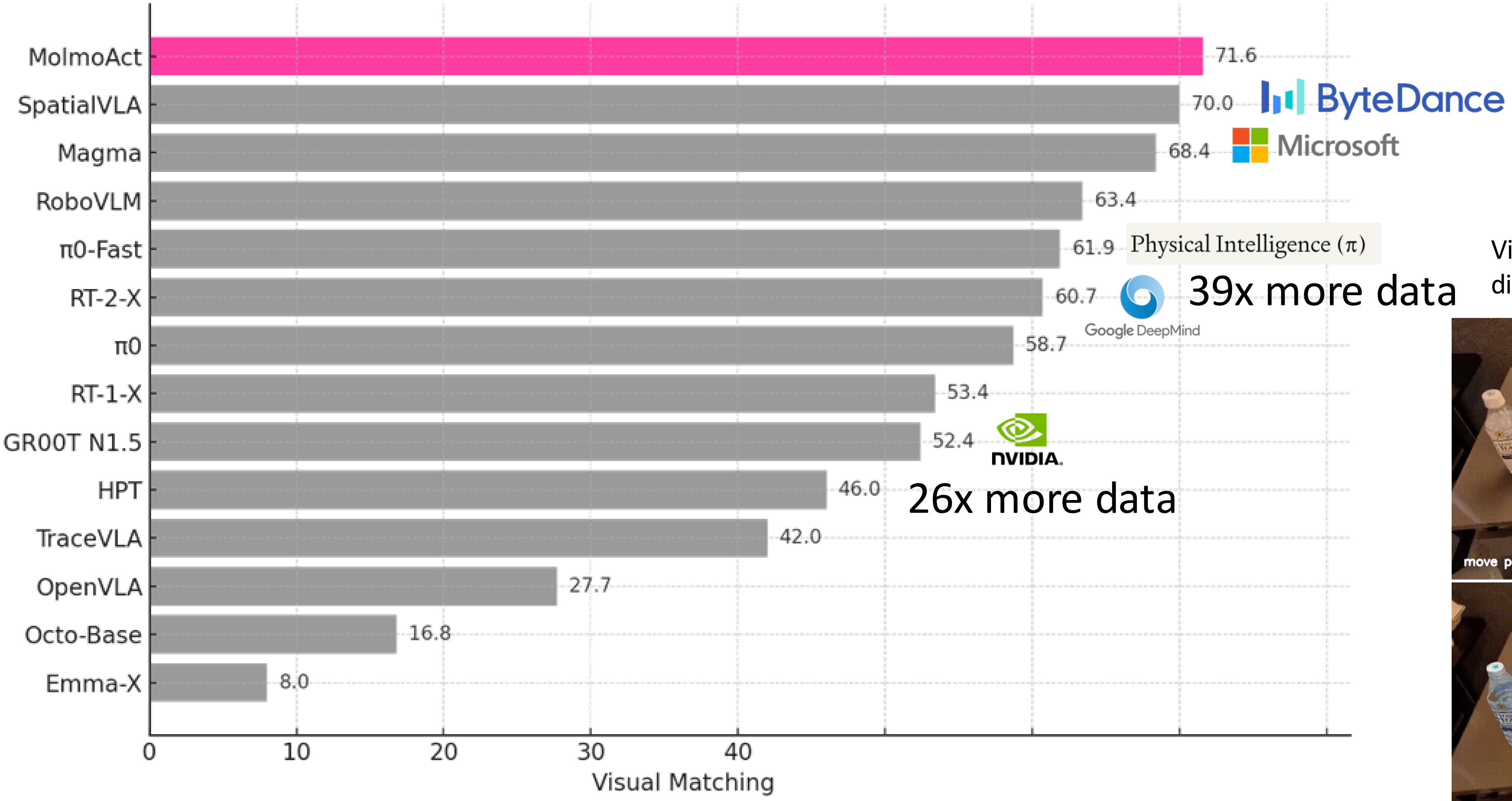
MolmoAct Inference

Pre-training Evaluations (SimplerEnv)

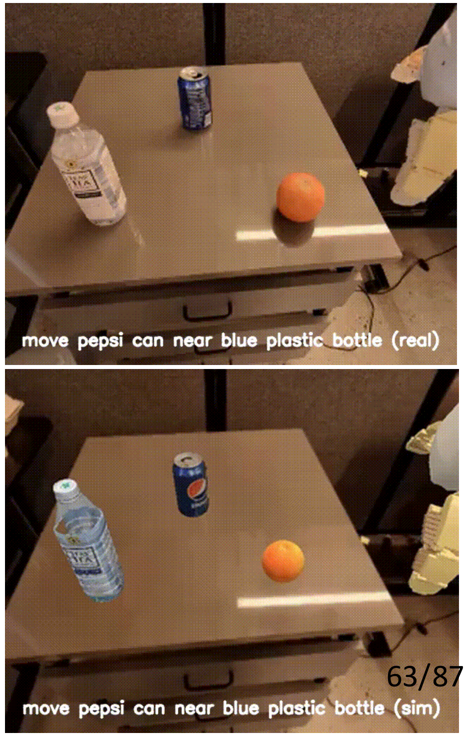
Visual Matching (In-distribution)

Variant Aggregation (Out-of-distribution)

Pre-training evaluation with MolmoAct



Visual Matching (In-distribution)



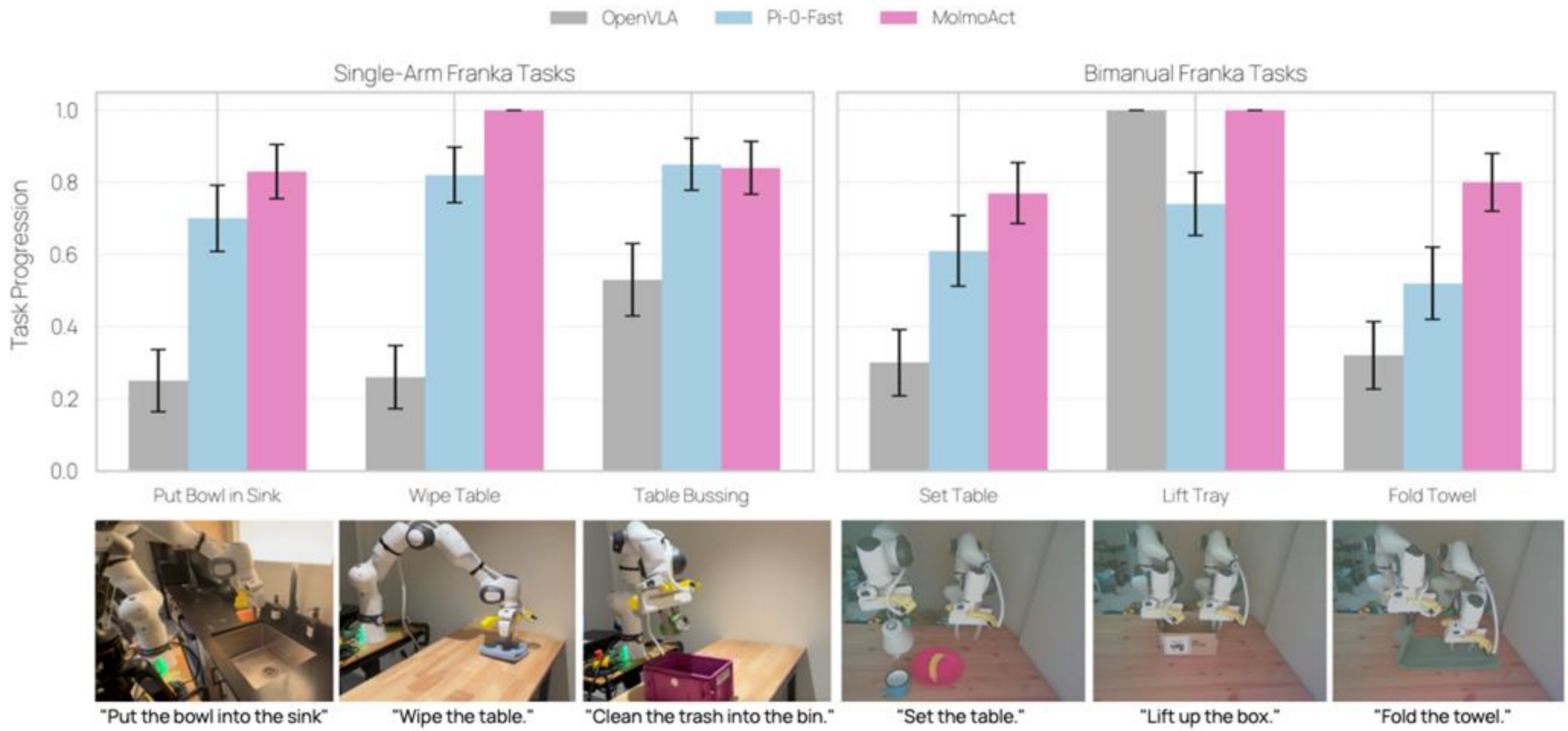
Post-training Evaluations (LIBERO)

SOTA in Post-training Evaluations (LIBERO)*

*compared to other autoregressive models

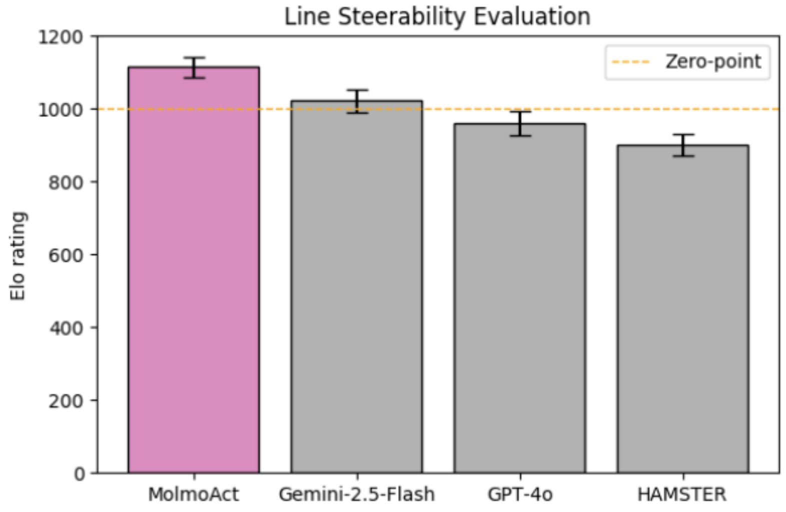
Baseline	Spatial	Object	Goal	Long	Avg
TraceVLA (Zheng et al., 2024)	84.6%	85.2%	75.1%	54.1%	74.8%
Octo-Base (Team et al., 2024b)	78.9%	85.7%	84.6%	51.1%	75.1%
OpenVLA (Kim et al., 2024)	84.7%	88.4%	79.2%	53.7%	76.5%
SpatialVLA (Qu et al., 2025)	88.2%	89.9%	78.6%	55.5%	78.1%
CoT-VLA (Zhao et al., 2025)	87.5%	91.6%	87.6%	69.0%	83.9%
NORA-AC (Hung et al., 2025)	85.6%	89.4%	80.0%	63.0%	79.5%
WorldVLA (Cen et al., 2025)	87.6%	96.2%	83.4%	60.0%	79.1%
π_0-FAST (Black et al.)	96.4%	96.8%	88.6%	60.2%	85.5%
ThinkAct (Huang et al., 2025)	88.3%	91.4%	87.1%	70.9%	84.4%
MolmoAct-7B-D	87.0%	95.4%	87.6%	77.2%	**86.6%**

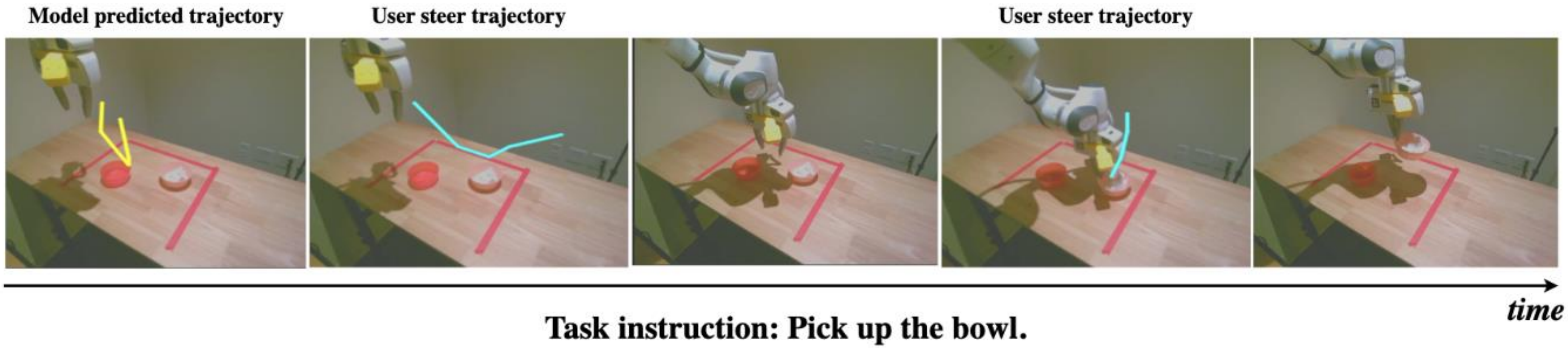
Post-training Evaluations (Real-world)



MolmoAct outperforms π0-FAST by an average of 10% in task progression on single-arm tasks and by 22.7% on bimanual tasks.

Steerability -> allow users to interpret and guide robot behavior





Prioritizing perception

Perceptual tests for VLMs
[ECCV 2024]

Sketching for perceptual reasoning
[NeurIPS 2024] [CVPR 2025]

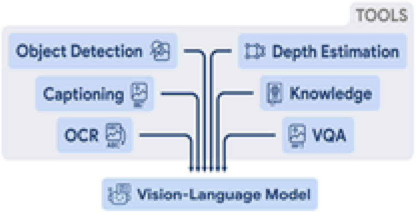
Distilling perceptual capabilities
[ACL 2023] [CVPR 2024] [CVPR 2025]

Enabling robots to sketch
[ArXiv 2025]

BLINK
Visual tasks beyond language descriptions

Most fundamental vision capabilities are still out of reach

Enabling sketching:
visual chain of thought



How can we distill from specialist models into generalist VLMs?

MolmoAct

Complete open Action Reasoning model for robotics

Thank you

Ranjay Krishna

BLINK: Multimodal Large Language Models Can See but Not Perceive
Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, Ranjay Krishna
ECCV 2024

Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models
Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, Ranjay Krishna
NeurIPS 2024

Visual Program Distillation: Distilling Tools and Programmatic Reasoning into Vision-Language Models
Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, Ariel Fuxman
CVPR 2025

Perception Tokens Enhance Visual Reasoning in Multimodal Language Models
Mahtab Bigverdi, Zelun Luo, Cheng-Yu Hsieh, Ethan Shen, Dongping Chen, Linda G. Shapiro, Ranjay Krishna
CVPR 2025

MolmoAct: Action Reasoning Models that can Reason in Space
Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, Ranjay Krishna
ArXiv 2025