Large Language Models

Large Multi-modal Models (LMMs)

M. Soleymani
Sharif University of Technology
Fall 2023

Outline

- Instruction-following LMMs
- Multi-modal agents
- LLMs as planner in embodied agents

Tasks without Instruction

Input	Output
Old habits die hard.	ترک عادت موجب مرض است.
Welcome to the Large Language Models course at Sharif University of Technology! This course delves into the fascinating world of LLMs, a pivotal area of artificial intelligence that has transformed the field of natural language processing (NLP). Over the semester, we will explore the foundations and practical applications of these cutting-edge models.	*Sharif University's LLM course: explore foundations and applications of LLMs*

Task instructions are implicit in the data.

Hard to generalize to new tasks in zero-shot manner.

Instruction

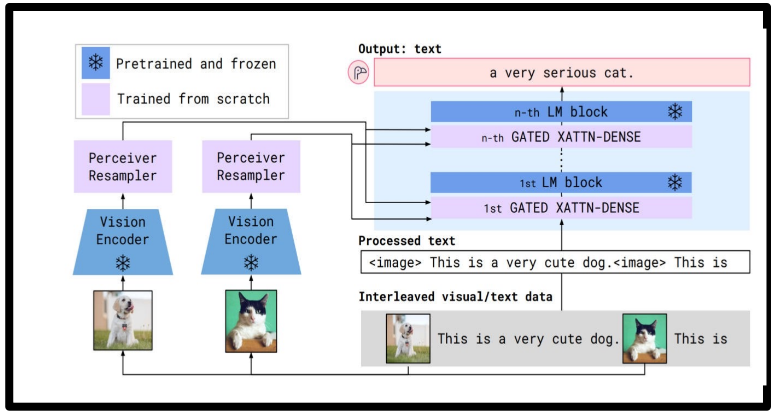
Instruction	Input	Output
Translate English into Farsi	*Old habits die hard.*	ترک عادت موجب مرض است.
Summarize in just 10 words to make the message even more brief.	*Welcome to the Large Language Models course at Sharif University of Technology! This course delves into the fascinating world of LLMs, a pivotal area of artificial intelligence that has transformed the field of natural language processing (NLP). Over the semester, we will explore the foundations and practical applications of these cutting-edge models.*	*Sharif University's LLM course: explore foundations and applications of LLMs*

Instruction Tuning: Transfer to new tasks

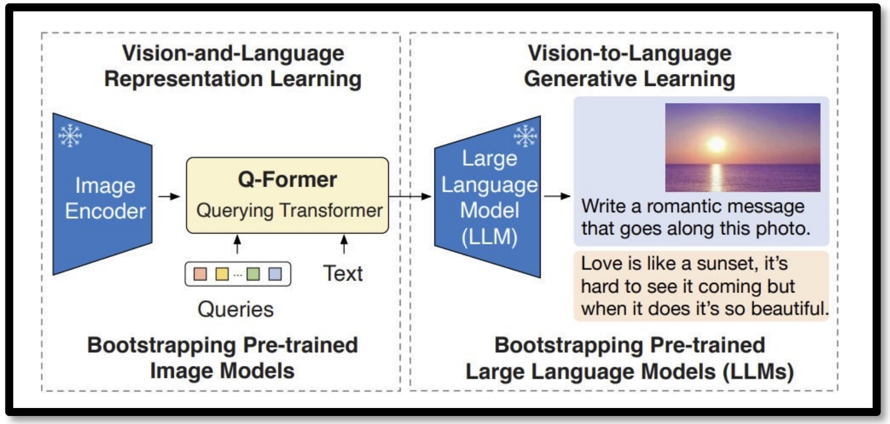
Instruction	Input	Output
Summarize in Farsi to make the message more brief.	*Welcome to the Large Language Models course at Sharif University of Technology! This course delves into the fascinating world of LLMs, a pivotal area of artificial intelligence that has transformed the field of natural language processing (NLP). Over the semester, we will explore the foundations and practical applications of these cutting-edge models.*	درس LLMs دانشگاه شریف: کاوش پایه‌ها و کاربردهای LLMs

Example of LMMs

Flamingo



BLIP2



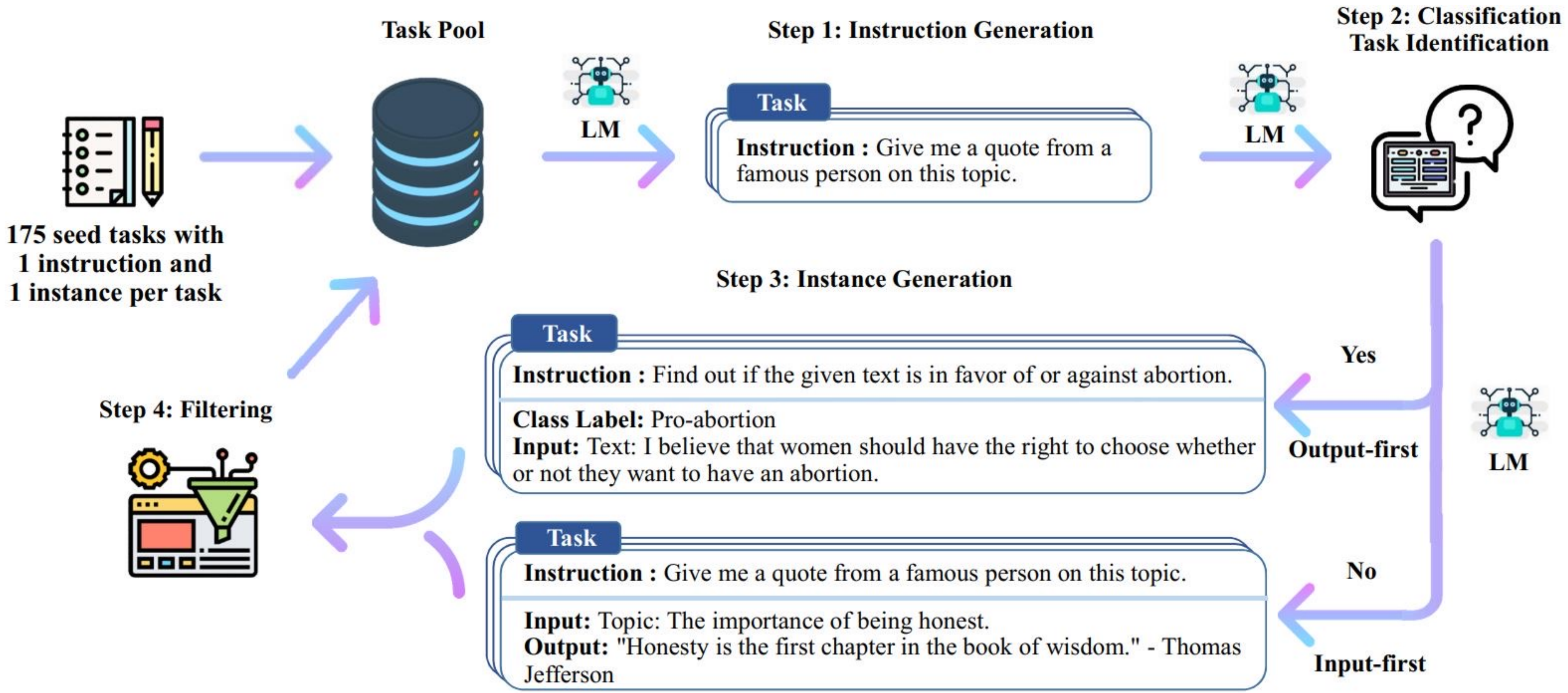
Language Model	Pre-trained: 70B Chinchilla	Pre-trained: FLAN-T5/OPT
Connection Module	Perceiver Resampler Gated Cross-attention + Dense	Q-Former: Lightweight Querying Transformer
Vision Encoder	Pre-trained: Nonrmalizer-Free ResNet (NFNet)	Contrastive pre-trained: EVA/CLIP

Instruction Tuning

- Instruction-following data is expensive to collect

- What about a human-machine collecting?
  - More scalable

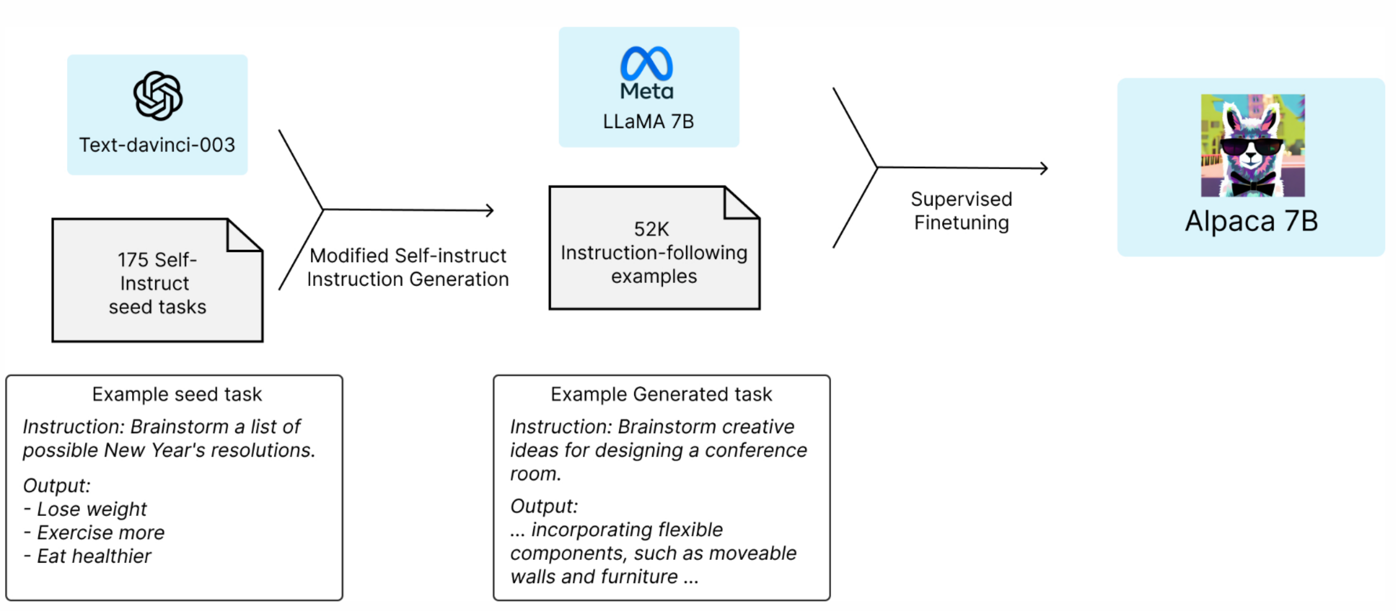
Self-Instruct for Instruction Tuning of LLMs



Wang et al., SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions, ACL 2023

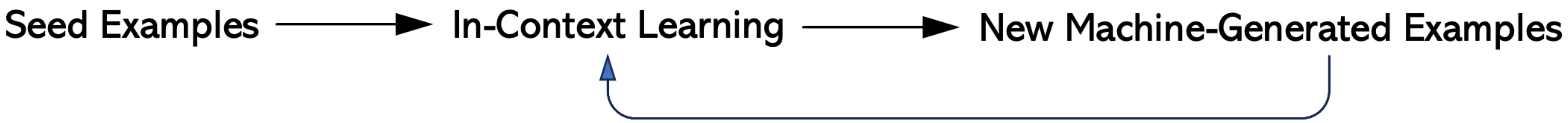
Self-Instruct for Instruction Tuning of LLMs

- Facilitates the preparation of instruction-following examples by employing a semi-automated generation of them
- Example: Alpaca produces 52K high-quality instruction-following demonstrations (using text-davinci-003) to fine-tune LLaMA

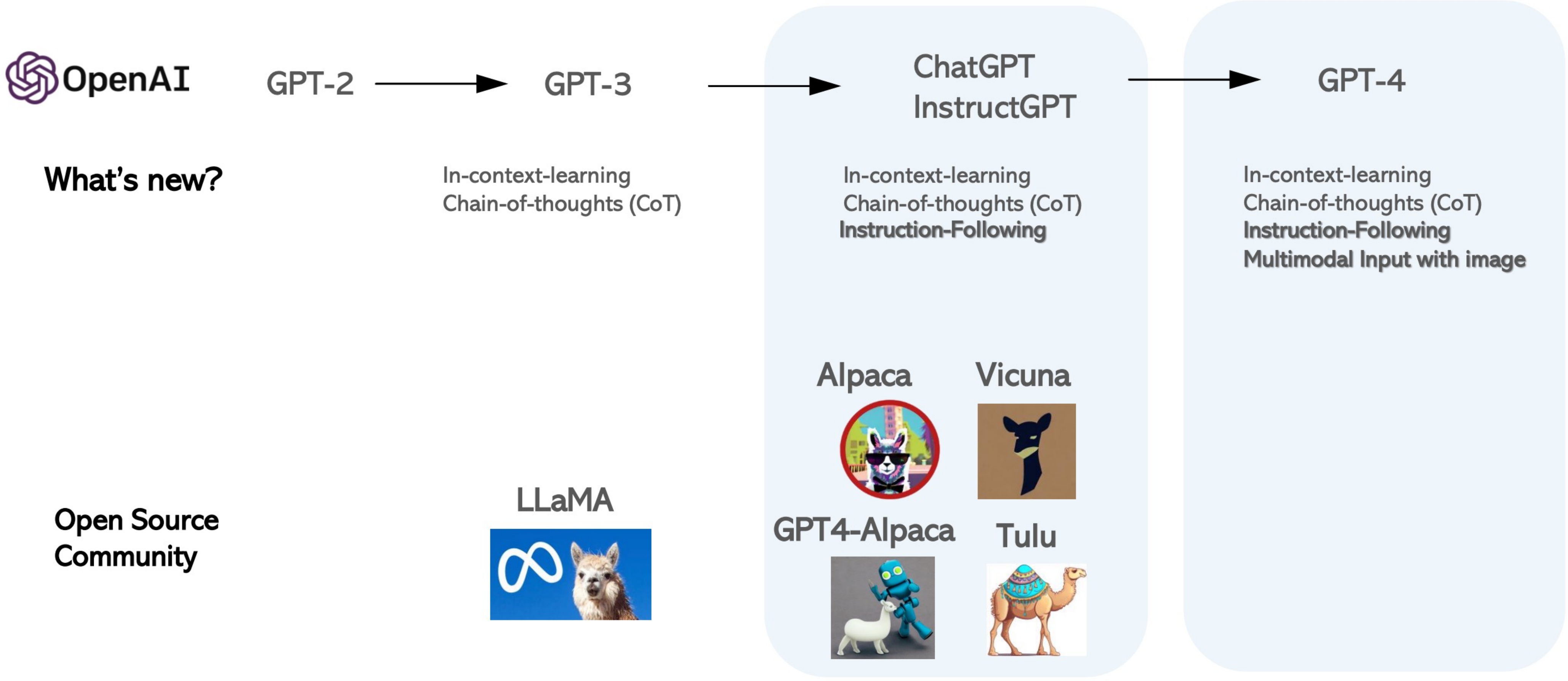


Taori et al., "Alpaca: A Strong, Replicable Instruction-Following Model", 2023

Self-instruct with Strong LLMs



Self-instruct in Open-source LLMs



Source: Chunyuan Li's Slides. CVPR 2023 Tutorial

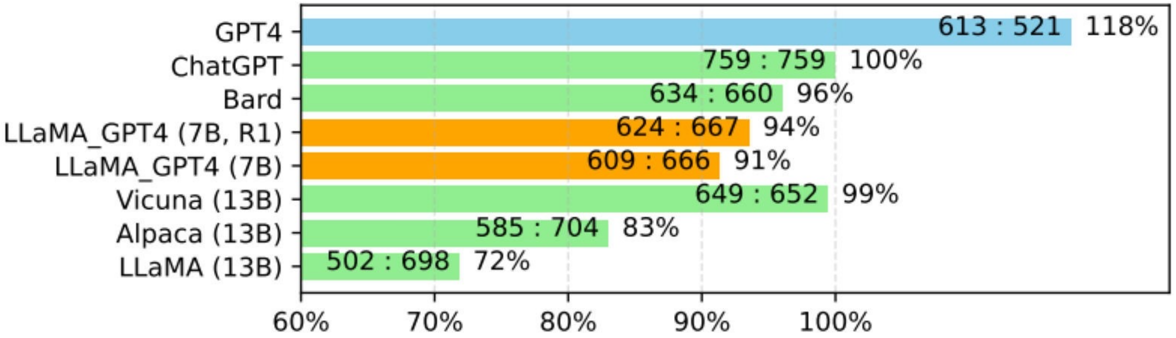
Comparison of Chatbots by GPT-4

Evaluation Metric: Ask GPT-4 to rate the two model responses (1-10), then compute the ratio, i.e. relative score

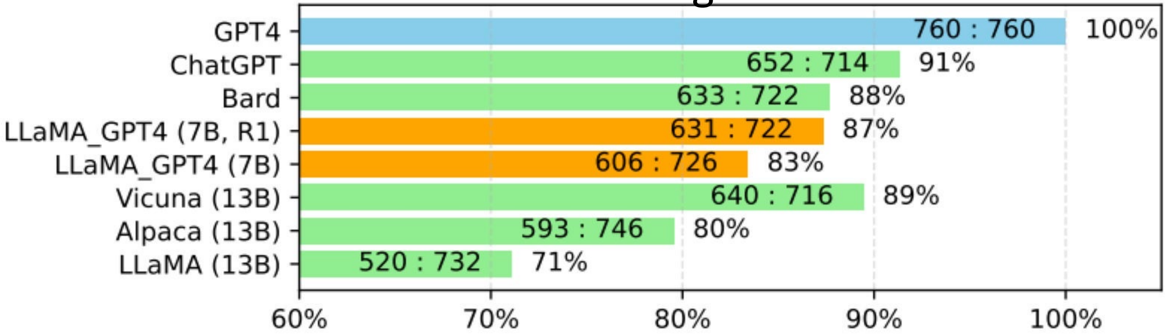
Findings:
A Very Consistent Evaluation Metric !

Opensourced Chatbots mimicked commercial ones

All chatbots against ChatGPT



All chatbots against GPT-4



Source: Chunyuan Li's Slides, CVPR 2023 Tutorial

MultiModal GPT-4

- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild
- How can we build Multimodal GPT-4 like models?

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.

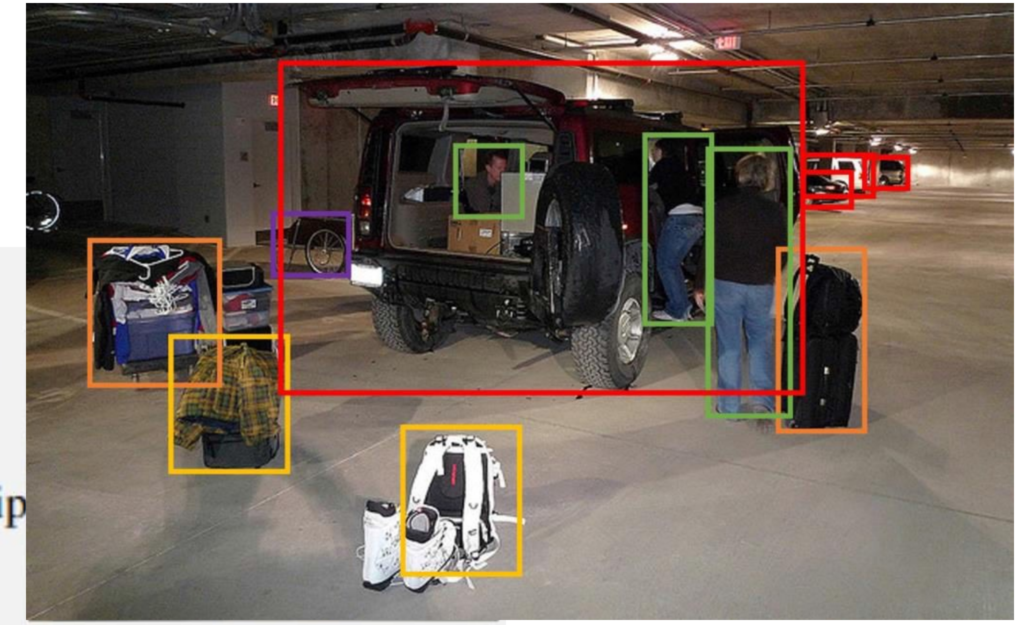
GPT-4 visual input example, Chicken Nugget Map:

User: Can you explain this meme?

GPT-4: 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.

GPT-assisted Visual Instruction Data Generation

- Rich Symbolic Representations of Images
- In-context-learning with a few manual examples
- → Text-only GPT-4

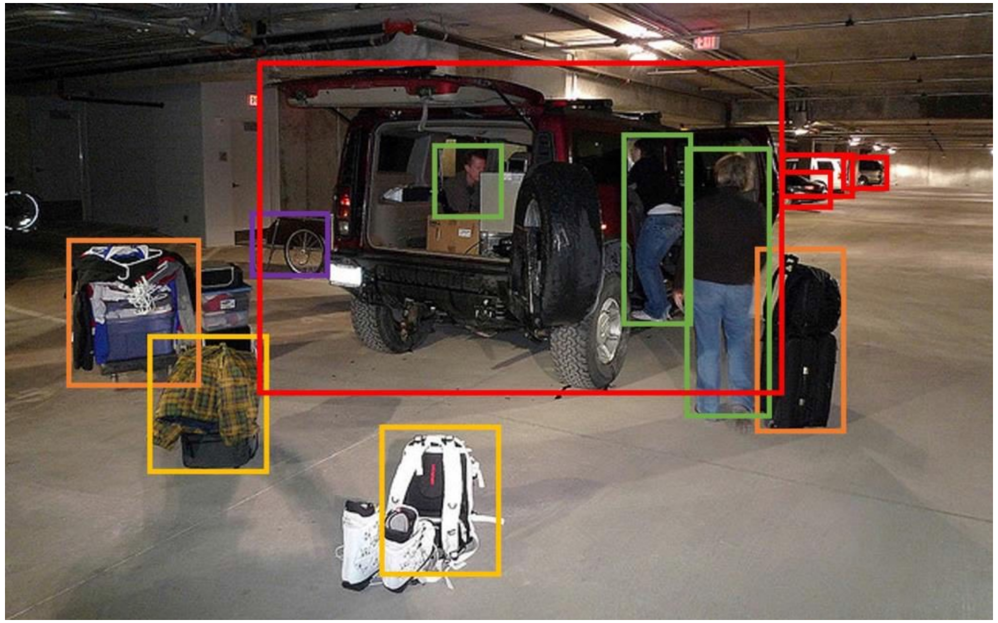


Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.
Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], person: [0.63, 0.222, 0.686, 0.516], person: [0.444, 0.233, 0.487, 0.34], backpack: [0.384, 0.696, 0.485, 0.914], backpack: [0.755, 0.413, 0.846, 0.692], suitcase: [0.758, 0.413, 0.845, 0.69], suitcase: [0.1, 0.497, 0.173, 0.579], bicycle: [0.282, 0.363, 0.327, 0.442], car: [0.786, 0.25, 0.848, 0.322], car: [0.783, 0.27, 0.827, 0.335], car: [0.86, 0.254, 0.891, 0.3], car: [0.261, 0.101, 0.787, 0.626]

GPT-assisted Visual Instruction Data Generation

Three type of instruction-following responses

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV).
Question: Where is the vehicle parked?
Answer: The vehicle is parked in an underground parking area, likely in a public garage.
Question: What are the people in the image doing?
Answer: The people are trying to fit all of their luggage into the SUV, likely preparing for a trip.
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip.
In the parking area, various luggage items are scattered around the vehicle. There are two backpacks, one located near the left rear wheel and the other closer to the right side of the vehicle. Additionally, there are two suitcases, one on the right side of the car and another further away near the center of the parking area. A bicycle can also be seen on the left side of the vehicle.
Other cars are parked around the main SUV, with one car positioned behind it and slightly to the left, another behind and slightly to the right, and the third car further behind on the right side.
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings to accommodate. They might have to strategize and arrange the luggage efficiently to ensure that everything fits properly into the vehicle. Additionally, they need to consider the comfort of the passengers and visibility while driving, so the placement of the luggage must not obstruct the driver's view or make the passengers uncomfortable during the trip.



Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Example: Instructions for Detailed Image Description

- "Describe the following image in detail"
- "Provide a detailed description of the given image"
- "Give an elaborate explanation of the image you see"
- "Share a comprehensive rundown of the presented image"
- "Offer a thorough analysis of the image"
- "Explain the various aspects of the image before you"
- "Clarify the contents of the displayed image with great detail"
- "Characterize the image using a well-detailed description"
- "Break down the elements of the image in a detailed manner"
- "Walk through the important details of the image"
- "Portray the image with a rich, descriptive narrative"
- "Narrate the contents of the image with precision"
- "Analyze the image in a comprehensive and detailed manner"
- "Illustrate the image through a descriptive explanation"
- "Examine the image closely and share its details"
- "Write an exhaustive depiction of the given image"

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

LLaVA: Large Language-and-Vision Assistant

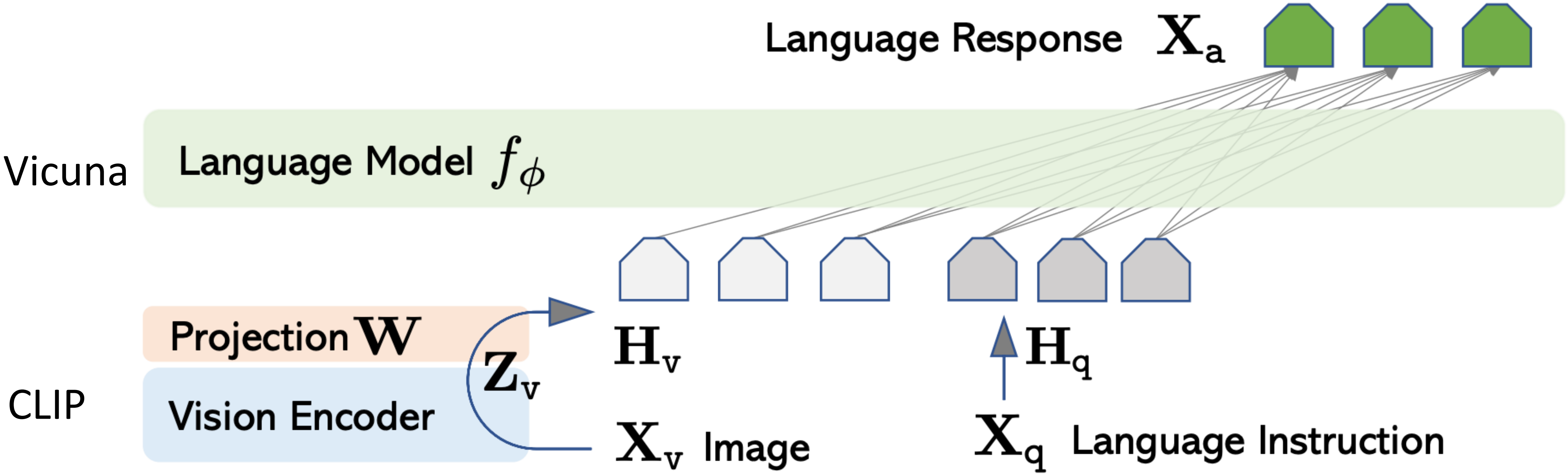
- The available multimodal instruction-following data is limited
  - time-consuming collection
  - less well-defined when human crowd-scouring is considered.

- LLaVA is the first visual instruction-following method

- LLaVA fine-tunes the whole LLM on 150K high-quality multi-modal instruction data generated by GPT-4.
  - expanding the original captions of COCO with handwritten seed instructions using GPT-4 to provide descriptions and multi-round conversations.

Liu et al., “Visual Instruction Tuning”, NeurIPS 2023

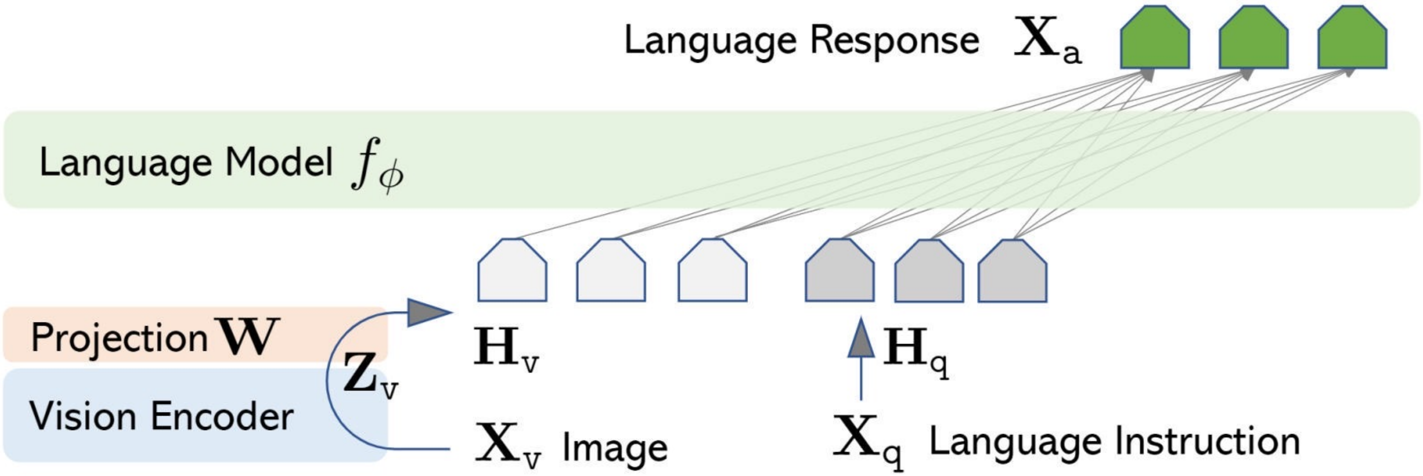
LLaVA: Architecture



The vision encoder converts input images into features
Linear projection layer convert these features into a space compatible with the LLM.

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

LLaVA: Training



Two-stage Training

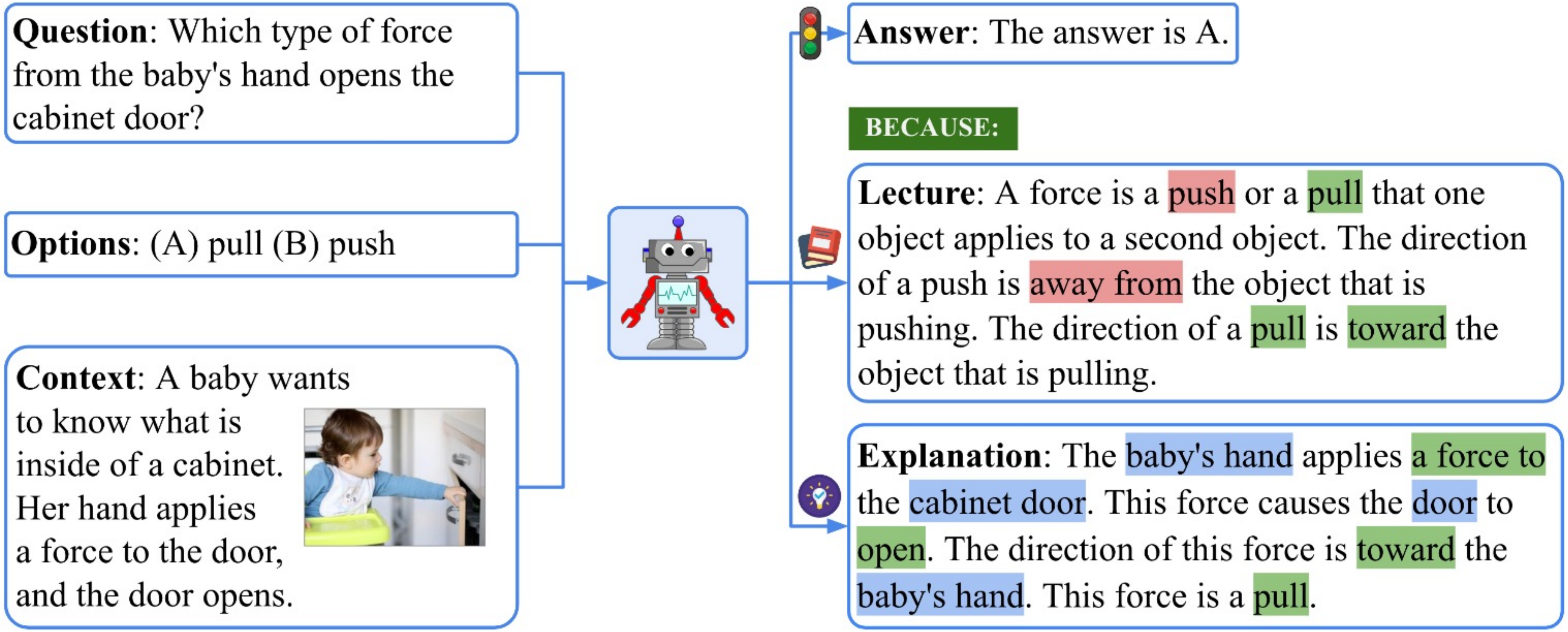
- **Stage 1: Pre-training for Vision-language Alignment**.

  Only the projection matrix is updated, based on a subset of CC3M.

- **Stage 2: Fine-tuning End-to-End.** Both the projection matrix and LLM are finetuned on curated dataset
  - **Visual Chat**: Generated multimodal instruction data for daily user-oriented applications.
  - **Science QA**: Multimodal reasoning dataset for the science domain.

Liu et al., “Visual Instruction Tuning”, NeurIPS 2023

ScienceQA



Lu et al., "Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering", NeurIPS 2022

Strong Visual Reasoning Ability

Example 1: Extreme Ironing

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 [34]	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	What is unusual about this image?
BLIP-2	a man is sitting on the back of a yellow cab
User	What is unusual about this image?
OpenFlamingo	The man is drying his clothes on the hood of his car.

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Strong Visual Reasoning Ability

Example 2: Chicken Nugget Map

Visual input example, Chicken Nugget Map:

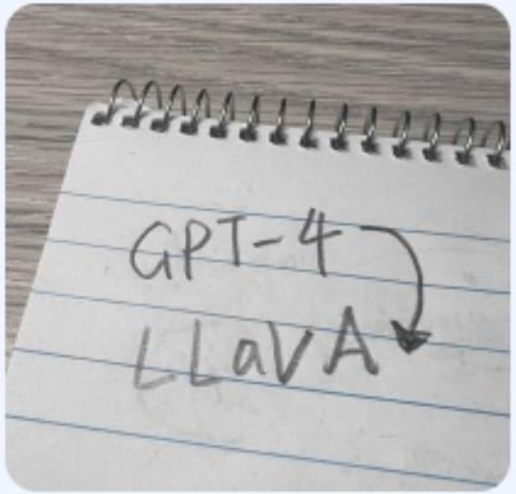
User	Can you explain this meme in detail?
LLaVA	The meme in the image is a creative and humorous take on food, with a focus on chicken nuggets, as the center of the universe. The meme begins with the phrase "Sometimes I just look at pictures of the Earth from space and I marvel at how beautiful it all 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.

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Strong OCR Ability

Example 3: LLaVA

what is written in the photo?



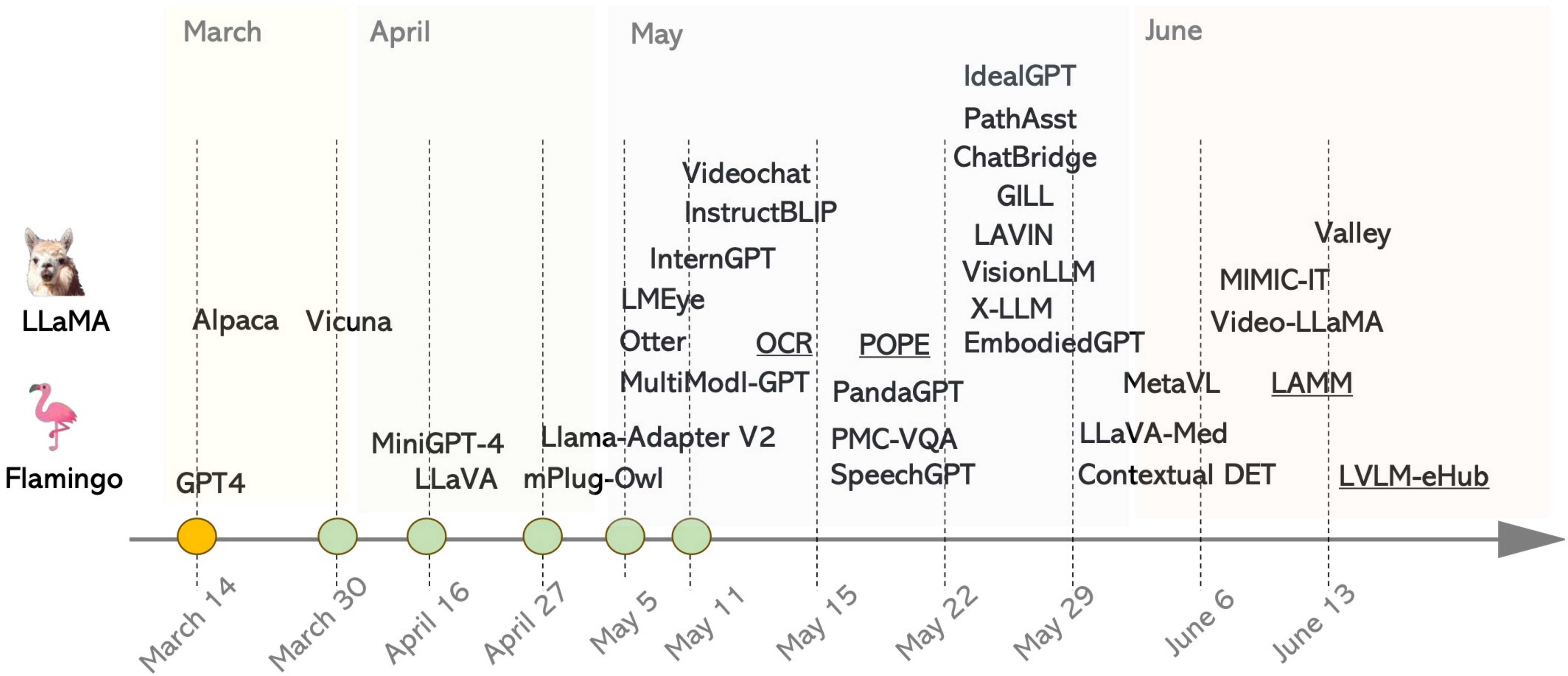
In the photo, there is written "Llava" which is presumably an abbreviation for a location or other identifier.

Method	Subject NAT	SOC	LAN	Context Modality TXT	IMG	NO	Grade G1-6	G7-12	Average
Representative & SoTA methods with numbers reported in the literature									
Human [30]	90.23	84.97	87.48	89.60	87.50	88.10	91.59	82.42	88.40
GPT-3.5 [30]	74.64	69.74	76.00	74.44	67.28	77.42	76.80	68.89	73.97
GPT-3.5 w/ CoT [30]	75.44	70.87	78.09	74.68	67.43	79.93	78.23	69.68	75.17
LLaMA-Adapter [55]	84.37	88.30	84.36	83.72	80.32	86.90	85.83	84.05	85.19
MM-CoT$_{Base}$ [57]	87.52	77.17	85.82	87.88	82.90	86.83	84.65	85.37	84.91
MM-CoT$_{Large}$ [57]	95.91	82.00	90.82	95.26	88.80	92.89	92.44	90.31	91.68
Results with our own experiment runs									
GPT-4	84.06	73.45	87.36	81.87	70.75	90.73	84.69	79.10	82.69
LLaVA	90.36	95.95	88.00	89.49	88.00	90.66	90.93	90.90	90.92
LLaVA+GPT-4 (complement)	90.36	95.50	88.55	89.05	87.80	91.08	92.22	88.73	90.97
LLaVA+GPT-4 (judge)	91.56	96.74	91.09	90.62	88.99	93.52	92.73	92.16	**92.53**

Table 6: Results (accuracy %) on Science QA dataset. Question classes: NAT = natural science, SOC = social science, LAN = language science, TXT = text context, IMG = image context, NO = no context, G1-6 = grades 1-6, G7-12 = grades 7-12.

Liu et al., "Visual Instruction Tuning", NeurIPS 2023

Emerging Topics



Source: Chunyuan Li's CVPR 2023 Tutorial

Mini-GPT4

- Mini-GPT4 follows BLIP-2's architecture but replaces the language decoder with Vicuna, which better supports longer responses and multi-round conversations.

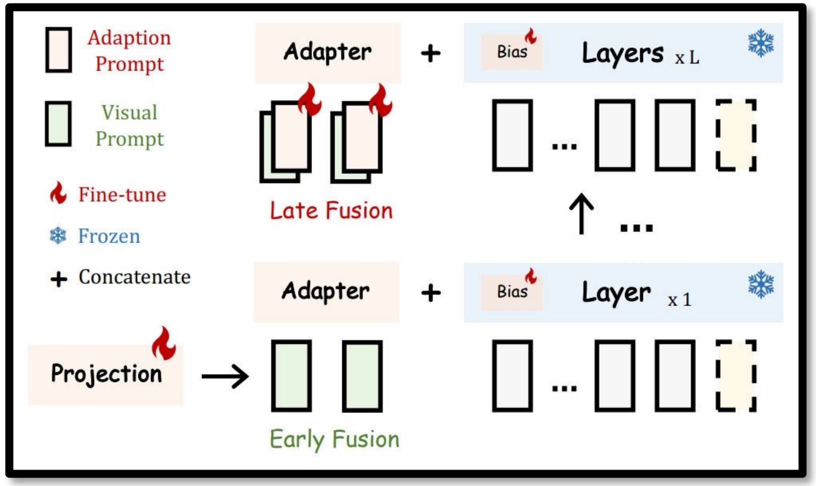
- MiniGPT-4 connects a frozen visual encoder and an LLM by pre-training on 134 million image-text pairs
  - MiniGPT4 constructs its instruction following dataset by combining Conceptual Caption, SBU, and LAION with handwritten instruction templates.

- Then improves the model's performance by further fine-tuning on a well-aligned image-text dataset.

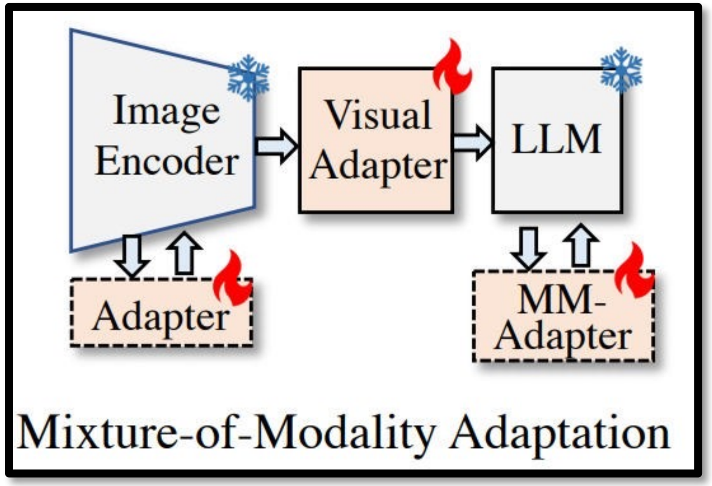
Zhu et al., "MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models", 2023

Main Categories of Instruction Tuned Models

- **Efficient Adaptation**: LLaMA-Adapter V2, LAVIN
  - LLaMA-Adapter V2: Finetuning 65B LLaMA for 24 hours on a single GPU, reaching 99.3% of the performance level of ChatGPT



LLaMA-Adapter V2: 14M parameters



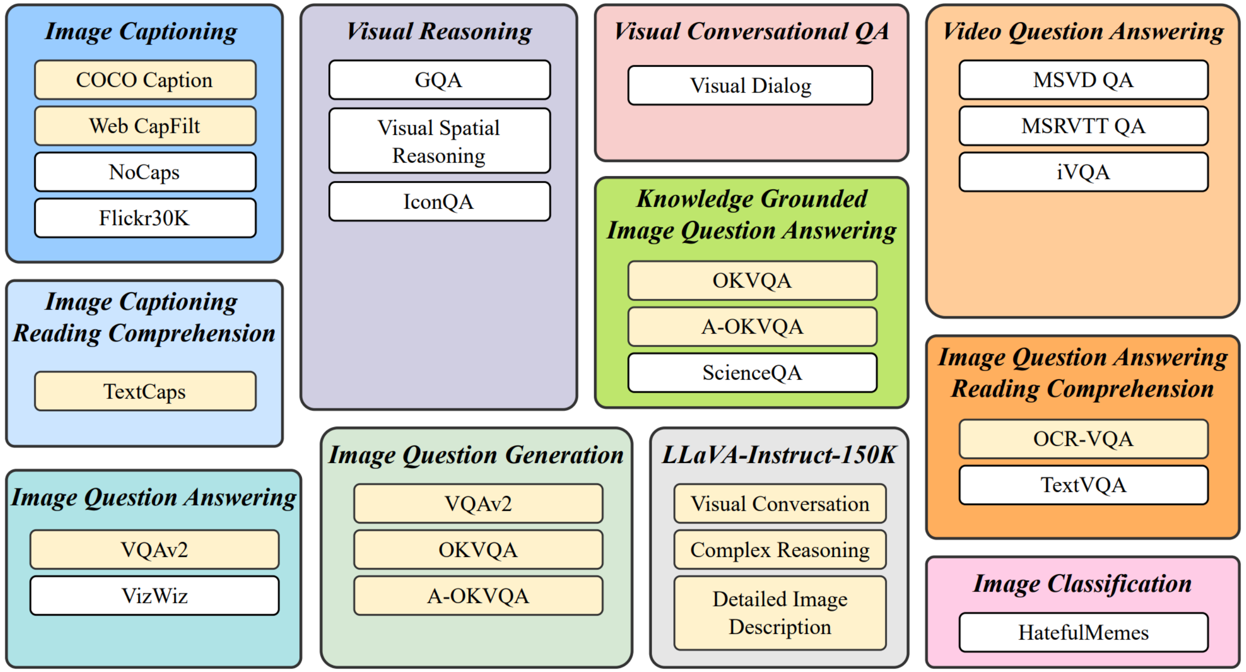
LAVIN: 3.8M parameters

QLoRA: Efficient Finetuning of Quantized LLMs

Source: Chunyuan Li's CVPR 2023 Tutorial

Main Categories of Instruction Tuned Models

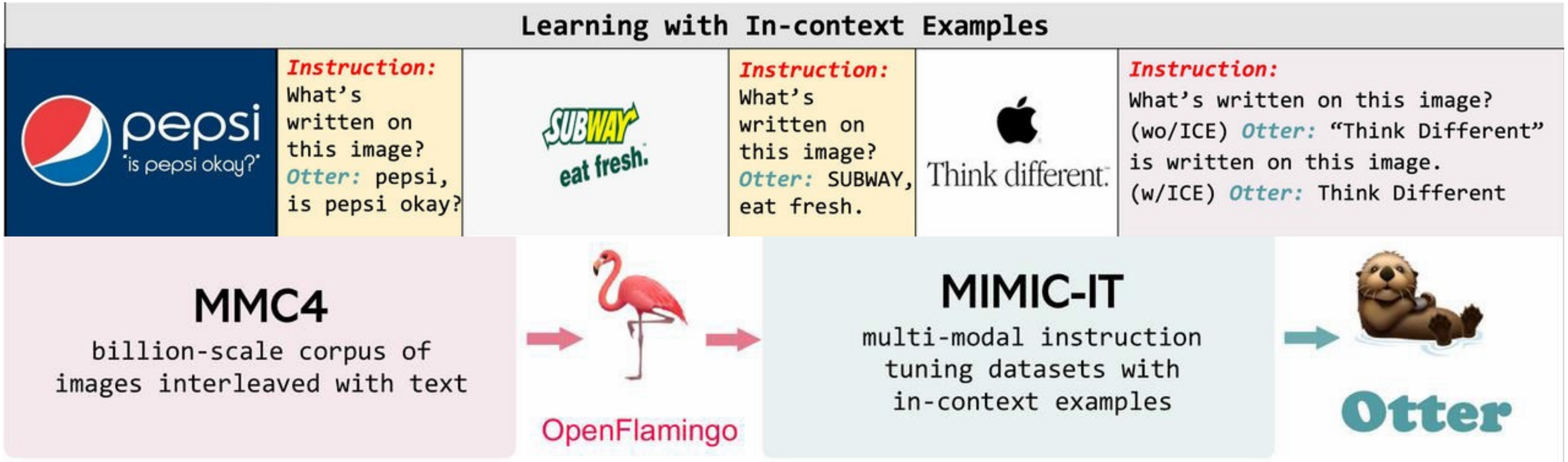
- Efficient Adaptation: LLaMA-Adapter V2, LAVIN
- **Multitask Instruct with Established Datasets**: InstructBLIP, mPlug-Owl, MultiInstruct, Multimodal GPT, InstructViT



Source: Chunyuan Li's CVPR 2023 Tutorial

Main Categories of Instruction Tuned Models

- Efficient Adaptation: LLaMA-Adapter V2, LAVIN
- Multitask Instruct with Established Datasets: InstructBLIP, mPlug-Owl, MultiInstruct, Multimodal GPT, InstructViT
- **Multimodal In-Context Learning**: OpenFlamingo, Otter/MIMIC-IT, M3IT, MetaVL

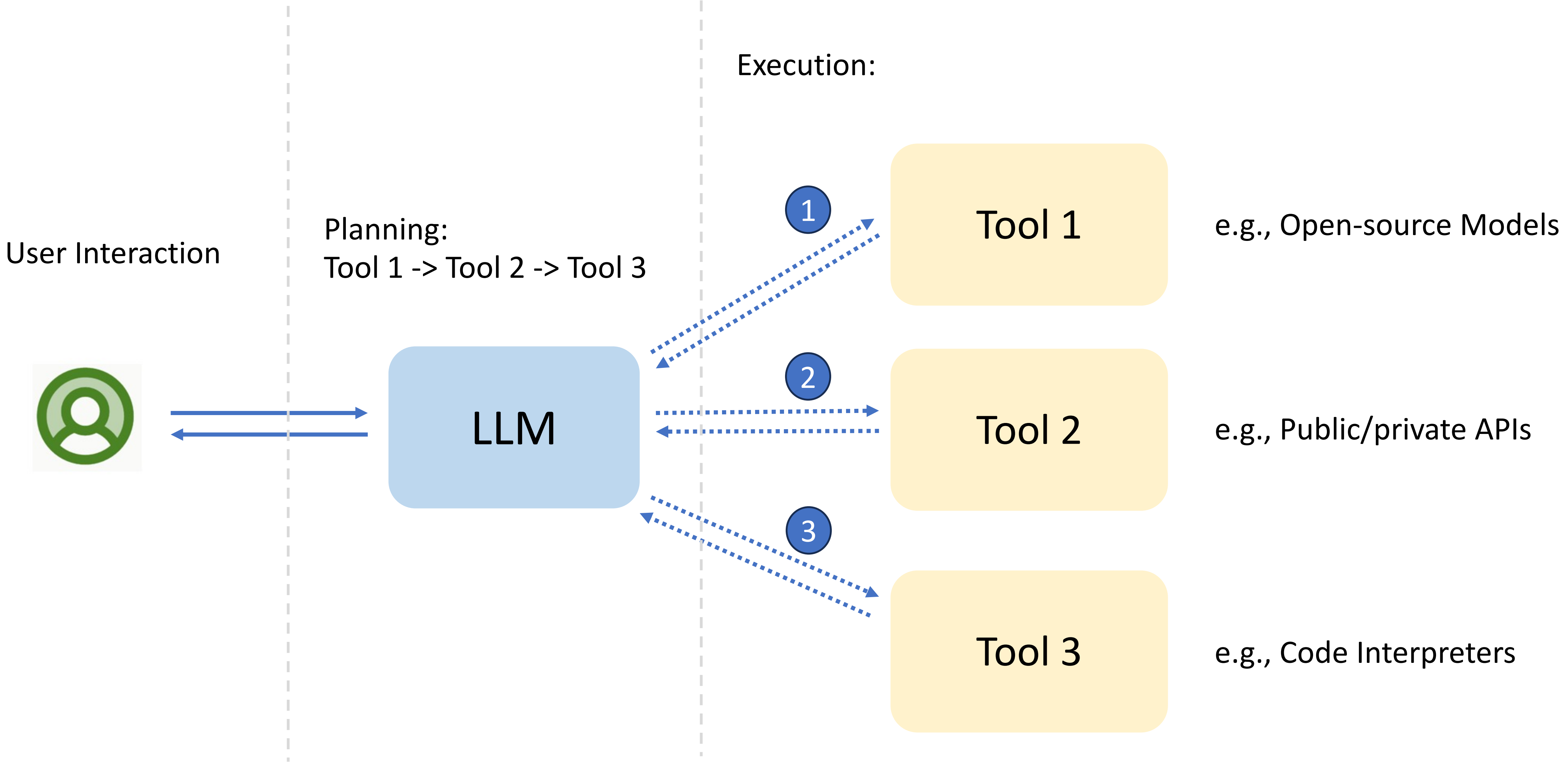


Source: Chunyuan Li's CVPR 2023 Tutorial

Adapting LMMs to Novel Tasks

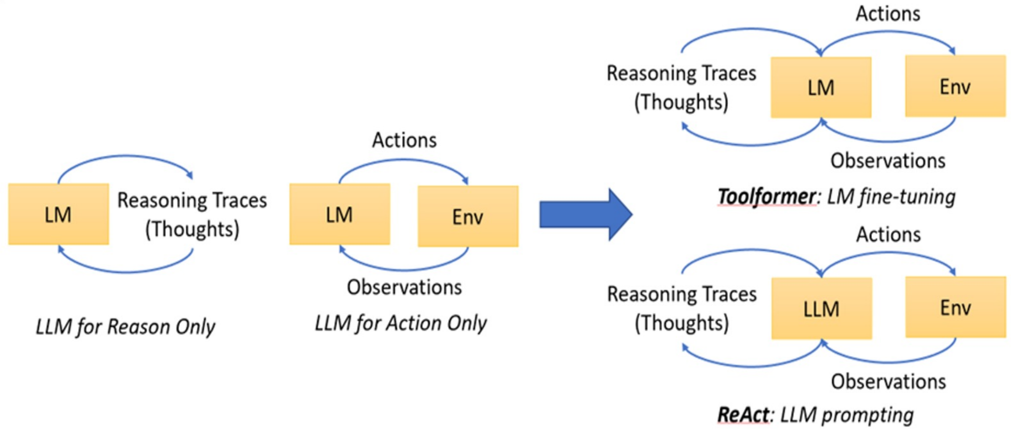
- End-to-End Trainable Models Perspective: multi-modal foundation models
  - Commercial: GPT-4v, Google's PaLM-E, Baidu's ERNIE, …
  - Academic: LLaVA, Mini-GPT4, …
- System design perspective: Multi-modal agents
  - VisualChatGPT, HuggingGPT, Cola, XGPT, MM-REACT, ViperGPT, …

New Paradigm: Standalone models on finite scenarios => Chaining tools for open problems



Chaining Paradigm (NLP): Standalone models on finite scenarios => Chaining tools for open problems

- Text-only Agents
  - ReAct, Toolformer



Model	Tool Use						Skill Dimension				
	Size						Image	Web	Know.	Math	Table
Lila [37]	1	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
PoT [6]	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
Code4Struct [51]	1	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗
PAL [9]	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
MathPrompter [17]	2	✓	✗	✗	✗	✓	✗	✗	✗	✓	✗
ART [41]	4	✓	✗	✗	✓	✓	✗	✓	✗	✓	✗
Toolformer [47]	5	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗
WebGPT [38]	10	✓	✗	✗	✓	✗	✗	✓	✗	✗	✗

- LangChain

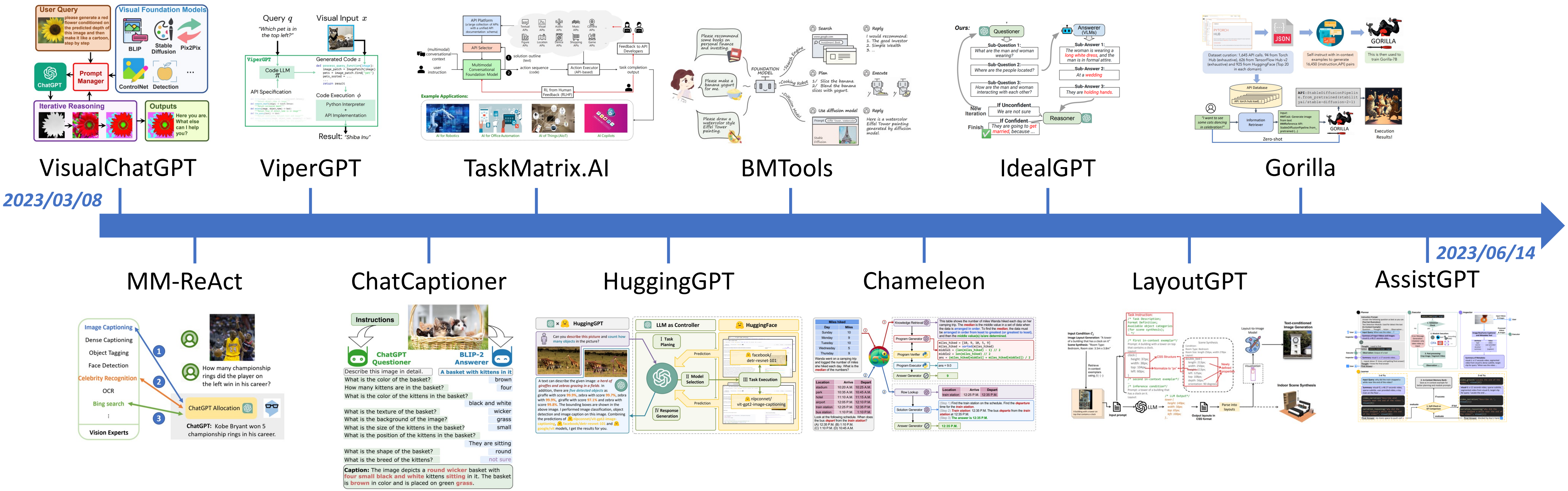
- Synergistically chaining and allocating NLP tools with LLMs

[1] Yao, Shunyu, et al. *"React: Synergizing reasoning and acting in language models."* Oct. 2022.
[2] Schick, Timo, et al. *"Toolformer: Language models can teach themselves to use tools."* Feb. 2023.
[3] Harrison Chase. *"Langchain. https://langchain.readthedocs.io"* 2023.

Linjie Li's slides, CVPR 2023 Tutorial

New Paradigm: Standalone models on finite scenarios => Chaining tools for open problems

- Multimodal Agents

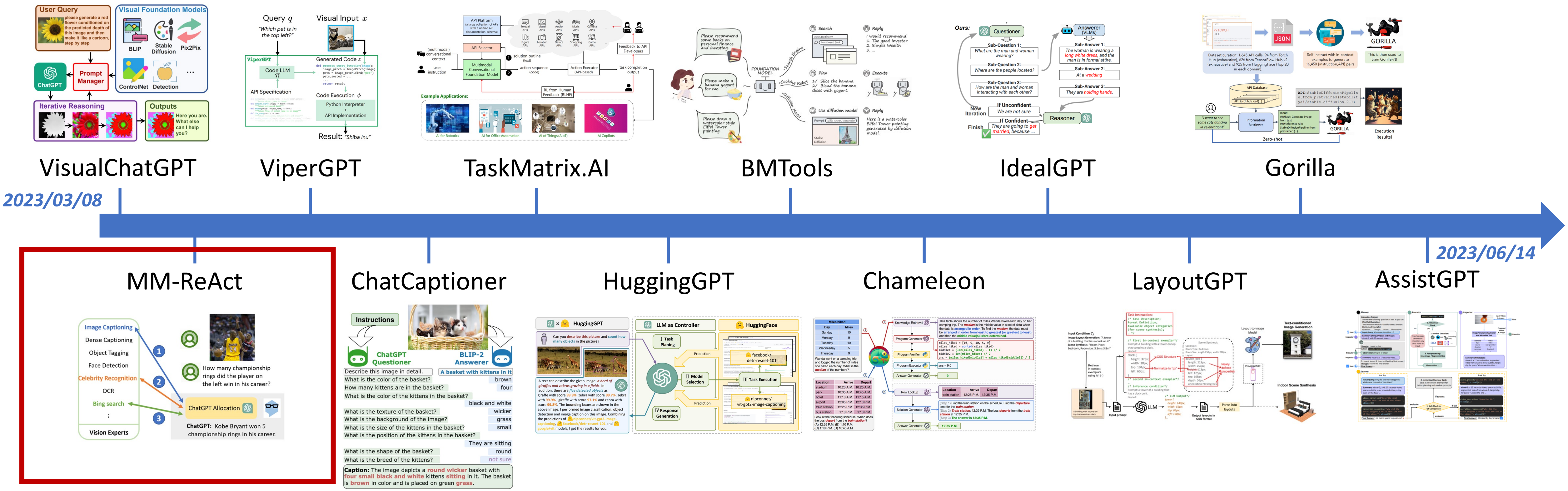


- Multimodal experts as tools for advanced multimodal reasoning and action

Linjie Li's slides, CVPR 2023 Tutorial

New Paradigm: Standalone models on finite scenarios => Chaining tools for open problems

- Multimodal Agents



- Multimodal experts as tools for advanced multimodal reasoning and action

Linjie Li’s slides, CVPR 2023 Tutorial

MM-REACT

- Leverages LLMs' high-level planning abilities to allocate various vision experts rather than solely using LLMs for text generation conditioned on visual inputs

- encourages ChatGPT to show the thought (reasoning) process to highlight why an external tool is required

Yang et al., "MM-REACT : Prompting ChatGPT for Multimodal Reasoning and Action", arXiv 2023

MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action

https://multimodal-react.github.io/

Input

Output

Chat as Interface

"What should I do after kneading the dough, before dividing it into pieces? For how long?"

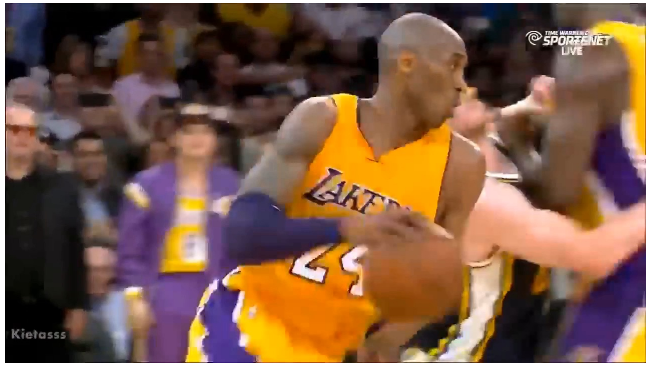
After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

You should put the dough in a warm place for 1 hour.

"Can you change the squirrel in the image to a mouse?"

"I want to make a short video to highlight the long deuce by Kobe if there is any in the video.

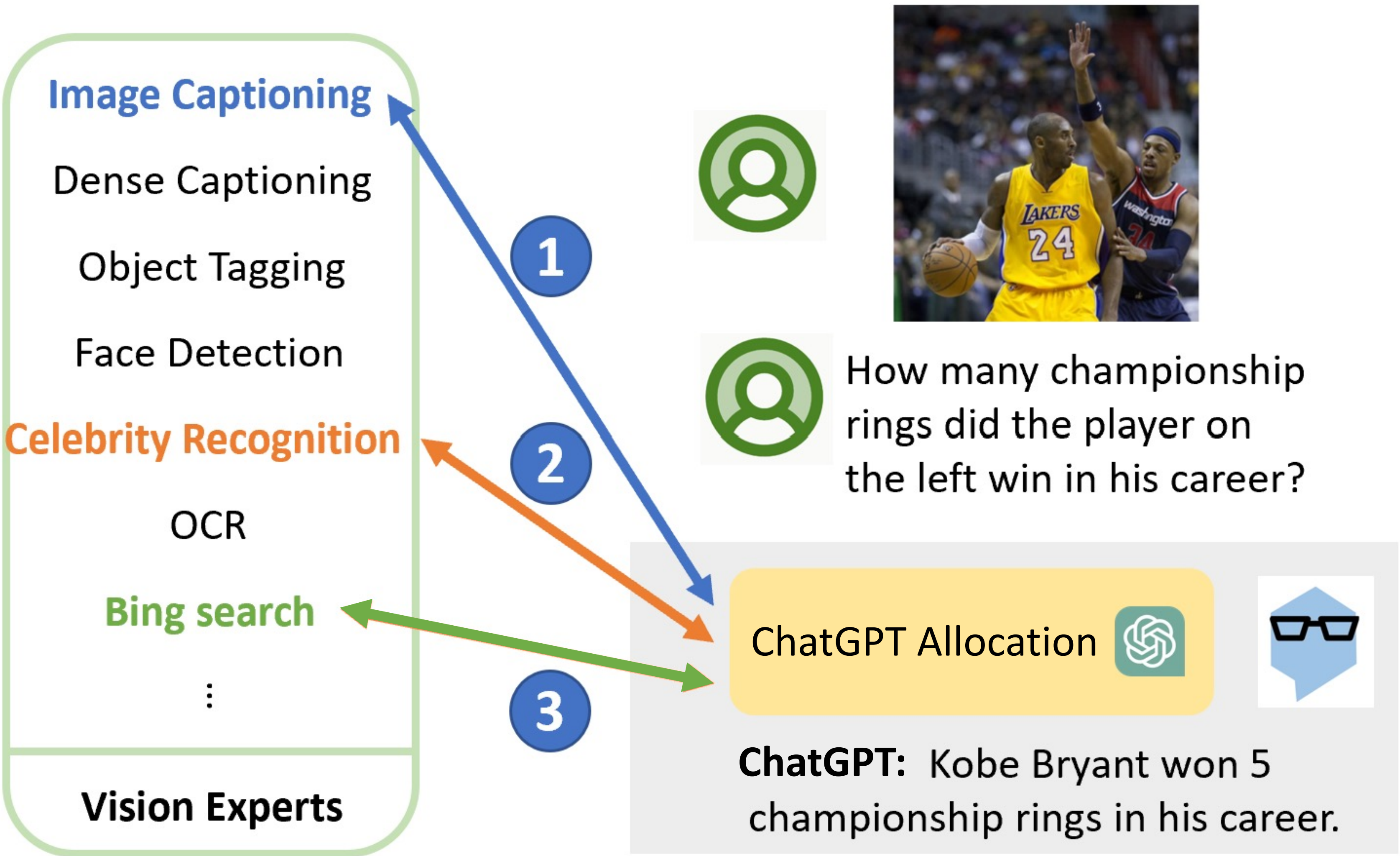
Chat as Interface



A 6-min video about the last NBA game of Kobe Bryant

MM-ReAct Design

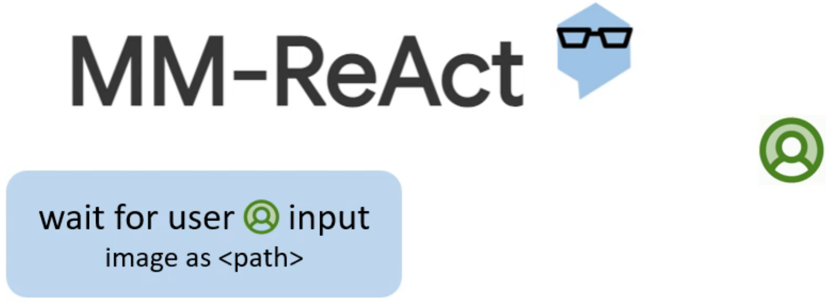
https://multimodal-react.github.io/



MM-ReAct is a system paradigm that composes numerous vision experts with ChatGPT for multimodal reasoning and action.

MM-ReAct Design

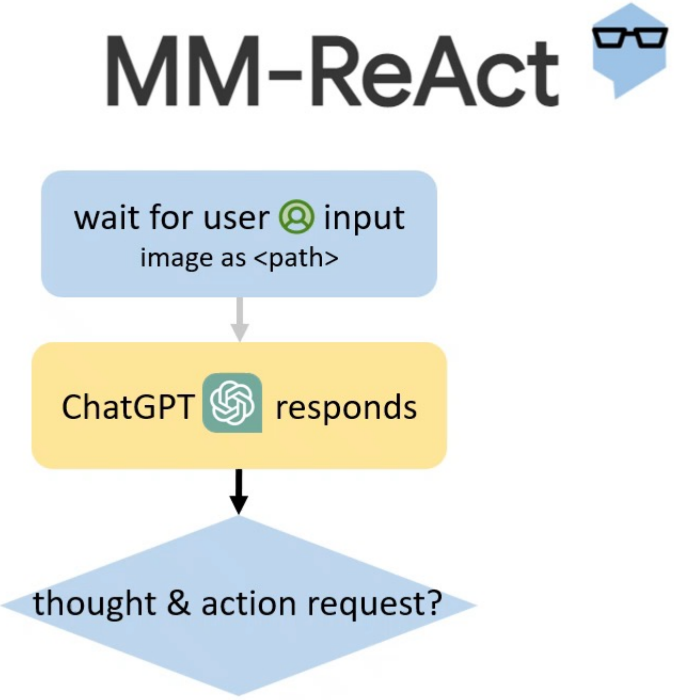
https://multimodal-react.github.io/

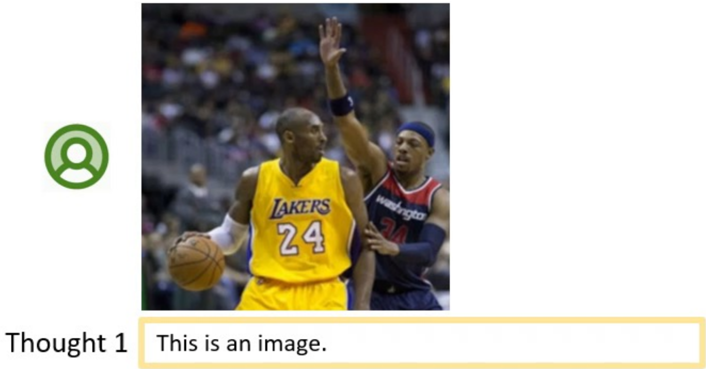


- To enable the image as input, we simply use the file path as the input to ChatGPT. The file path functions as a placeholder, allowing ChatGPT to treat it as a black box.

MM-ReAct Design

https://multimodal-react.github.io/

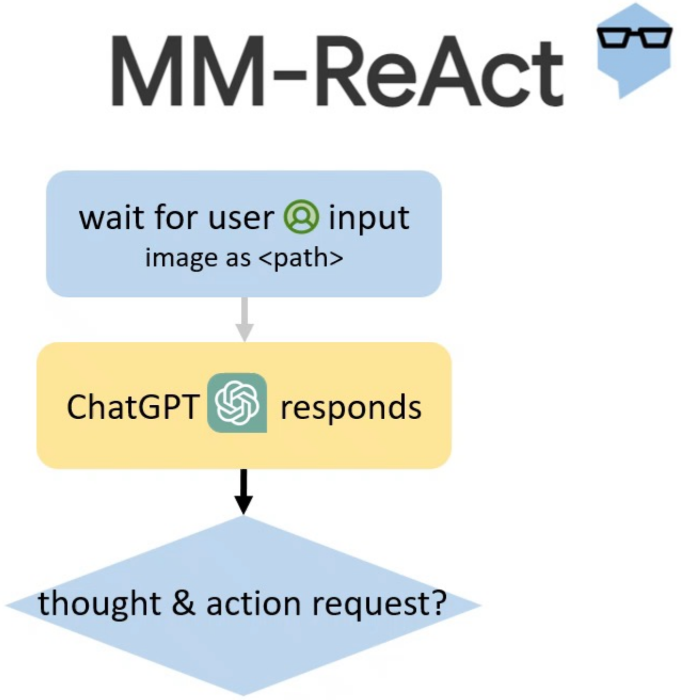


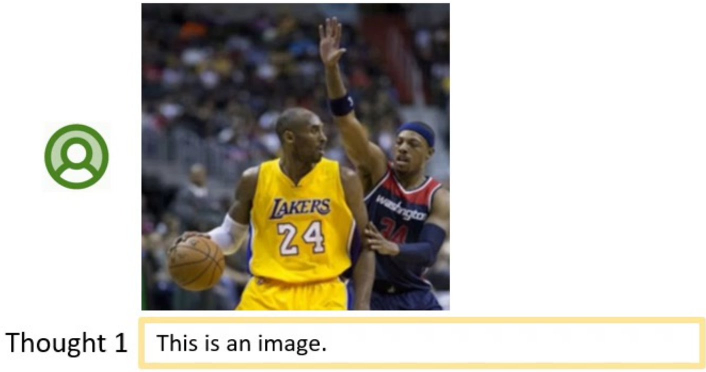


- Whenever a specific property, such as celebrity names or box coordinates, is required, ChatGPT is expected to seek help from a specific vision expert to identify the desired information.
- The expert output is serialized as text and combined with the input to further activate ChatGPT.

MM-ReAct Design

https://multimodal-react.github.io/

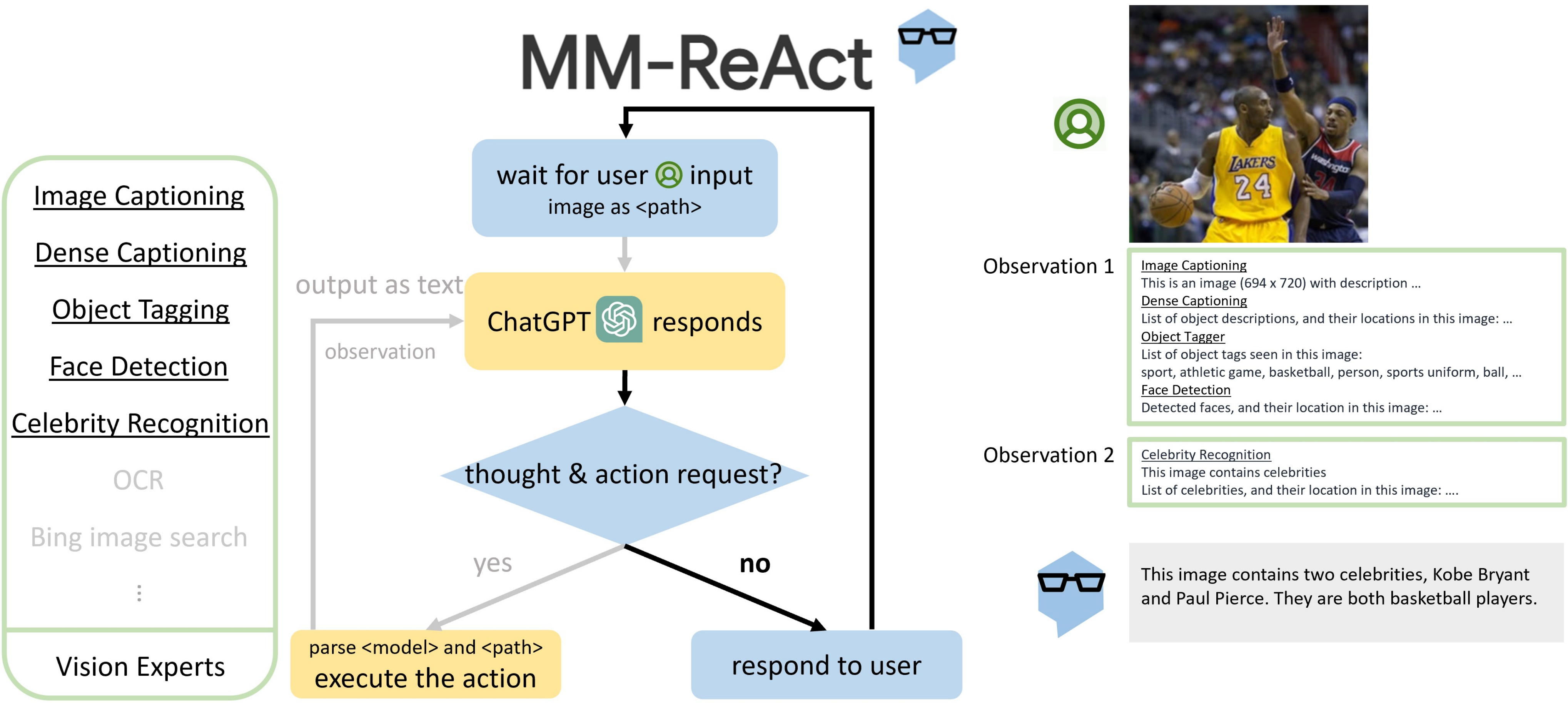




- Whenever a specific property, such as celebrity names or box coordinates, is required, ChatGPT is expected to seek help from a specific vision expert to identify the desired information.
- The expert output is serialized as text and combined with the input to further activate ChatGPT.

MM-ReAct Design

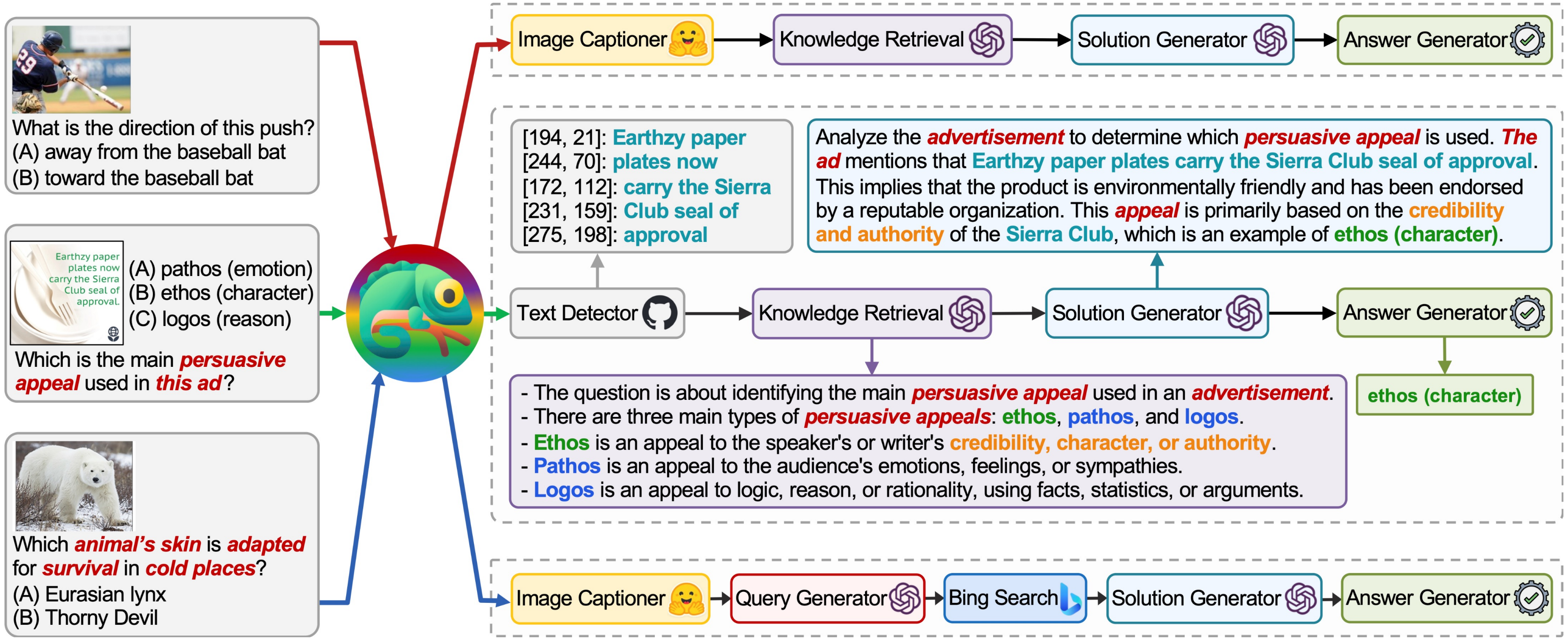
https://multimodal-react.github.io/



- If no external experts are needed, the response is directly returned to the user.

Extensibility: Plug-and-play (Adding More Tools)

Chameleon: Plug-and-Play Compositional Reasoning with Large Language Models



Model	Tool Use						Skill Dimension					Inference & Extension		
	Size						Image	Web	Know.	Math	Table	Composition	Planning	Plug-n-Play
MM-ReAct [56]	>10	✓	✗	✗	✓	✗	✓	✓	✓	✓	✓	✓	word match	✓
Visual ChatGPT [55]	>10	✓	-	-	✗	✗	✓	✗	✗	✗	✗	✓	natural lang.	✓
ViperGPT [49]	>10	✓	-	-	✗	✗	✓	✗	✓	✓	✗	✓	program	✓
VisProg [12]	>10	✓	-	-	✗	✓	✓	✗	✗	✗	✗	✓	program	✓
HuggingGPT [48]	>10	✓	✓	✗	✗	✗	✓	✗	-	✗	-	✓	natural lang.	✓
Chameleon (ours)	>10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	natural lang.	✓

Linjie Li's slides, CVPR 2023 Tutorial

Discussion

- LLMs -> Advanced Multimodal Systems

	Pros	Cons
Instruction Tuning	• End-to-end model that directly interprets rich semantics in multimodal inputs	• Require data curation and training => more expensive • Limited data => limited capabilities in certain scenarios (e.g., OCR)
Tool Using	• No training needed • Can leverage abundant off-the-self models/APIs/code interpreters as tools	• No training => Failures in invoking the right tool • Weak domain experts => weak performance

- Can we use LMMs (e.g., LLaVa) as the tool allocator?
- If so, what capabilities would require a tool? And what can be solved by instruction tuning?

Embodied Agents: Language-driven Robotics

- Embodied agents can follow natural language commands to do different tasks as well as learn to do new tasks quickly
- High-level planning, perceptual feedback, and low-level control are among sub-tasks of an embodied agent

Planning and Embodied Agents

- Zero-shot planners (Huang et al., ICML 2022)
- SayCan (Ahn et al., 2022)
- Inner Monolouge (Huang et al., CoRL 2022): Embodied reasoning through planning with language models
- LLM-Planner (Song et al., ICCV 2023)
- PaLM-E (Driess et al., 2023)

LLM as a High-level Planner

- High-level planning: Uses LLMs to decompose abstract instructions into a sequence of subgoals executable by an agent (i.e., actionable steps)
  - Cook a potato and put it on the plate =>

    [Navigation potato, Pickup potato, Navigation microwave, ...]

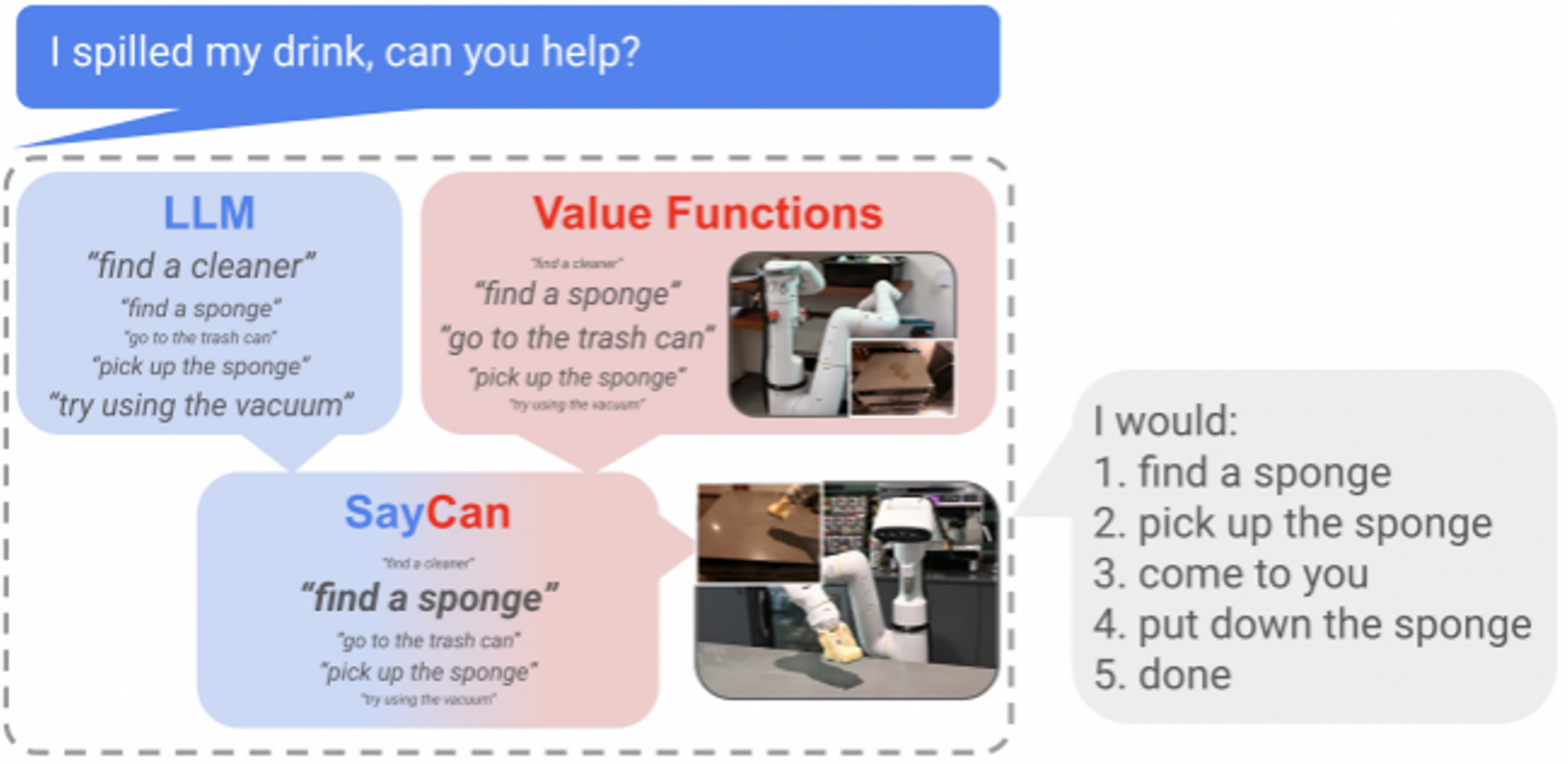
- Low-level planner can be trained more easily: agent can execute short-horizon skills from a library of previously trained policies trained with RL or BC

Do As I Can, Not As I Say: Grounding Language in Robotic Affordances

[1] Michael Ahn*, Anthony Brohan*, Noah Brown*, Yevgen Chebotar*, Omar Cortes*, Byron David*, Chelsea Finn*, Chuyuan Fu†, Keerthana Gopalakrishnan*, Karol Hausman*, Alex Herzog†, Daniel Ho†, Jasmine Hsu*, Julian Ibarz*, Brian Ichter*, Alex Irpan*, Eric Jang*, Rosario Jauregui Ruano*, Kyle Jeffrey*, Sally Jesmonth*, Nikhil J Joshi*, Ryan Julian*, Dmitry Kalashnikov*, Yuheng Kuang*, Kuang-Huei Lee*, Sergey Levine*, Yao Lu*, Linda Luu*, Carolina Parada*, Peter Pastor†, Jornell Quiambao*, Kanishka Rao*, Jarek Rettinghouse*, Diego Reyes*, Pierre Sermanet*, Nicolas Sievers*, Clayton Tan*, Alexander Toshev*, Vincent Vanhoucke*, Fei Xia*, Ted Xiao*, Peng Xu*, Sichun Xu*, Mengyuan Yan†, Andy Zeng*

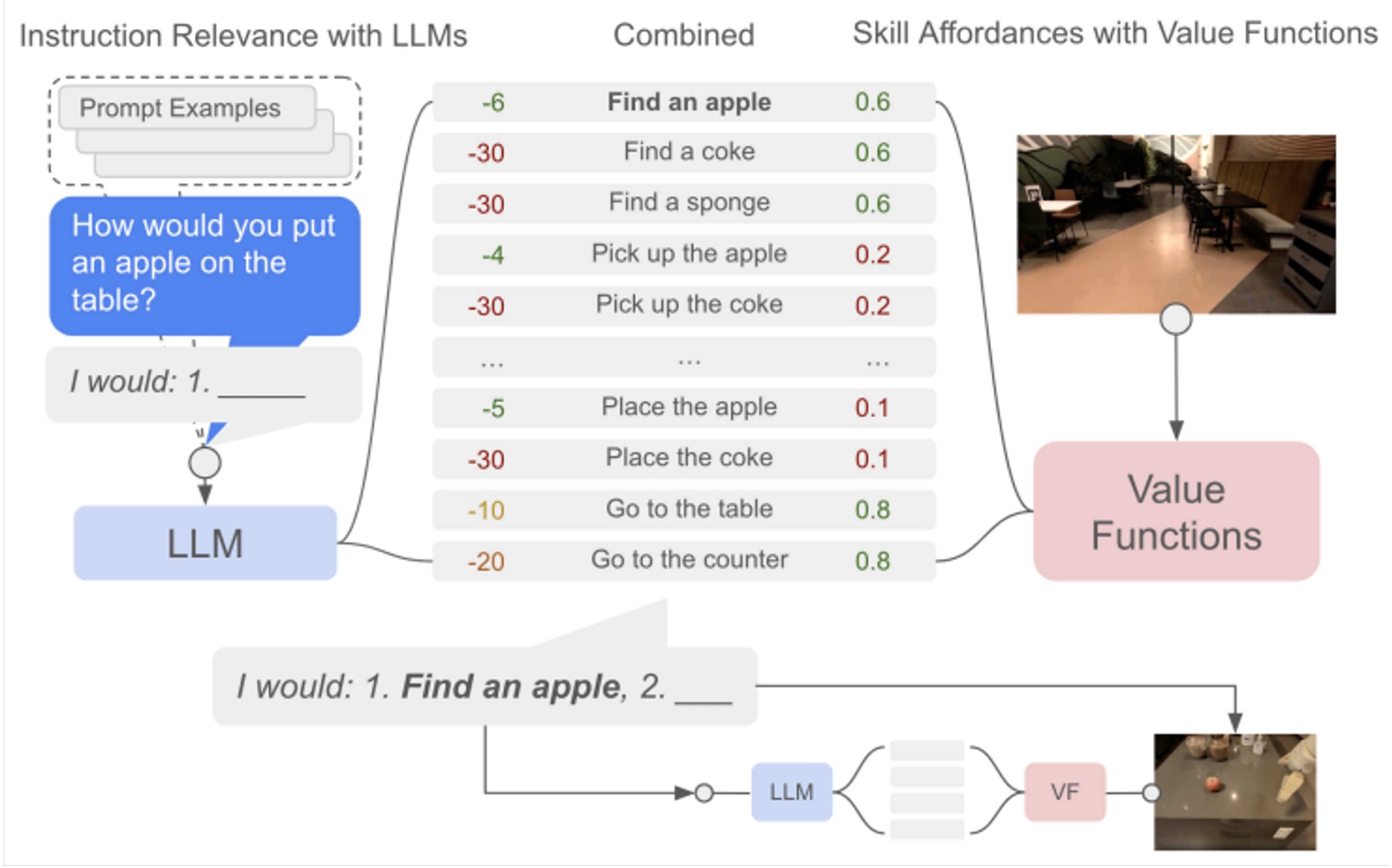
*Robotics at Google, †Everyday Robots

SayCan one of the pioneering work on using LLMs for embodied instruction following



Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances", arXiv 2022

Ranking admissible skills → multiplying each candidate action's probability under LLMs with the action's value function



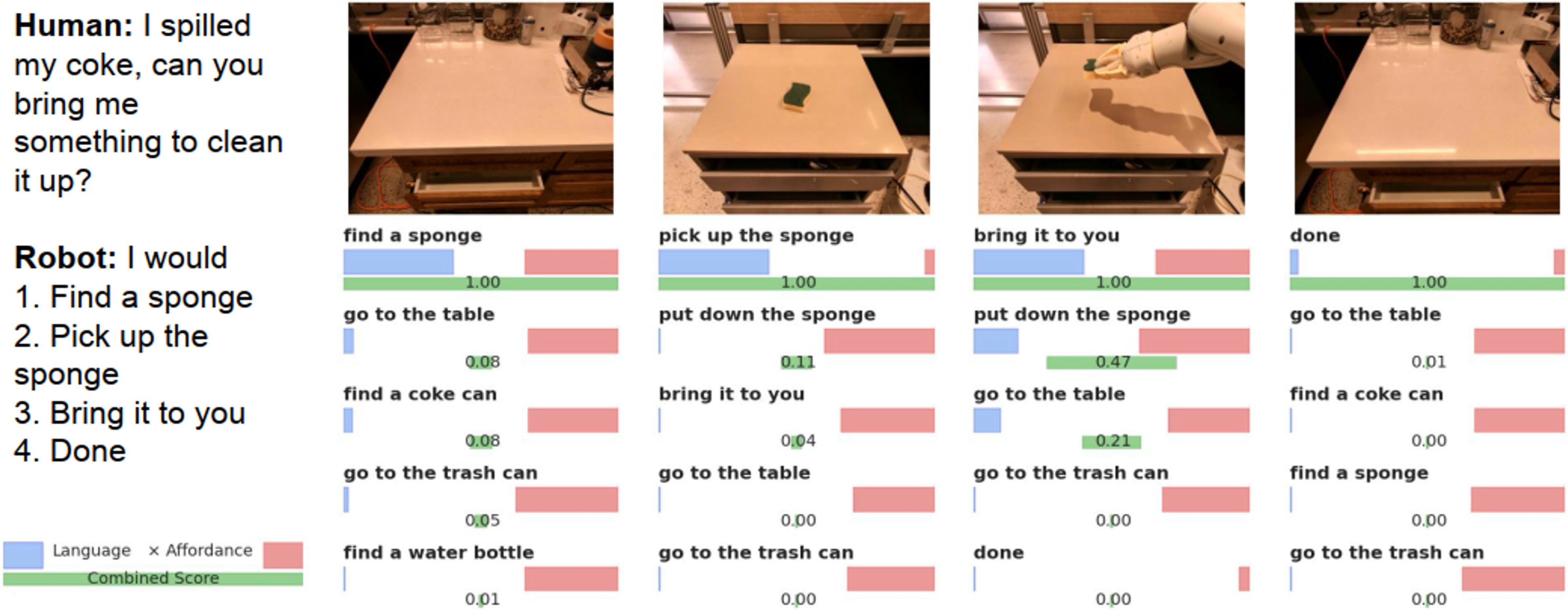
Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances", arXiv 2022

Algorithm 1 SayCan

Given: A high level instruction i, state s_0, and a set of skills Π and their language descriptions ℓ_Π

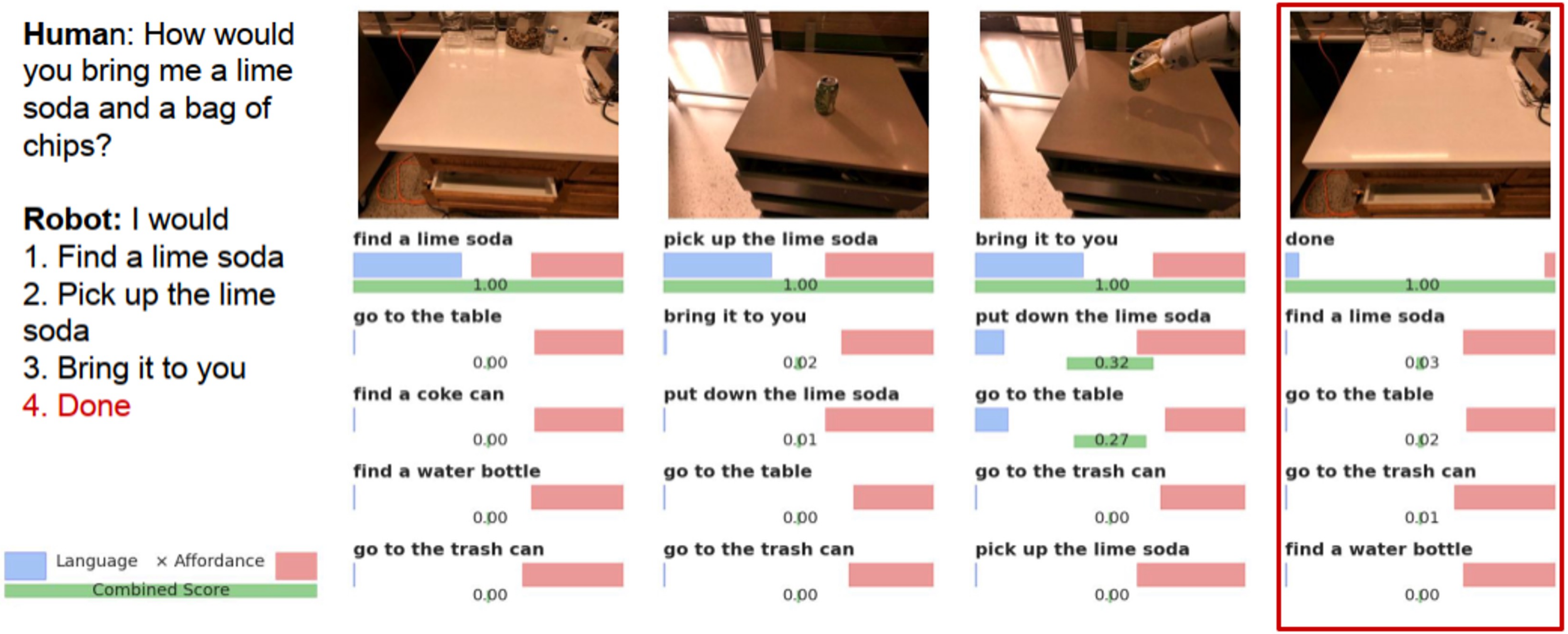
$n = 0, \pi = \emptyset$
while $\ell_{\pi_{n-1}} \neq$ "done" **do**
  $\mathcal{C} = \emptyset$
  for $\pi \in \Pi$ and $\ell_\pi \in \ell_\Pi$ **do**
    $p_\pi^{\text{LLM}} = p(\ell_\pi | i, \ell_{\pi_{n-1}}, ..., \ell_{\pi_0})$ ▷ Evaluate scoring of LLM
    $p_\pi^{\text{affordance}} = p(c_\pi | s_n, \ell_\pi)$ ▷ Evaluate affordance function
    $p_\pi^{\text{combined}} = p_\pi^{\text{affordance}} p_\pi^{\text{LLM}}$
    $\mathcal{C} = \mathcal{C} \cup p_\pi^{\text{combined}}$
  end for
  $\pi_n = \arg\max_{\pi \in \Pi} \mathcal{C}$
  Execute $\pi_n(s_n)$ in the environment, updating state s_{n+1}
  $n = n + 1$
end while

Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances", arXiv 2022



SayCan assumes that each proposed step is executed successfully by the agent.

Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances", arXiv 2022



65% of errors was the planner error and 35% affordance error

Ahn et al., "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances", arXiv 2022

Inner Monologue: Embodied Reasoning through Planning with Language Models

Wenlong Huang[†], Fei Xia[†], Ted Xiao[†], Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, Brian Ichter

Robotics at Google, [†] equal contribution and alphabetically listed

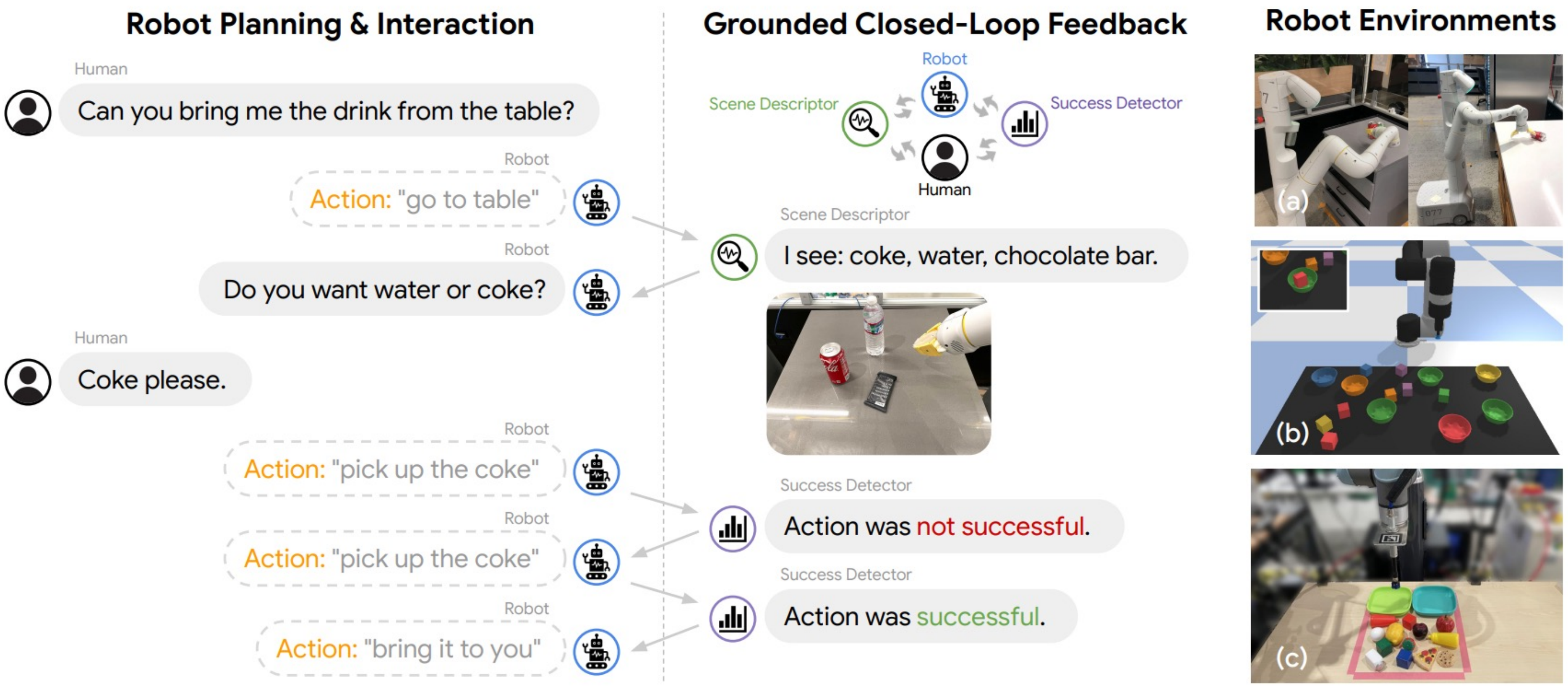
Project website: https://innermonologue.github.io

Inner Monologue

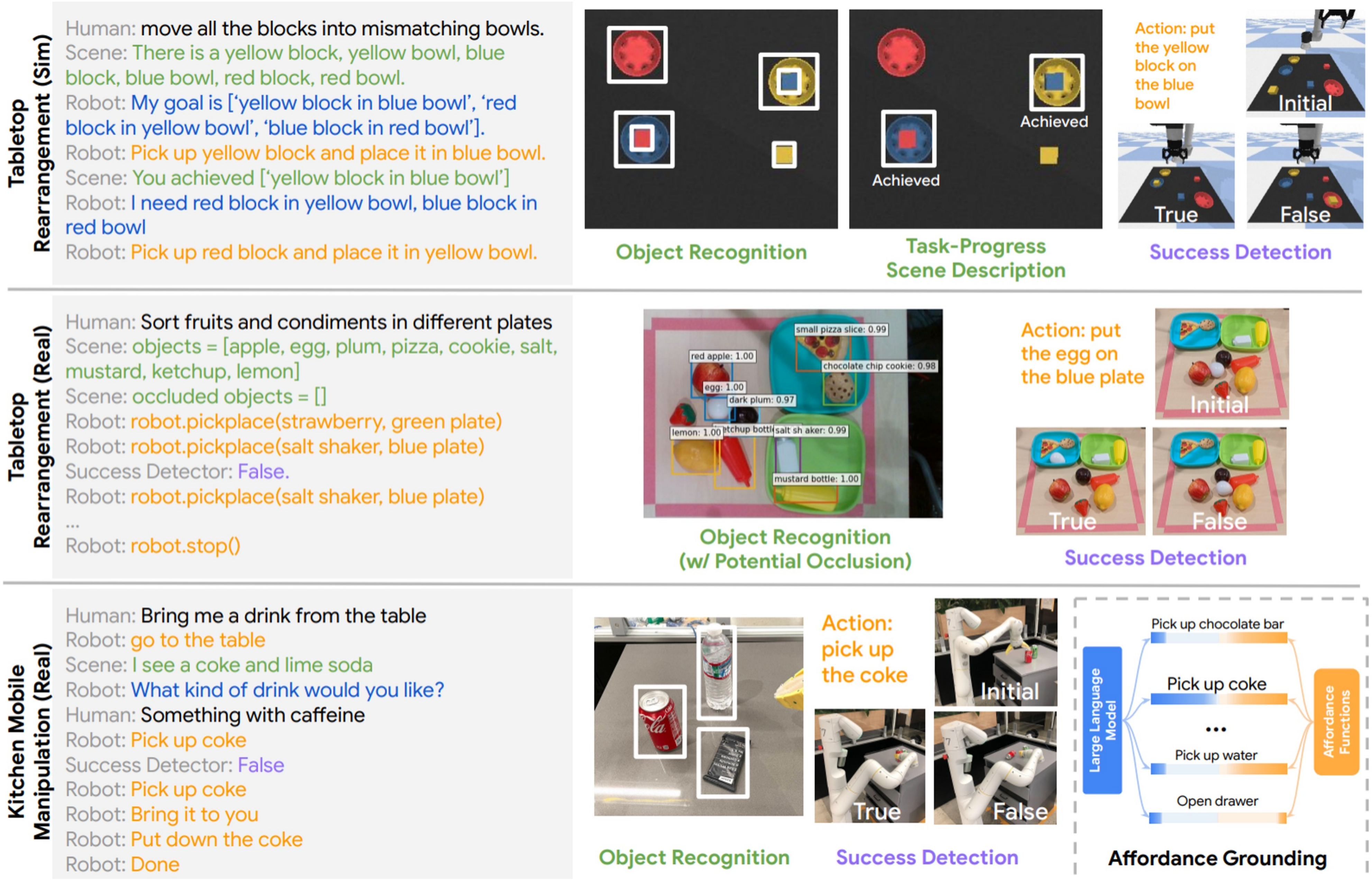
- SayCan may not be robust in handling intermediate failures
- Inner Monologue explores ways to incorporate grounded feedback from the environment into LLM

Huang et al., “Inner Monologue: Embodied Reasoning through Planning with Language Models”, CoRL 2022

Inner-Monolouge: Closed-loop Feedback



Huang et al., "Inner Monologue: Embodied Reasoning through Planning with Language Models", CoRL 2022



Huang et al., "Inner Monologue: Embodied Reasoning through Planning with Language Models", CoRL 2022

Task Family	SayCan	+Inner Monologue *Success*	+Inner Monologue *Object + Success*
No Disturbances			
Manipulation	50.0%	62.5%	**75.0%**
Mobile Manipulation	50.0%	50.0%	**75.0%**
Drawers	83.3%	83.3%	**100.0%**
With Disturbances			
Manipulation	12.5%	25.0%	**33.3%**
Mobile Manipulation	0.0%	25.0%	**75.0%**
Drawers	0.0%	**44.4%**	**44.4%**
Total	30.8%	48.7%	**60.4%**

Table 3: Averaged success rate across 120 evaluations on several task families in our real-world mobile manipulation environment. We consider a standard setting and adversarial setting with external human disturbances. In all cases, LLM-informed embodied feedback is shown to be effective in improving robustness of the system, especially when low-level policies are prone to failures.

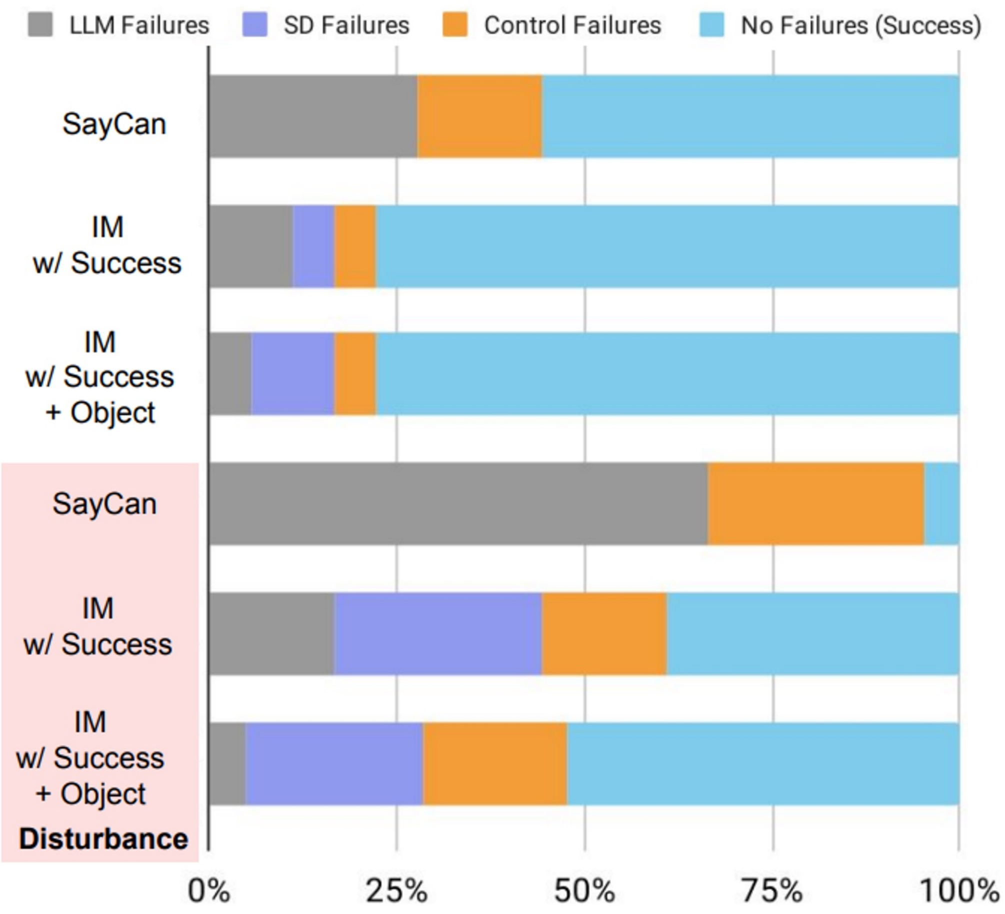
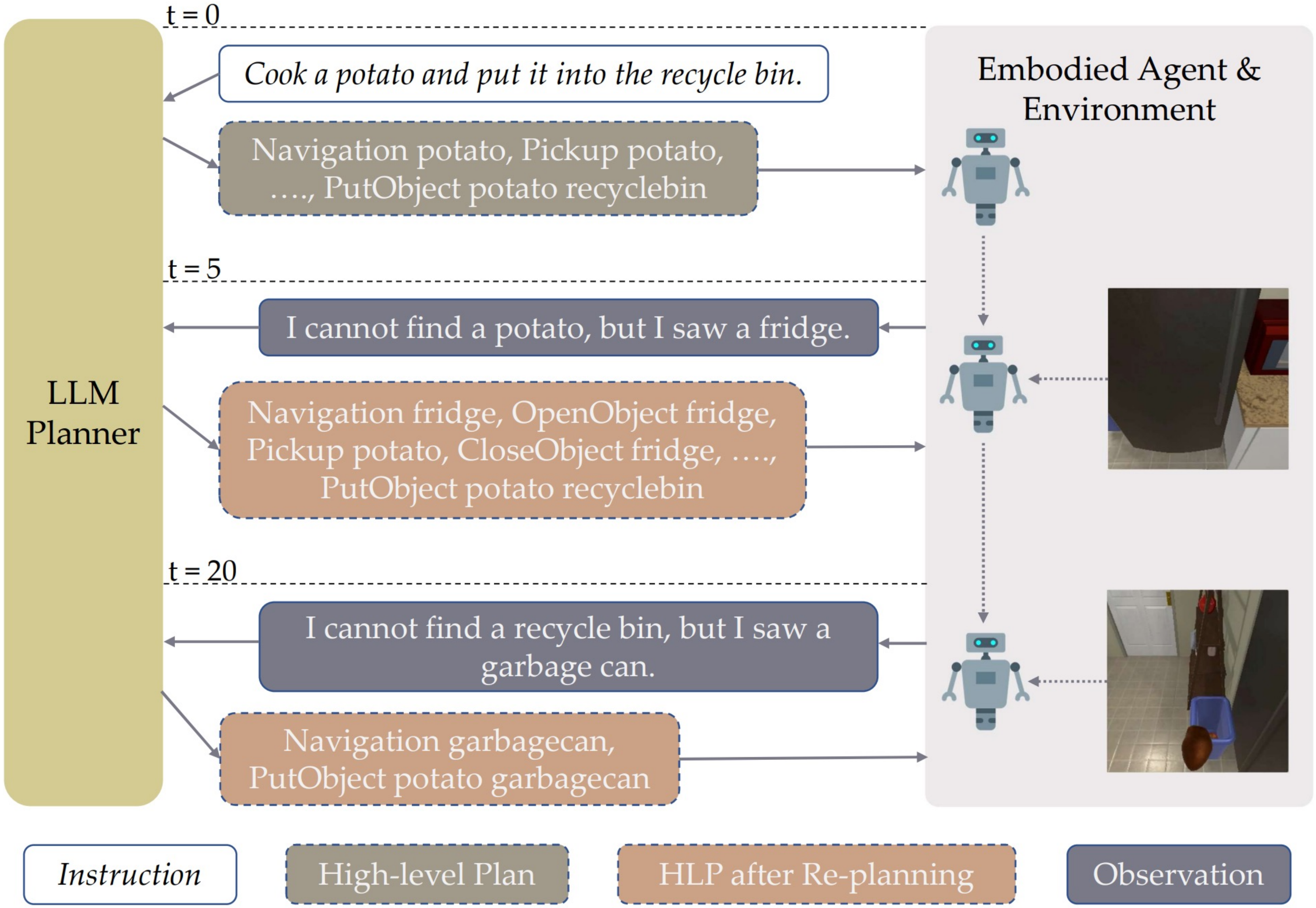


Figure 4: Failure causes on 120 evaluations. When disturbances are added (red), only the Inner Monologue variants consistently complete the instructions.

Huang et al., "Inner Monologue: Embodied Reasoning through Planning with Language Models", CoRL 2022

LMM-Planner



Song et al., "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models", ICCV 2023

LMM-Planner

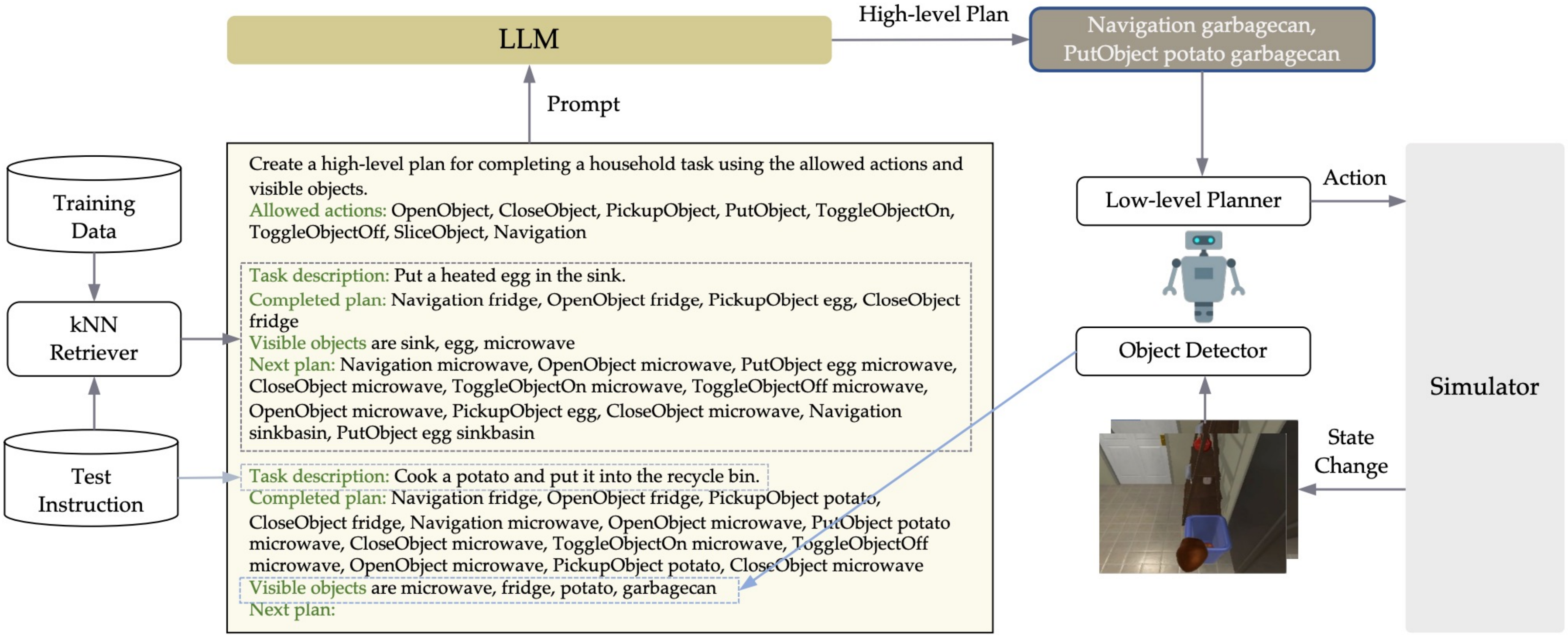
- LLM-Planner **dynamically adjust** the plan from LLMs based on **environmental perception**

- Whenever agent has taken too many steps for the current subgoal or has made too many failed attempts, it replans
  - prompts the LLM again to generate a new continuation of the plan

- A novel grounded replanning algorithm to empower LLM-Planner with physical grounding.

Song et al., "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models", ICCV 2023

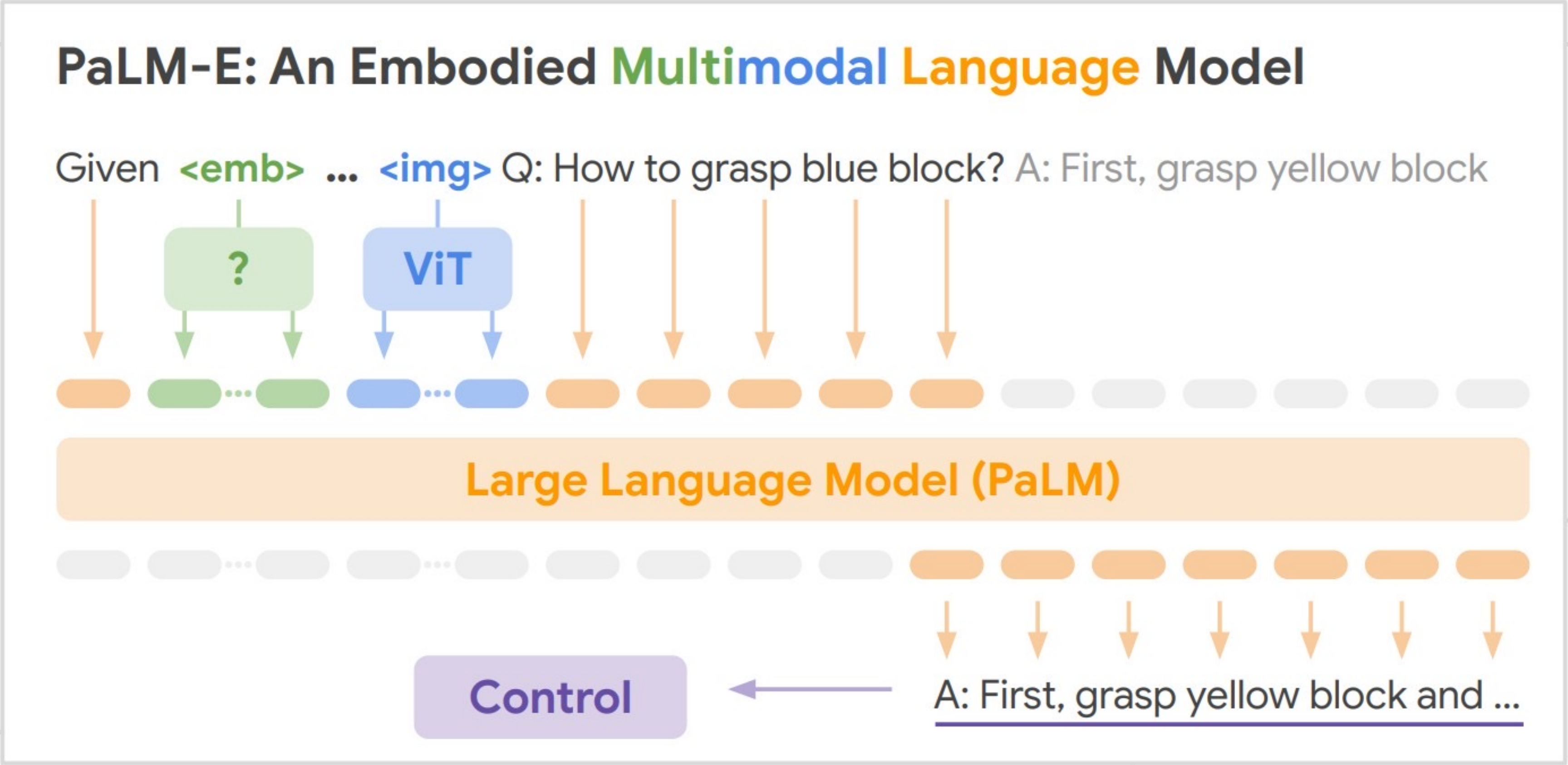
LMM-Planner



Song et al., "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models", ICCV 2023

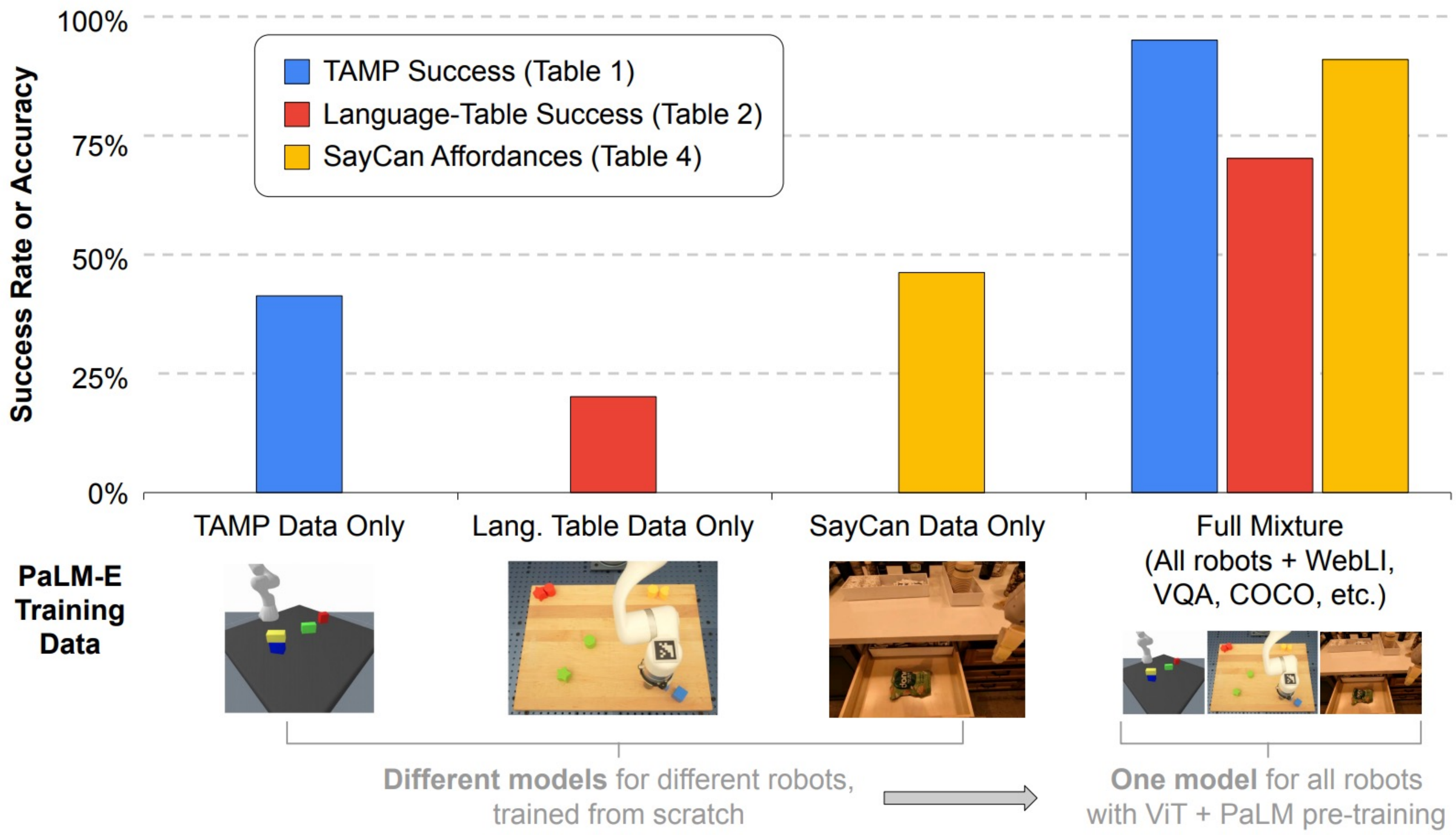
PaLM-E: An LMM for embodied AI

PaLM-E is a generative model producing text based on multi-model sentences as input



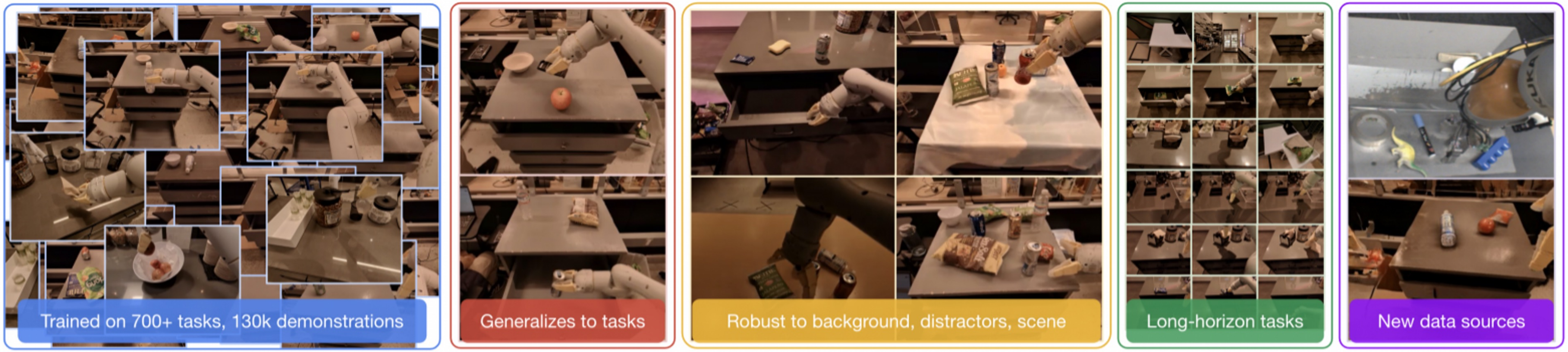
https://palm-e.github.io/#demo

Driess et al., "PaLM-E: An Embodied Multimodal Language Model", 2023



Driess et al., "PaLM-E: An Embodied Multimodal Language Model", 2023

RT-1



(b) RT-1's large-scale, real-world training (130k demonstrations) and evaluation (3000 real-world trials) show impressive generalization, robustness, and ability to learn from diverse data.

Conclusion

- Instruction tuning to improve LMMs capabilities
- LLMs as planner for chaining tools as another way
- Embodied Agents for language-driven robotics by the above two ways

Questions