



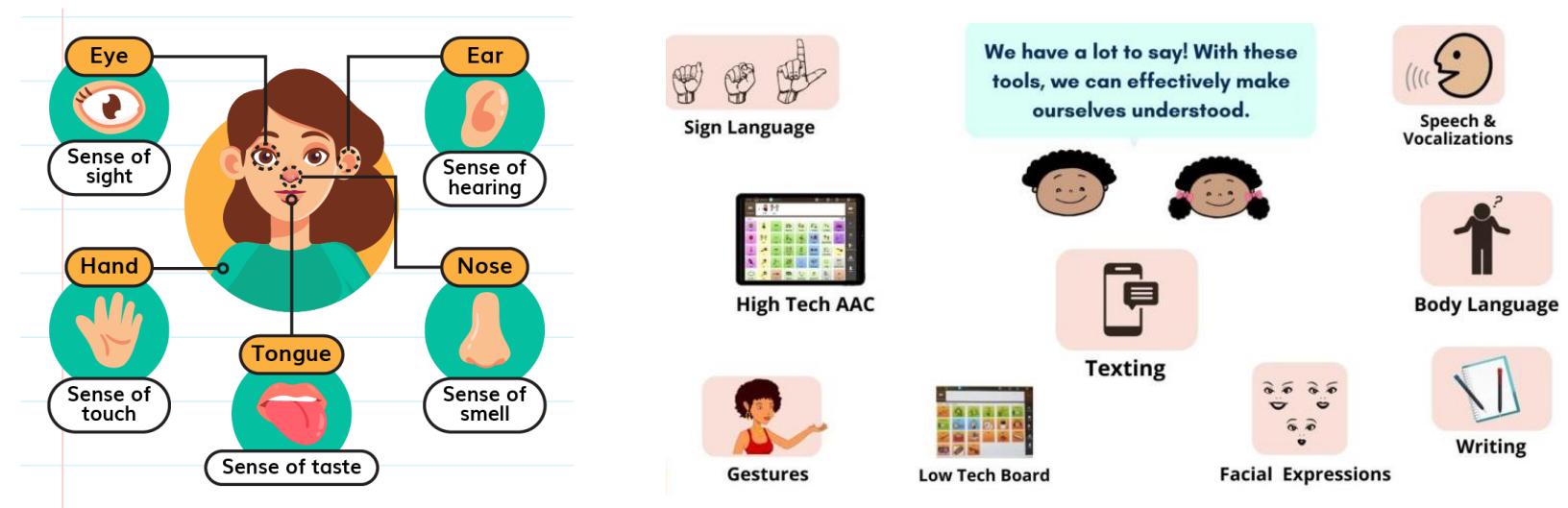
# Connecting Language to the World

CSCI 601-471/671 (NLP: Self-Supervised Models)

<https://self-supervised.cs.jhu.edu/sp2024/>

# Connecting Language to the World

- So far, we have focused on only “language” as our primary focus.
- But there are other modalities in which humans communicate with the world.



# Connecting LMs to the World: Chapter Plan

---

1. Connecting vision and language
2. Generative vision-language models
3. Other modalities [speech, audio, ...]
4. From language to action

**Chapter goal:** Getting comfortable with thinking about extending LLMs to other modalities of the world and their limits.



# Connecting vision and language

# Computer Vis

- 1960s - First computer vis

MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
PROJECT MAC

Artificial Intelligence Group  
Vision Memo. No. 100.

July 7, 1966

## THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

# Computer Vision Abridged History

- 1960s - First computer vision projects (MIT summer project)
- 2000s - Emergence of “tasks” and benchmarking in computer vision

PASCAL Visual Object Classes (2005-2012)



[https://en.wikipedia.org/wiki/List\\_of\\_datasets\\_in\\_computer\\_vision\\_and\\_image\\_processing](https://en.wikipedia.org/wiki/List_of_datasets_in_computer_vision_and_image_processing)



Caltech-101 (2003)

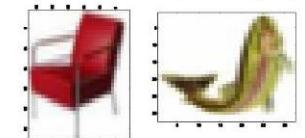


Caltech-256 (2007)



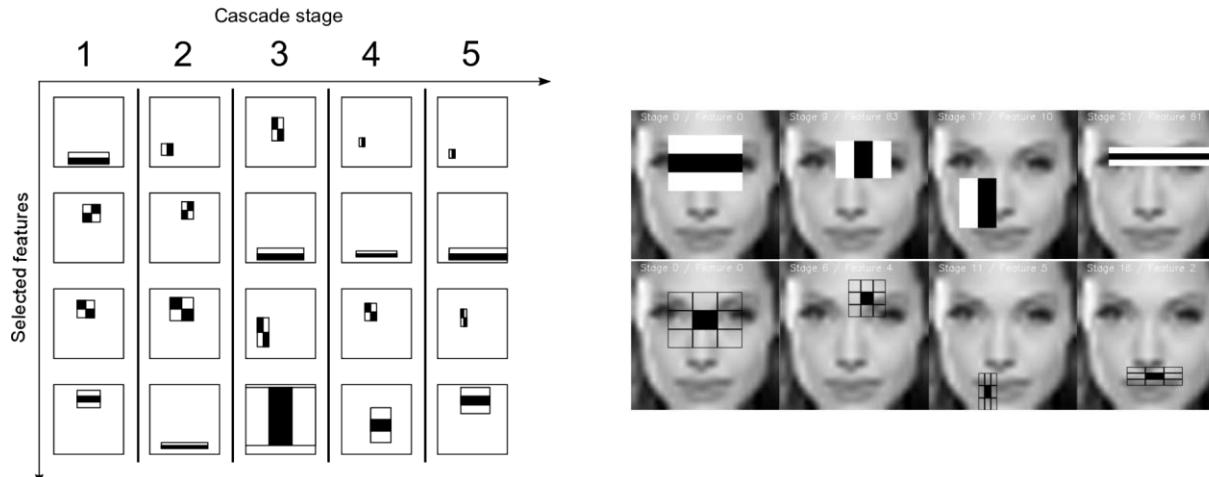
ImageNet (2009)

ImageNet (2009)  
(subset of tiny images)



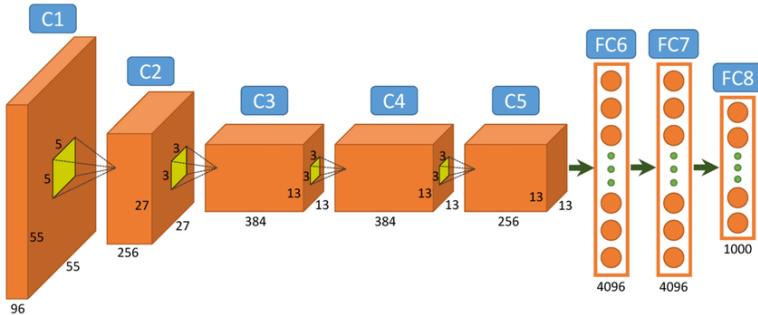
# Computer Vision Abridged History

- 1960s - First computer vision projects (MIT summer project)
- 2000s - Emergence of “tasks” and benchmarking in computer vision
- 2000s - Shallow classifiers and feature engineering (e.g., Viola & Jones algorithm)

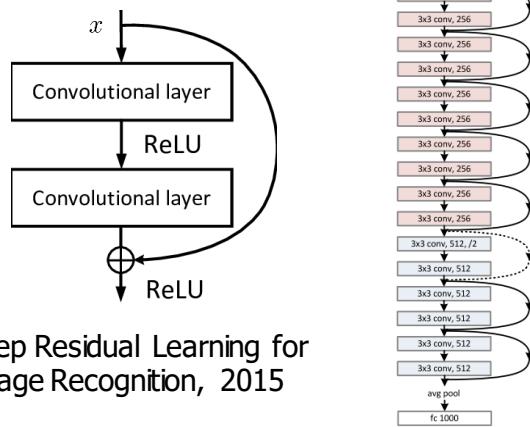


# Computer Vision Abridged History

- 1960s - First computer vision projects (MIT summer project)
  - 2000s - Emergence of “tasks” and benchmarking in computer vision
  - 2000s - Shallow classifiers and feature engineering
  - 2012 - Deep Learning revolution:
    - Success of Convolutional neural nets in ImageNet



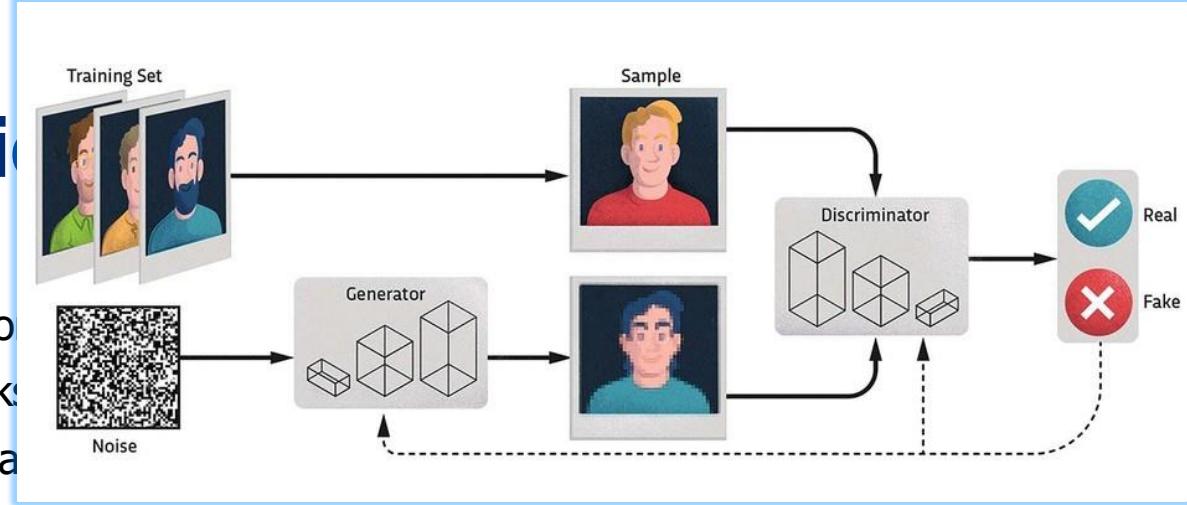
# ImageNet Classification with Deep Convolutional Neural Networks, 2012



# Deep Residual Learning for Image Recognition, 2015

# Computer Vision

- 1960s - First computer vision
- 2000s - Emergence of “task”
- 2000s - Shallow classifiers a
- 2012 - Deep Learning revolution:
  - Success of Convolutional neural nets in ImageNet
  - Unification of architectures
  - Rise of image generation (VAEs, GANs, etc.)



# Computer Vision Abridged History

---

- 1960s - First computer vision projects (MIT summer project)
- 2000s - Emergence of “tasks” and benchmarking in computer vision
- 2000s - Shallow classifiers and feature engineering
- 2012 - Deep Learning revolution:
  - Success of Convolutional neural nets in ImageNet
  - Unification of architectures
  - Rise of image generation (VAEs, GANs, etc.)
- 2020s - Era of Vision Transformer
  - Stronger connection to language
  - Better generative models
  - Further unification of models and tasks

This is where we begin!

# Let's Consider Images – How to Encode?

---



# Vision Transformers

---

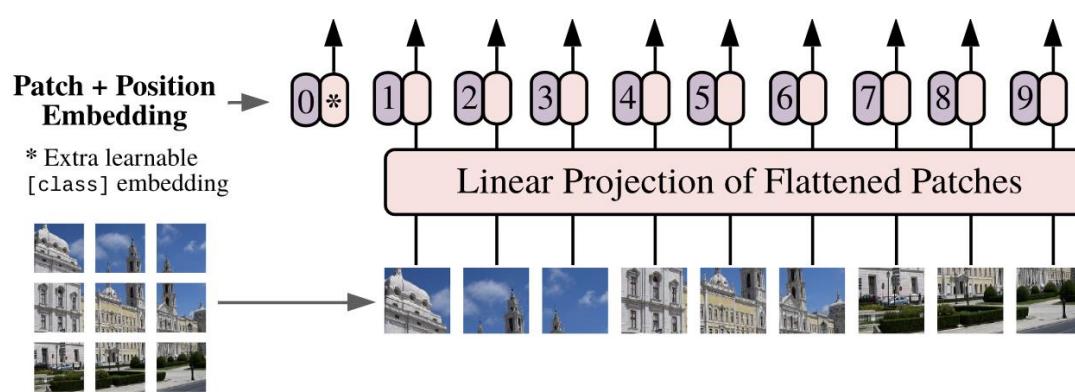
## Patch + Position Embedding

\* Extra learnable [class] embedding

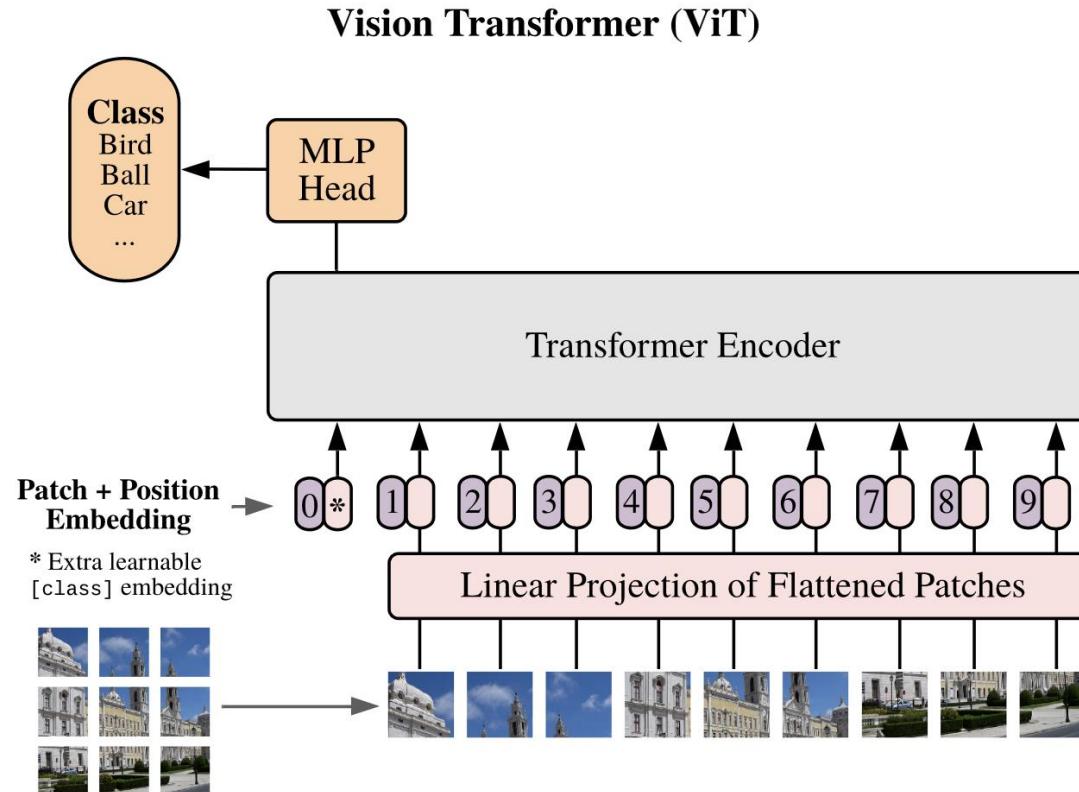


tokenize" the image by cutting it into patches of  $16px^2$ , and treating each patch as a token, e.g. embedding it into input space

# Vision Transformers



# Vision Transformers



## QUESTION

- TODO: how to train it

# What about paired image-text – How to Encode?



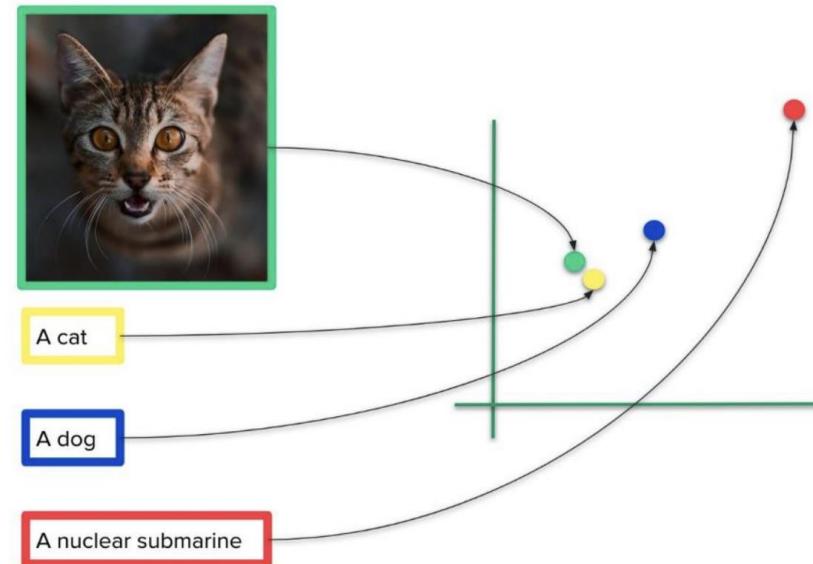
Basilica of St. John Lateran



House with Chimeras – Kiev

# What about paired image-text – How to Encode?

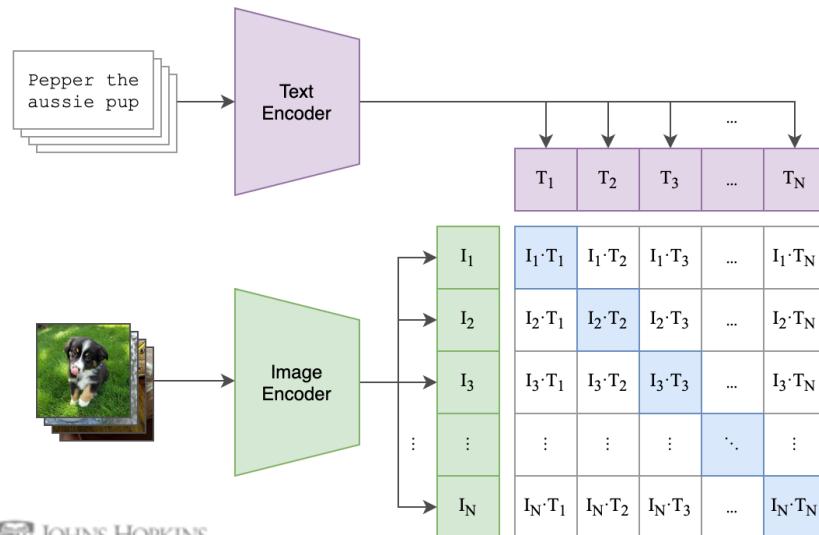
- The idea is to create a space to represent both semantics of language and image.
- Then, create a model that can align semantically-equivalent text and images nearby.



# Contrastive Language–Image Pre-training (CLIP)

- Training: simple **contrastive loss** between image-text pairs
  - Learning image representations from web-scale noisy text supervision

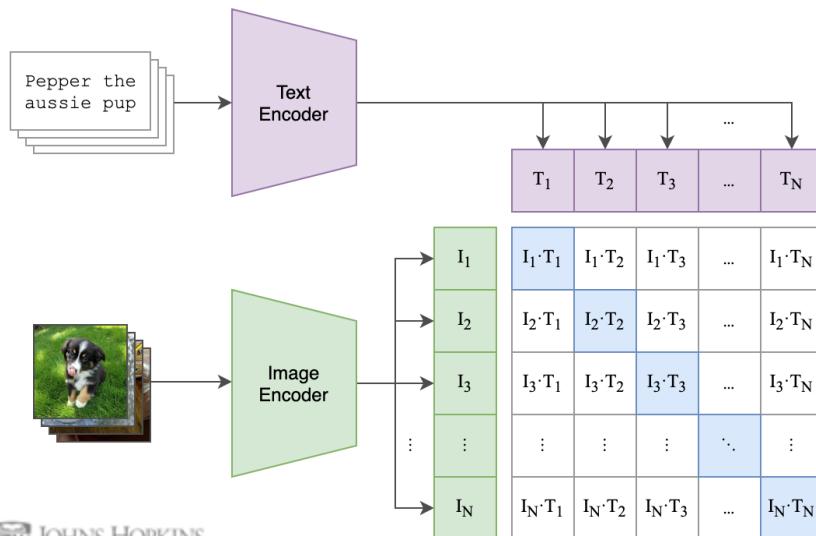
## (1) Contrastive pre-training



# Contrastive Language–Image Pre-training (CLIP)

- Given a batch of  $N$  (image, text) pairs, predict which of the  $N \times N$  possible (image, text) pairings across a batch occurred.

## (1) Contrastive pre-training



```
# image_encoder - ResNet or Vision Transformer
# text_encoder - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, 1]        - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t              - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T) # [n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

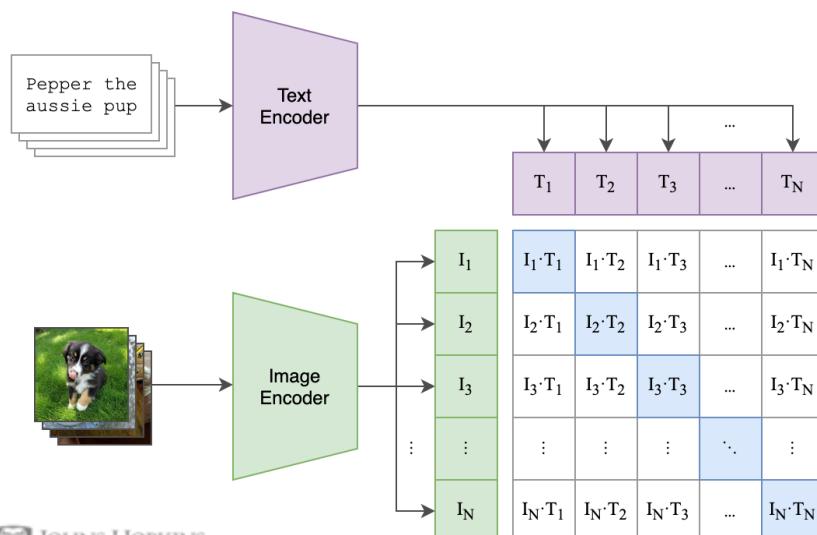
# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss = (loss_i + loss_t)/2
```

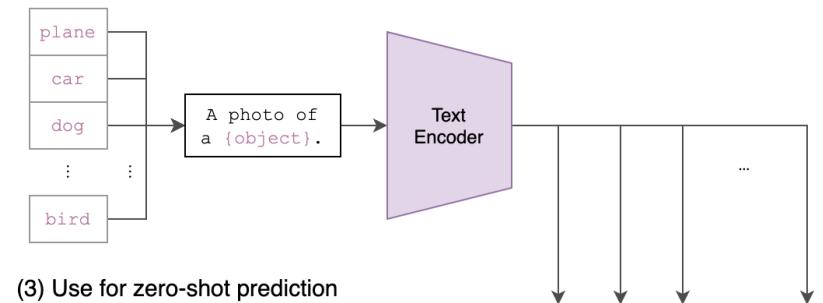
# What can CLIP do?

- Downstream: **zero-shot image** classification and image-text retrieval
  - Image classification can be reformatted as a retrieval task via considering the semantics behind label name.

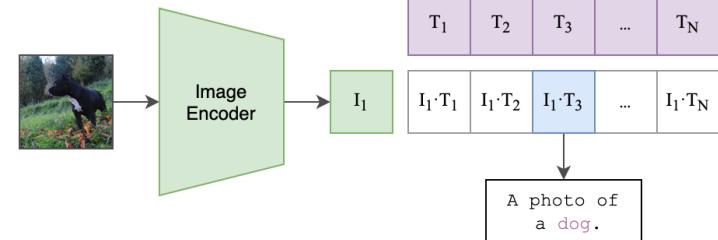
(1) Contrastive pre-training



(2) Create dataset classifier from label text



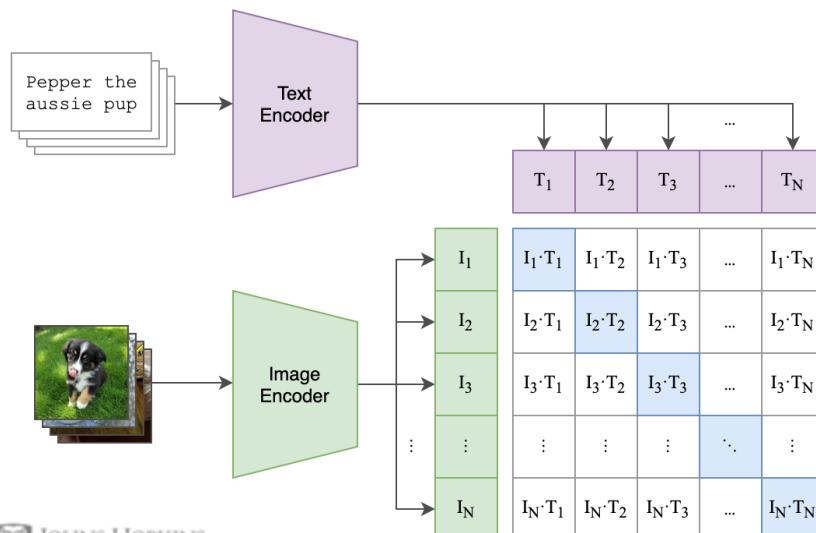
(3) Use for zero-shot prediction



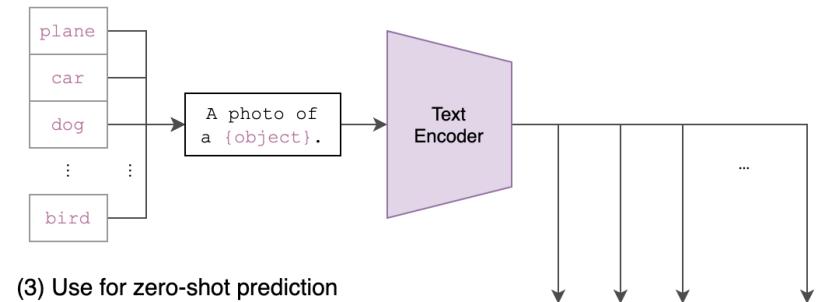
# What can CLIP do?

- Consider classifying photos of dogs vs cats - for each image, check if CLIP predicts text description "a photo of a dog" or "a photo of a cat" is more likely to be paired with it.

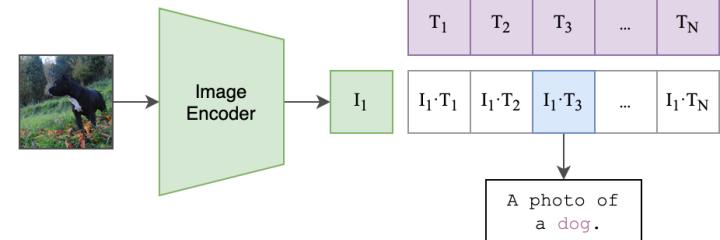
(1) Contrastive pre-training



(2) Create dataset classifier from label text



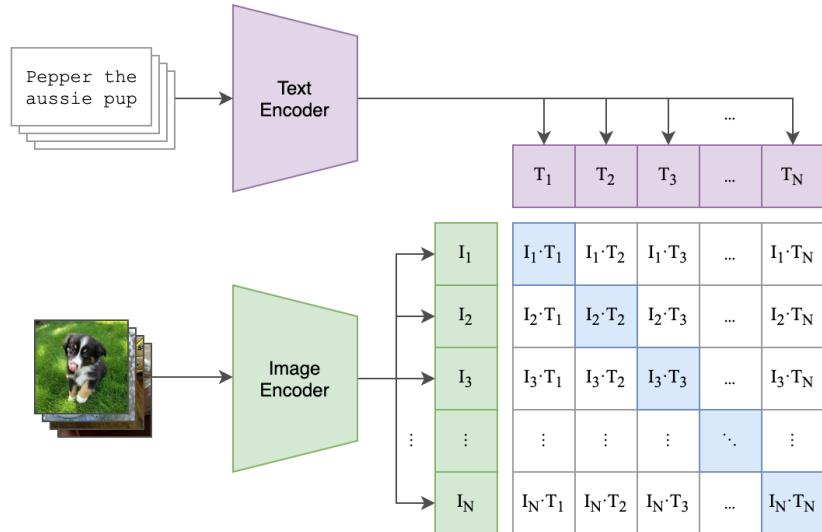
(3) Use for zero-shot prediction



# What can't CLIP do?

- No generation capabilities
- Prompting / In-Context Learning
  - Few-shot captioning

(1) Contrastive pre-training



# What can't CLIP do?

<https://huggingface.co/openai/clip-vit-large-patch14>

- No generation capabilities
- Prompting / In-Context Learning
  - Few-shot captioning
- CLIP is not perfect.



Possible class names (comma-separated)

too many fingers, not enough fingers, just the right amount of fingers

Compute

Computation time on cpu: 0.291 s

not enough fingers

0.461

just the right amount of fingers

0.339

too many fingers

0.200

</> JSON Output

Maximize

# What can't CLIP do?

<https://huggingface.co/openai/clip-vit-large-patch14>

- No generation capabilities
- Prompting / In-Context Learning
  - Few-shot captioning
- CLIP is not perfect.



Possible class names (comma-separated)

too many fingers, not enough fingers, just the right amount of fingers

Compute

Computation time on cpu: 0.251 s

just the right amount of fingers 0.412

too many fingers 0.316

not enough fingers 0.272

# What can't CLIP do?

- No generation capabilities
- Prompting / In-Context Learning
  - Few-shot captioning
- CLIP is not perfect.

<https://huggingface.co/openai/clip-vit-large-patch14>



Possible class names (comma-separated)

black and white cat, black and white dog, brown cat, brown dog

Compute

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.625 s

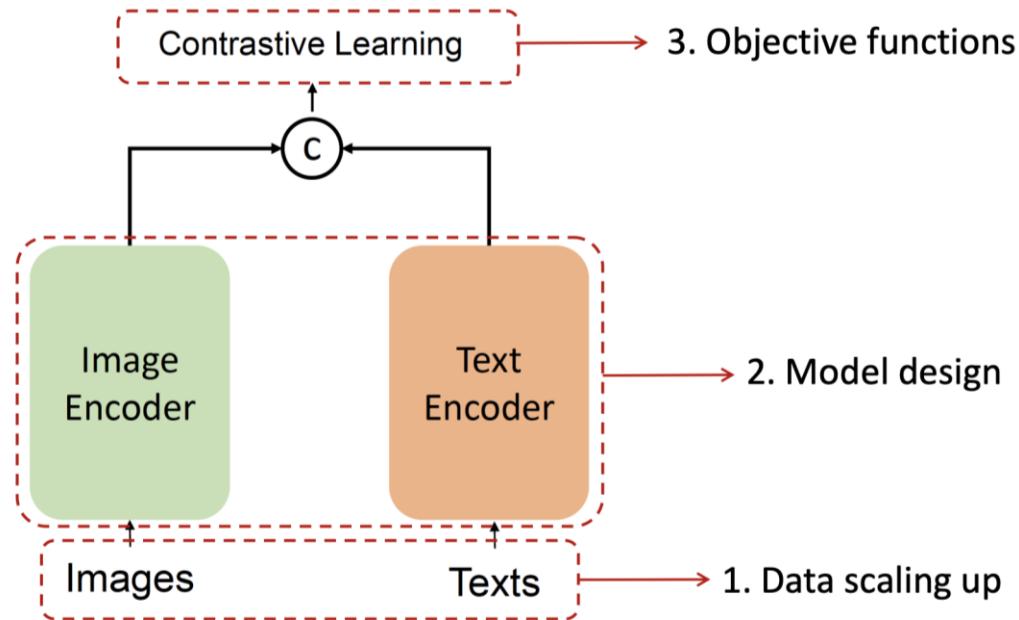
black and white dog	0.460
black and white cat	0.262
brown cat	0.181
brown dog	0.097

</> JSON Output

Maximize

# What happened after CLIP?

- Ton of follow-up work on improve its design:



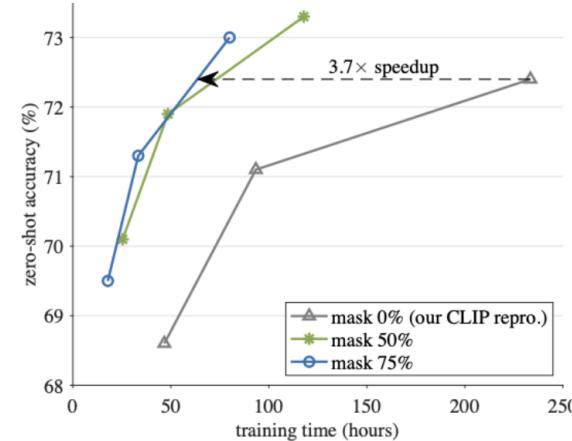
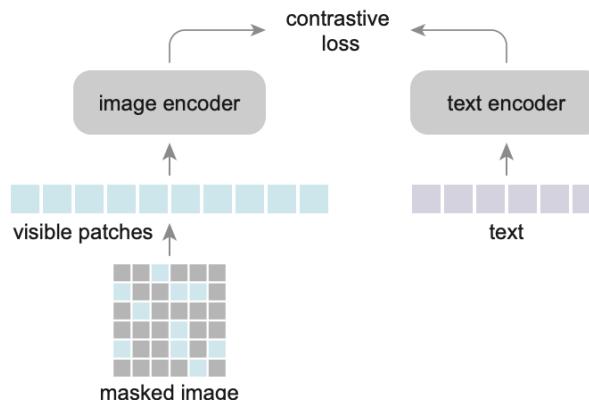
# What happened after CLIP?

- Open-source reproduction and scaling up
  - Open-source model: OpenCLIP
  - Pre-training on LAION-5B dataset

	Data	Arch.	ImageNet	VTAB+	COCO
CLIP [55]	WIT-400M	L/14	75.5	55.8	61.1
Ours	LAION-2B	L/14	75.2	54.6	71.1
Ours	LAION-2B	H/14	<u>78.0</u>	<u>56.4</u>	<u>73.4</u>

# What happened after CLIP?

- Open-source reproduction and scaling up
- Modifying the training process
  - A notable example is FLIP
    - Randomly masking out image patches
    - Does not hurt the performance, yet it improves the training efficiency



Scaling Language-Image Pre-training via Masking, 2022

# What happened after CLIP?

- Open-source reproduction and scaling up
- Modifying the training process
- Adding more modalities:
  - ImageBind: linking 7 modalities

<https://facebookresearch.github.io/ImageBind>

ImageBind: One Embedding Space To Bind Them All, 2023

1) Cross-Modal Retrieval

Audio	Images & Videos			Depth	Text
					"A fire crackles while a pan of food is frying on the fire." "Fire is crackling then wind starts blowing." "Firewood crackles then music..."
					"A baby is crying while a toddler is laughing." "A baby is laughing while an adult is laughing." "A baby laughs and something..."

2) Embedding-Space Arithmetic

Dog + Waves = Egret

3) Audio to Image Generation

Dog → Dog  
Engine → Fire truck  
Fire → Fire  
Rain → Rain

# What happened after CLIP?

---

- Open-source reproduction and scaling up
- Modifying the training process
- Adding more modalities:
- Generative models:
  - Text → Image
  - Image, Text → Image
  - Image, Image → Text
  - ...

Forthcoming

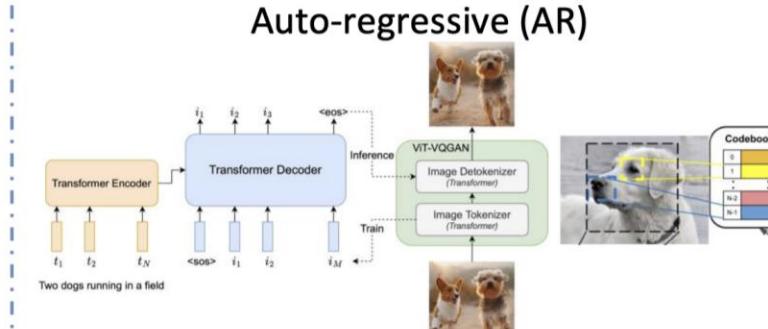
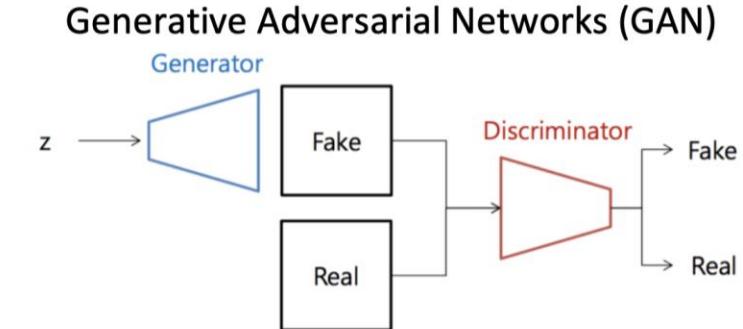
# Summary

---

- The same computational architecture (e.g., Transformer) can represent different modalities.
- One can build models that embed different modalities in the same space.
- Next: generative vision-language models.

# Generative Vision-Language Models

# Image Generation Toolkit



# Image Generation Toolkit: Diffusions

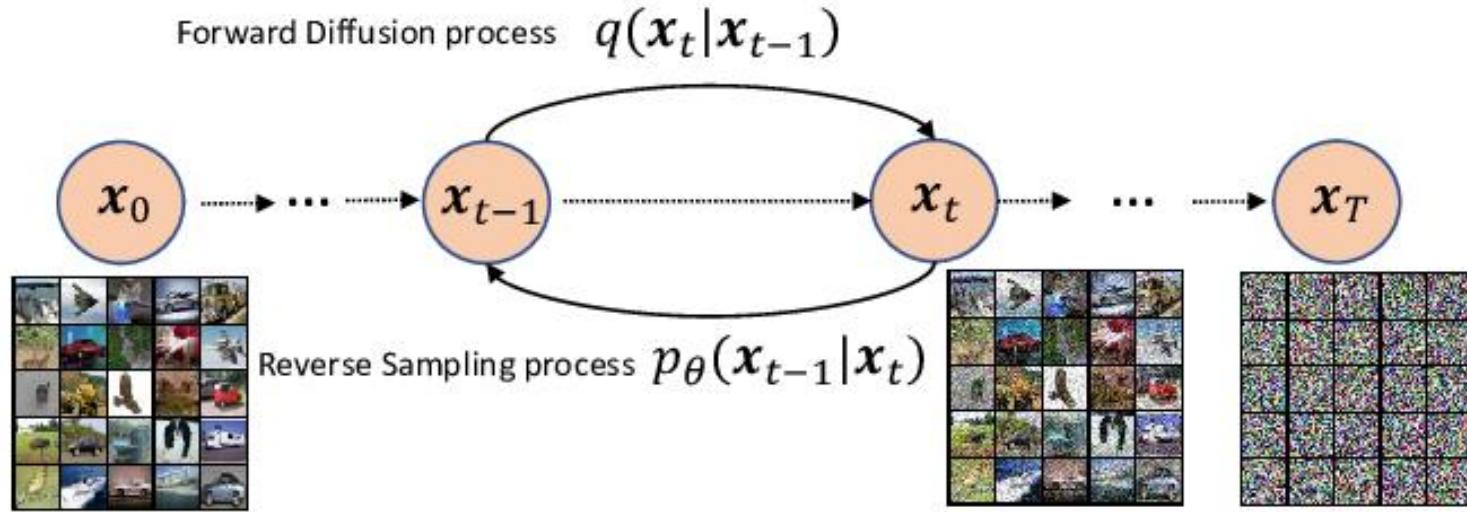
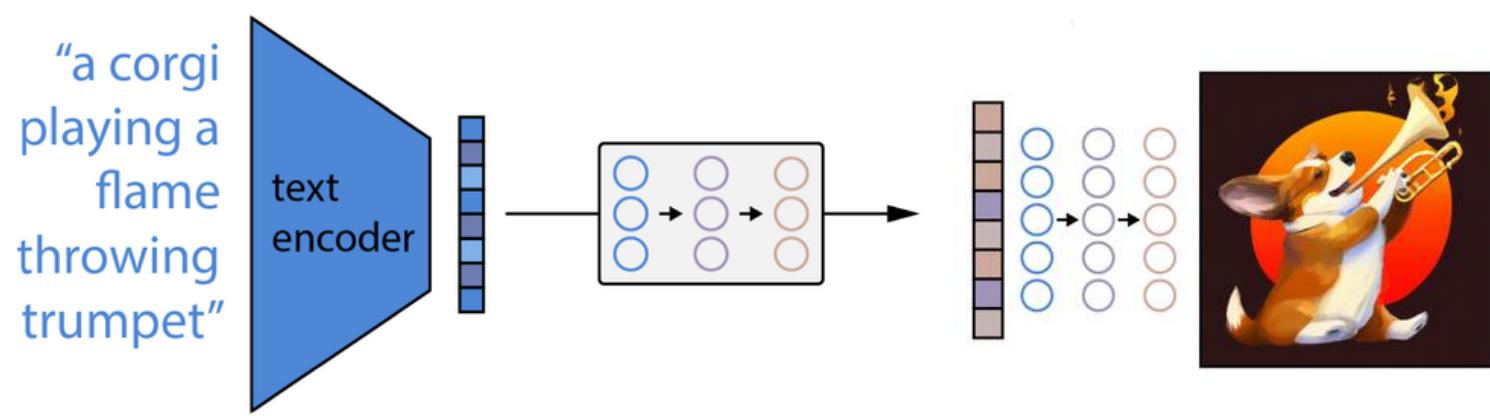


Image generation is out of scope for us. You can learn more by taking a computer vision class or watching the online tutorials. For example:

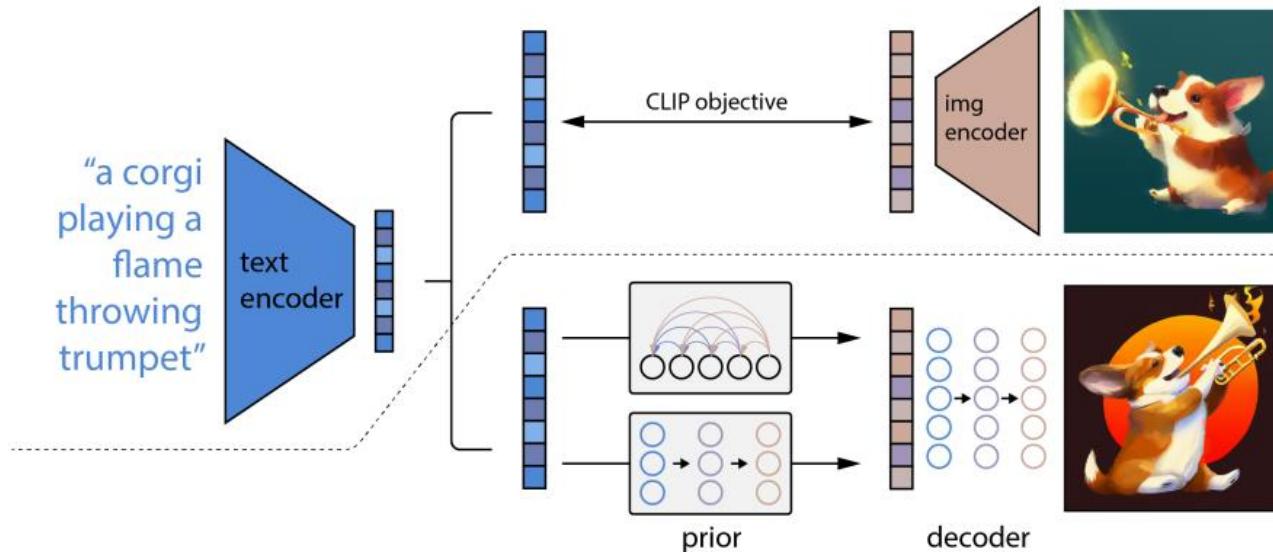
<https://cvpr2022-tutorial-diffusion-models.github.io/>



# Text to Image Generation

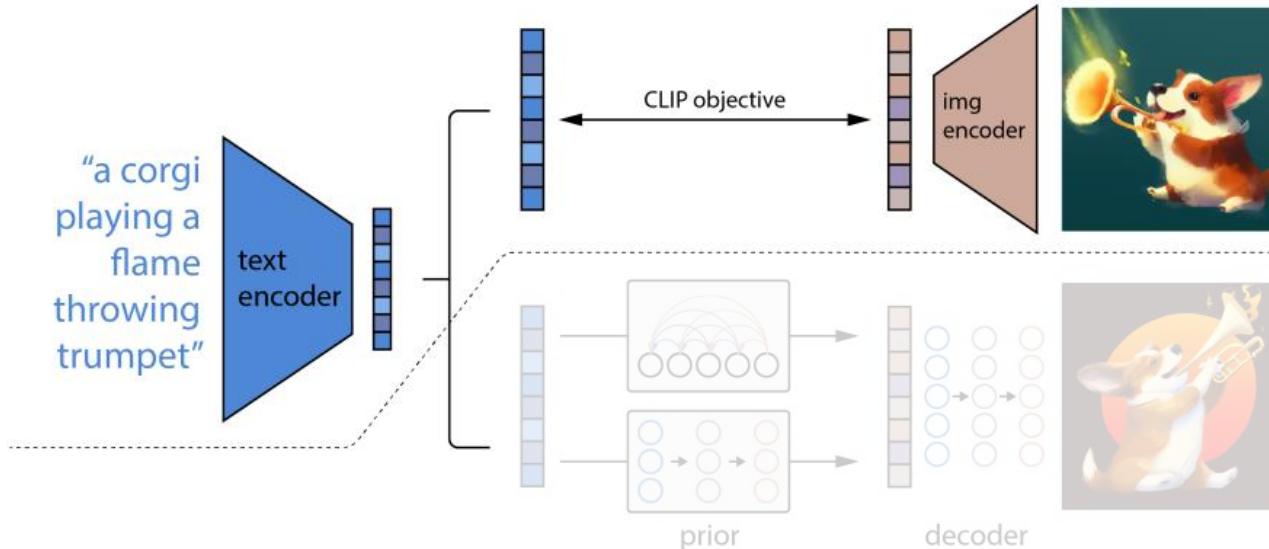


# DALL-E



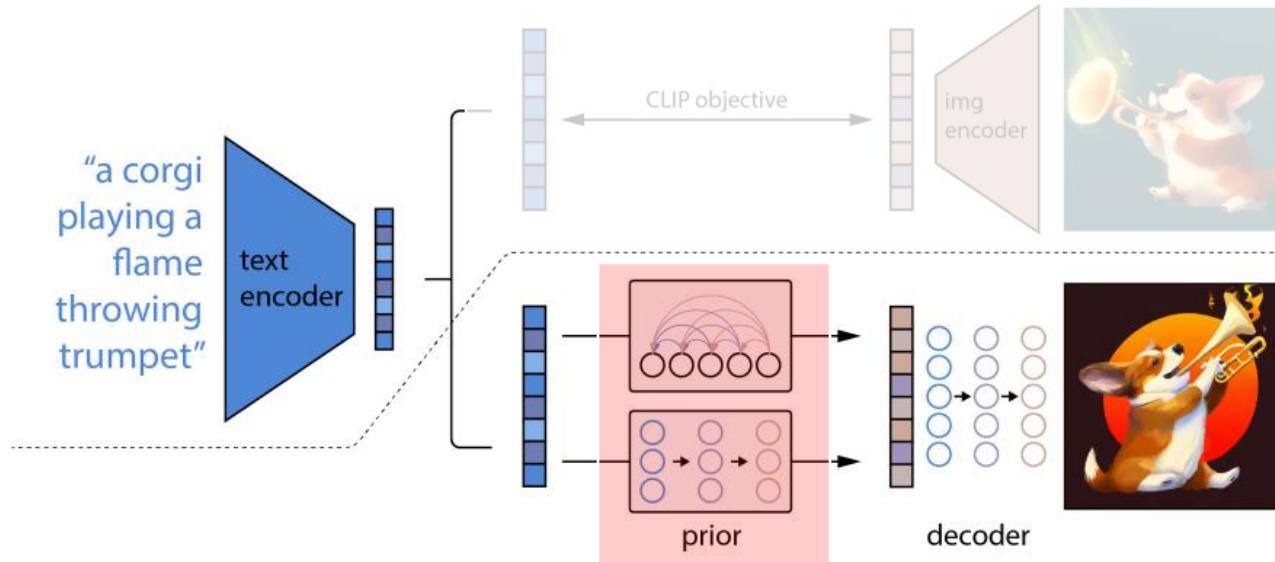
# DALL-E

- DALL-E is built on top of a pre-trained CLIP model.
  - This frozen model provides the representations of text and images.



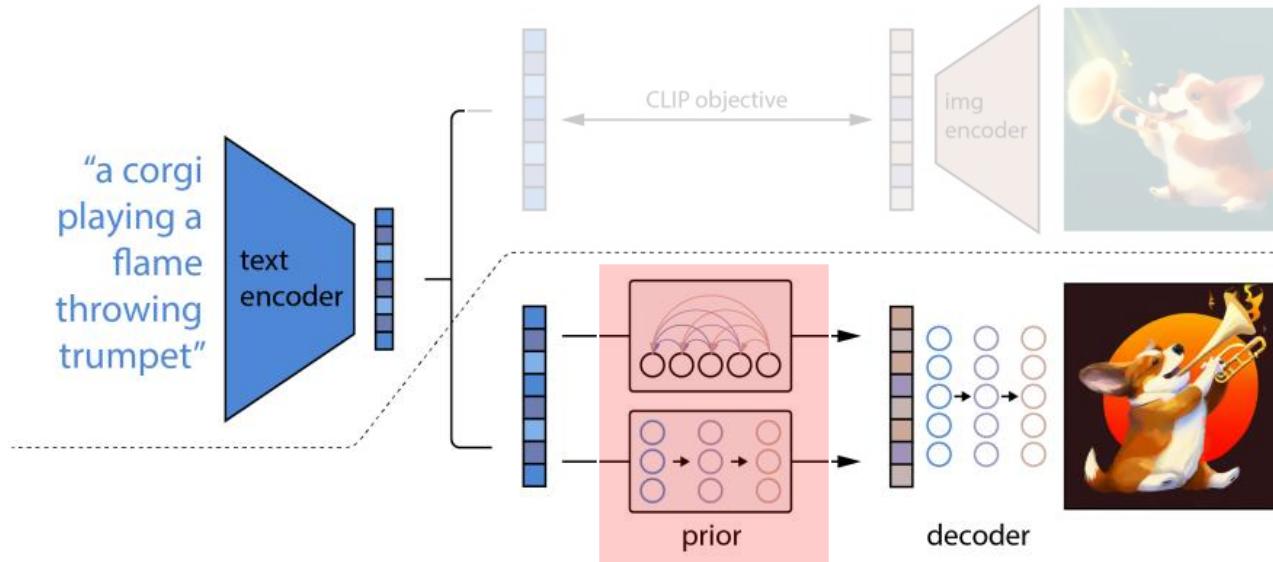
# DALL-E

- First, a text prompt is input into a **text encoder** that is trained to map the prompt to a representation space.



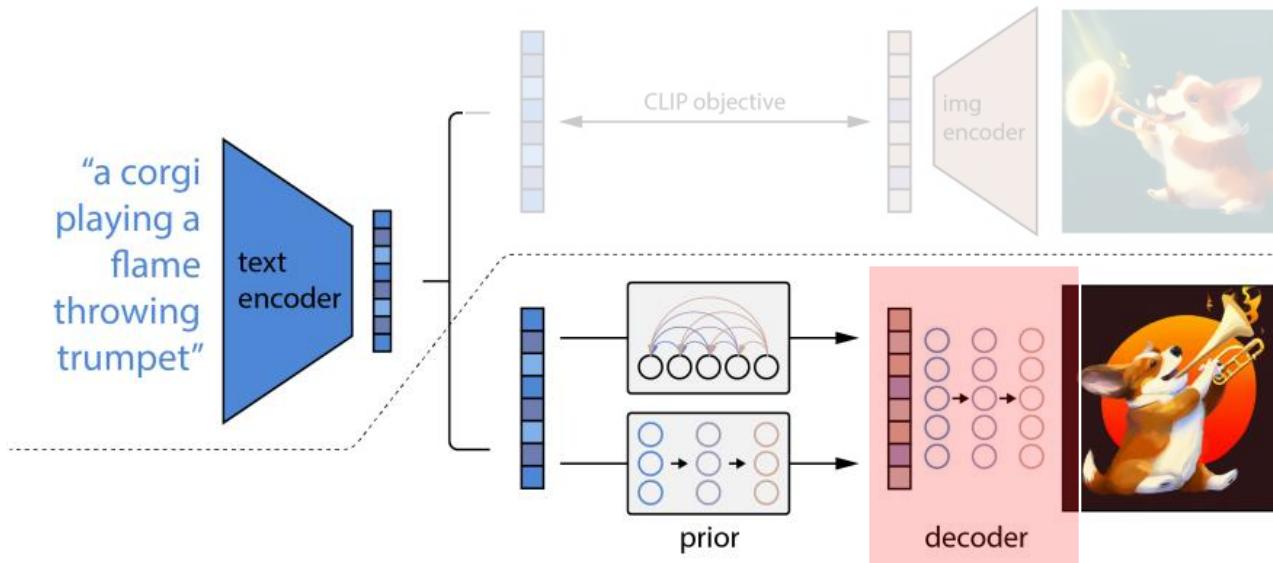
# DALL-E

- A **prior** maps the text encoding to a corresponding **image encoding** that captures the semantic information of the prompt contained in the text encoding.



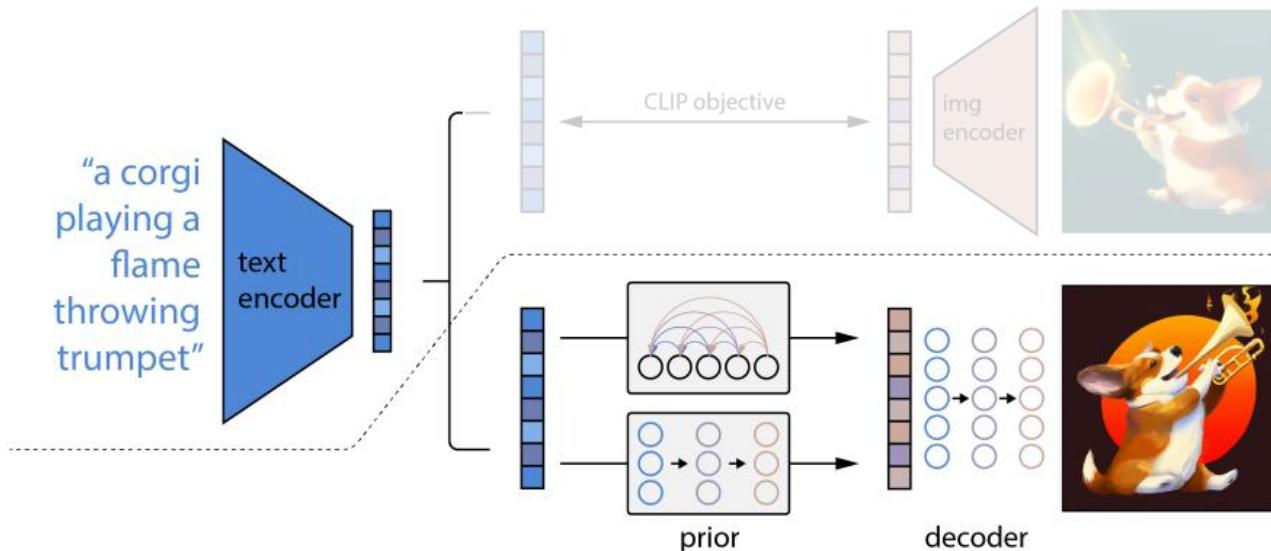
# DALL-E

- Finally, an **image decoder** stochastically generates an image which is a visual manifestation of this semantic information.



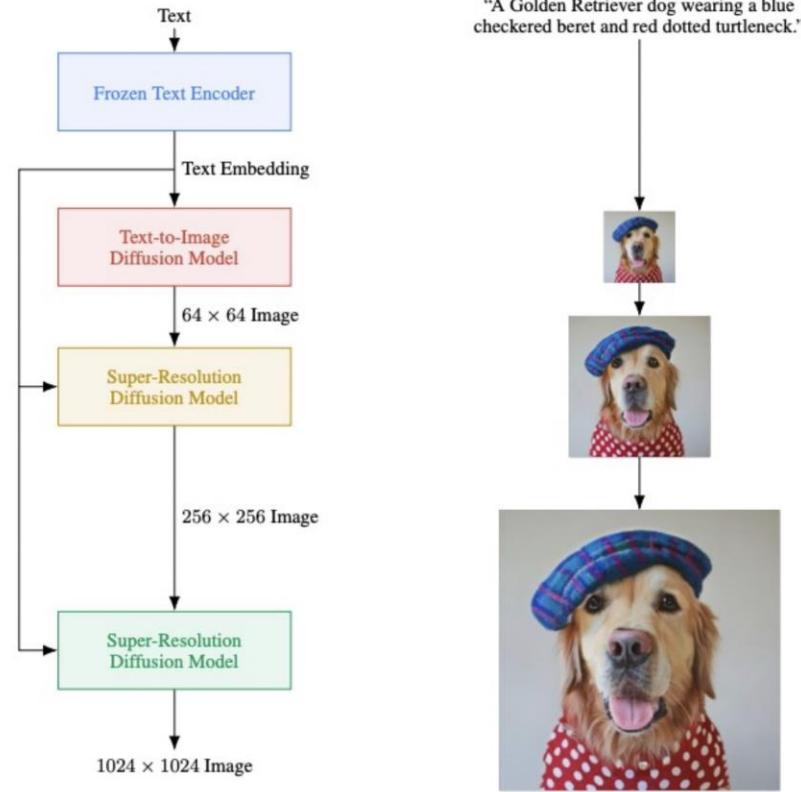
# DALL-E

- [Ignoring some details] Both modules are implemented using “diffusion models”.



# Imagen

- Simpler than DALL-E
- Key components:
  - Frozen language model providing text embeddings to all diffusion models.
    - Human raters prefer T5-XXL as the text encoder over CLIP encoder.
  - Cascaded diffusion models providing effective way to generate high-quality images.



# Imagen





Generate an image of a dancing elephant



Images

Videos

Free

Perspectives

App

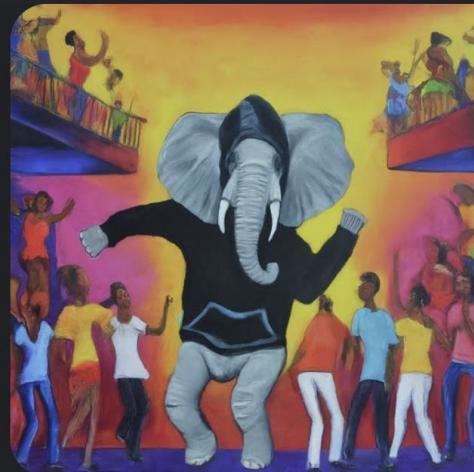
News

Shopping

Books

Maps

Generative AI is experimental. Learn more



*Generative imagery may display inaccurate, misleading or offensive images that do not represent Google's views*

[Images](#)[Perspectives](#)[Videos](#)[Shopping](#)[Books](#)[News](#)[Maps](#)[Flights](#)[Finance](#)

Generative AI is experimental. Learn more :



Generative AI is experimental. Learn more



# A ton of work on $T \rightarrow V$

- Text to video



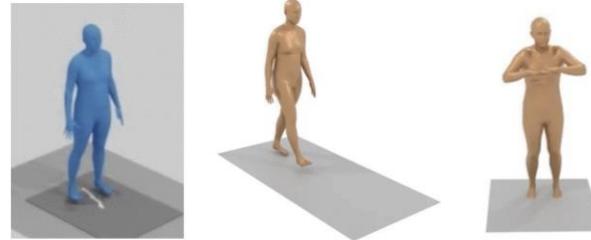
Text-to-Video generation: "a panda is playing guitar on times square"

- Text to 3D shapes

"A living room with a lit furnace, couch, and cozy curtains, bright lamps that make the room look well-lit."



- Text to motions/navigation





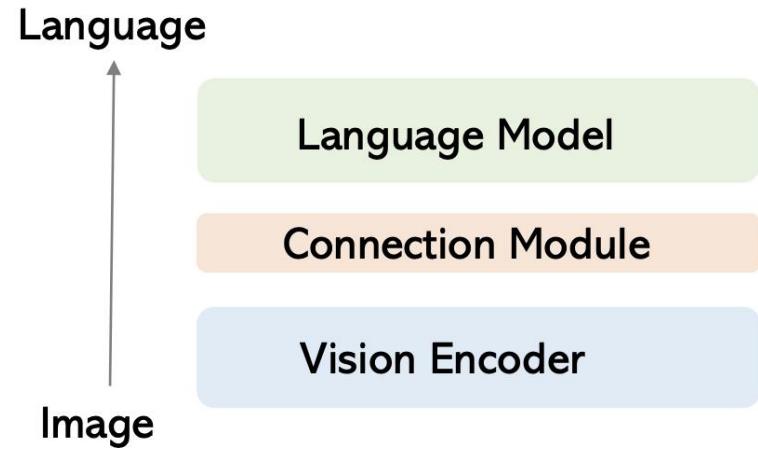
What about models that generate **text**?



# Image-to-Text Generative Models

- Model architectures:
  - Pre-trained image encoders
  - Pre-trained language models
  - Modules (to be trained) to connect the two modalities

A dog lying on the grass next to a frisbee





# LLaVA Model

Sometimes I just look at pictures of the earth from space and I marvel at how beautiful it all is.



User  
LLaVA

Can you explain this meme in detail?

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.

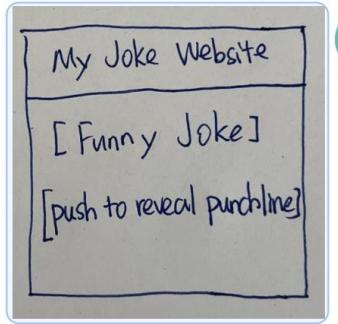
User  
GPT-4 [36]

Can you explain this meme?

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.

<https://arxiv.org/pdf/2304.08485.pdf>

# LLaVA Model



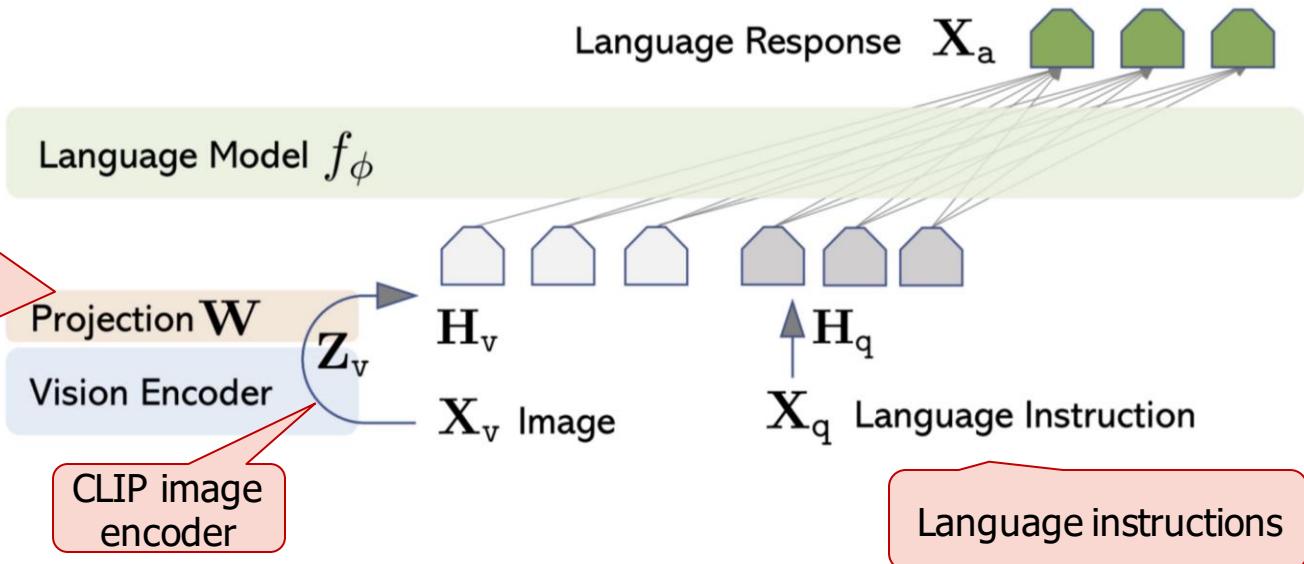
Turn this mock-up into an interactive website using html/js, replace the joke with a real joke.



```
<!DOCTYPE html>
<html>
<head>
<title>My Joke Website</title>
<style>
body { background-color: white; font-family: Arial, sans-serif; }
.container { display: flex; flex-direction: column; align-items: center; justify-content: center; height: 100vh; }
h1 { font-size: 36px; margin-bottom: 20px; }
p { font-size: 20px; }
button { background-color: blue; color: white; border: none; padding: 10px 20px; font-size: 18px; cursor: pointer; }
</style>
</head>
<body>
<div class="container">
<h1>My Joke Website</h1>
<p id="punchline">Welcome to my joke website. Push the button to reveal the punchline!</p>
<button onclick="revealPunchline()">Push me!</button>
</div>
<script>
function revealPunchline() {
    document.getElementById("punchline").innerHTML = "The punchline is: 'Why was the math book sad? Because it had too many problems.'";
}
</script>
</body>
</html>
```

# LLaVA Architecture

This projection matrix maps the embedded image to the same dimension as the LLM input (word embeddings).



# LLaVA: GPT4-assisted data generation

- **The raw data:** image captioning datasets
  - Consists of images, captions and bounding boxes of objects

## Context type 1: Captions

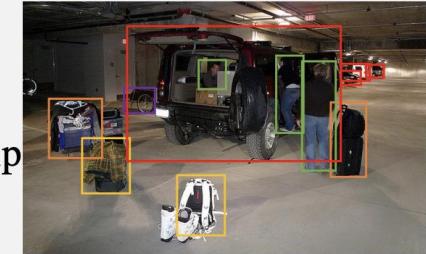
A group of people standing outside of a black vehicle with various luggage.  
Luggage surrounds a vehicle in an underground parking area

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

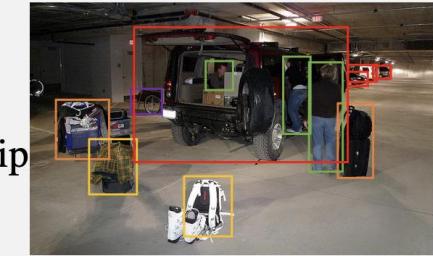
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>



- **Augmentation:** it uses GPT4 to augment these with detailed captions.

## 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], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

## Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

## 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. ...<omitted>

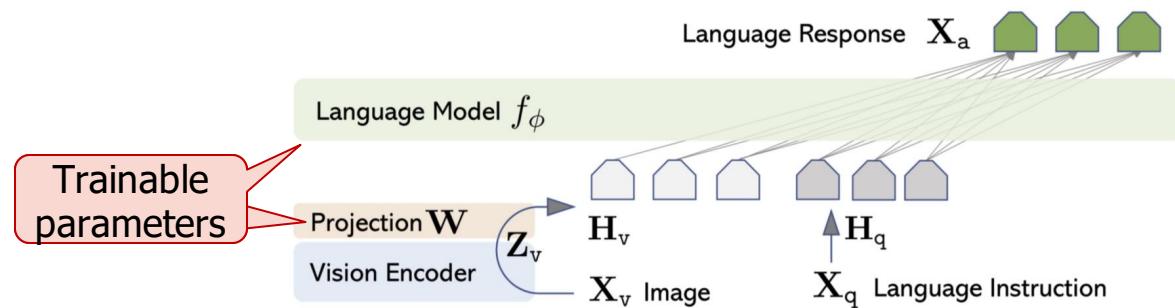
## Response type 3: complex reasoning

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 ...<omitted>

# LLaVA Architecture: Training

- **Step 1:** Feature alignment — aligning the representation of Vision Encoder and LLM
  - Both Vision Encoder and LLM are kept frozen.
  - The only training parameter is  $W$  (the projection matrix)
- **Step 2:** End-to-end fine-tuning
  - Vision Encoder is kept frozen. The training params are  $W$  and LLM.



<https://arxiv.org/pdf/2304.08485.pdf>

Visual Instruction Tuning, 2023

# Many open-source models ...

---



- BLIP/BLIP2 [Li et al. 2023]
- Instruct-BLIP [Dai et al. 2023]
- MiniGPT4 [Zhu et al. 2023]
- CoCa [Yu et al. 2022]
- Shikra [Chen et al. 2023]
- LLaVA 1.5 [Liu et al. 2023]
- ...

# Multi-modal GPT4

- Model: unknown
- Strong zero-shot visual understanding & reasoning capability
- How do we build opensource models that are as good as GPT4?

## GPT-4 visual input example, Extreme Ironing:

User      What is unusual about this image?

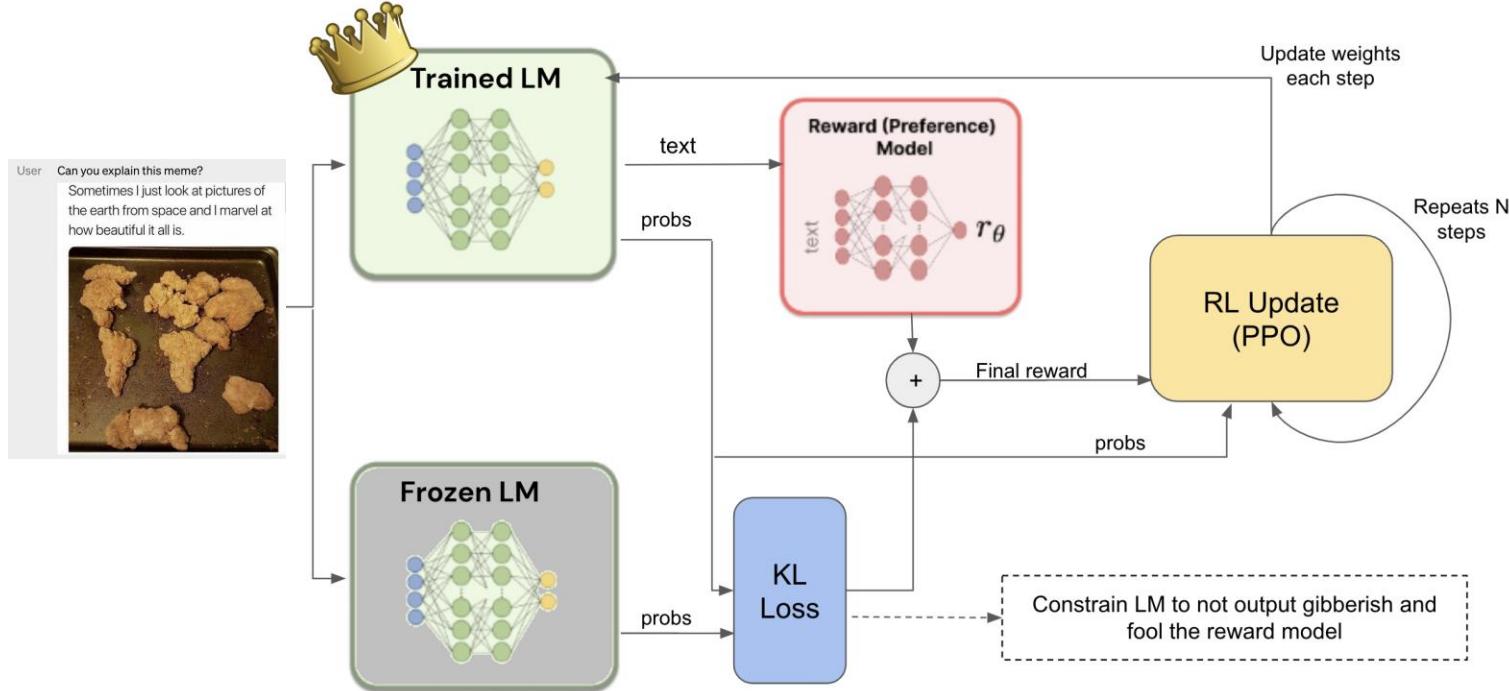


Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

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.

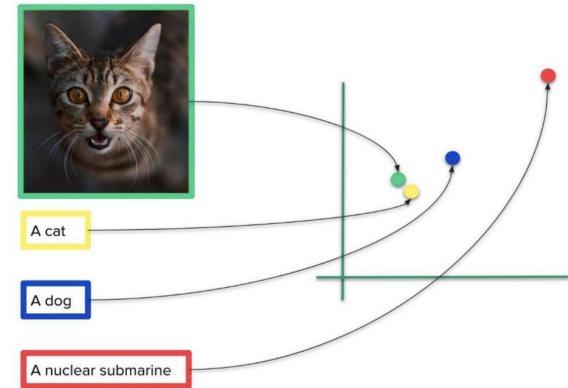
# Multimodal - RLHF



# Summary

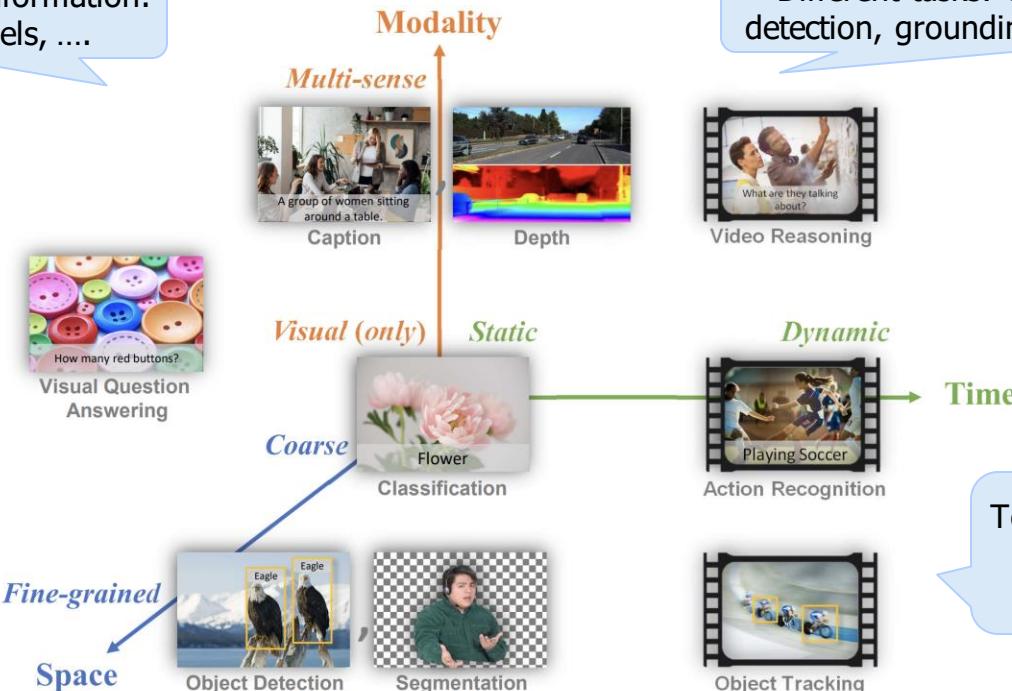
---

- There are massive amounts of unimodal data in the world.
  - We can build strong unimodal self-supervised models.
- Multi-modal models: the key challenge is **aligning** the representations.
- Many recent successes but a lot of challenges remain.
  - Making the models efficient.
  - Further scaling up — data scaling and model scaling
  - Better alignment of the representations — more natural/richer signals for alignment.



# Challenges for Future Multi-modal Models

Different modalities of information:  
Images, regions, pixels, ....

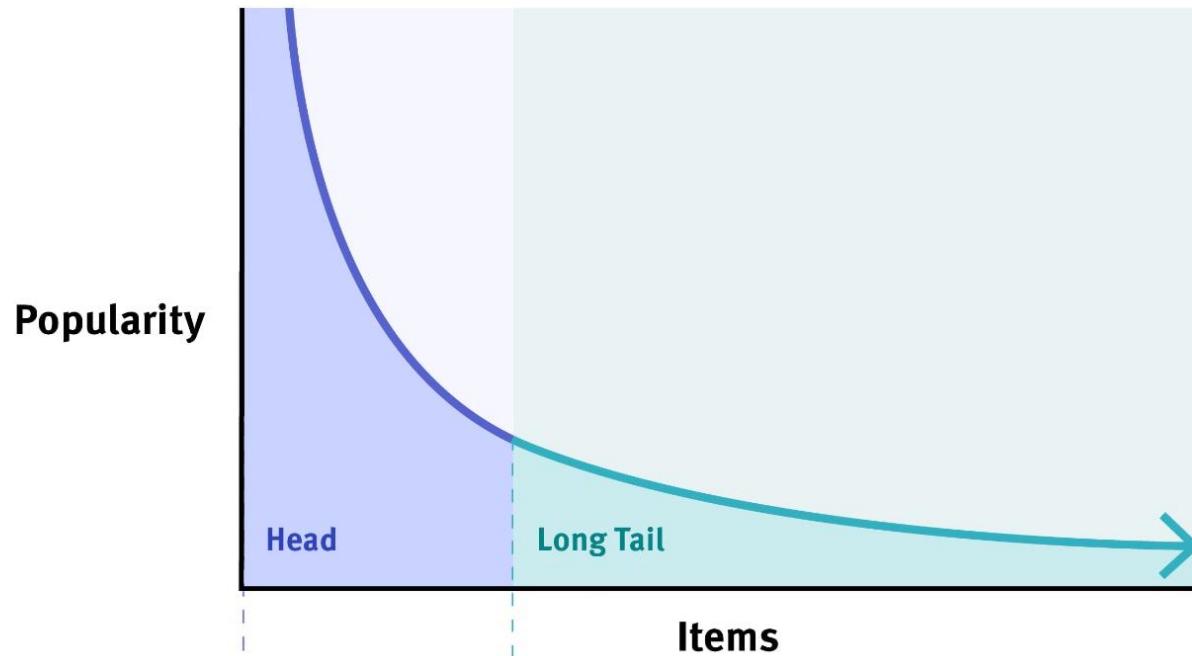


# Challenges for Future Multi-modal Models

---



# The long tail: *Most things are infrequent*





# JOHNS HOPKINS

## WHITING SCHOOL *of* ENGINEERING

# Other Modalities

## [Speech, Audio, ... TBD]

# Evolution of Modeling Paradigm

Task-specific Modeling

Training on small-scale,  
well-annotated data

*Instruction-following Models*

Pre-training on large-scale,  
noisy data

Generalist Modeling

Pre-training on XX..XLarge-scale,  
noisy data

*Instruction tuning* on small-scale,  
pseudo-labeling data

NLP: Chat-GPT, Alpaca, Vicuna, ...  
VL: LLaVa, MiniGPT4, Otter, ...

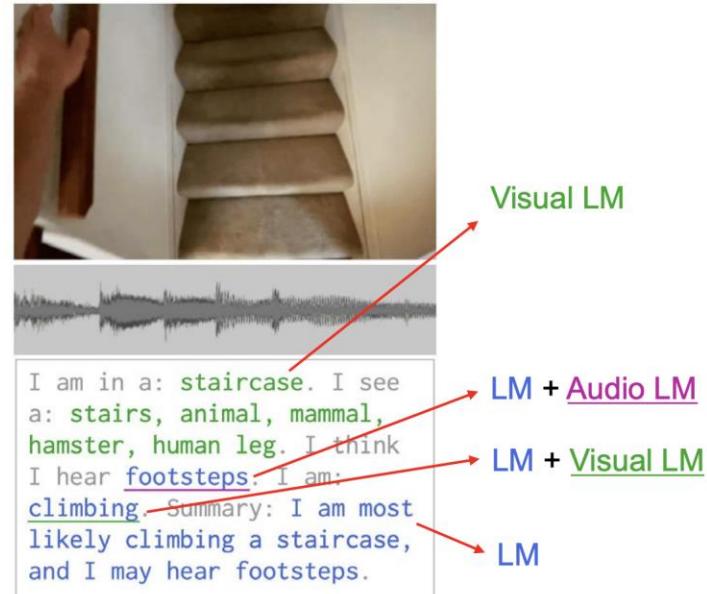
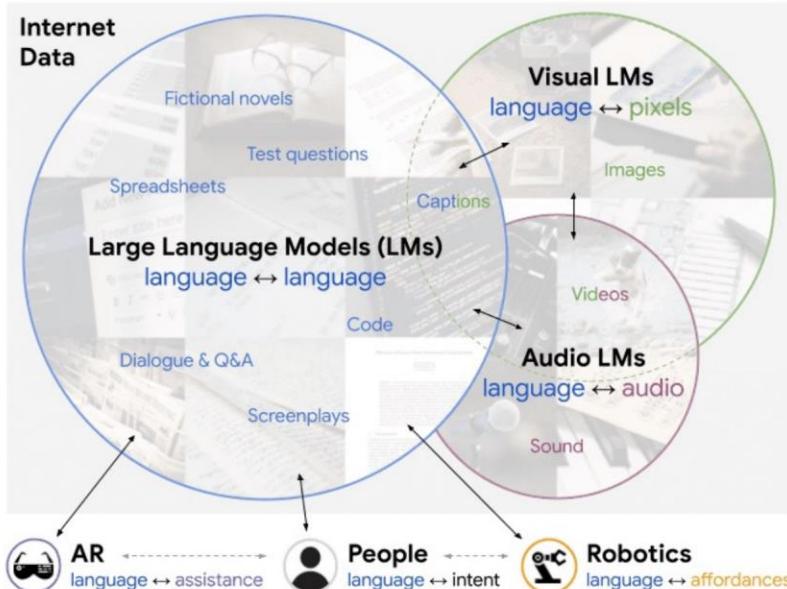
Zero-shot or In-context Few-shot with A few  
examples as demonstration

LLMs: GPT3, PaLM, LLaMa, ...  
LMMs: Flamingo, PaLM-E, GPT-4, ...

# Socratic: Composing Multi-modality w/ LLM



- A modular framework in which multiple pretrained models may be composed zero-shot through language without training.



Summarize ego-centric videos.

# Socratic: Composing Multi-modality w/ LLM



- The model works well on **vision-language tasks** such as image captioning, it can also **parse & generate robot instructions** from free form human language.

Human defines when to use which model.

Method	BLEU-4	METEOR	CIDEr	SPICE	ROUGE-L
*ClipCap [45]	40.7	30.4	152.4	25.2	60.9
†MAGIC [61]	11.4	16.4	56.2	11.3	39.0
ZeroCap [62]	0.0	8.8	18.0	5.6	18.3
SMs 0-shot (ours)	6.9	15.0	44.5	10.1	34.1
SMs 3-shot (ours)	<b>18.3</b>	<b>18.8</b>	<b>76.3</b>	<b>14.8</b>	<b>43.7</b>

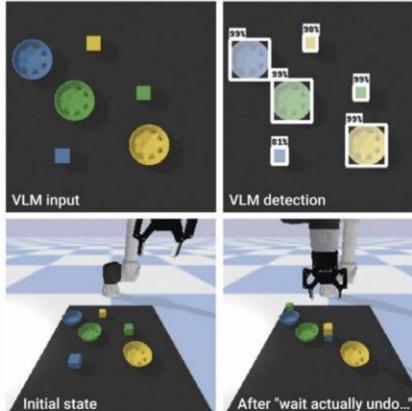
COCO  
Captions

\* finetuned on full training set with image-text pairs.

† finetuned on unpaired training set, zero-shot on image-text pairs.

Visual LM  
User Instruction  
LM

```
objects = ["green block", "blue block", "yellow block", "green bowl", "blue bowl", "yellow bowl"]
# move all the blocks to different corners.
Step 1. robot.pick_and_place("green block", "top left corner")
Step 2. robot.pick_and_place("blue block", "top right corner")
Step 3. robot.pick_and_place("yellow block", "bottom left corner")
# now move the blue block to the middle.
Step 1. robot.pick_and_place("blue block", "middle")
# stack the blocks on top of each other.
Step 1. robot.pick_and_place("yellow block", "blue block")
Step 2. robot.pick_and_place("green block", "yellow block")
# wait actually undo that last step.
Step 1. robot.pick_and_place("green block", "top left corner")
# put the yellow block in the bowl you think it best fits.
Step 1. robot.pick_and_place("yellow block", "yellow bowl")
# ok now sort the remaining blocks in the same way.
Step 1. robot.pick_and_place("blue block", "blue bowl")
Step 2. robot.pick_and_place("green block", "green bowl")
```

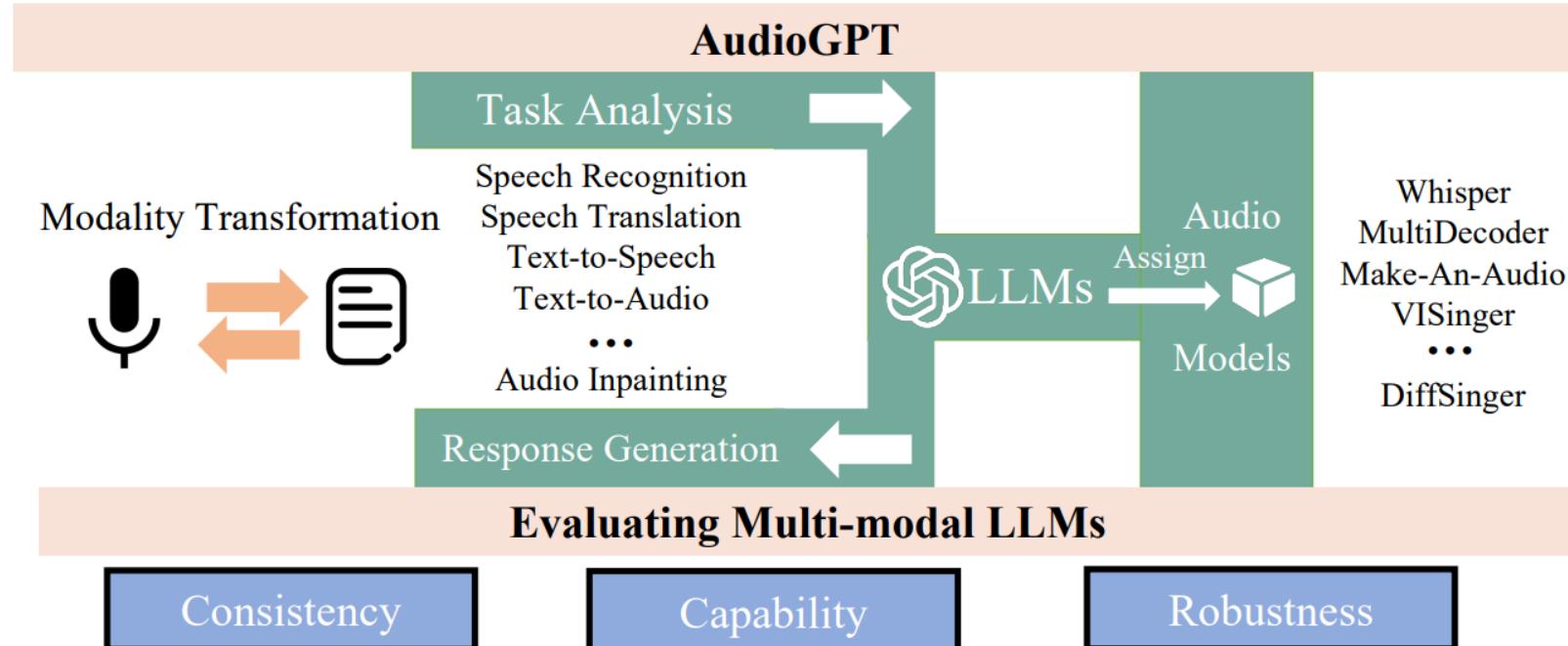


344

Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, Zeng et al, ICLR 2023

# Applications : Audio Understanding & Generation

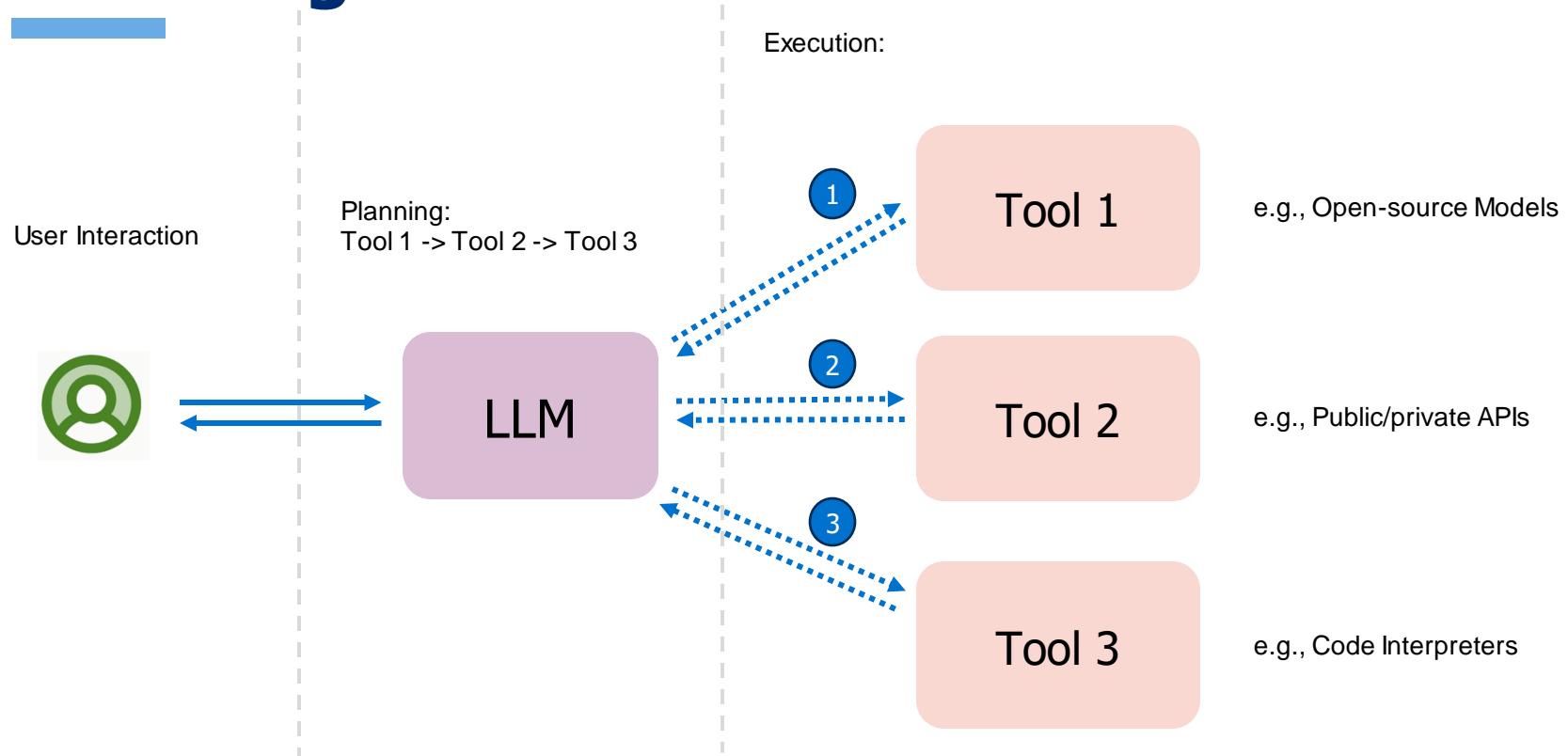
AudioGPT: Understanding and Generating Speech, Music, Sound, and Talking Head





# Language Models as Agents

# Tool-Augmented LLMs





# Enabling LLMs to do tool-use



# MM-ReAct: Capabilities

## Visual Math and Text Reasoning

What is "a" in the first equation?  
What is y in the last equation?

$5 + a = 7$	$40 + y = 45$
$10 + y = 15$	$39 + a = 41$
$20 + a = 22$	$24 + y = 29$
$48 + a = 50$	$15 + a = 17$
$10 + a = 12$	$34 + y = 39$

The value of "a" in the first equation " $5 + a = 7$ " is 2.

The value of "y" in the last equation " $34 + y = 39$ " is 5.

What is funny about it?



The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.

## Spatial/Coordinate Understanding

Where is the frisbee?  
What is the relative position of the frisbee and the man?



The frisbee is located at coordinates (63,119,133,177).

The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man..

## Visual Planning and Prediction

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

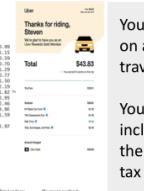


After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

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

## Multi-Image Reasoning

How much in total did I spend on travel?  
How much in total did I pay for taxes?



You spent \$1299.90 on airfare and \$43.83 on an Uber ride for a total of \$1343.73 on travel.

You paid \$323.23 in total for taxes. This includes \$321.90 for taxes and fees for the flight purchase and \$1.33 for sales tax for the Windmill Family Restaurant.

## Multi-Hop Document Understanding

Excluding the regular cab, what is the average total fueling cost?



The average total fueling cost excluding the Ford F150 (Regular Cab) is \$76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. ( $\$82.56 + \$70.55\right) / 2 = \$76.55$ .

## Open-World Concept Understanding

What is the brand of these drink? Which one is unsweetened?



The brand of these drinks is "BETTER THAN MILK". The unsweetened one is "BETTER THAN MILK organic OAT DRINK".

Who are the celebrities in this image?



Keira Knightley and Natalie Portman are the celebrities in this image.

Is it safe to eat this mushroom?



Morel mushrooms are generally considered safe to eat when properly cooked.

## Video Summarization/Event Localization

Can you breakdown this video tutorial to step-by-step instructions, with start and end time?



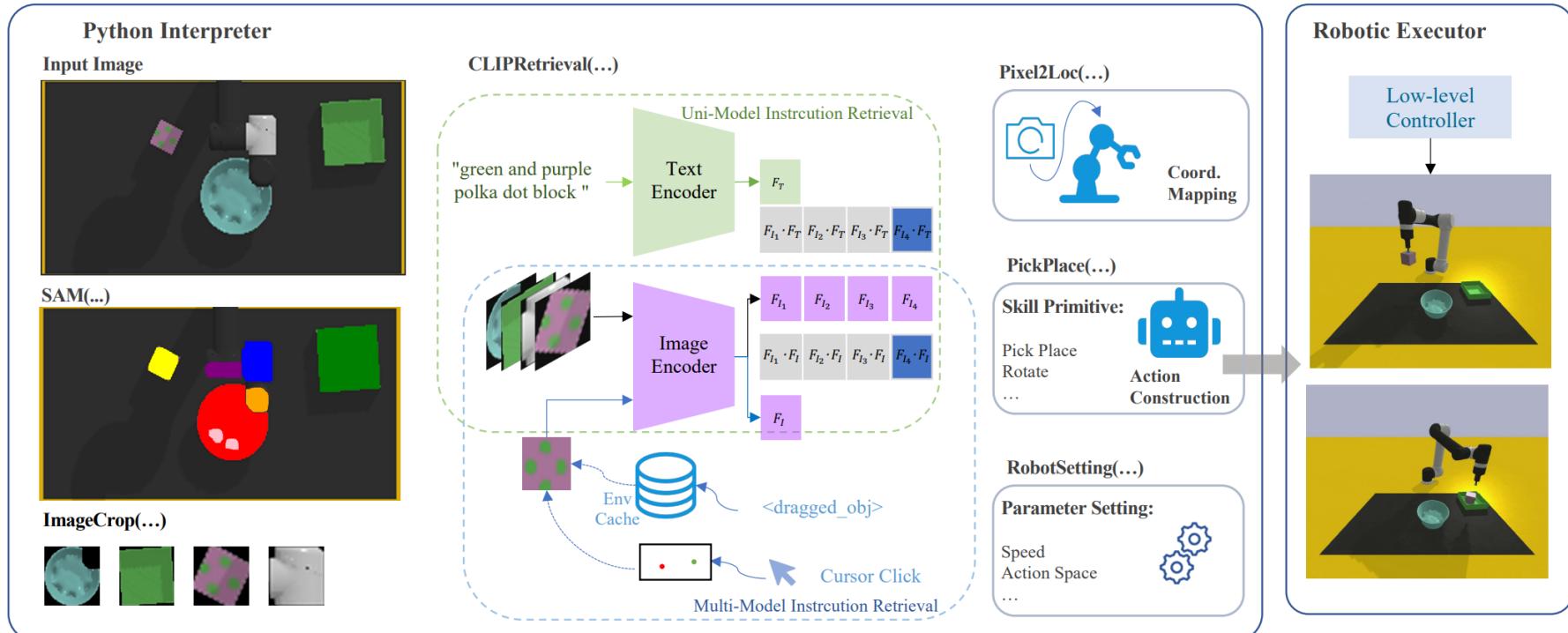
Here's a step-by-step breakdown of the tutorial with start and end time.



... Add Lettuce    Add Tomato    Add Bacon ...

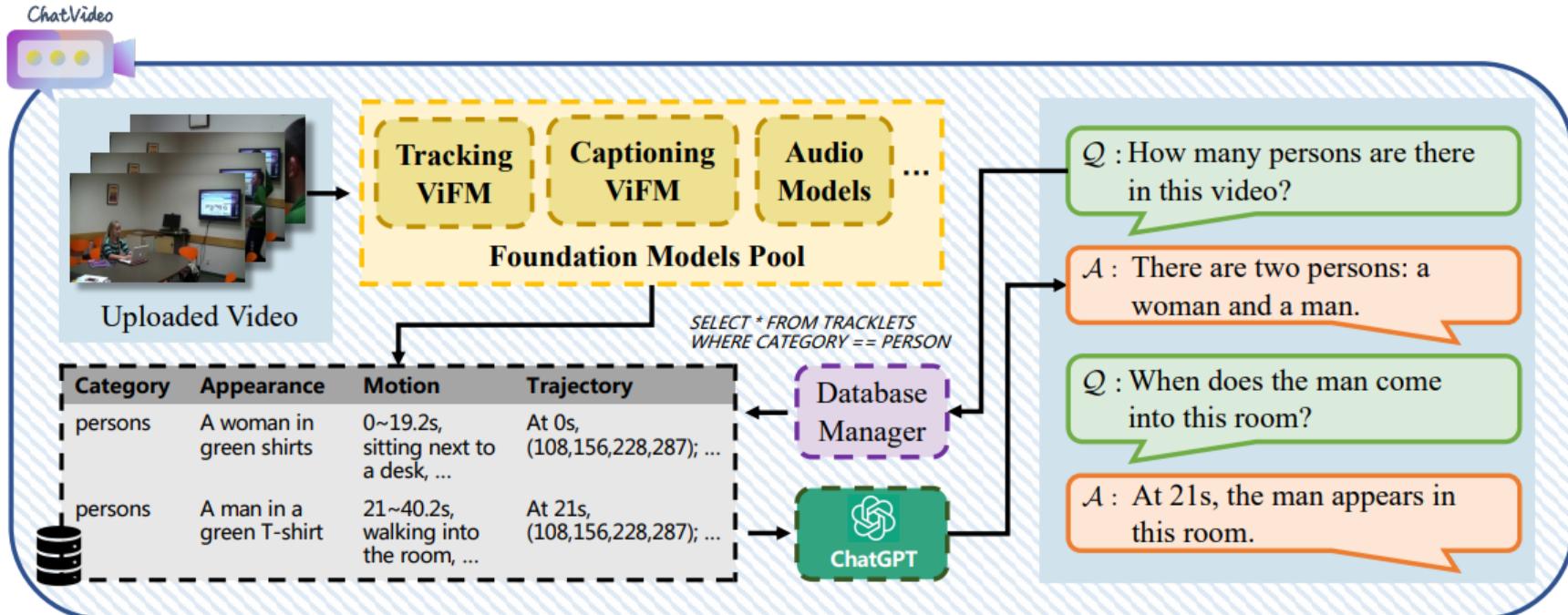
# Applications : Robotics

Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions with Large Language Model



# Applications: Multi-Channel Video Understanding

ChatVideo: A Tracklet-centric Multimodal and Versatile Video Understanding System



# What is embodied AI?



Aiming at the creation of an **embodied agent** (e.g., a robot) which learns, through **interaction and exploration**, to creatively solve challenging tasks within its environment.



- See: perceive their environment through vision or other senses.
- Talk: hold a natural language dialog grounded in their environment.
- Listen: understand and react to audio input anywhere in a scene.
- Act: navigate and interact with their environment to accomplish goals.
- Reason: consider and plan for the long-term consequences of their actions.

# SayCan: Grounding Language in Affordances



We need to ground the language model in tasks that are feasible within a specific real-world context.



# SayCan: Grounding Language in Affordances



Instruction Relevance with LLMs

How would you put  
an apple on the  
table?

I would: 1. \_\_\_\_\_

LLM

Combined

-6	Find an apple	0.6
-30	Find a coke	0.6
-30	Find a sponge	0.6
-4	Pick up the apple	0.2
-30	Pick up the coke	0.2
...	...	...
-5	Place the apple	0.1
-30	Place the coke	0.1
-10	Go to the table	0.8
-20	Go to the counter	0.8

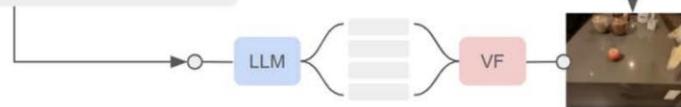
Task Affordances with Value Functions



Value  
Functions

Note that the LM  
scores options  
instead of doing  
generation

I would: 1. **Find an apple**, 2. \_\_\_\_\_



# Combining SayCan with Chain of Thought Prompting



Chain-of-thought prompting can improve examples with **negation**.

---

#### CHAIN-OF-THOUGHT PLANNING ROLLOUT WITH PALM-SAYCAN.

**Human:** Can you bring a fruit-flavored drink without caffeine?

**Explanation:** The user has asked for a drink that is fruit-flavored and does not have caffeine, I will bring the lime soda.

**Robot:** 1. find a lime soda, 2. pick up the lime soda, 3. bring it to you, 4. put down the lime soda, 5. done

**Human:** Bring me the spicy chips.

**Explanation:** The user has asked for chips that are spicy, I will bring the Jalapeno chips.

**Robot:** 1. find the jalapeno chips, 2. pick up the jalapeno chips, 3. bring it to you, 4. put down the jalapeno chips, 5.done

**Human:** The snack is too small, bring me something more filling.

**Explanation:** The user has asked for a snack that is more filling, I will bring the multigrain chips.

**Robot:** 1. find the multigrain chips, 2. pick up the multigrain chips, 3. bring it to you, 4. put down the multigrain chips, 5. done

---

Table 4: Chain-of-thought planning rollout with PaLM-SayCan. The highlighted part is the chain of thought generated by PaLM-SayCan.

# PaLM-E: An Embodied Multimodal Language Model



Encoding embodied observations as language tokens.

## Mobile Manipulation



Human: Bring me the rice chips from the drawer. Robot: 1. Go to the drawers, 2. Open top drawer. I see 3. Pick the green rice chip bag from the drawer and place it on the counter.

## Visual Q&A, Captioning ...



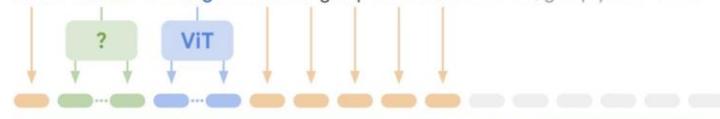
Given Q: What's in the image? Answer in emojis.  
A: 🍎🍊🍇🍏🍐🍒.



Describe the following   
A dog jumping over a hurdle at a dog show.

## PaLM-E: An Embodied Multimodal Language Model

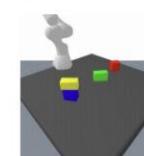
Given **<emb>** ... **<img>** Q: How to grasp blue block? A: First, grasp yellow block



## Control

A: First, grasp yellow block and ...

## Task and Motion Planning



Given **<emb>** Q: How to grasp blue block?  
A: First grasp yellow block and place it on the table, then grasp the blue block.

## Tabletop Manipulation

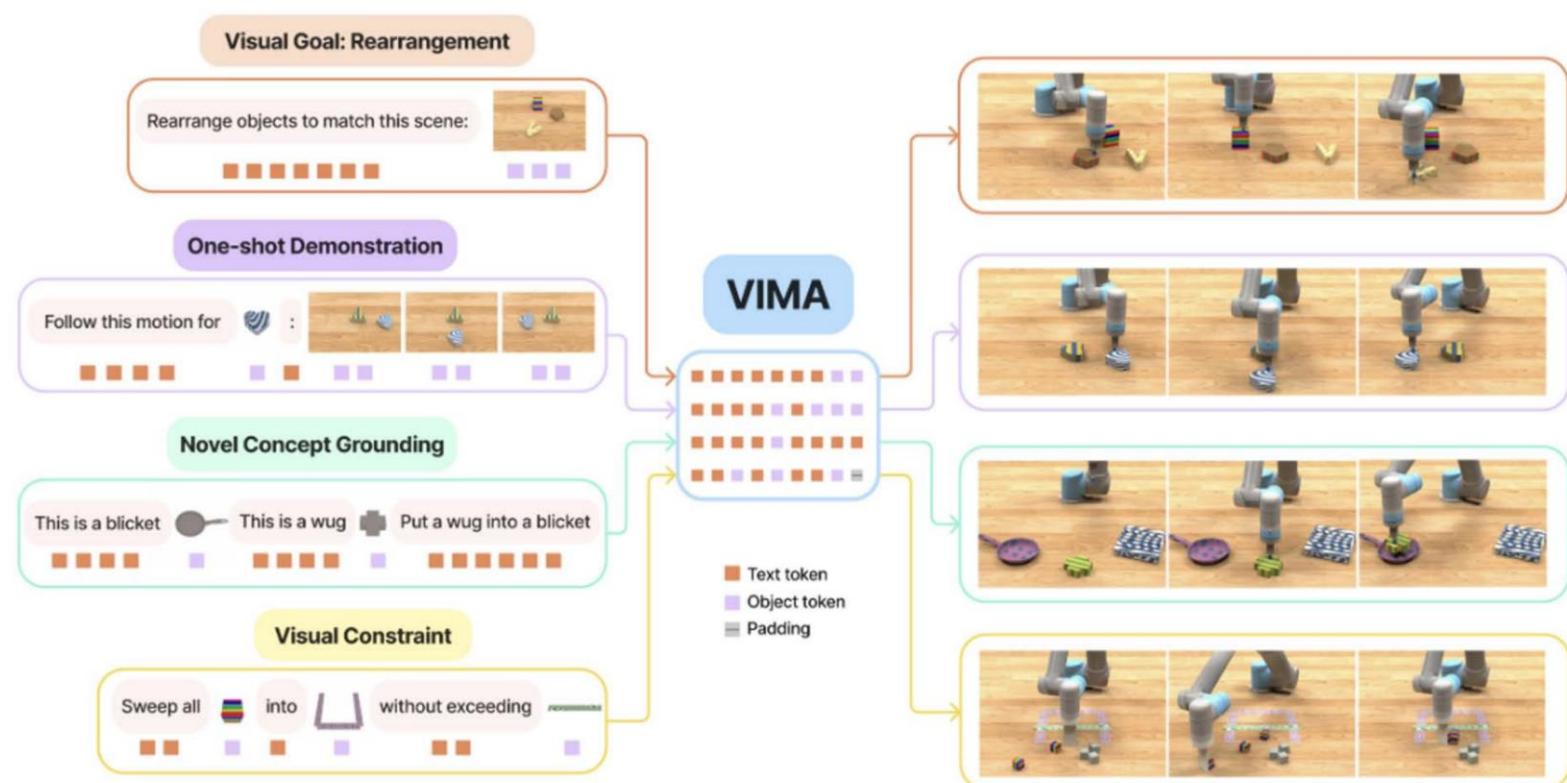


Given Task: Sort colors into corners.  
Step 1. Push the green star to the bottom left.  
Step 2. Push the green circle to the green star.

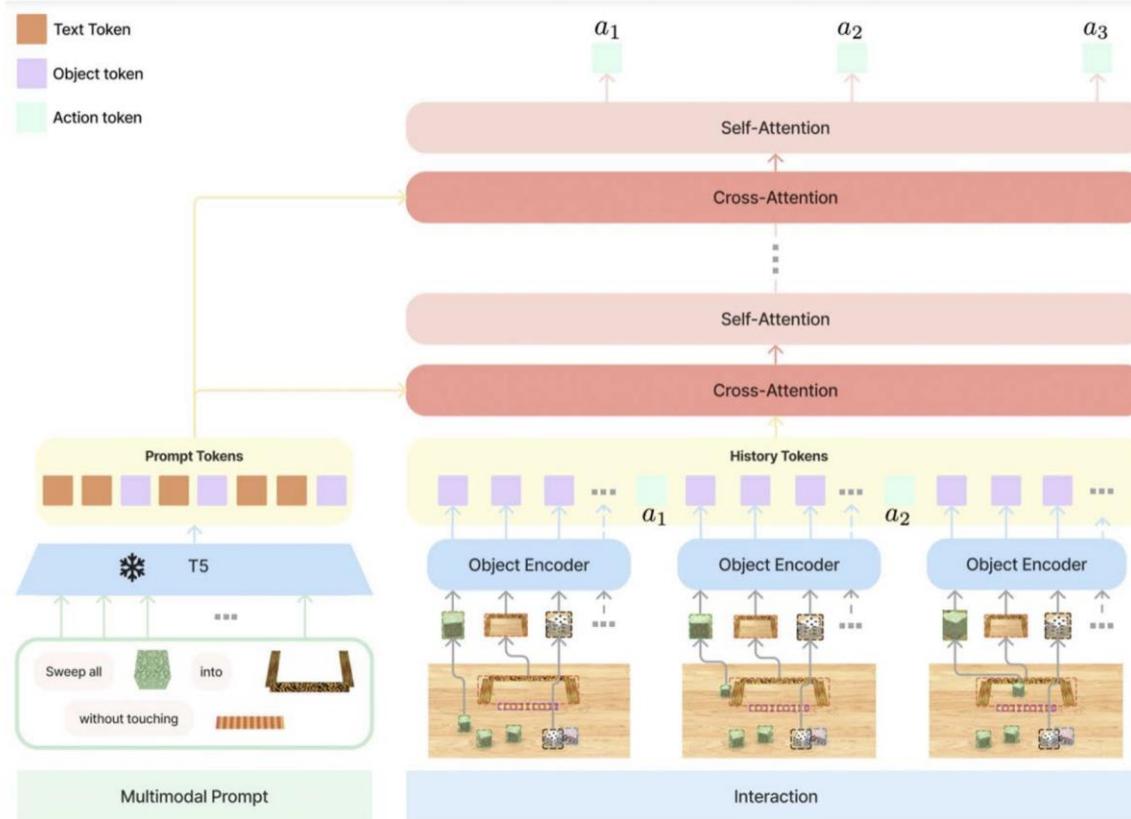
## Language Only Tasks

Q: Miami Beach borders which ocean? A: Atlantic. Q: What is  $372 \times 18$ ? A: 6696. Q: Write a Haiku about embodied LLMs. A: Embodied language. Models learn to understand. The world around them.

# VIMA: Robot Manipulation with Multimodal Prompts



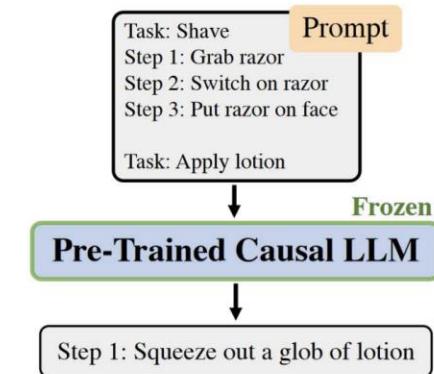
# VIMA: Robot Manipulation with Multimodal Prompts



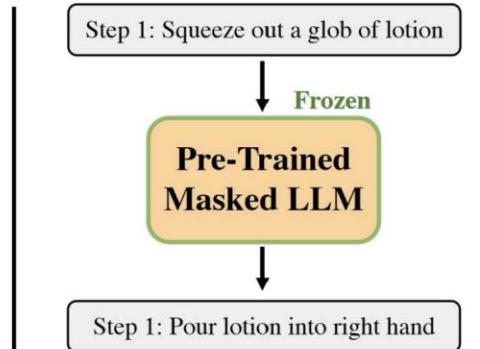
# Language models as zero-shot planners: Extracting actionable knowledge for embodied agents



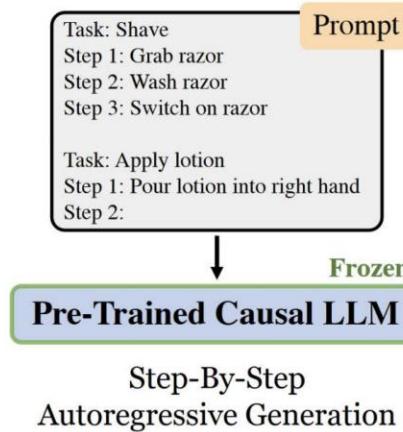
- Key Idea: Large language models learn rich procedural knowledge and such knowledge could be extracted.



Zero-Shot Planning via Causal LLM



Translation to Admissible Action



Step-By-Step  
Autoregressive Generation

# Language models as zero-shot planners: Extracting actionable knowledge for embodied agents



- Impressive results.
- Challenge: verification; groundability to real-world videos.

<p><b>Human</b></p> <p><b>Task: Throw away paper</b></p> <p>Step 1: Walk to home office Step 2: Walk to desk Step 3: Find desk Step 4: Turn to desk Step 5: Find chair Step 6: Sit on chair Step 7: Find check Step 8: Grab check Step 9: Squeeze check Step 10: Stand up Step 11: Walk to trashcan Step 12: Put check on trashcan</p>	<p><b>GPT-3 175B</b></p> <p><b>Task: Throw away paper</b></p> <p>Step 1: Walk to home office Step 2: Walk to wastebasket Step 3: Find wastebasket Step 4: Grab wastebasket Step 5: Walk to desk Step 6: Drop paper in wastebasket</p>	<p><b>Translated GPT-3 175B</b></p> <p><b>Task: Throw away paper</b></p> <p>Step 1: Walk to home office Step 2: Walk to table Step 3: Find table Step 4: Turn to table Step 5: Find paper Step 6: Grab paper Step 7: Walk to trashcan Step 8: Open trashcan Step 9: Put paper on trashcan Step 10: Close trashcan</p>	<p><b>Task: Complete Amazon Turk Surveys</b></p> <p>Walk to Home Office      Sit on Chair Switch on Computer      Look at Computer</p>
<p><b>GPT-2 1.5B</b></p> <p><b>Task: Brush teeth</b></p> <p>Step 1: Go to bathroom</p>	<p><b>Codex 12B</b></p> <p><b>Task: Brush teeth</b></p> <p>Step 1: Walk to bathroom Step 2: Walk to sink Step 3: Find toothbrush Step 4: Pick up toothbrush Step 5: Put toothbrush in mouth Step 6: Move brush around mouth for two minutes Step 7: Spit out toothpaste and brush into sink Step 8: Turn on water in sink and rinse brush for one minute Step 9: Turn off water in sink and return brush to cupboard</p>	<p><b>Translated Codex 12B</b></p> <p><b>Task: Brush teeth</b></p> <p>Step 1: Walk to bathroom Step 2: Open door Step 3: Walk to sink Step 4: Put pot on sink Step 5: Put brush on toothbrush Step 6: Turn to toothpaste Step 7: Put toothpaste on toothbrush Step 8: Put teeth on toothbrush</p>	<p><b>Task: Get Glass of Milk</b></p> <p>Walk to Kitchen      Open Fridge Grab Milk      Close Fridge</p>

# From Language to Actions

# Limits in LLMs

- Language reasoning (e.g., math, logics, common sense, ...)



GPT-4

Explain the process of  $3452 \times 253$

$$\begin{array}{r} 3452 \\ \times 253 \\ \hline \end{array}$$

$$10356 \quad (3452 \times 3)$$

**6904**  $(3452 \times 5, \text{ shifted one position to the left})$

**+3452**  $(3452 \times 2, \text{ shifted two positions to the left})$

$$\hline 873256$$

So,  $3452 \times 253 = 873256$ .

**(correct answer: 873356)**

# Limits in LLMs

- Language reasoning (e.g., math, logics, common sense, ...)

Track the same group of students over time



GPT-4

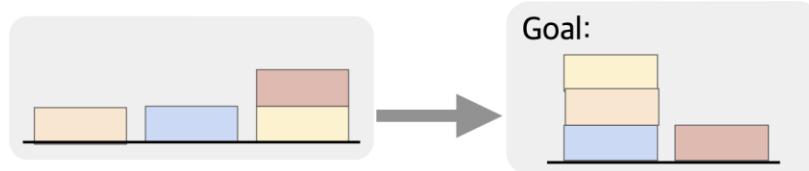
Here is a plan for surveying the mental health of undergraduate students on our campus. This year, we will ask freshmen; next year, sophomores; the year after, juniors; and the year after that, seniors. Are there any problems with this plan?

- ...
  1. **Longitudinal Data and Cohort Effects:** This approach doesn't track the same group of students over time...
  2. ...
  - ...

# Limits in LLMs

- Embodied reasoning: planning

Blocksworld: How to move the blocks to the goal state?



GPT-4

## Invalid Action!

The yellow block is still under the red one.

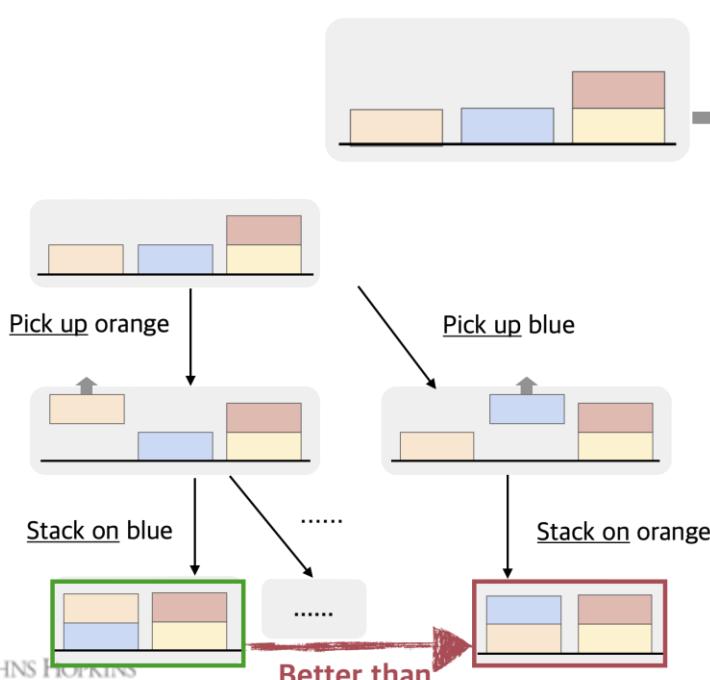
1. Pick up the orange block.
2. Stack it on the blue block.
3. Pick up the yellow block. X
4. Stack it on the orange block.
5. Pick up the red block.
6. Put it on the table.

LLMs: Autoregressive plan generation

# Limits in LLMs

- Embodied reasoning: planning

Blocksworld: How to move the blocks to the goal state?



Goal:



Human: **model-based** planning

- Internal **world model**
- Forward **simulation** of alternative plans

# Limits in LLMs

- Building *socially* intelligent systems requires **social reasoning**

AI Drivers



AI Coworkers



AI Assistants



AI Companions



AI Teachers



# Limits in LLMs

---



Does this person need help?

- (1) Language reasoning
- (2) Embodied reasoning
- (3) Social reasoning

# Limits in LLMs

- Social reasoning: understanding hum

Social reasoning also requires a **world model**:  
Humans' actions are constrained by physics



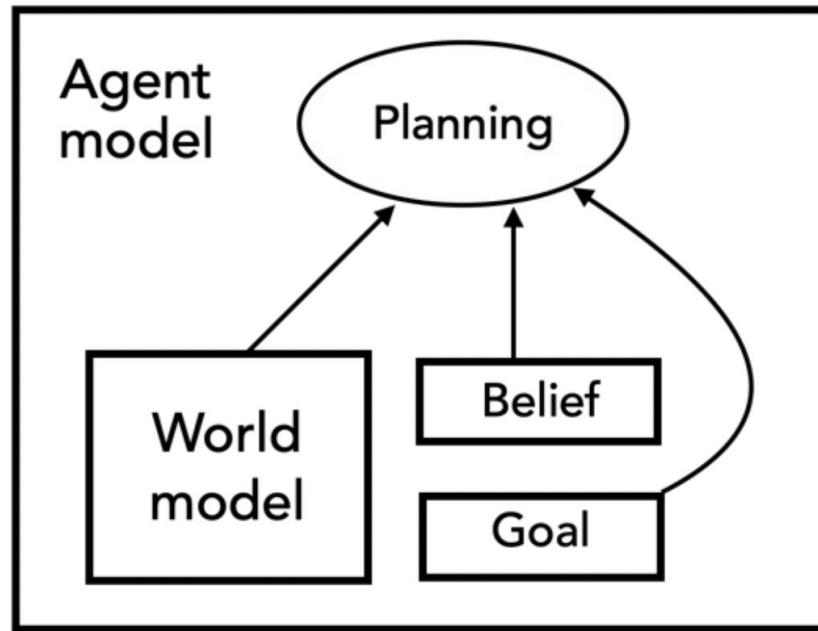
Does this person need help?



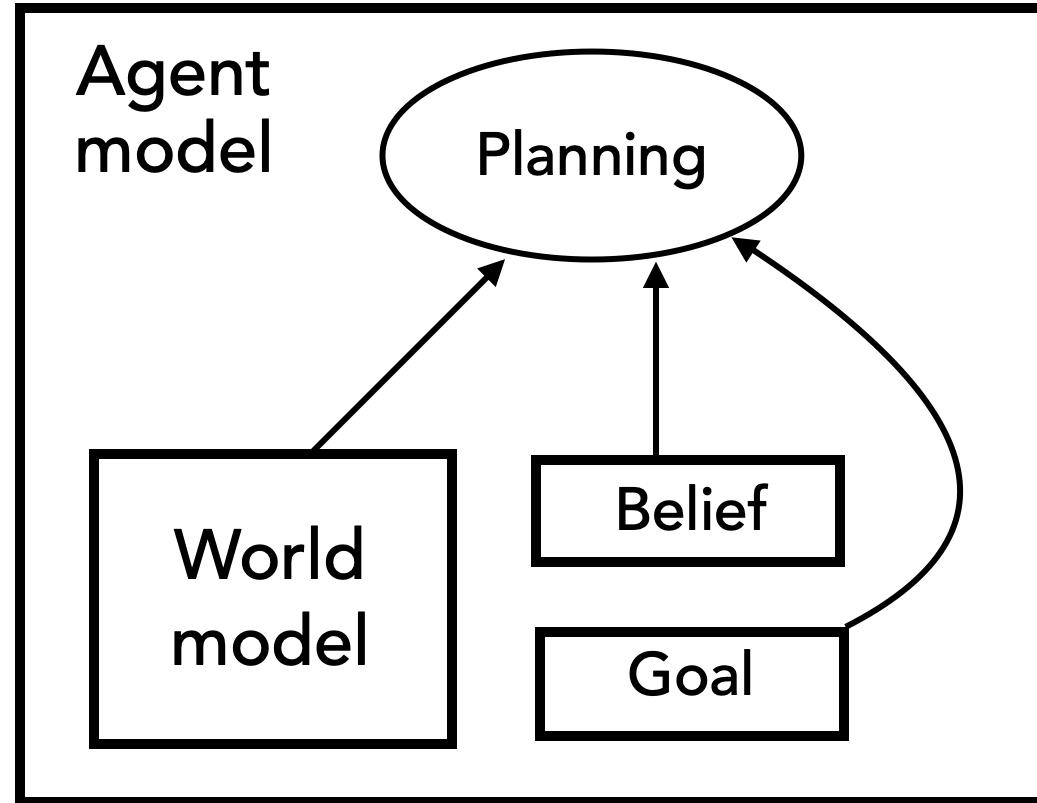
GPT-4V

It looks like the person in the image is carrying a heavy load of boxes, which might suggest they are in the process of moving. While I can't determine the actual need for help without context or the ability to interact with the situation, generally speaking, someone carrying a heavy load might appreciate an offer of assistance. It's always kind to offer help if you see someone who might be struggling or if the task at hand seems like a lot for one person to handle.

## The minimum definition of an agent model



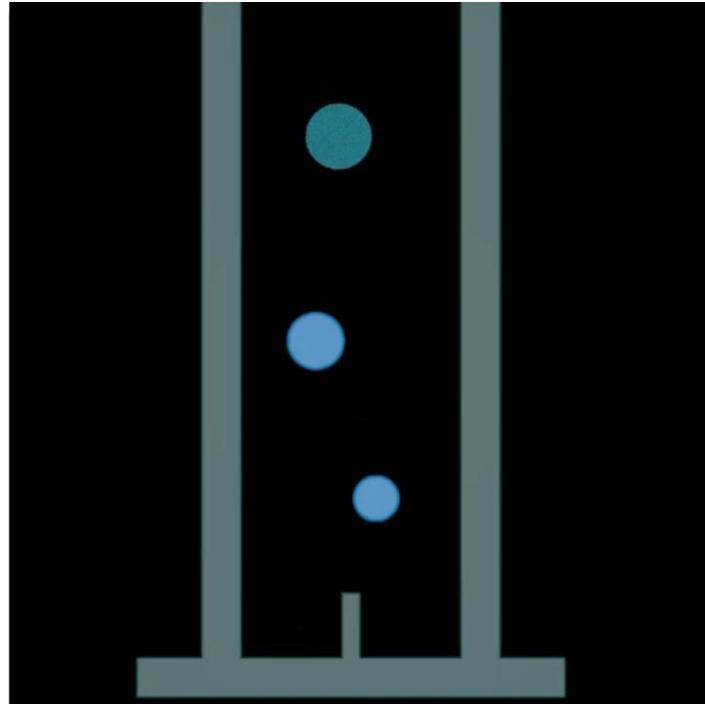
# World models and agent models



# World models in humans

---

- Predicting  $d_\text{y}$

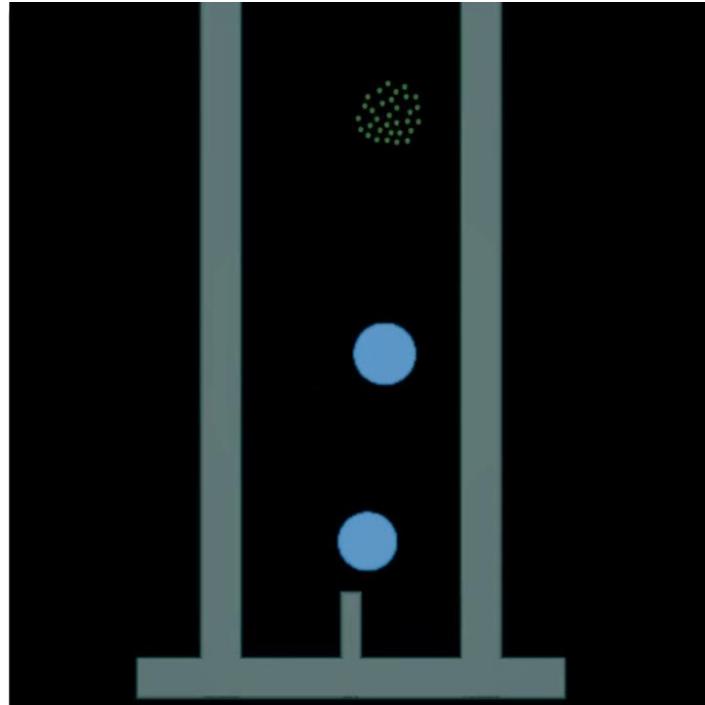


Kubricht et al. (2017)

# World models in humans

---

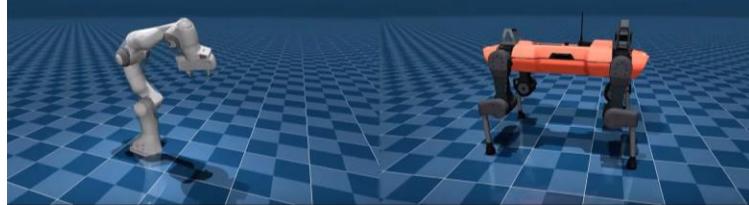
- Predicting  $d$



Kubricht et al. (2017)

# Simulators as world models

- Physics engines and simulated simulators



MuJoCo

AI2-THOR



iGibson 2.0

Todorov et al. (2012)

Habitat 2.0



Li et al. (2021)

Kolve et al. (2017)  
ThreeDWorld



Gan et al. (2020)

# Simulators as world models

- Embodied simulators + synthetic humans  
VirtualHome 2.0



Puig et al. (2021)



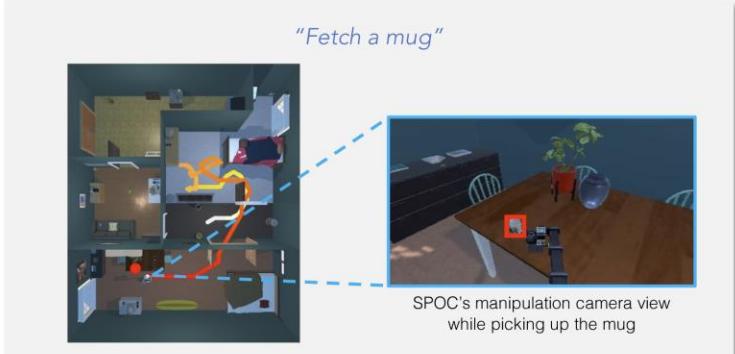
Puig et al. (2023)

# Simulators as world models

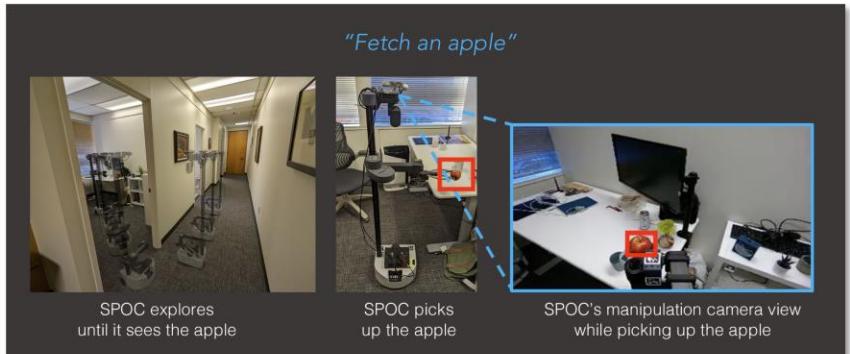
Simulator



Real World



SPOC's manipulation camera view while picking up the mug



SPOC explores until it sees the apple

SPOC picks up the apple

SPOC's manipulation camera view while picking up the apple

Ehasani et al. (2023)

# Video prediction for robot planning

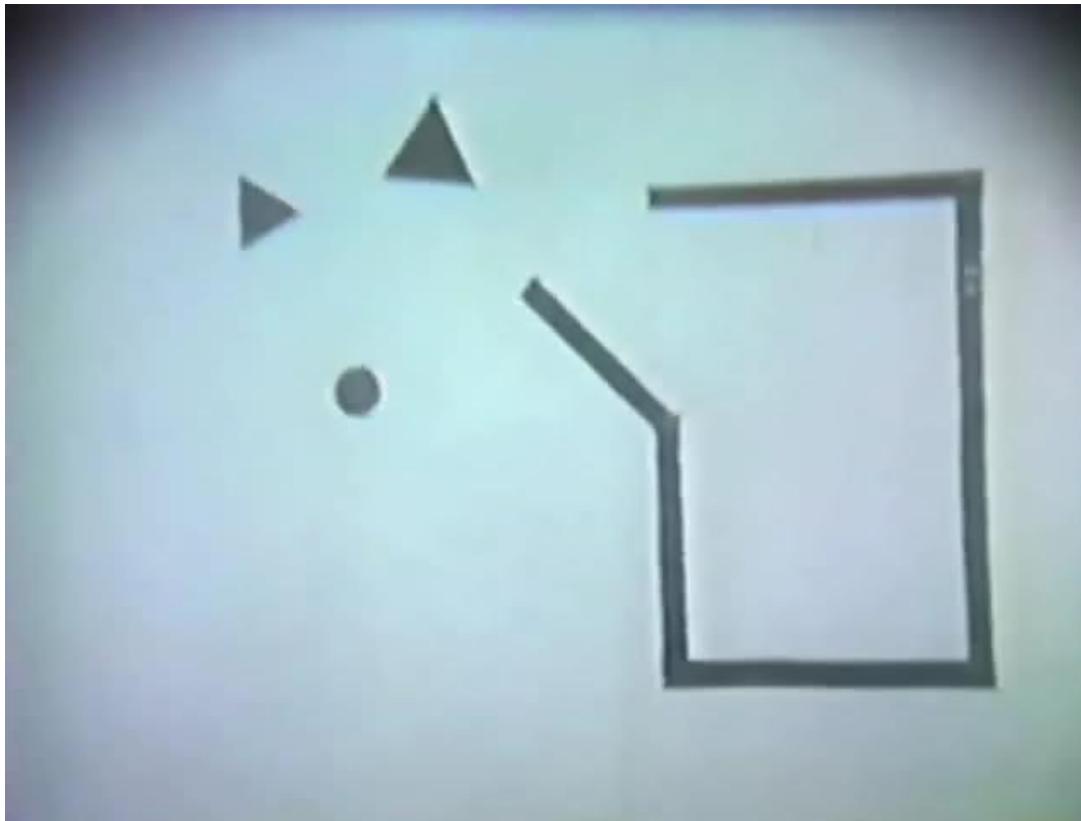
Simulating long sequence of robot executions.

Step 1:



# Humans represent agents differently from objects

---



Heider & Simmel (1944)

# Humans represent agents differently from objects

Strengths

strong, weak

Goals

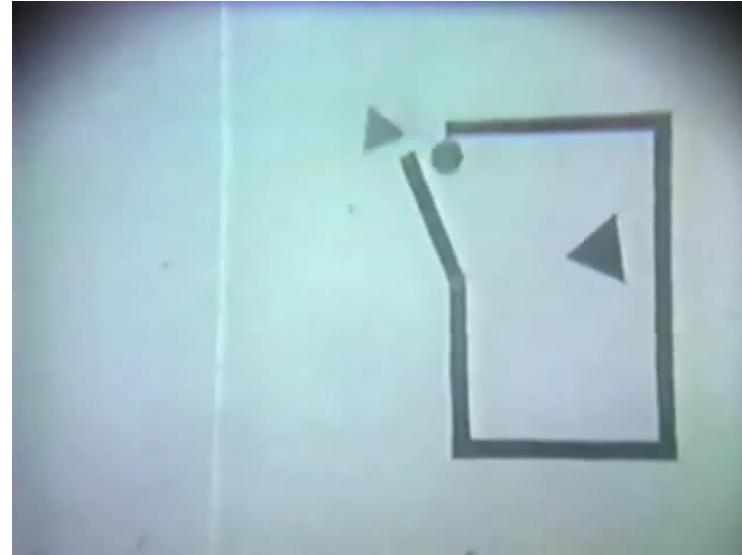
helping, hurting, escaping

Relationships

friends, enemies

Moral judgment

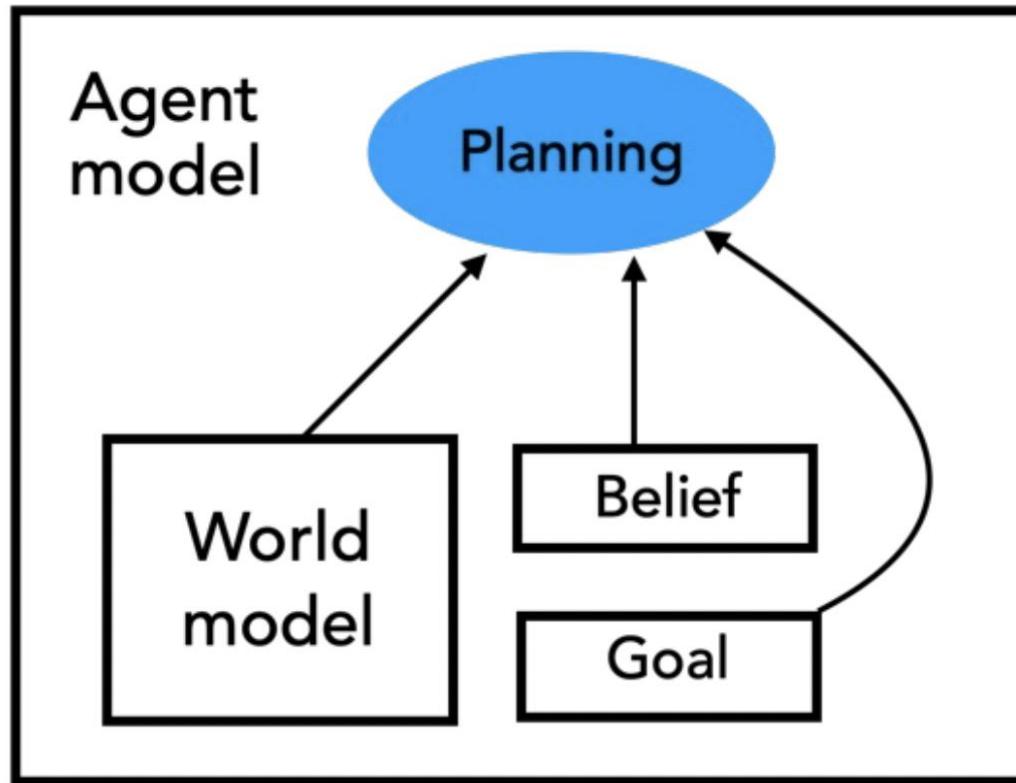
good guy, bully



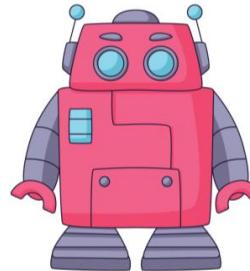
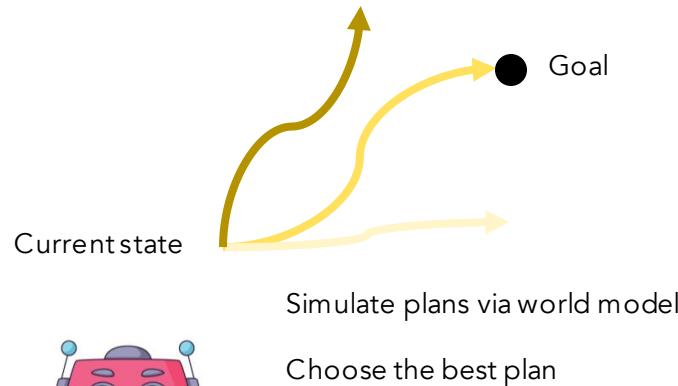
(size / velocity / angle...)

A big triangle moves back and forth, while a small triangle and a small circle rotate 360°...

# Language models as (autoregressive) planners

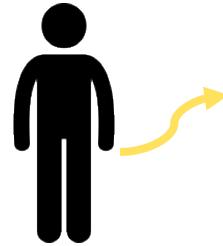


# Level-0 agent models for embodied tasks



# Level-1 agent models for social reasoning

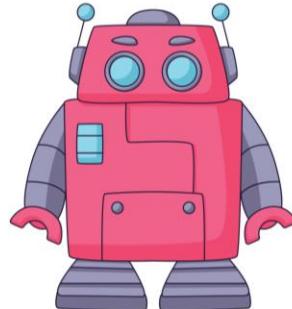
---



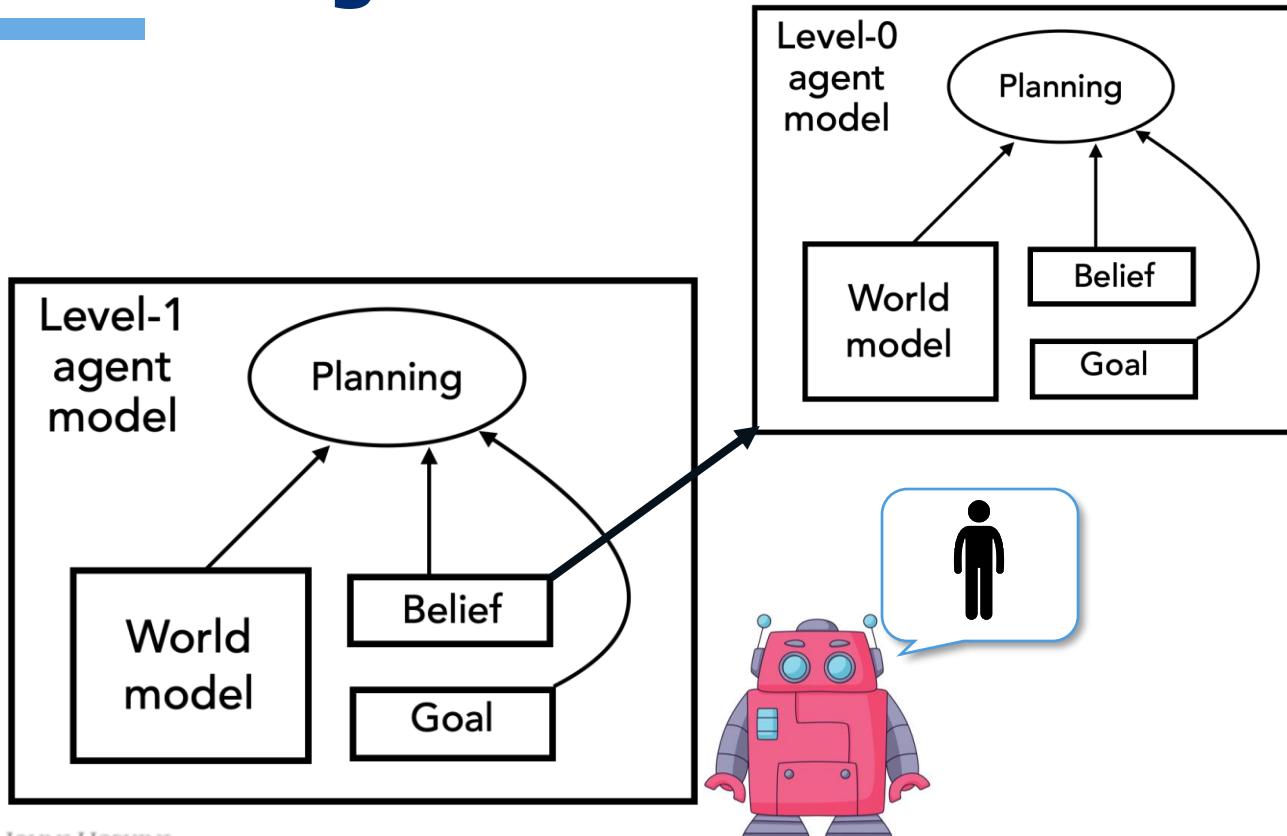
Goal: Office or coffee shop?



An observer



# Level-1 agent models for social reasoning

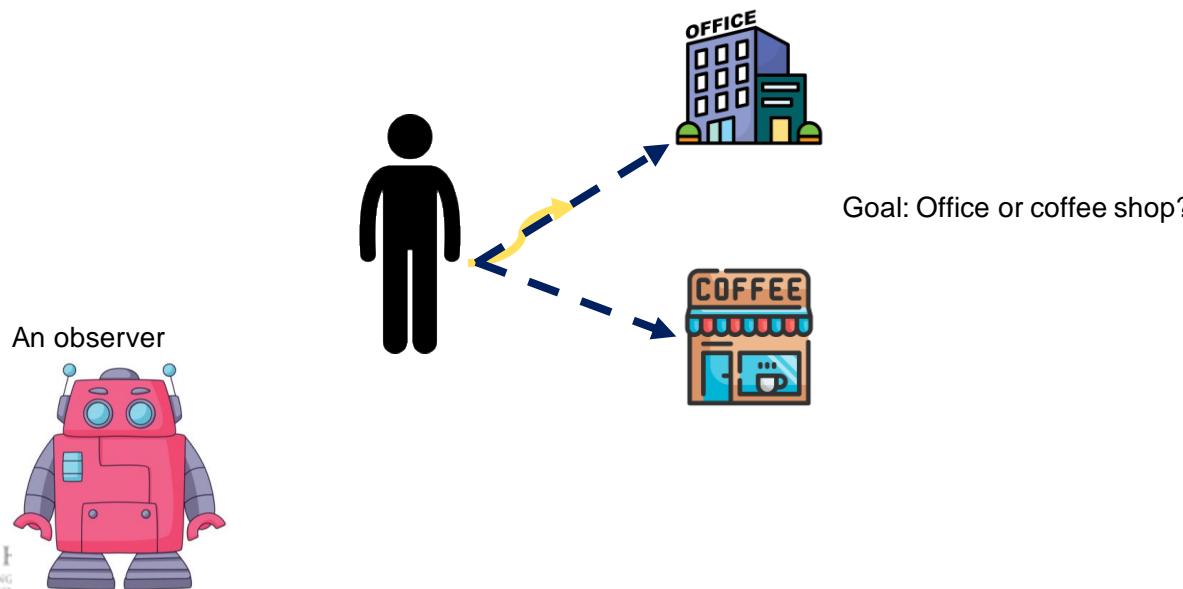


# Level-1 agent models for social reasoning

## Model-based Theory of Mind

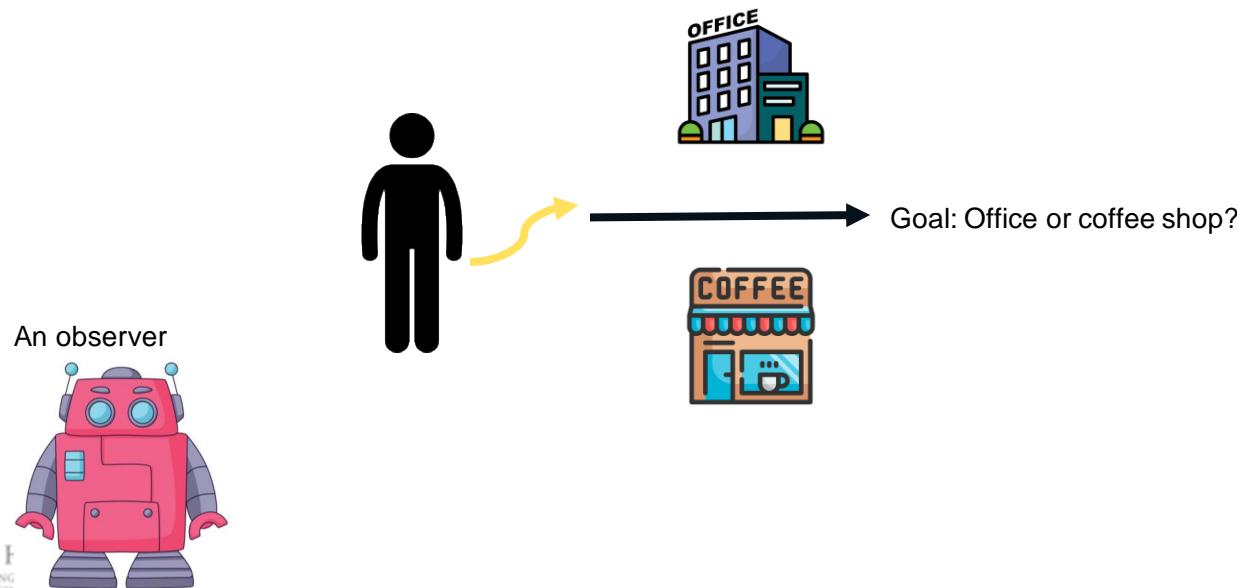
$$P(\text{mind}|\text{state, actions}) \propto P(\text{actions}|\text{state, mind})P(\text{mind})$$

Level-0 agent model



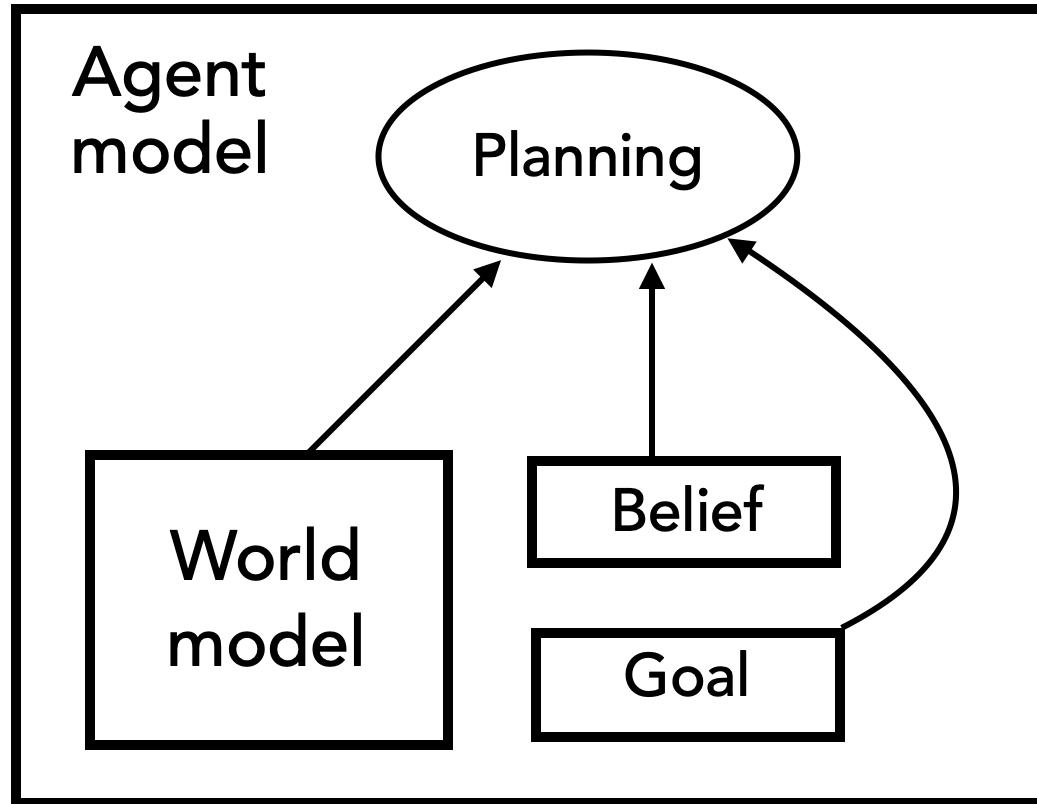
# Social reasoning without agent models

Model-free Theory of Mind: directly maps actions to mental state

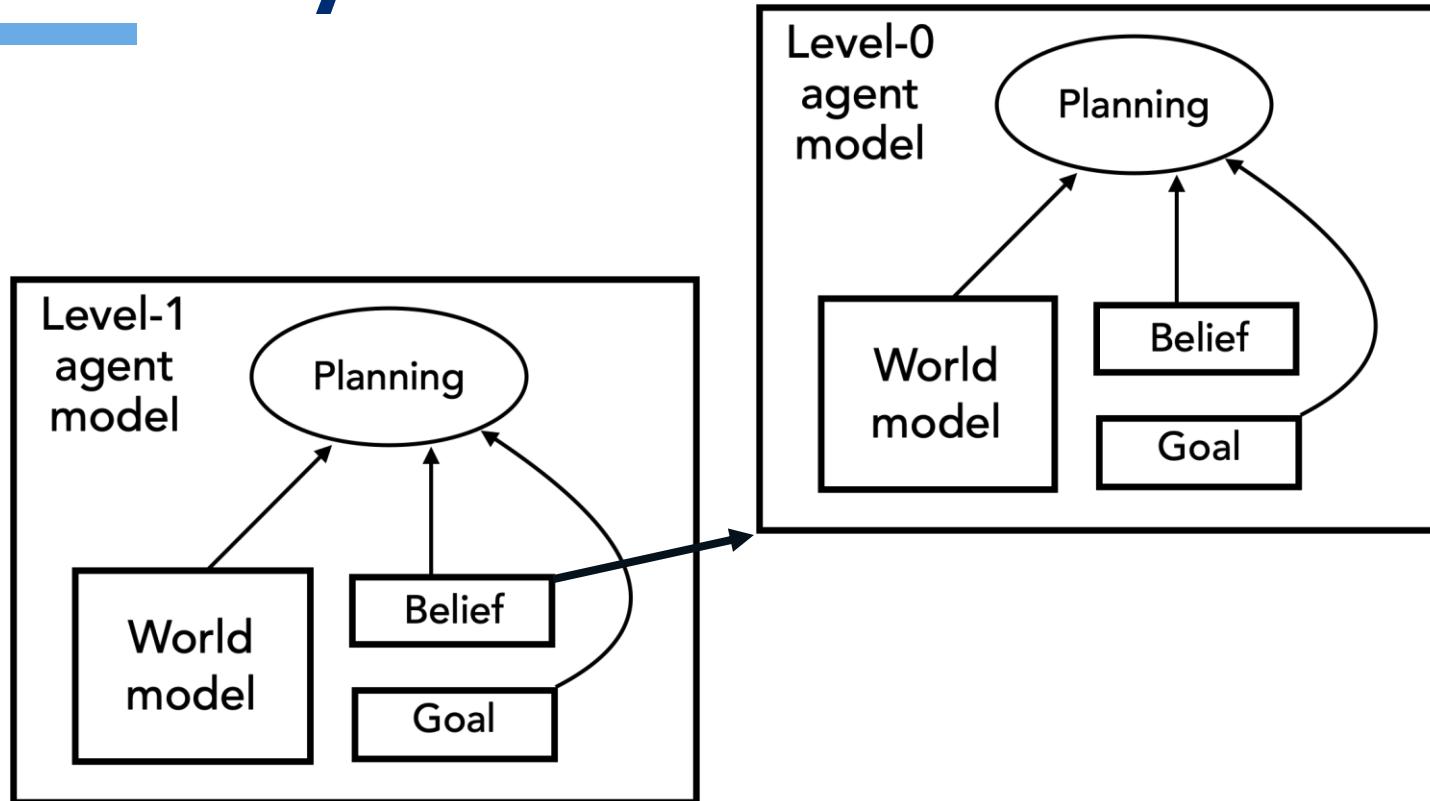


# Summary

---

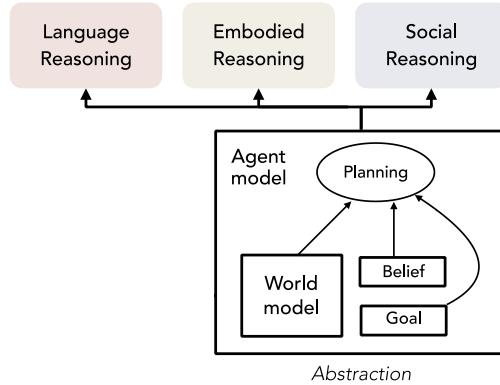


# Summary

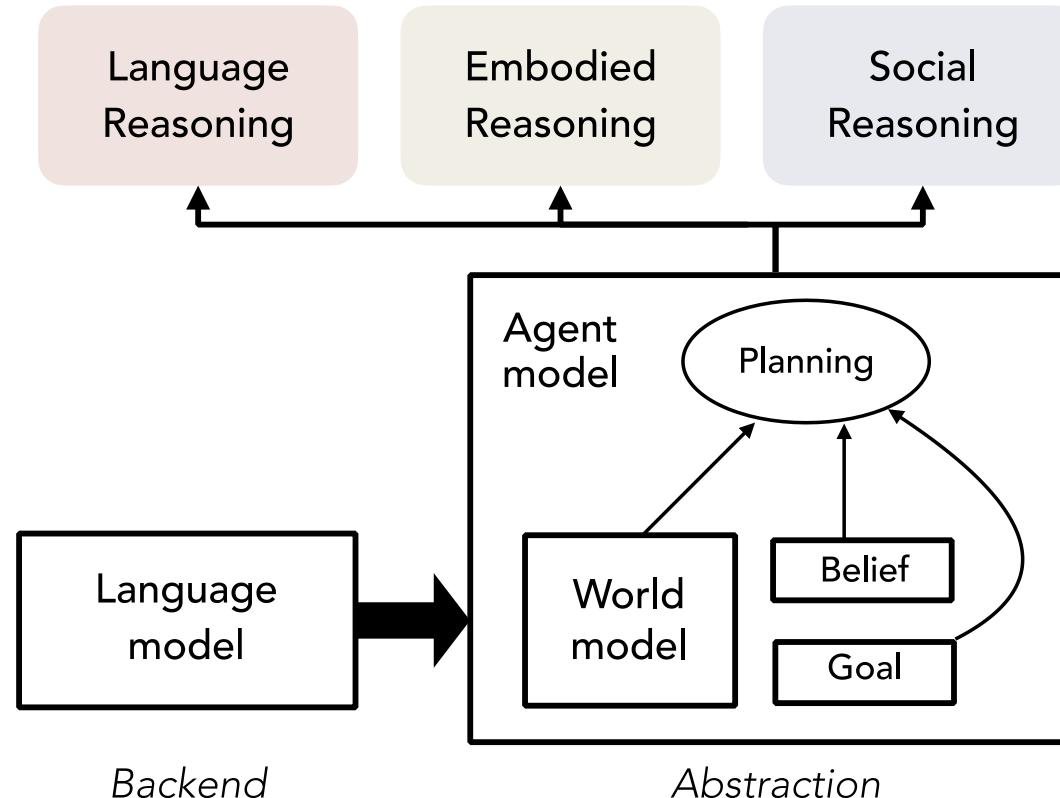


# Summary

- Model-based reasoning
  - Data-efficient
  - Generalizes to new scenarios well
- Traditionally, world models and agent models are all domain specific
- Can we leverage the open-endedness of LMs to construct world and agent models?
- Language models as backend
- Questions?



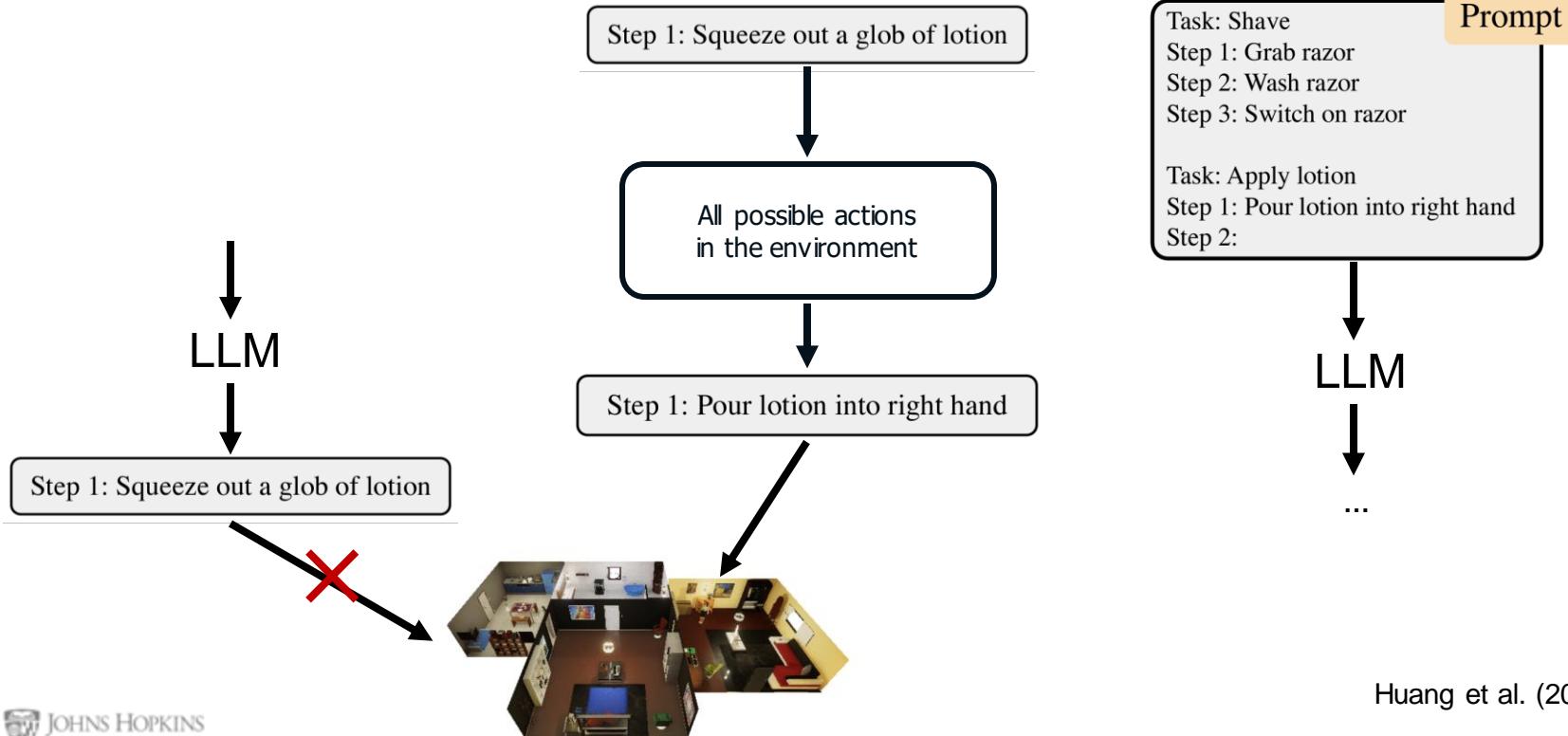
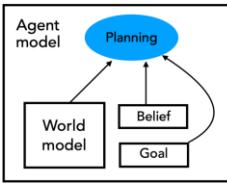
# World models and agent models with LMs as backend



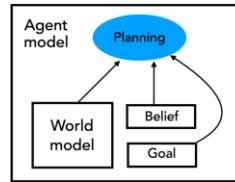
# Language models as (autoregressive) planners

---

# Language models as (autoregressive) planners



# Language models as (autoregressive) planners



Browse Internet

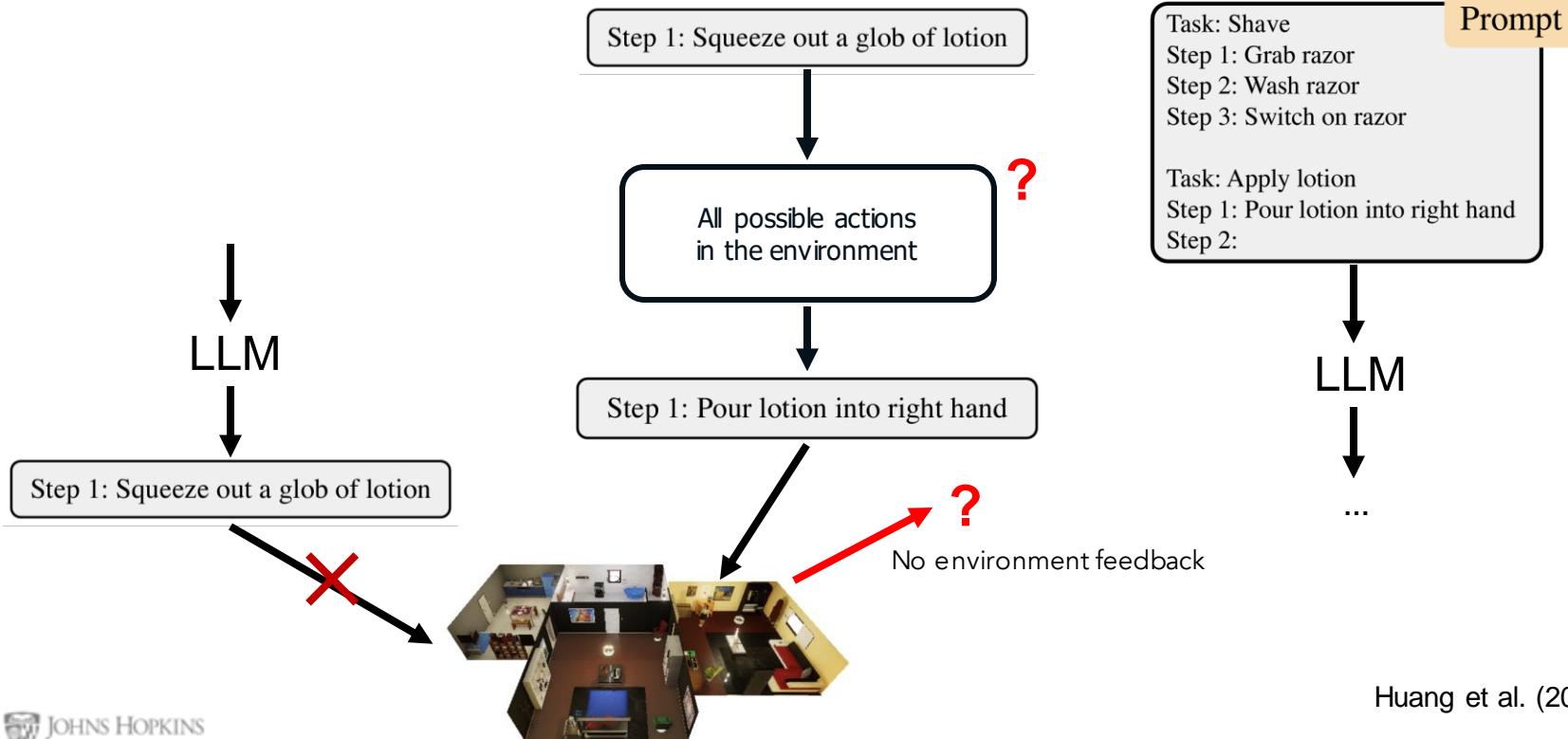
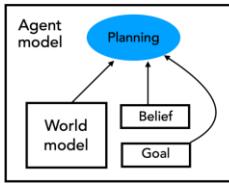


Organize Closet

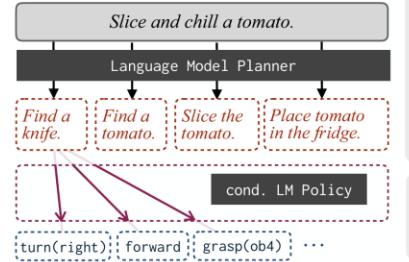
Empty Dishwasher

Huang et al. (2022)

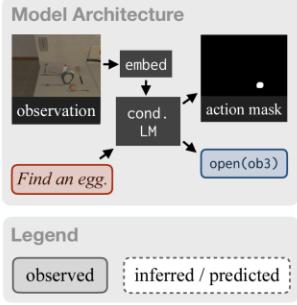
# Language models as (autoregressive) planners



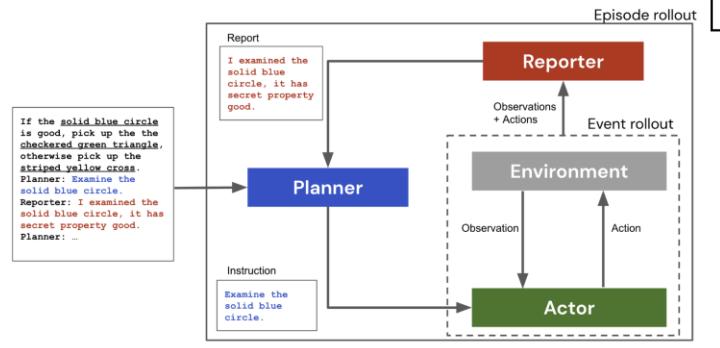
# Interactive planner



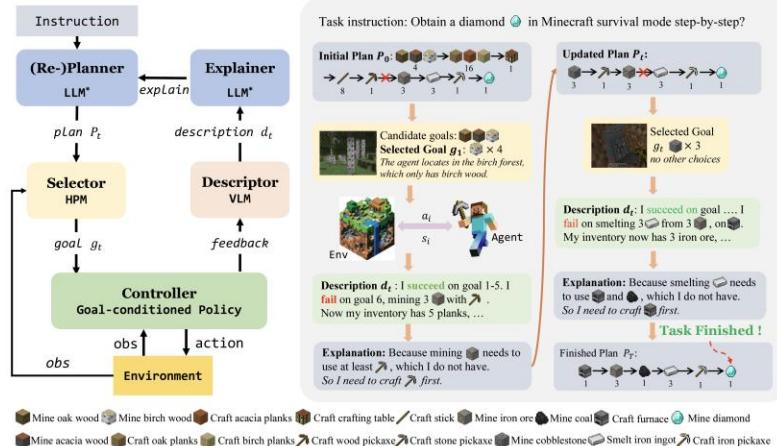
Sharma et al. (2021)



Li et al.  
(2022)

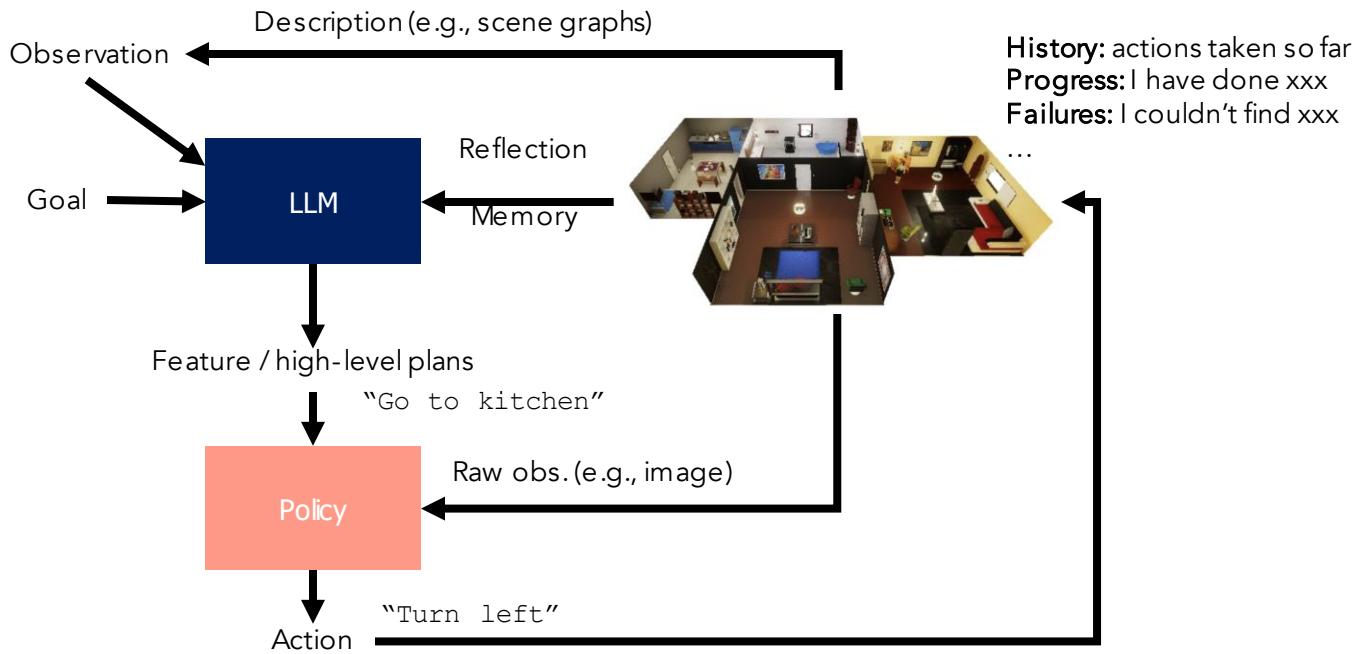
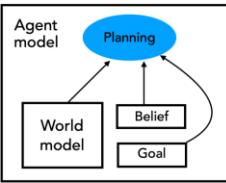


Dasgupta et al. (2022)



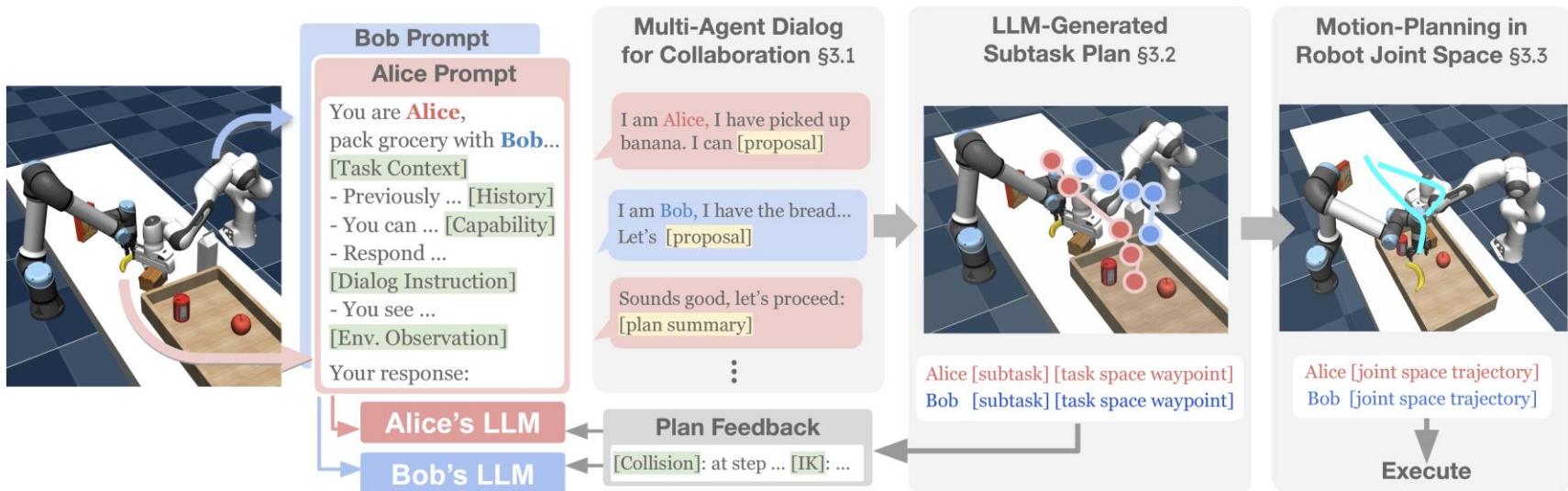
Wang et al. (2023)

# Interactive planner



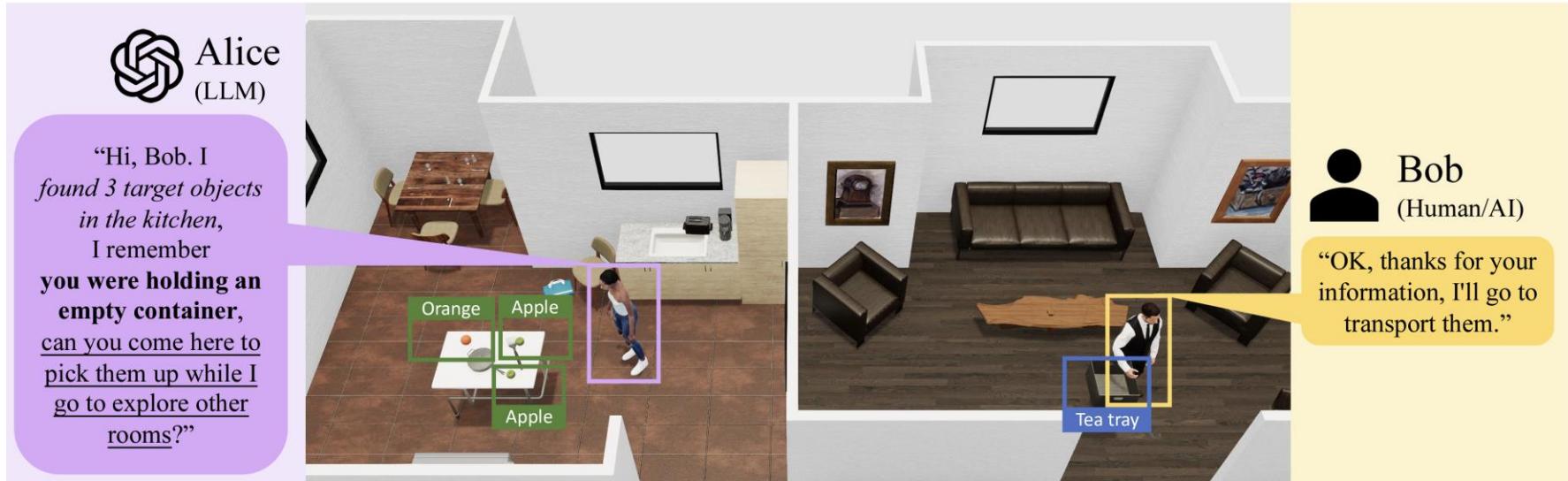
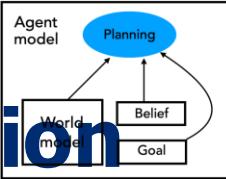
# Multi-agent planning and communication

- Multi-agent collaboration



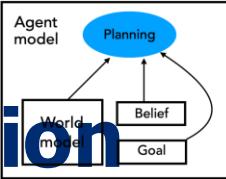
Mandi et al. (2023)

# Multi-agent planning and communication

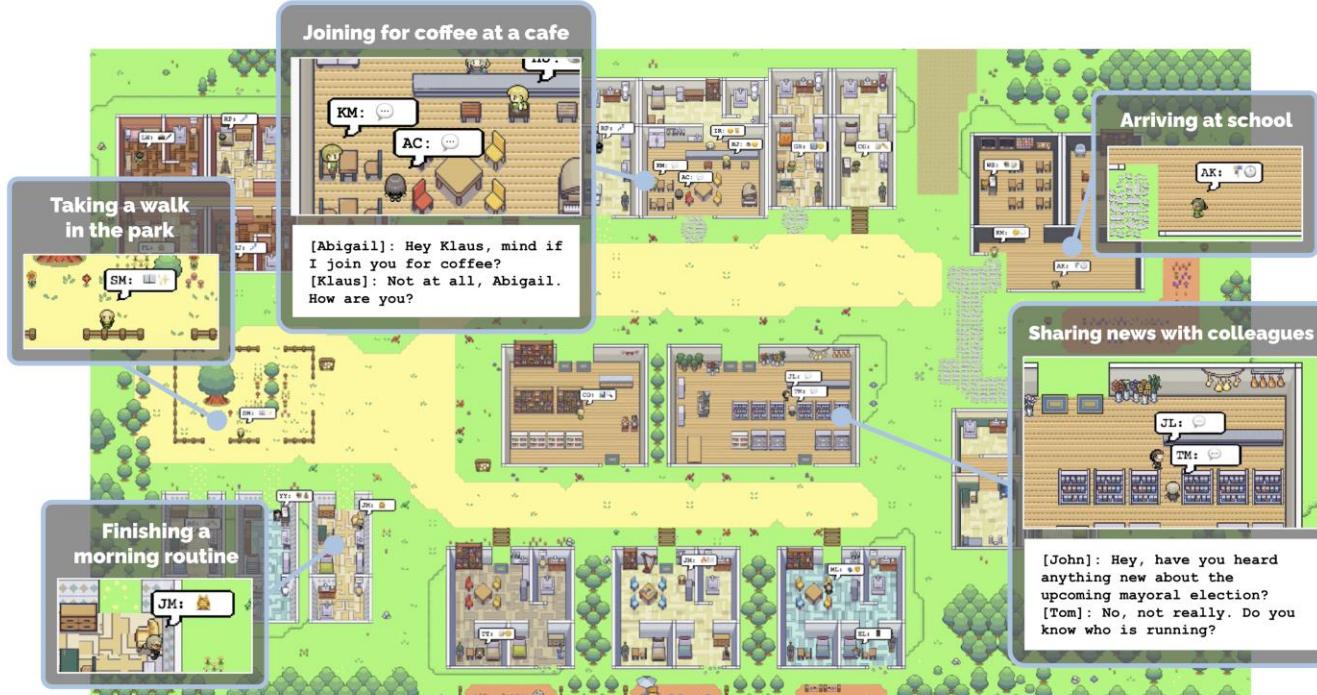


Zhang et al. (2023)

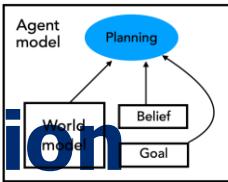
# Multi-agent planning and communication



- Social



Park et al. (2023)



# Multi-agent planning and communication

- Social interaction between *multiple* simulated agents

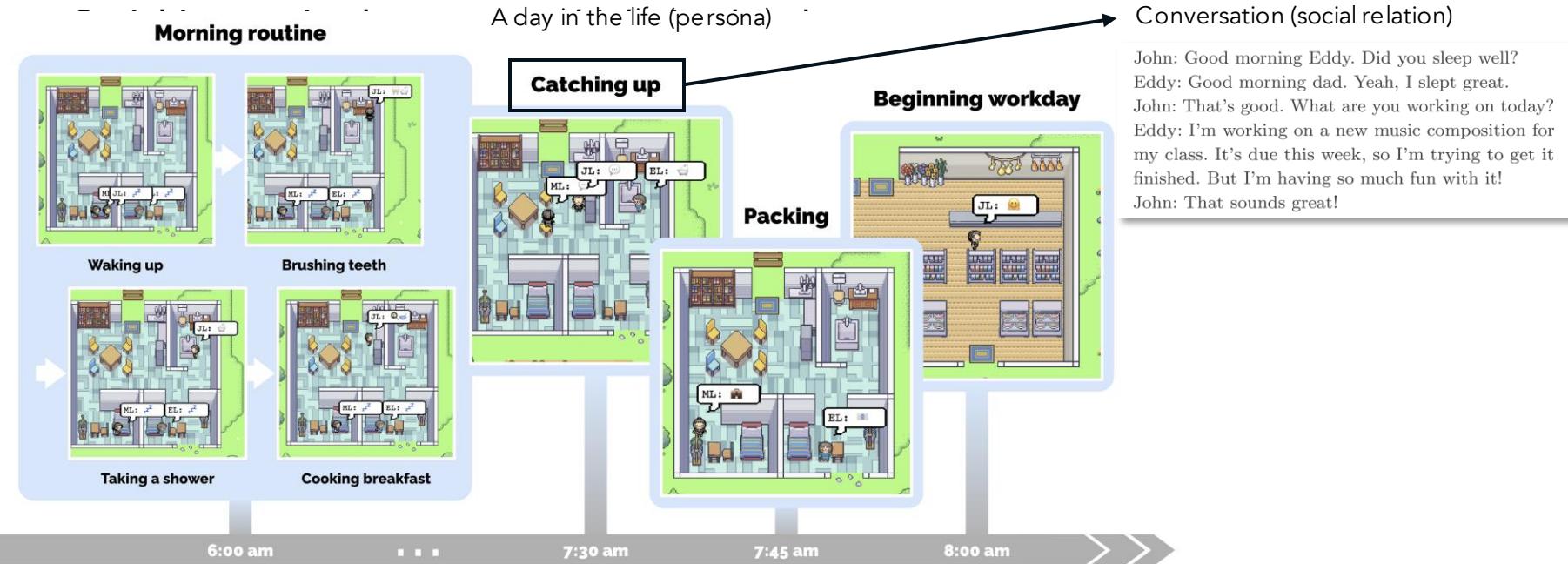
- Persona

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people.

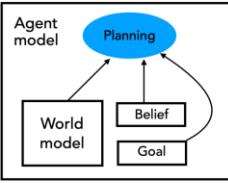
- Social relationships with other agents

...  
John Lin is living with his wife, Mei Lin, who is a college professor, and on, Eddy Lin, who is a student studying music theory

# Multi-agent planning and communication



# Limitation: increasing complexity



LLM

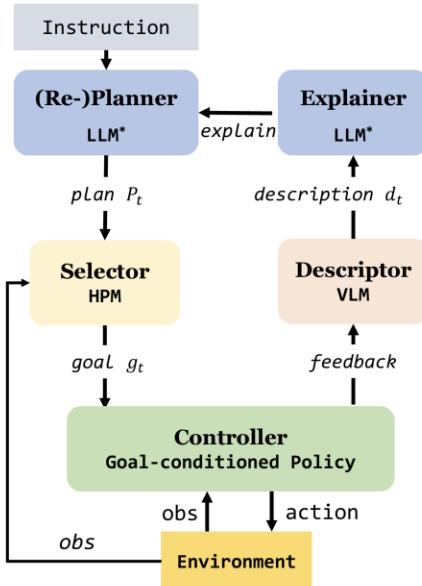
Step 1: Squeeze out a glob of lotion

In Minecraft

Huang et al. (2022)

Obtain diamond: 0%

5 modules



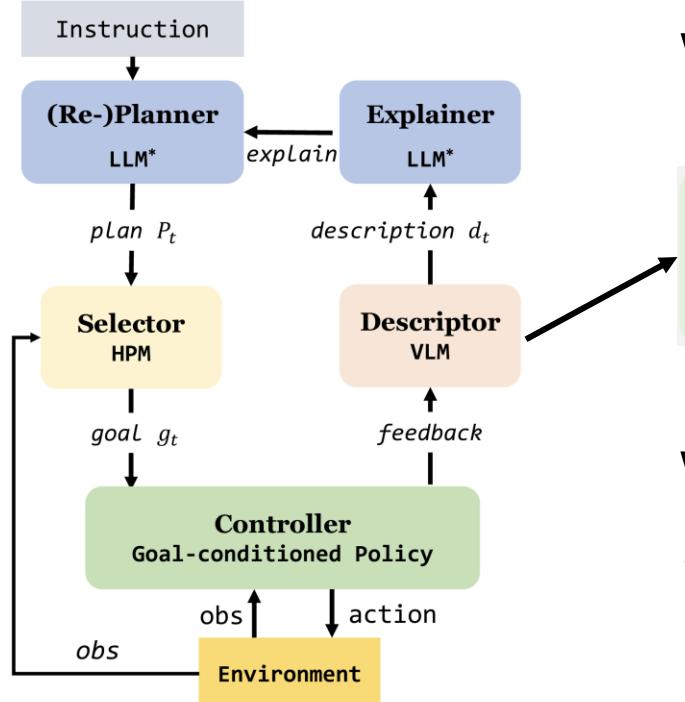
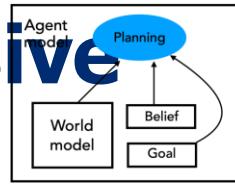
Wang et al. (2023)

Obtain diamond: 9.2%

# Lacking explicit modeling of other components

---

# Model-based planning vs. autoregressive generation



Without a world model:  
Repeated trial-and-error

Description  $d_t$ : I **succeed** on goal 1-5. I **fail** on goal 6, mining 3 with .  
Now my inventory has 5 planks, ...

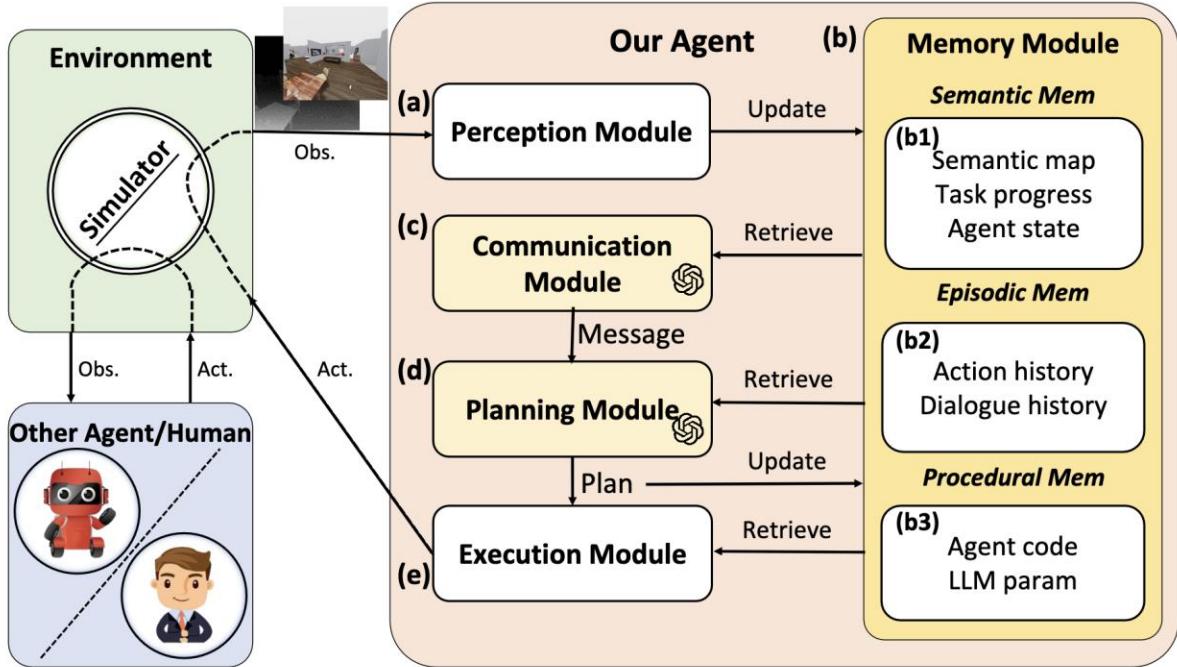
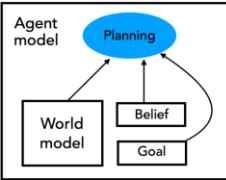
With a world model:  
Simulate outcomes of actions

Mine oak wood Mine birch wood Craft acacia planks

Mine acacia wood Craft oak planks Craft birch plank:



# Limitation: increasing complexity



~~A manual implementation of belief for a specific domain~~  
General belief model that updates belief of the state consistent with the observation history

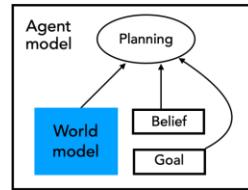
# Language models as belief?

---

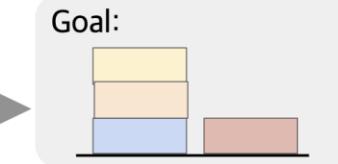
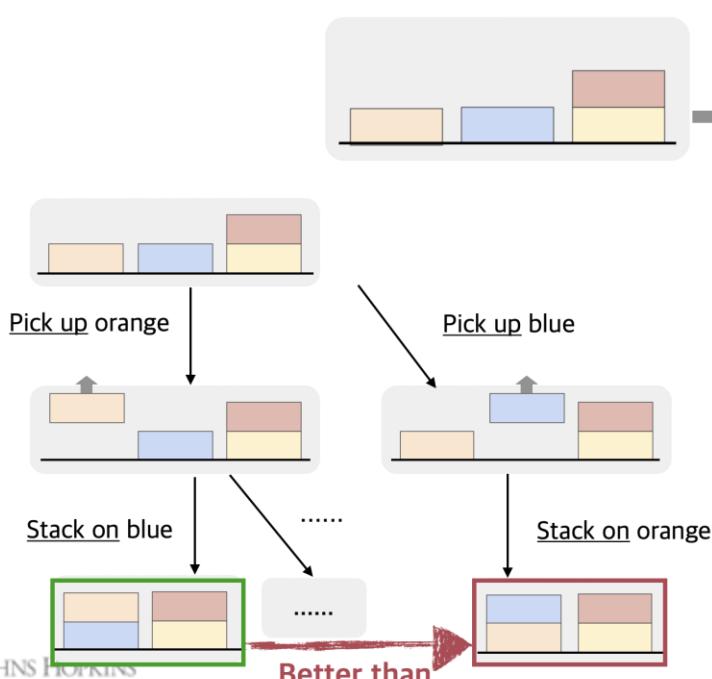
# Language models as world models

---

# Language models as world models

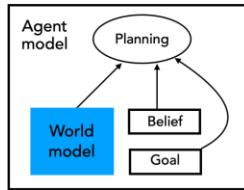


Blocksworld: How to move the blocks to the goal state?



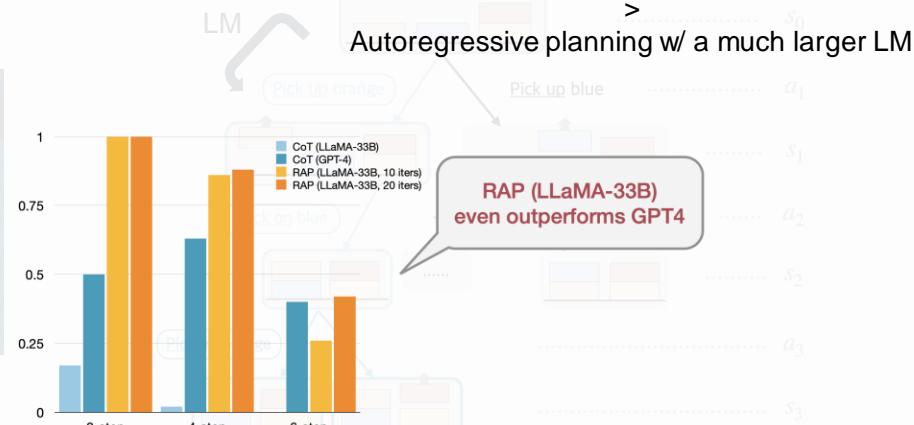
Human: **model-based** planning

- Internal **world model**
- Forward **simulation** of alternative plans



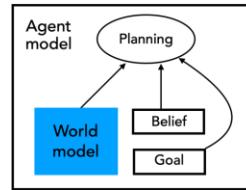
# Language models as world models

- Reasoning-via-Planning (RAP), Hao et al. (2023)
- Monte Carlo Tree Search (MCTS):



Using LMs to construct WMs for embodied reasoning

# Language models as world models

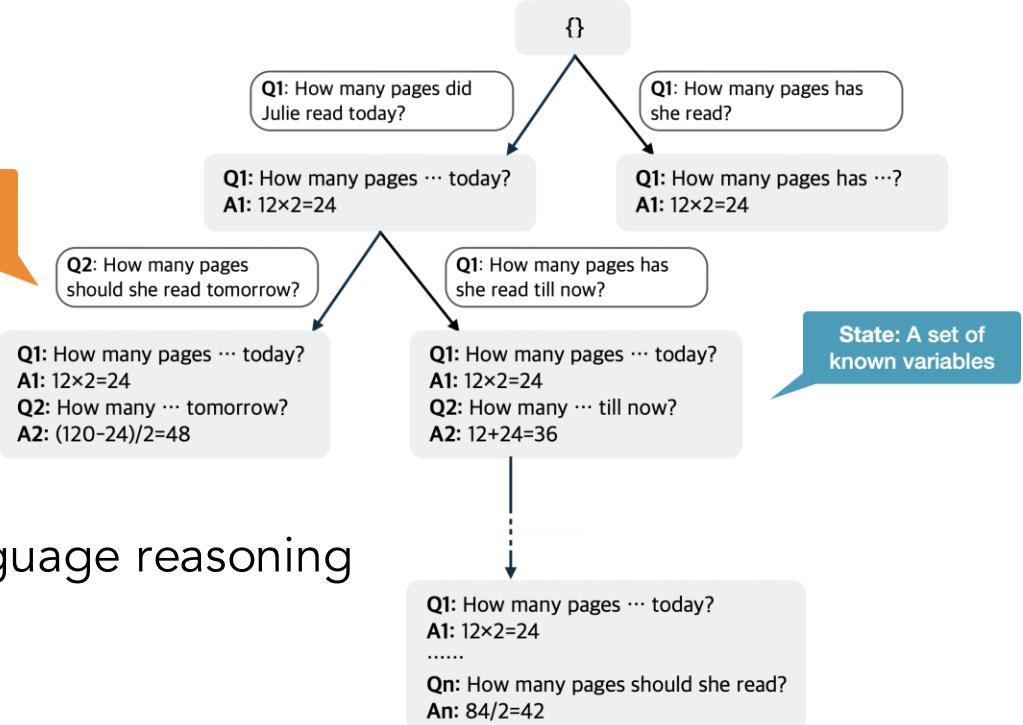


## DAD for solving math problems

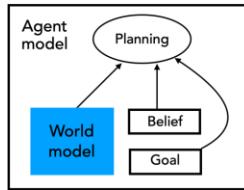
Question:

Julie is reading a 120-page book.  
Yesterday ... 12 pages  
Today ... twice as many pages as yesterday  
Tomorrow ... half of the remaining pages  
**How many pages should she read?**

Action: a sub-question for an unknown variable



Using LMs to construct WMs for language reasoning



# Language models as world models

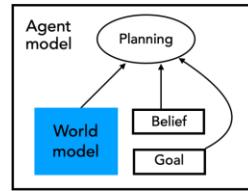
- Why can LMs simulate the world?

Knowledge of the world encoded in the training text

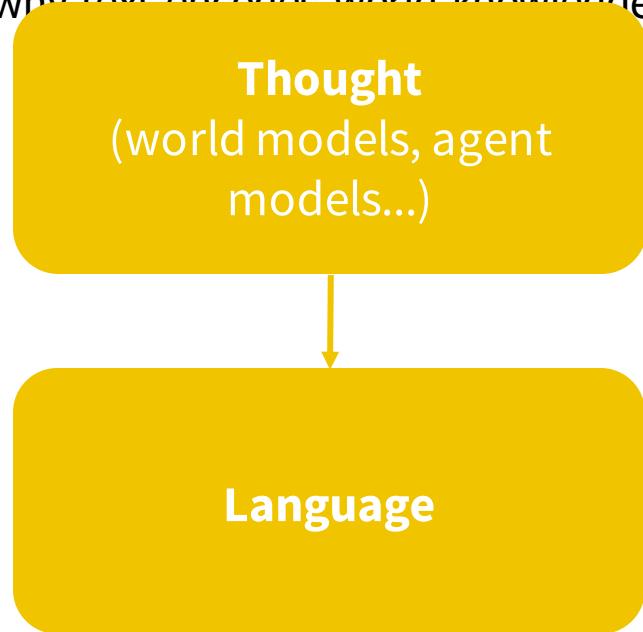
*"If a wine glass falls onto the ground, it will break."*

*"If a basketball falls onto the ground, it will bounce back."*

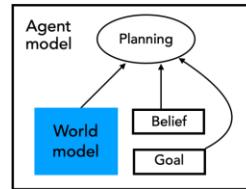
# Language models as world models



- Language of thoughts (why text encodes world knowledge?)



# Language models as world models



- Language of thoughts

**Thought**  
(world models, agent  
models...)

Probabilistic programs

From Word Models to World Models:  
Translating from Natural Language to the  
Probabilistic Language of Thought

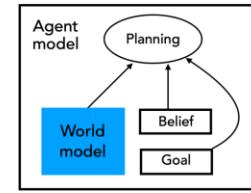
Lionel Wong<sup>1\*</sup>, Gabriel Grand<sup>1\*</sup>, Alexander K. Lew<sup>1</sup>, Noah D. Goodman<sup>2</sup>, Vikash K. Mansinghka<sup>1</sup>, Jacob Andreas<sup>1</sup>, Joshua B. Tenenbaum<sup>1</sup>

\*Equal contribution.

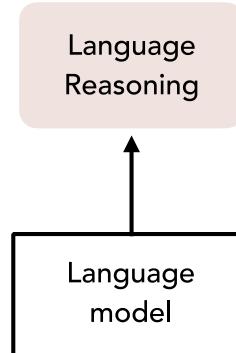
<sup>1</sup>MIT, <sup>2</sup>Stanford

Language

# Reasoning about the world using language



- Imagine a table with a red ball placed to the left of a blue ball. We can push the red ball and it hits the blue ball.
- Imagine that the red ball is pretty heavy. And the blue ball is fairly light.
- How fast does the blue ball move after the collision?



# Sampling a world model from text using LM

Text

Imagine a table with a red ball placed to the left of a blue ball. We can push the red ball and it hits the blue ball. Imagine that the red ball is pretty heavy. And the blue ball is fairly light.

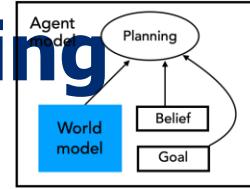
Probabilistic program

```
(define choose_shapes...)
(define get_initial_color ...)
(define choose_mass ...)
(define get_initial_x...) ...

(define generate-object
  (mem (lambda (obj-id) (list
    (pair 'object-id obj-id) (choose_shape obj-id)
    (choose_color obj-id) (choose_mass obj-id)...)))) ...

(define generate-initial-scene-state...) ...

(define simulate-physics (mem (lambda (scene total_t delta_t)
  (let check_collisions ...)
  (let generate_next_scene_state_at_time...))))
```



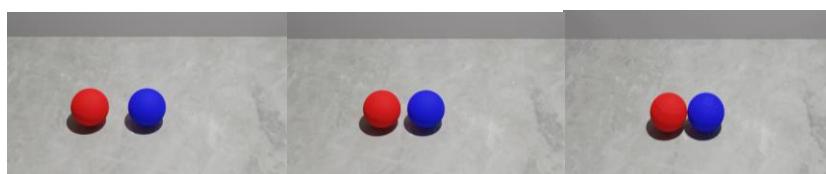
Attributes of objects

```
object-1: { color: red, shape: sphere, mass: 0.2, x: -3, v: 1.0,
            a: -0.05, force: 1.0 ... }
object-2: { color: blue, shape: sphere, mass: 3.0, x: 0, v: 0.0,
            a: 0.0, force: 0.0... }
```

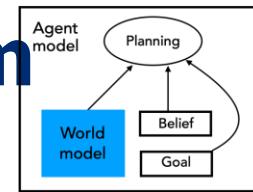
Physics simulation engine

<b>t=1</b> object-1: {..., x: -2.5, v: 0.95...}	<b>t=2</b> object-1: {..., x: -2.0, v: 0.9...}	...	<b>t=10</b> object-1: {..., x: 0.0, v: 0.01...}
---	--	-----	---

Graphics rendering engine

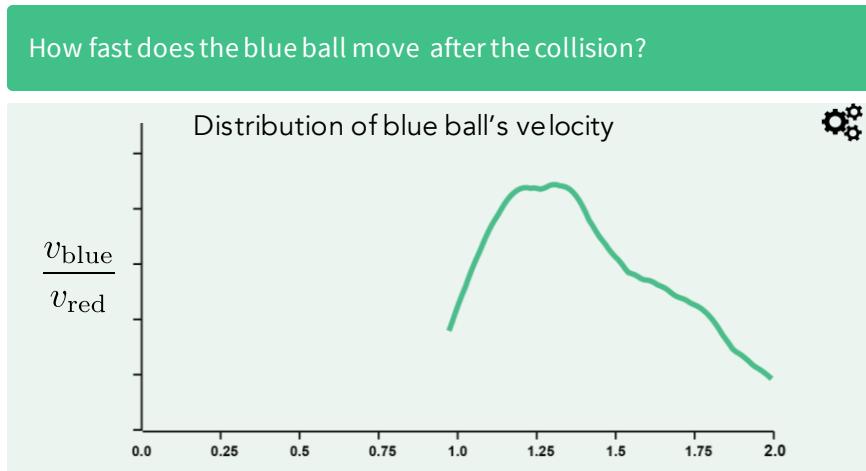


# Reasoning about physical scenes from language



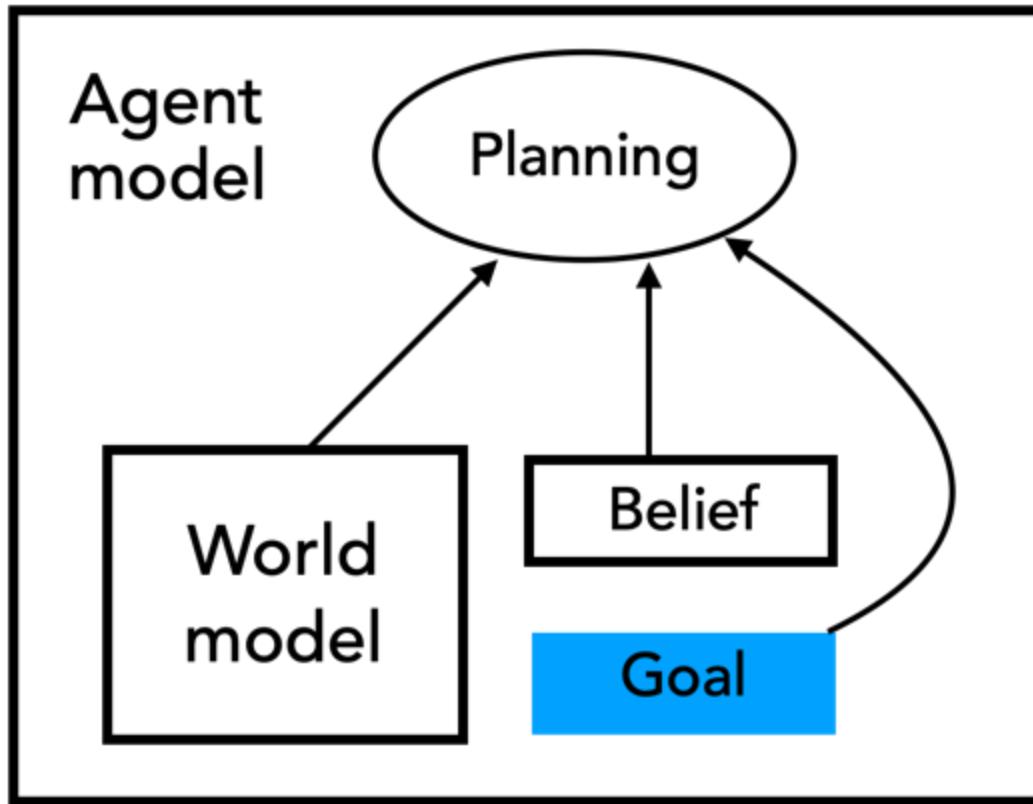
Using LMs to construct WMs via *probabilistic programs* for language reasoning

LM  
Text → World Model → simulate



Probabilistic reasoning → faster than the red ball's initial speed

# Language models as goal/reward?

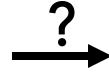


# Goal / reward specification can be difficult



User

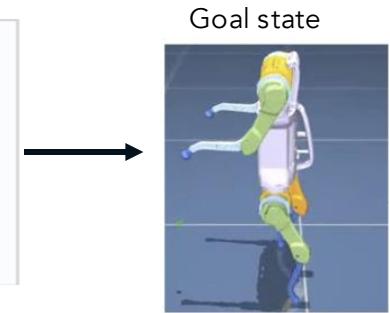
Make robot dog  
stand up on two feet.



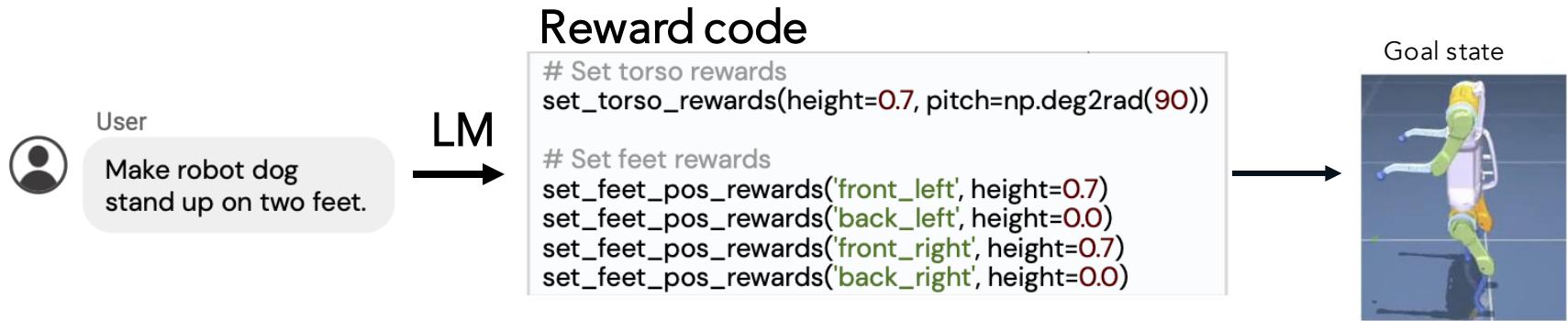
```
# Set torso rewards  
set_torso_rewards(height=0.7, pitch=np.deg2rad(90))  
  
# Set feet rewards  
set_feet_pos_rewards('front_left', height=0.7)  
set_feet_pos_rewards('back_left', height=0.0)  
set_feet_pos_rewards('front_right', height=0.7)  
set_feet_pos_rewards('back_right', height=0.0)
```

Reward written by experts

It is hard ground natural language to the physical world



# Language models as goal/reward



Yu et al. (2023)

# Language models as goal/reward

Why do we need reward at all?

Robot motion control is hard for LMs

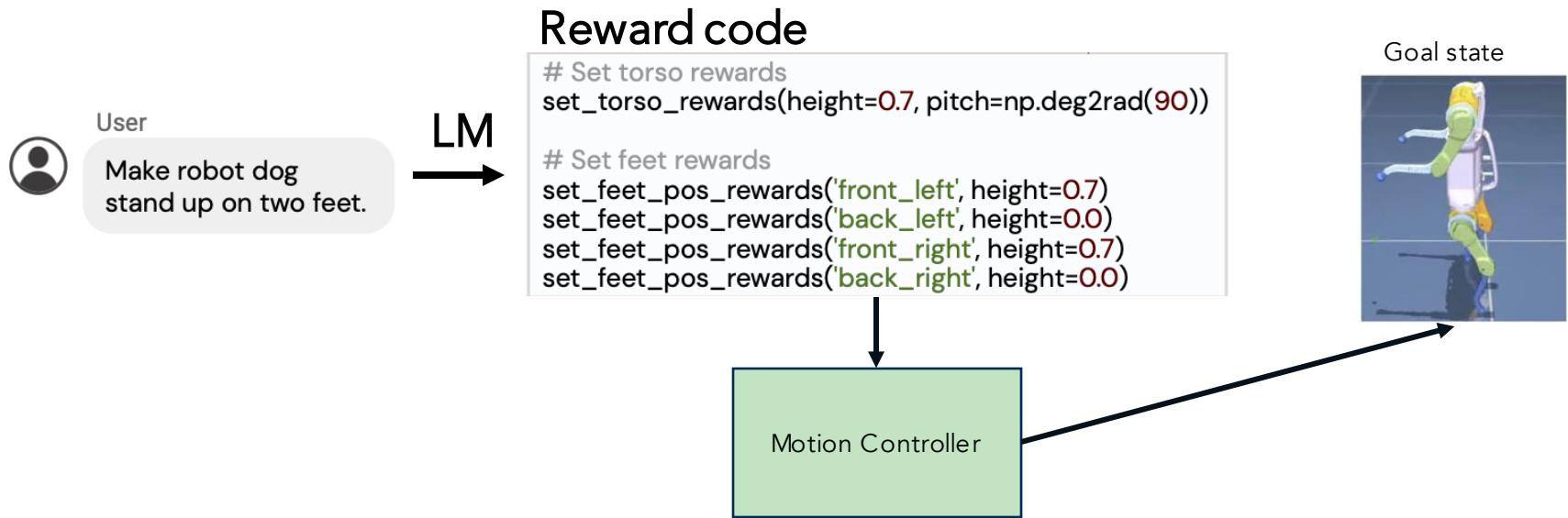


Natural language has a limited capacity in representing the world states and actions



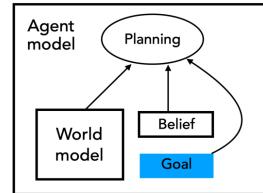
Yu et al. (2023)

# Language models as goal/reward



Yu et al. (2023)

# Language models as goal/reward



- Multi-round instructions to define rewards for complex tasks

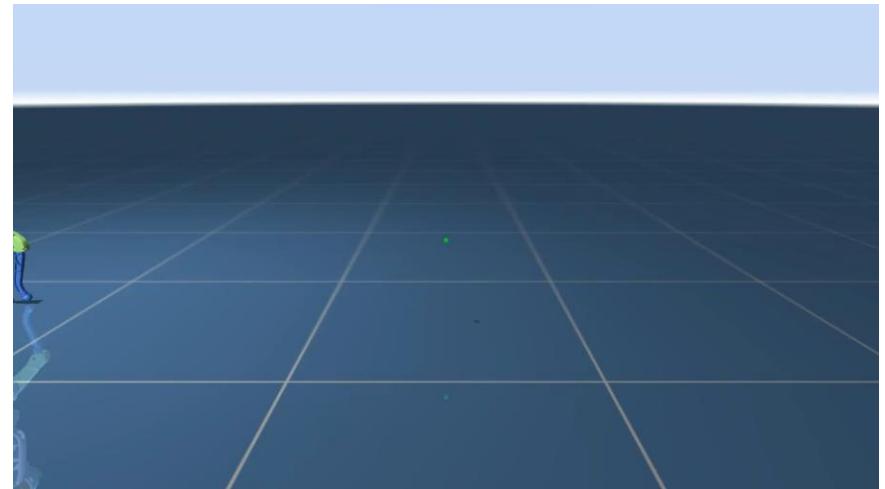
Reward of “moon walk”

Instruction 1: Make the robot stand upright on two back feet like a human.

Instruction 2: Good, you actually don't need to keep the front paws at certain height, just leave it to the controller.

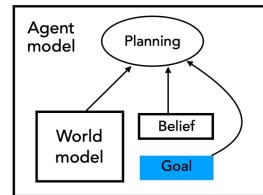
Instruction 3: Good, now make the robot do a moonwalk.

Instruction 4: Moon walk means the robot should walk backward while the feet swings as if they are moving forward. Correct your answer



Yu et al. (2023)

# Language models as goal/reward



- Multi-round instructions to define rewards for complex tasks

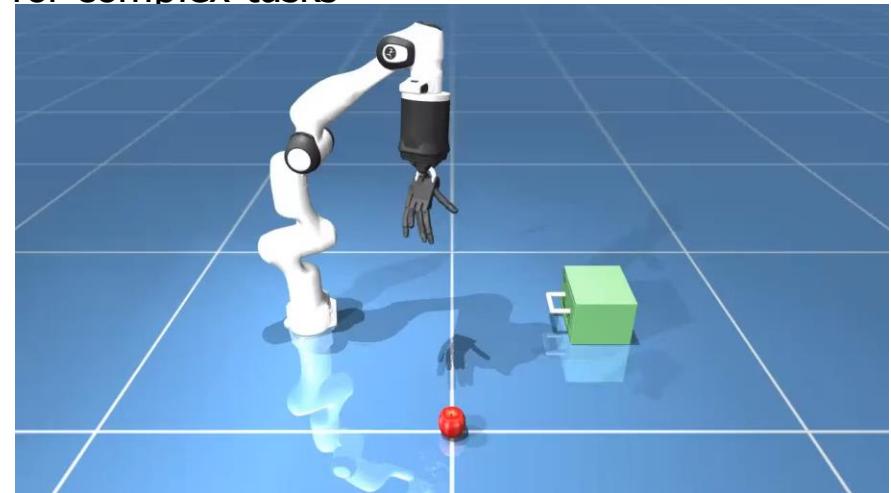
Rewards of a sequence of subtasks  
for a long-horizon task

Instruction 1: Open the drawer.

Instruction 2: Good, now put the apple inside the drawer  
while keep it open.

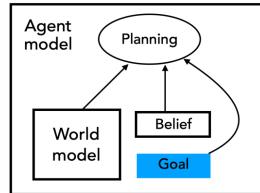
Instruction 3: Good, now release the apple and move hand  
away.

Instruction 4: Now close the drawer.



Yu et al. (2023)

# Language models as goal/reward



- Symbolic goals, e.g., PDDL (Planning Domain Definition Language) goals

Target domain PDDL

Language instruction

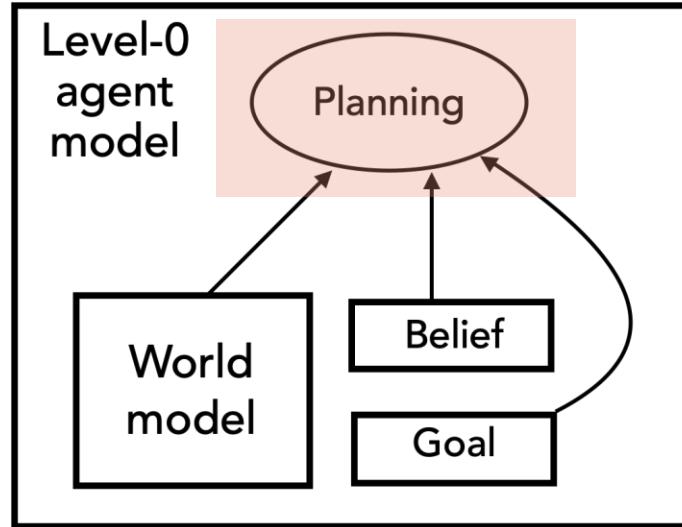
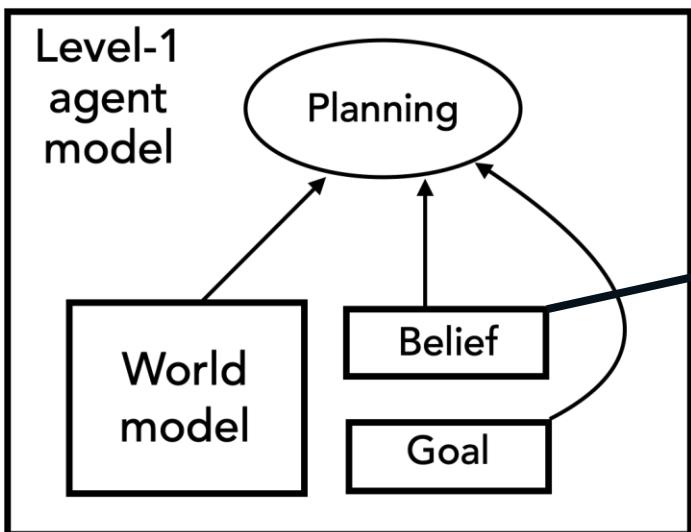
Q: (define (problem 54321)  
(:domain put\_task)  
(:objects  
    Laptop1 - object  
    Desk1 - receptacle  
    ...)  
(:init  
    (objectAtLocation Laptop1 Sofa1)  
    ...))  
{ Put a laptop in the drawer.  
Write the goal specification in PDDL:  
A:

PDDL goal

LM → (:goal  
                  (inReceptacle Laptop1 Drawer1))

Xie et al. (2023)

# Language models for social reasoning



$$P(\text{mind}|\text{state, actions}) \propto P(\text{actions}|\text{state, mind})P(\text{mind})$$

Model-based Theory of Mind via Language Models

# Language models for social reasoning



GPT-4V

(a) ... Since Jennifer is heading towards the cabinet which is said to contain a bag of chips, **but no mention of cupcakes, it suggests that Jennifer does not think there is a cupcake inside that cabinet.**

**Scene:** The microwave holds two cupcakes ... The cabinet is filled with a bag of chips ...

**Actions:** Jennifer heads towards the cabinet and is about to open it.

**Question:** If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?

- (a) Jennifer thinks that there isn't a cupcake inside the cabinet.
- (b) Jennifer thinks that there is a cupcake inside the cabinet.

Accuracy: 12%

# Language models for social reasoning



Model-based Theory of Mind via  
Language models (w/ GPT-J)

$$P(\text{mind}|\text{state, actions}) \propto P(\text{actions}|\text{state, mind})P(\text{mind})$$

Accuracy: 12% → 69%

**Scene:** The microwave holds two cupcakes ... The cabinet is filled with a bag of chips ...

**Actions:** Jennifer heads towards the cabinet and is about to open it.

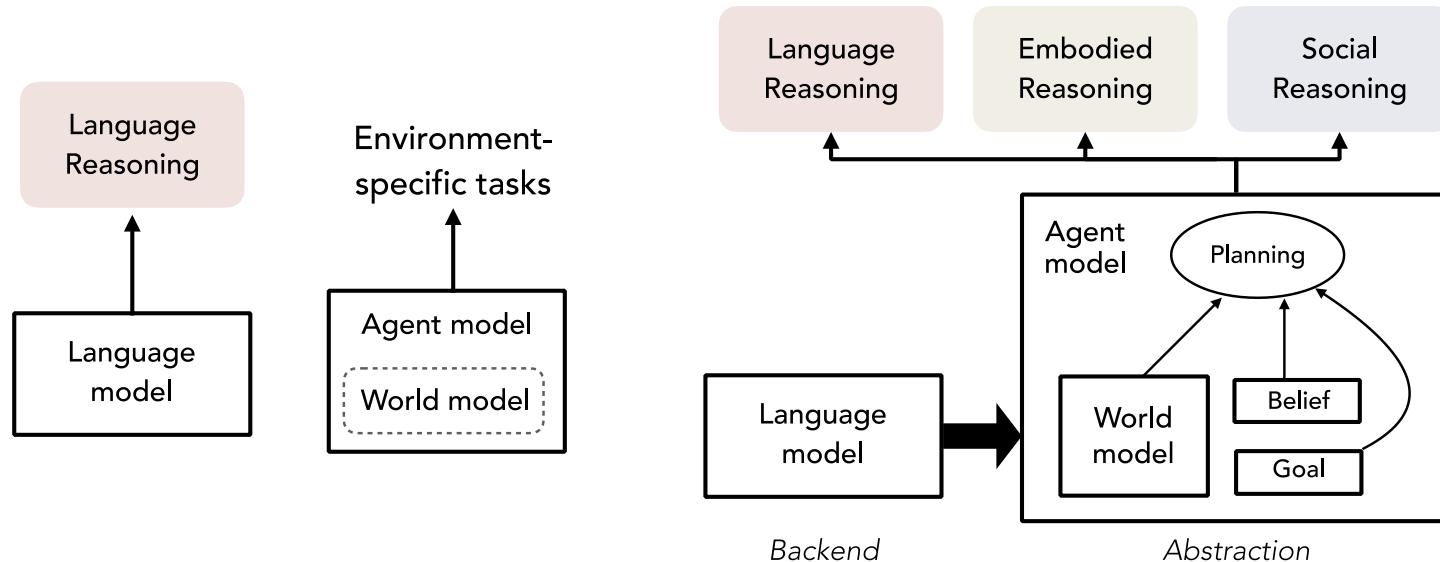
**Question:** If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?

- (a) Jennifer thinks that there isn't a cupcake inside the cabinet.
- (b) *Jennifer thinks that there is a cupcake inside the cabinet.*

Jin et al. (2023)

# Summary

- Language models as the backend for world models and agent models, supporting  
r Conventional paradigms LM backend: more general and open-ended



# Summary

---

- Language models as the backend for world models and agent models, supporting model-based reasoning
- While language alone is not sufficient for many tasks, language models can be connected with models operating on other modalities to achieve multimodal capacity
- There is still a need for manually crafted components (such as belief / memory modules) or conventional methods (such as classical planning)
- Enhancing the language model backend
  - Multimodality?
  - Single-model or modular design?
  - ...
- Questions?

# **“Agency” and Agents**



- <https://robotics-transformer2.github.io/assets/rt2.pdf>
- [https://web.eecs.umich.edu/~justincj/slides/eecs498/WI2022/598\\_WI2022\\_lecture18.pdf](https://web.eecs.umich.edu/~justincj/slides/eecs498/WI2022/598_WI2022_lecture18.pdf)
- <https://web.eecs.umich.edu/~justincj/teaching/eecs498/WI2022/schedule.html>



- [https://www.youtube.com/watch?v=EAd4oQtEJOM&ab\\_channel=HuggingFace](https://www.youtube.com/watch?v=EAd4oQtEJOM&ab_channel=HuggingFace)
- <https://sites.google.com/princeton.edu/cos597f>

# Text-image models Robotic environments

---

# Pixel-based Language Modeling

---

# Multi-modal Models

---

- ❖ Incorporate language, vision, etc?
- ❖ Other types of data?
- ❖ Do we build one unified model or different types of models?

<https://arxiv.org/pdf/2310.06114.pdf>

<https://arxiv.org/pdf/2310.13724.pdf>

[https://www.amazon.science/blog/teaching-household-robots-where-to-find-requested-objects?utm\\_source=email&utm\\_medium=newsletter&utm\\_campaign=science\\_news&utm\\_term=oct\\_2023&utm\\_content=teaching\\_household\\_robots&mkt\\_tok=MzIzLU1GUi01NjgAAAGPJky1b2j59IQMMuOIZ4uKO\\_1PdFji4BNRegPA3cDJICpPsZ6cYRyx21Ld4FnI6Gt9Id2P0Md8SxQTe4XTT\\_ZcykifL3JXjnaU5xeG](https://www.amazon.science/blog/teaching-household-robots-where-to-find-requested-objects?utm_source=email&utm_medium=newsletter&utm_campaign=science_news&utm_term=oct_2023&utm_content=teaching_household_robots&mkt_tok=MzIzLU1GUi01NjgAAAGPJky1b2j59IQMMuOIZ4uKO_1PdFji4BNRegPA3cDJICpPsZ6cYRyx21Ld4FnI6Gt9Id2P0Md8SxQTe4XTT_ZcykifL3JXjnaU5xeG)

- <https://arxiv.org/pdf/2310.10645.pdf>

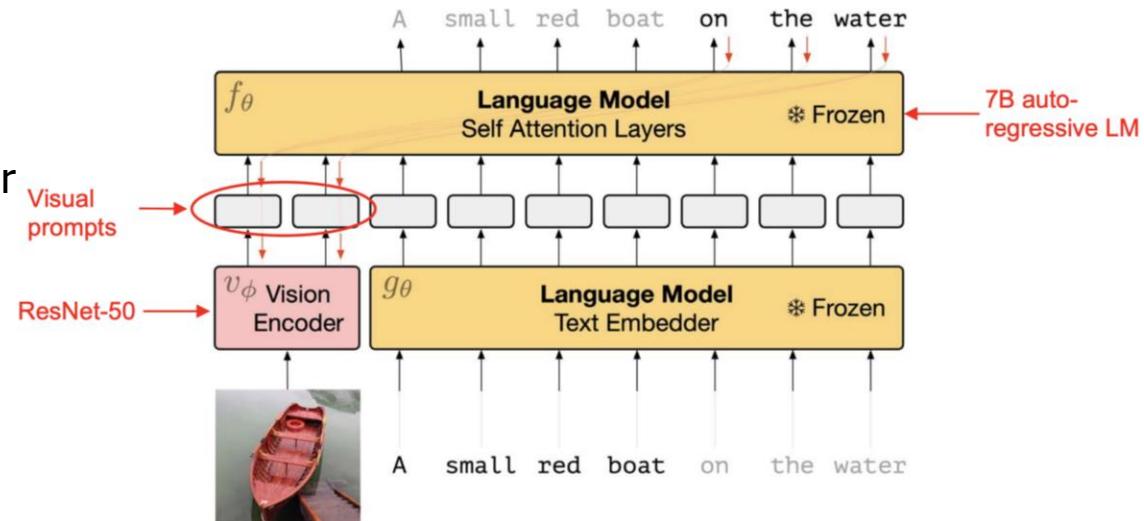
# Multimodal Prompting

---

- CLIP addresses ImageNet, but what if we introduce novel images / text / classes?
- DALL-E addresses controllable image generation, but still no ICL
- LLMs are transferrable to new tasks via prompting with examples
- Can we do the same thing in a multimodal setting, e.g. vision + language?

# Multimodal Few-Shot Learning with Frozen Language Models

- Preserves pretrained LM's ability by freezing it during cross-modal training
  - GPT-2-esque (7B)
  - Pretrained on C4
- Next, consider a vision encoder
  - Input images (e.g. WxHxC)
  - Output N tokens



# Multimodal Few-Shot Learning with Frozen Language Models

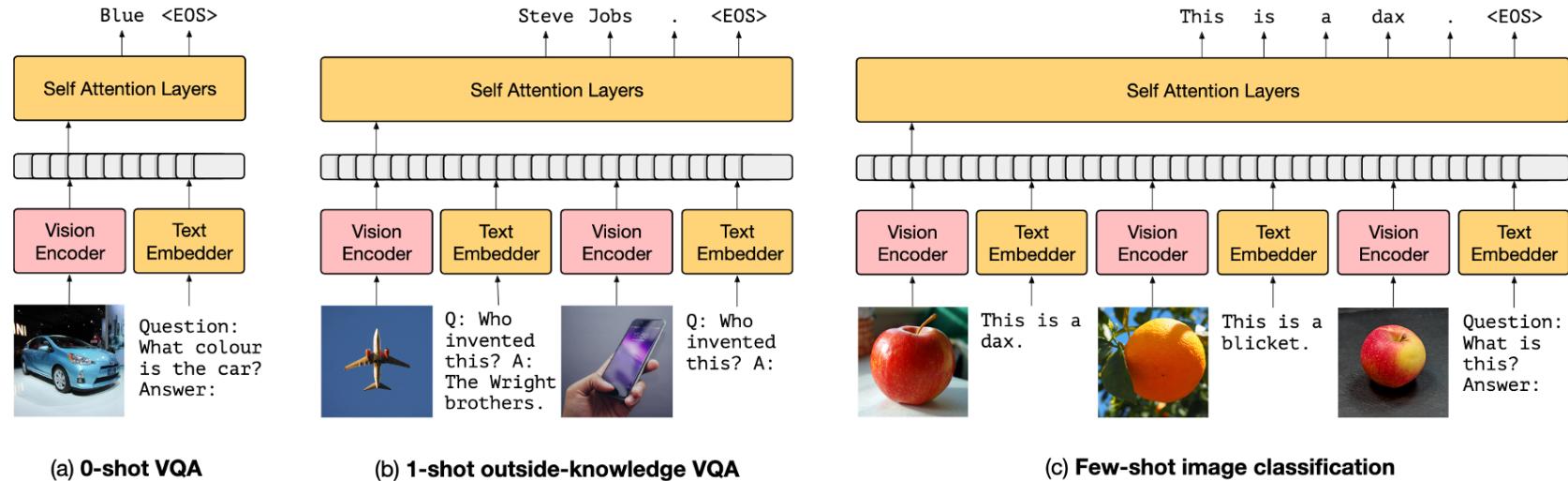
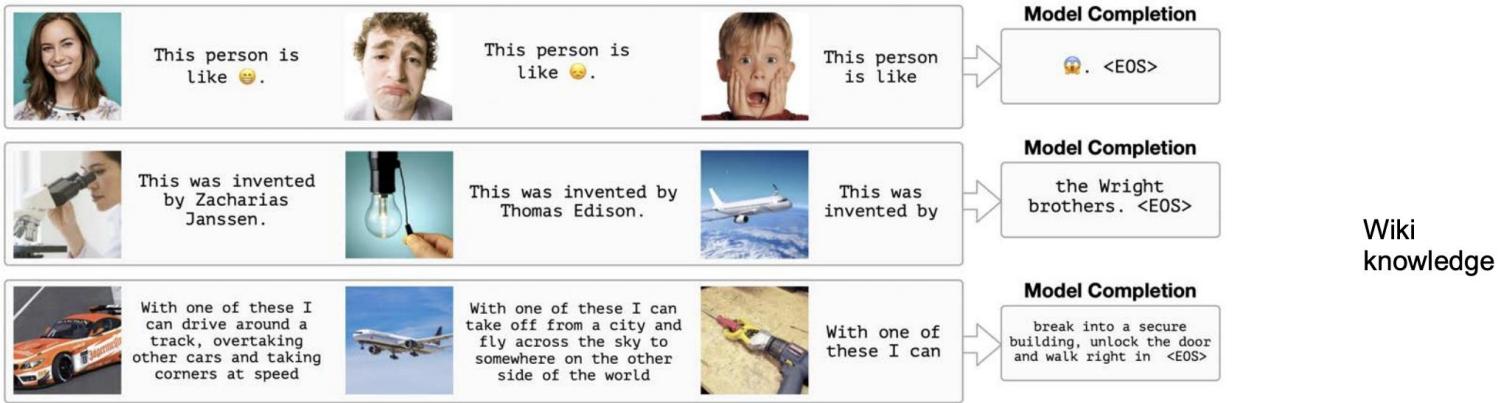


Figure 3: Inference-Time interface for *Frozen*. The figure demonstrates how we can support (a) visual question answering, (b) outside-knowledge question answering and (c) few-shot image classification via in-context learning.

# Multimodal Few-Shot Learning with Frozen Language Models



- Reasonably good zero/few-shot performance, but still underperform SOTA: limited multimodal data? (3M); LM is relatively small? (7B)

n-shot Acc.	n=0	n=1	n=4	$\tau$
<b>Frozen</b>	29.5	35.7	38.2	✗
<b>Frozen scratch</b>	0.0	0.0	0.0	✗
<b>Frozen finetuned</b>	24.0	28.2	29.2	✗
<b>Frozen train-blind</b>	26.2	33.5	33.3	✗

VQAv2

n-shot Acc.	n=0	n=1	n=4	$\tau$
<b>Frozen</b>	5.9	9.7	12.6	✗
<b>Frozen 400mLM</b>	4.0	5.9	6.6	✗
<b>Frozen finetuned</b>	4.2	4.1	4.6	✗
<b>Frozen train-blind</b>	3.3	7.2	0.0	✗

	n=0	n=1	n=4	$\tau$
<b>Frozen VQA</b>	48.4	—	—	✓
<b>Frozen VQA-blind</b>	39.1	—	—	✓
<b>Oscar [23]</b>	73.8	—	—	✓

OKVQA

Large gap w/  
SOTA

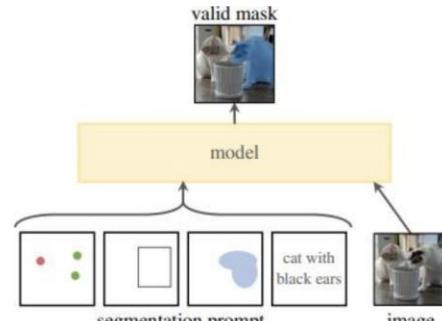
# A note on In-context learning

---

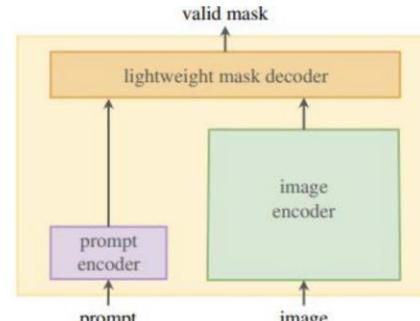
- TBD

# Interactive Interface for Vision

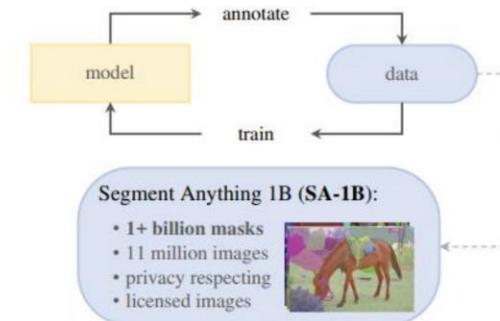
- **SAM: Segment Anything**
  - Promptable segmentation



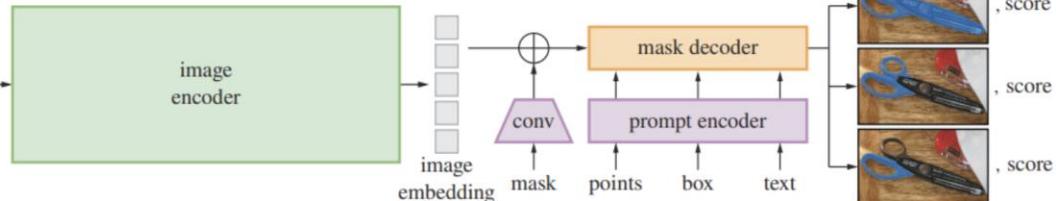
(a) Task: promptable segmentation



(b) Model: Segment Anything Model (SAM)

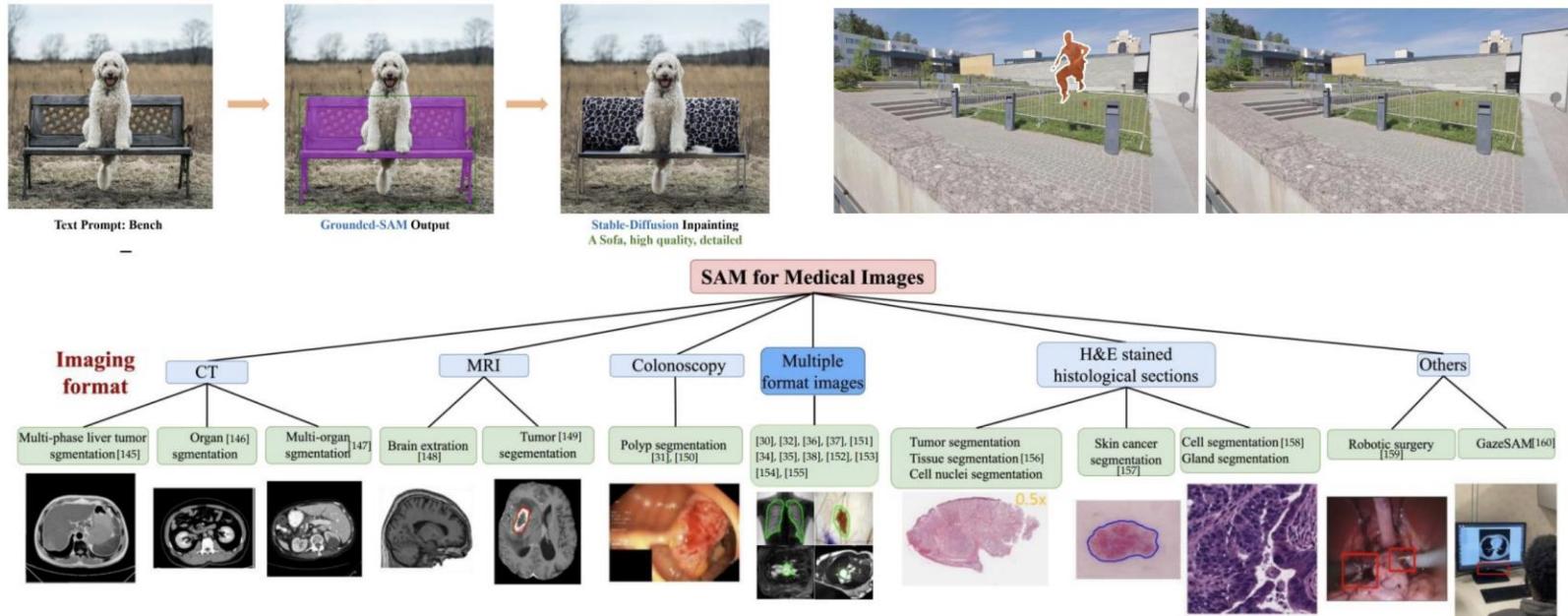


(c) Data: data engine (top) & dataset (bottom)



# Interactive Interface for Vision

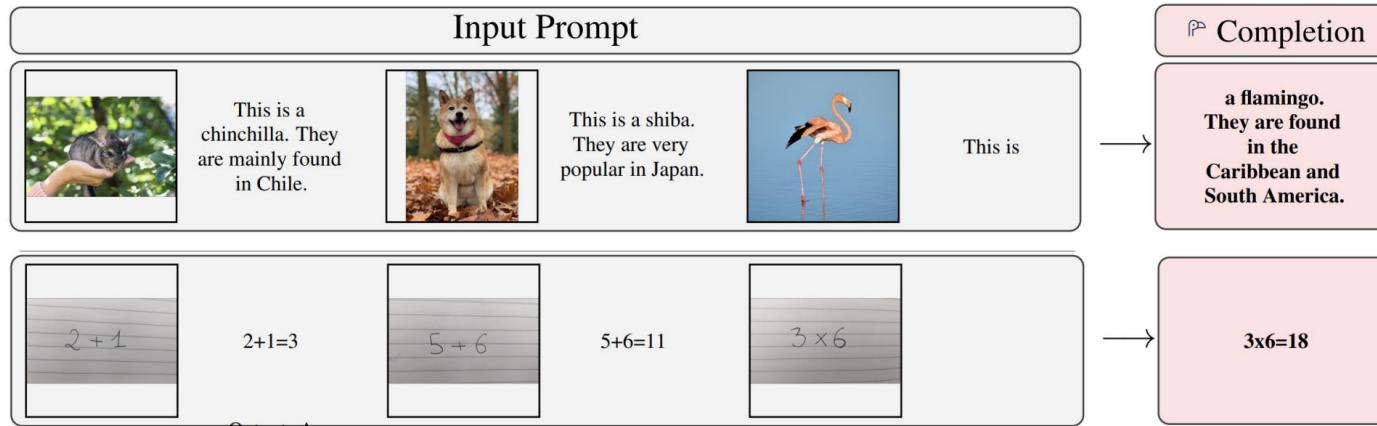
- **SAM:** Segment Anything



## Example 2: LMM with Interleaved Image-Text Data

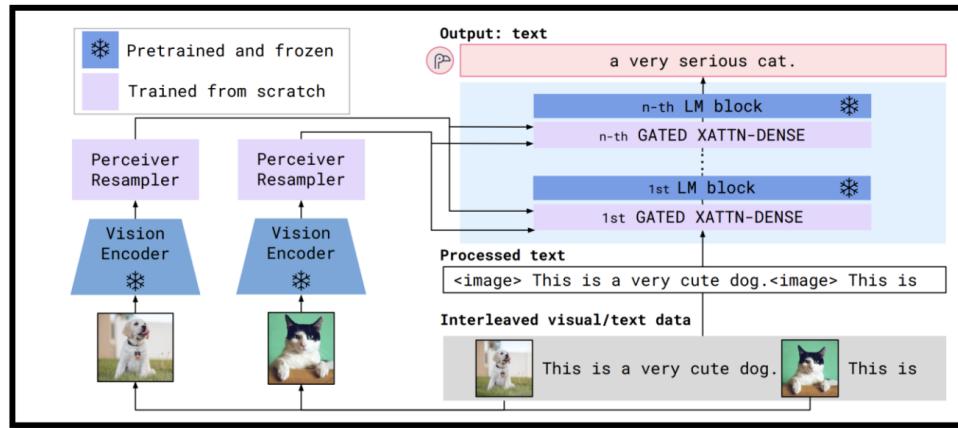
- Flamingo: Multimodal In-Context-Learning

Emerging Property



## Example 2: LMM with Interleaved Image-Text Data

- Flamingo:



Language Model

Connection Module

Vision Encoder

Pre-trained: 70B Chinchilla

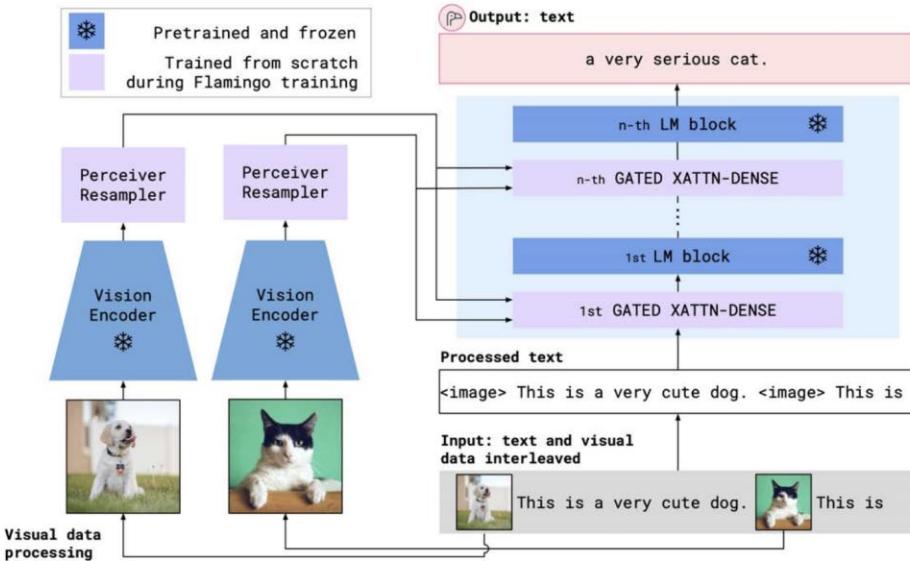
Perceiver Resampler  
Gated Cross-attention + Dense

Pre-trained: Nonnormalizer-Free ResNet (NFNet)

# Flamingo



- A frozen 70B pre-trained LM + a frozen pre-trained ResNet.
- Trained w/ image/video-text pairs, along with interleaved image-text data (M3W), which is important for in-context learning.



ALIGN: 1.8B +  
LTIP: 312M images  
VTP: 27M videos

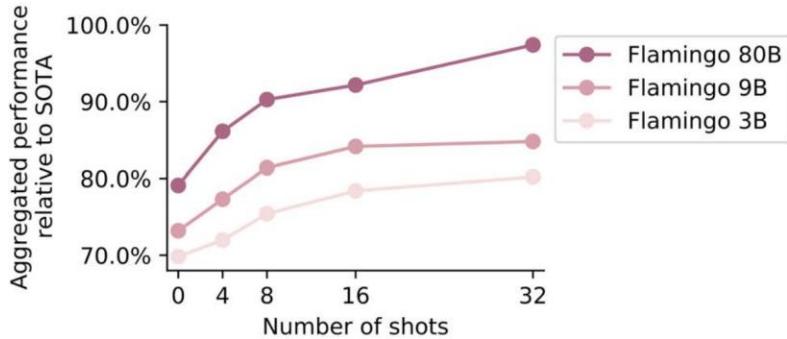


Multi-Modal Massive Web (M3W) dataset  
[ $N>1$ ,  $T=1$ ,  $H$ ,  $W$ ,  $C$ ]  
M3W: 43M webpages (185M images)

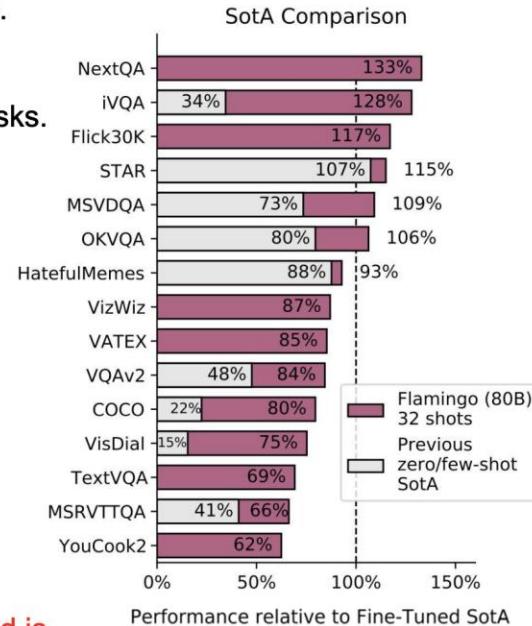
# Flamingo



- Left: larger model works better; more in-context examples helps.
- Right: thanks to larger model and more training data, the model achieves comparable or better results than SOTA on multiple tasks.



Expensive! The 80 model has 10B trainable parameters and is trained with 1536 TPUs for 15 days.

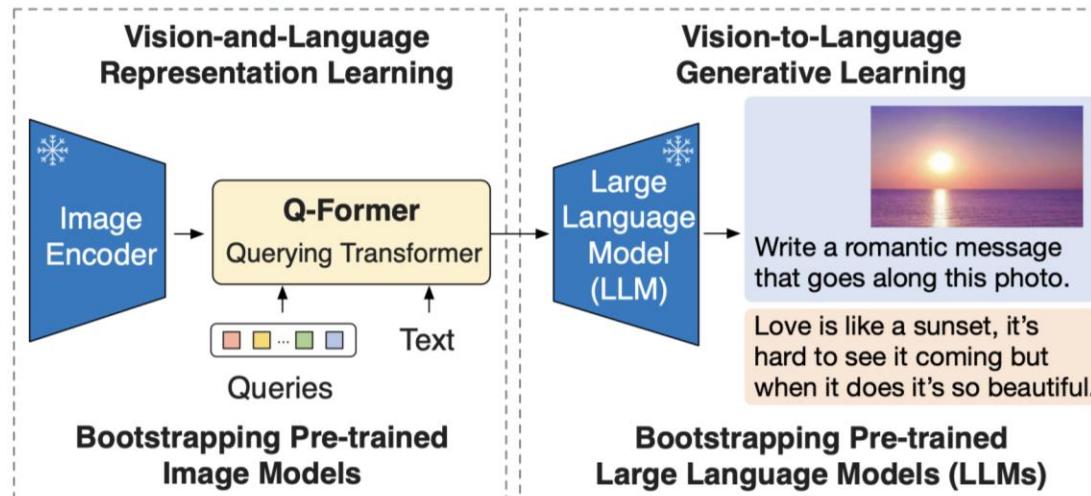


# BLIP2

Reduce your computational cost by not re-learning about language or images.

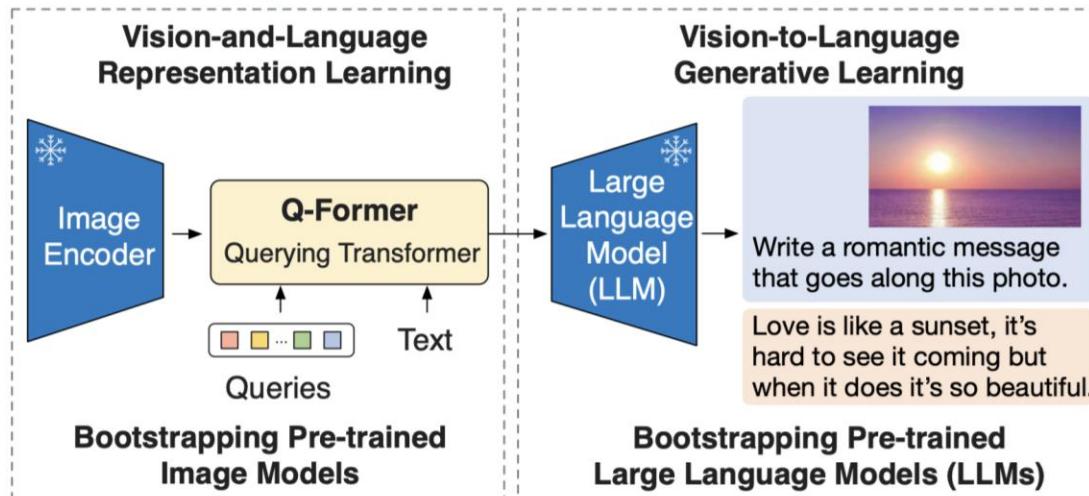
- Architecture:

- Pre-trained and frozen: (i) an image encoder; (ii) a language model
- Adapted to the task: light-weight modules (Q-Former) to connect (i) and (ii)



# BLIP2

- Trained on 129M image-text pairs.



# BLIP2:

Models	#Trainable Params	Open-sourced?	Visual Question Answering		Image Captioning		Image-Text Retrieval	
			VQAv2 (test-dev) VQA acc.	NoCaps (val) CIDEr	SPICE	TR@1	Flickr (test) IR@1	
BLIP (Li et al., 2022)	583M	✓	-	113.2	14.8	96.7	86.7	
SimVLM (Wang et al., 2021b)	1.4B	✗	-	112.2	-	-	-	
BEIT-3 (Wang et al., 2022b)	1.9B	✗	-	-	-	94.9	81.5	
Flamingo (Alayrac et al., 2022)	10.2B	✗	56.3	-	-	-	-	
BLIP-2	188M	✓	65.0	121.6	15.8	97.6	89.7	

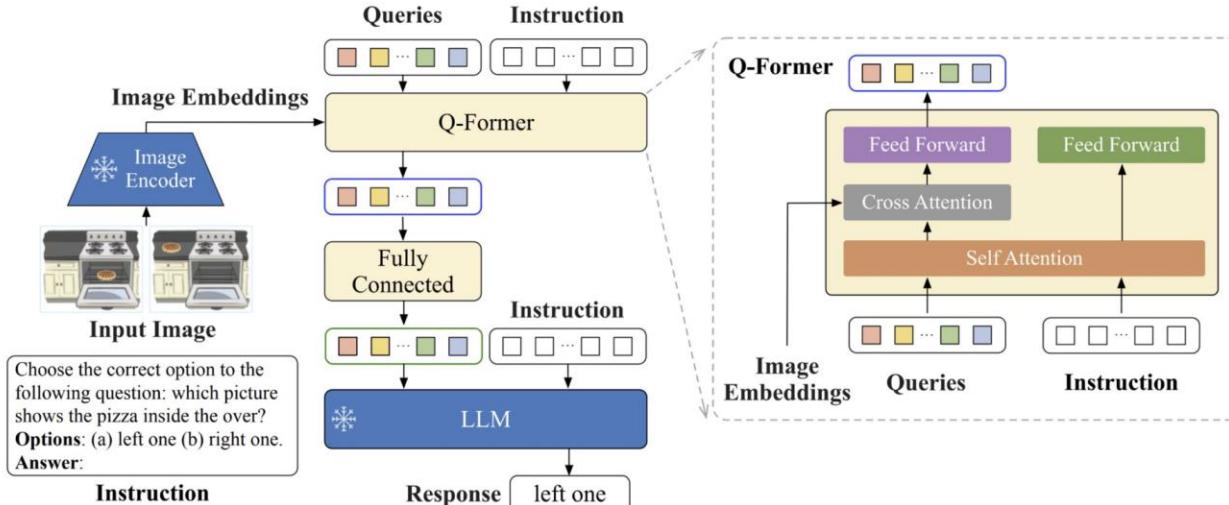
Results on zero-shot vision-language tasks.

# In Blip & InstructBLIP

salesforce

Q-Former is used to align the visual features into the LLM embedding space.

- BL

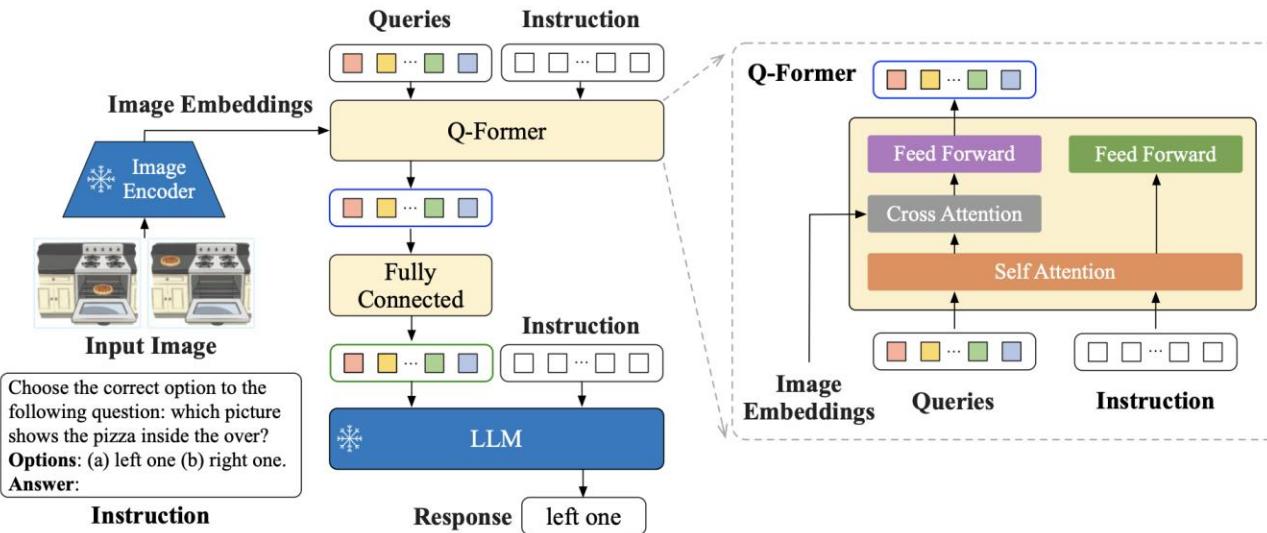


Credit: Wenliang Dai et al. Instructblip: Towards general-purpose vision language models with instruction tuning

# InstructBLIP



- **Architecture:** same as BLIP-2, except instruction text is added to Q-Former for instruction-aware visual feature extraction
- **Training:** BLIP-2 pre-training + Instruction Finetuning on 13 held-in datasets
- **Evaluation:** on both held-in and held-out datasets



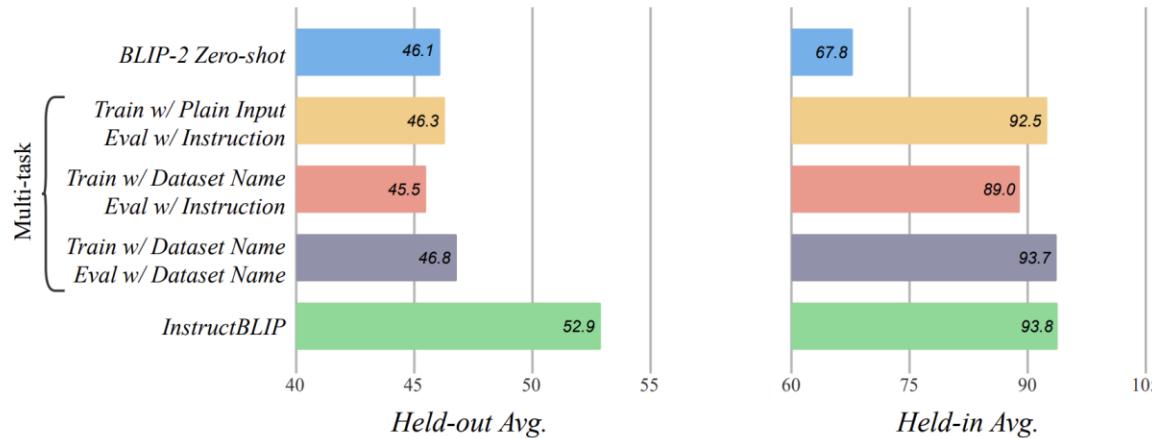
InstructBLIP Model Overview

# InstructBLIP: Instruction Tuning vs. Multi-task



- Instruction tuned model excels in unseen datasets and tasks

Strategy	Template (use VQAv2 dataset as an example)
Instruction Tuning	<ul style="list-style-type: none"><li>&lt;image&gt; Question: {question} Short answer:</li><li>&lt;image&gt; What is the answer to the following question? {question}</li><li>&lt;image&gt; Based on the image, respond to this question with a short answer: {Question}. Answer:</li><li>...</li></ul>
Multi-task	<ul style="list-style-type: none"><li>Plain text: {image} {question} → {answer}</li><li>Dataset Name: {image} [Visual question answering:VQAv2] {question} → {answer}</li></ul>



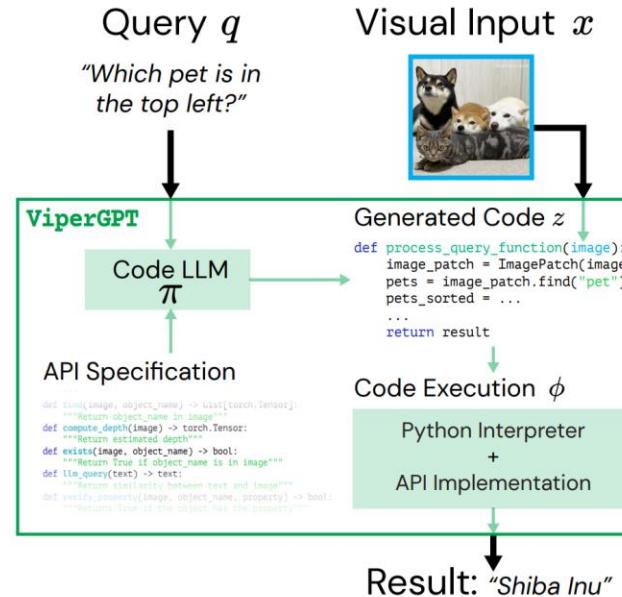
338

# Planning: Composing Tools via Code Generation

Visual Programming: Compositional visual reasoning without training



ViperGPT: Visual Inference via Python Execution for Reasoning



# Improving Accuracy in Tool Using: Self-Assessment

AssistGPT: A General Multi-modal Assistant that can Plan, Execute, Inspect, and Learn



# Envisioning the future

---

- More modalities — combinations of video (2D, 3D), text, code, etc.
- Large models and more efficient scaling
- More breath — more data and more types of data
- Interaction with physical world — models with hands and actuators
- Better personalization — these agents should serve your  needs
- Better human-machine teaming
  - CoPilot for coding
  - CoPilot for writing
  - ....
  - CoPilot for life!!!

- For coding: [lec19 \(princeton.edu\)](#)
- For VLMS: [Nov 21 Flamingo: A Visual Language Model for Few-Shot Learning \(princeton.edu\)](#)

# Apple MM model

- <https://arxiv.org/pdf/2403.09611.pdf>