Commonsense Reasoning in Natural Language Processing

CPSC 532V

Lecture 11: Multimodal Commonsense

Instructor: Vered Shwartz
Presented by:
Sahithya Ravi: sahiravi@cs.ubc.ca,
Aditya Chinchure : aditya10@cs.ubc.ca

Outline

Reasoning about vision and language

- Motivation
- Visual Commonsense Reasoning tasks
- Vision and language representations and models
- Open problems and future directions

Outline

Reasoning about vision and language

- **Motivation**
- Visual Commonsense Reasoning tasks
- Vision and language representations and models
- Open problems and future directions

Can you learn meaning only from text?

A monkey grabbed a plastic bag and jumped out the window of a moving bus.

Can you learn meaning only from text?

A monkey grabbed a plastic bag and jumped out the window of a moving bus.

💭 **Why did the monkey grab the bag?**	(Stealing food? Curious?)
💭 **How did it look while jumping?**	(Was it frantic, playful, or scared?)
💭 **What was inside the bag?**	(Food?)
💭 **What were the humans in the scene doing?**	(Chasing it? Ignoring it?)

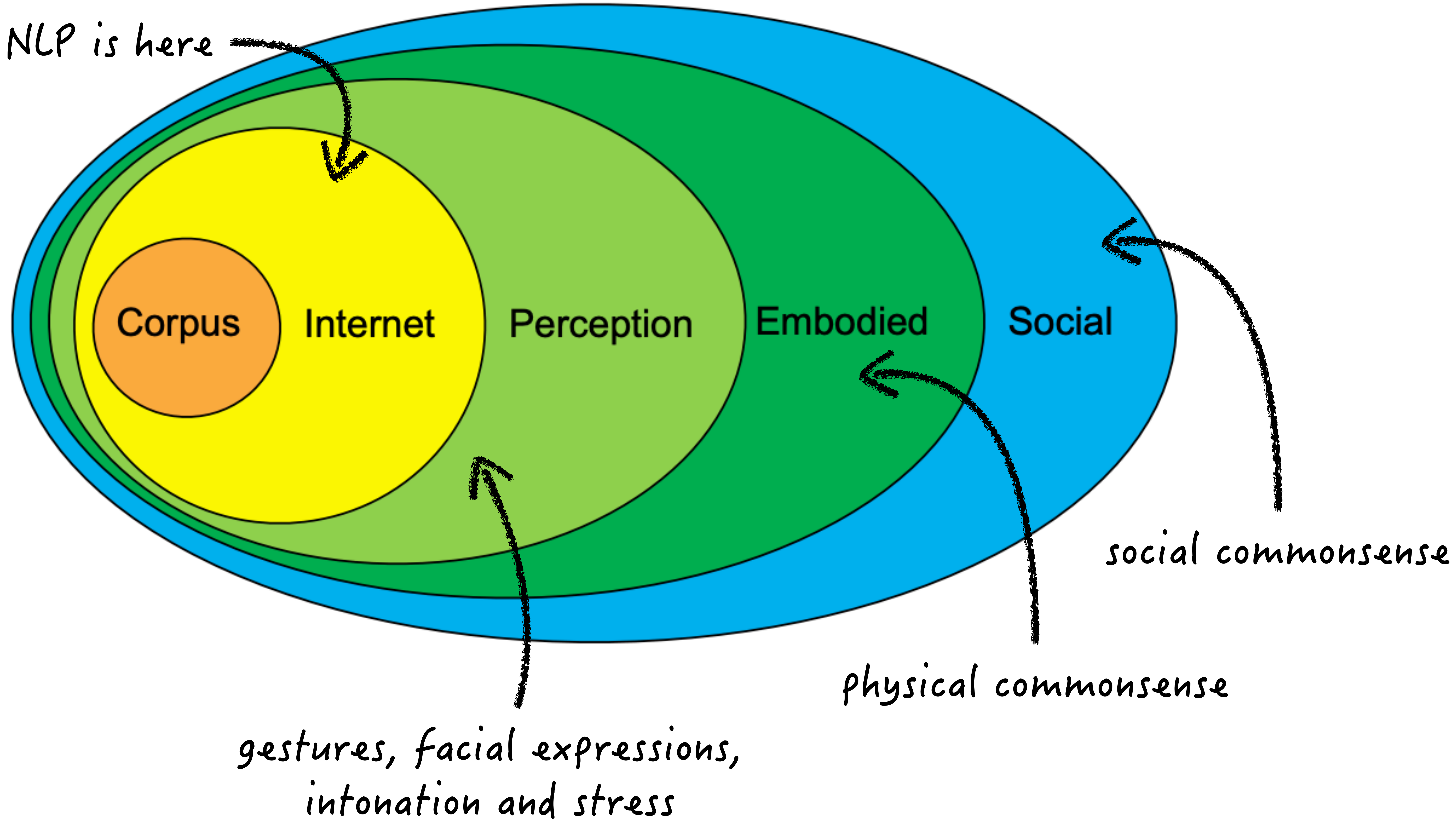
Can you learn meaning only from text?

A monkey grabbed a plastic bag and jumped out the window of a moving bus.

🤣😱

- 💭 **Why did the monkey grab the bag?**
- 💭 **How did it look while jumping?**
- 💭 **What was inside the bag?**
- 💭 **What were the humans in the scene doing?**

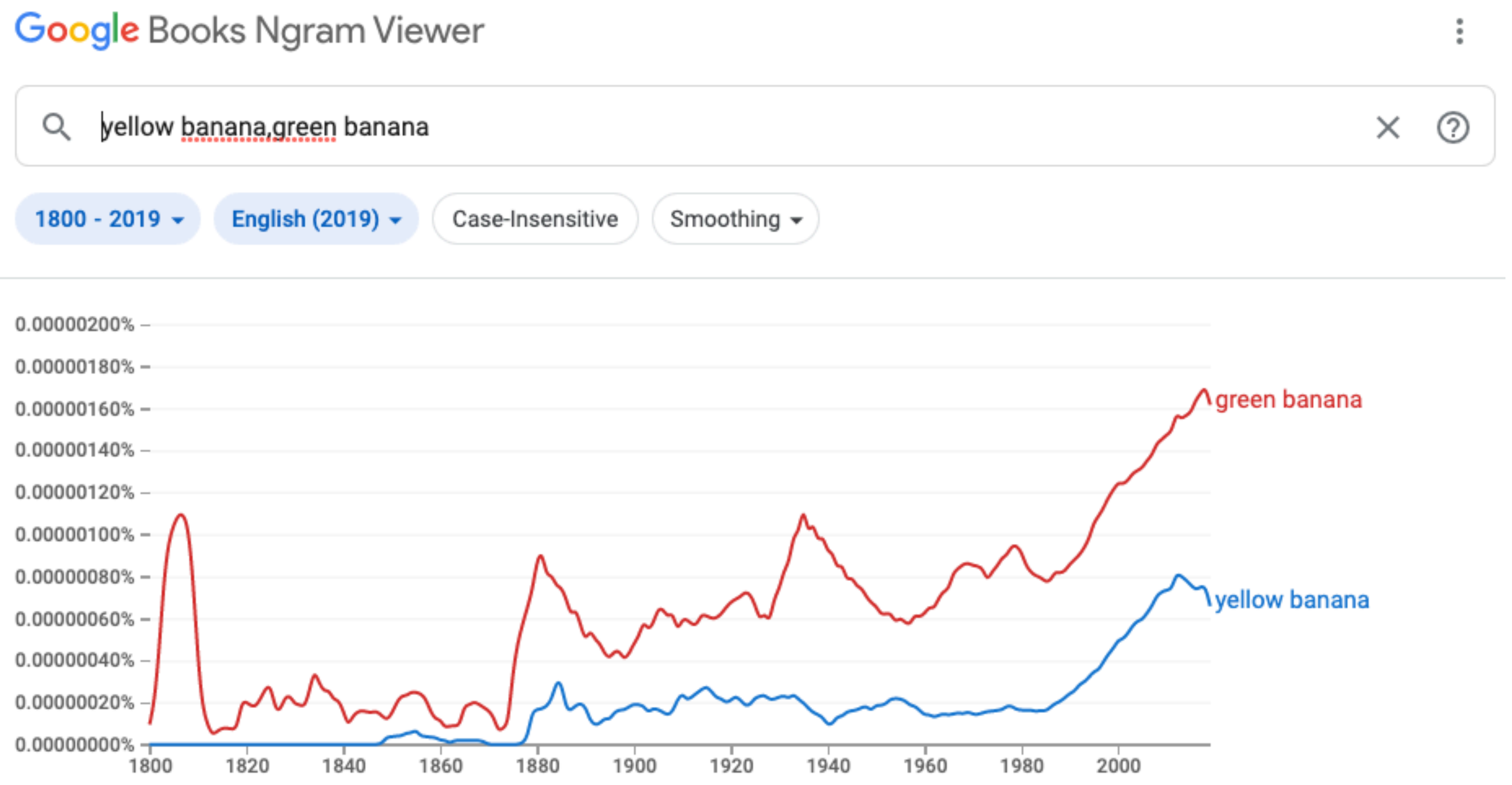
Can you learn meaning only from text?



Experience Grounds Language. Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, Joseph Turian. EMNLP 2020.

What is the colour of a banana?

Text has a predominant occurrence of green bananas...



Google Books Ngram Viewer

yellow banana,green banana

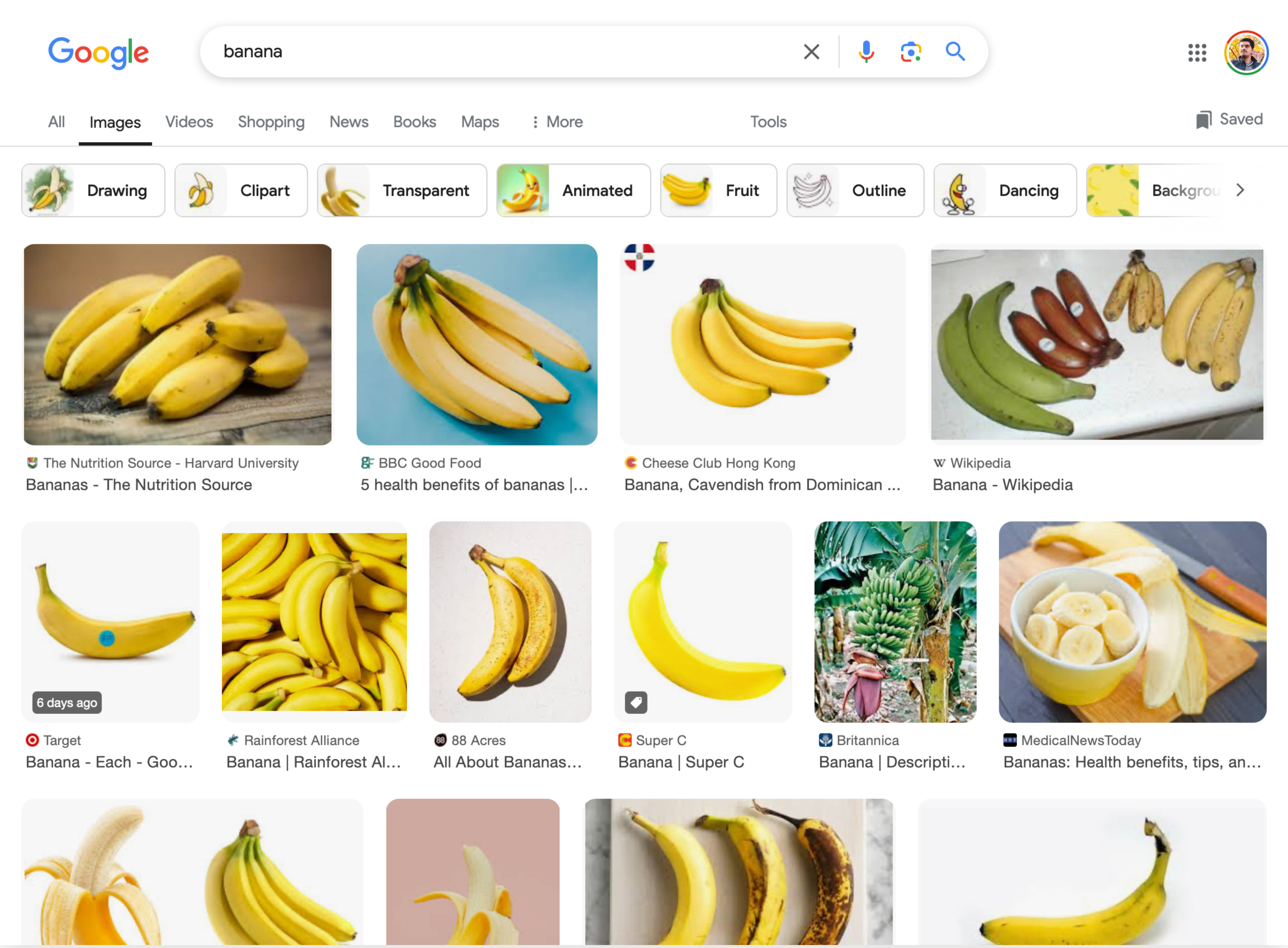
1800 - 2019 ▾ English (2019) ▾ Case-Insensitive Smoothing ▾

✘

Reporting bias
(Gordon and Van Durme, 2013)

What is the colour of a banana?

Text has a predominant occurrence of green bananas... but not in images

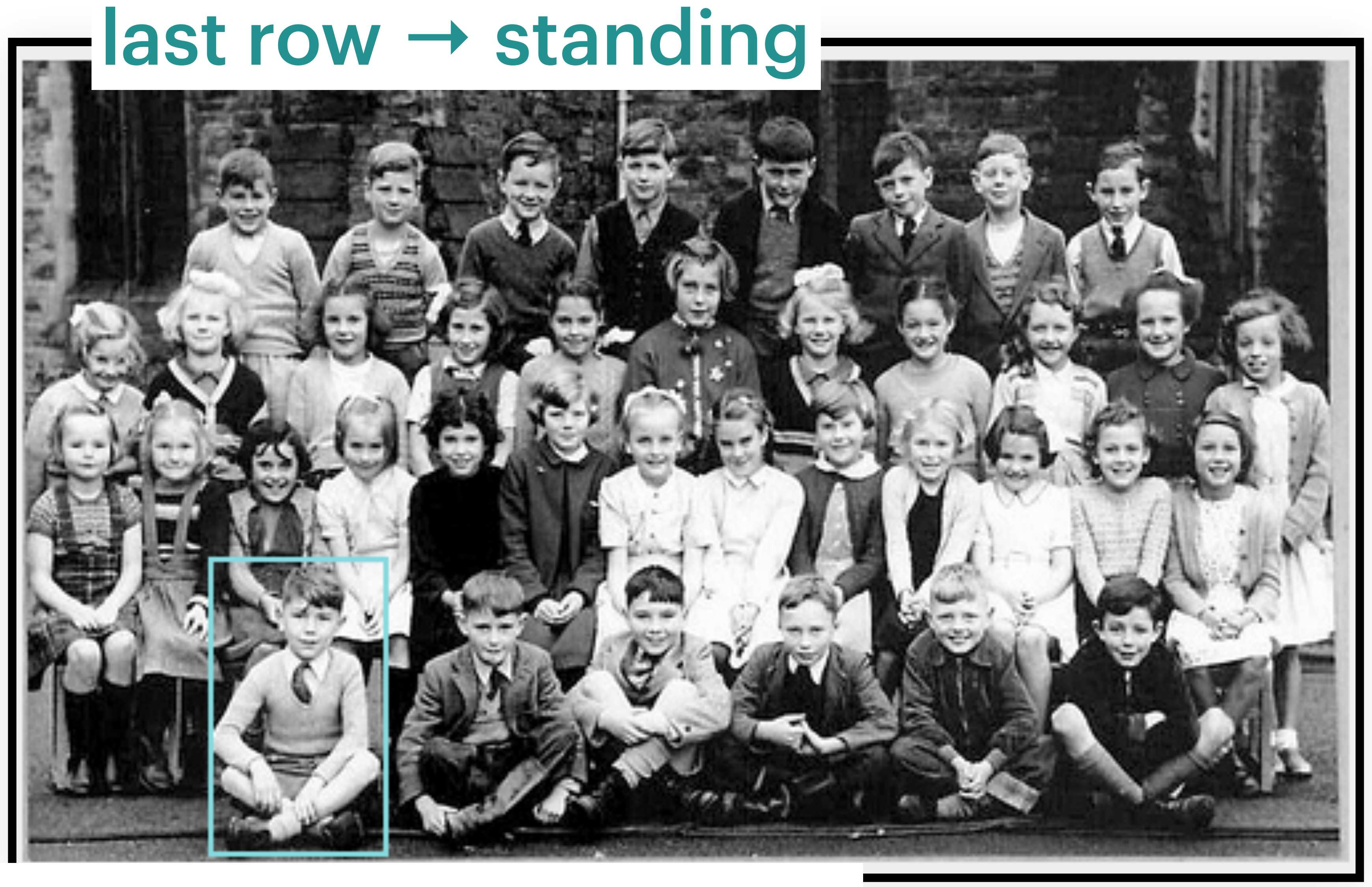


Acquiring Commonsense Knowledge

Sources of Knowledge

- 1️⃣ from language
- 2️⃣ from images
- 3️⃣ from video
- 4️⃣ from audio

Photograph of 6th grade students from the batch of 1995 at ABC School...

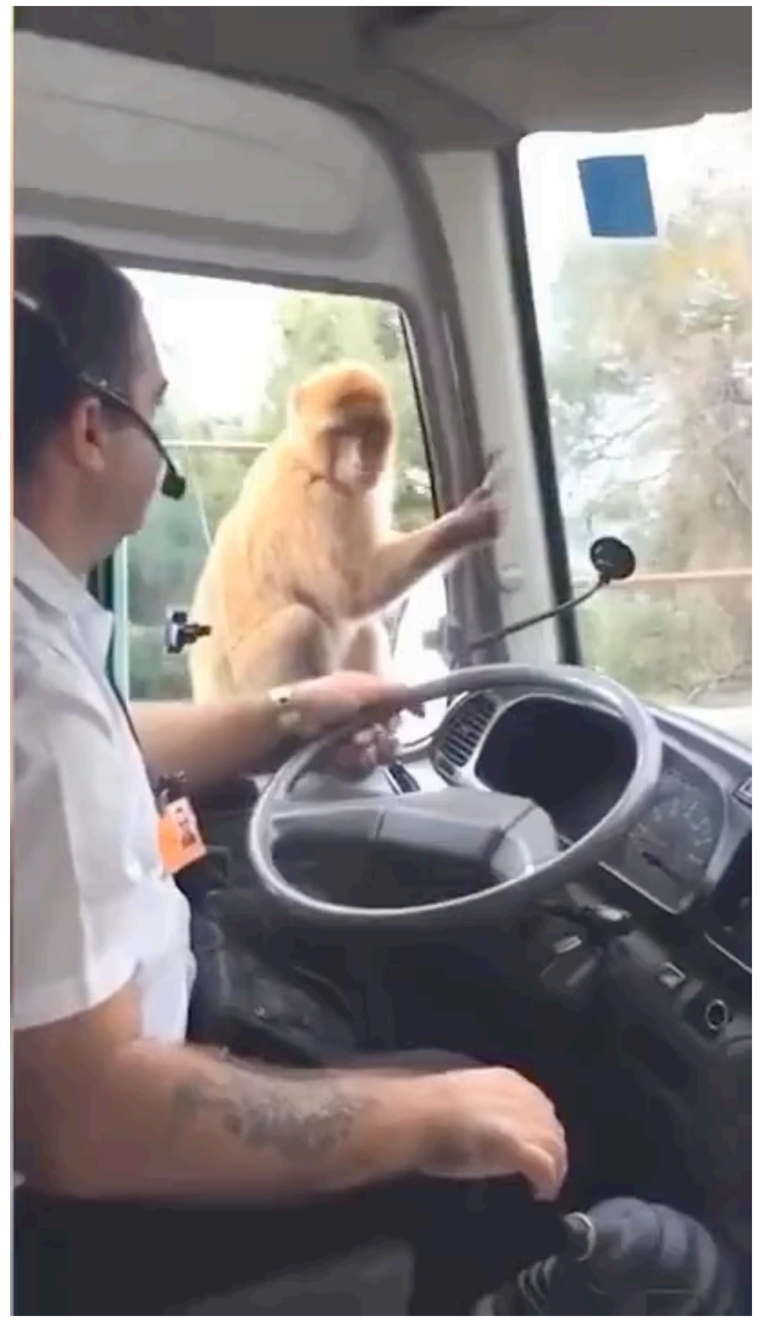


last row → standing

front row → cross-legged

Multimodal Learning!

Multimodal Model Skills?



- Perception - How many humans are there?
- Causal Reasoning – Why did the monkey jump?
- Temporal Understanding – What happened before and after?
- Physical Intuition – Could a monkey safely jump from a moving bus?
- Social & Commonsense Knowledge – Was the monkey stealing or playing?

Multimodal Models need to *see* 👁️, *interpret* 💡, and *reason* 🧐

Outline

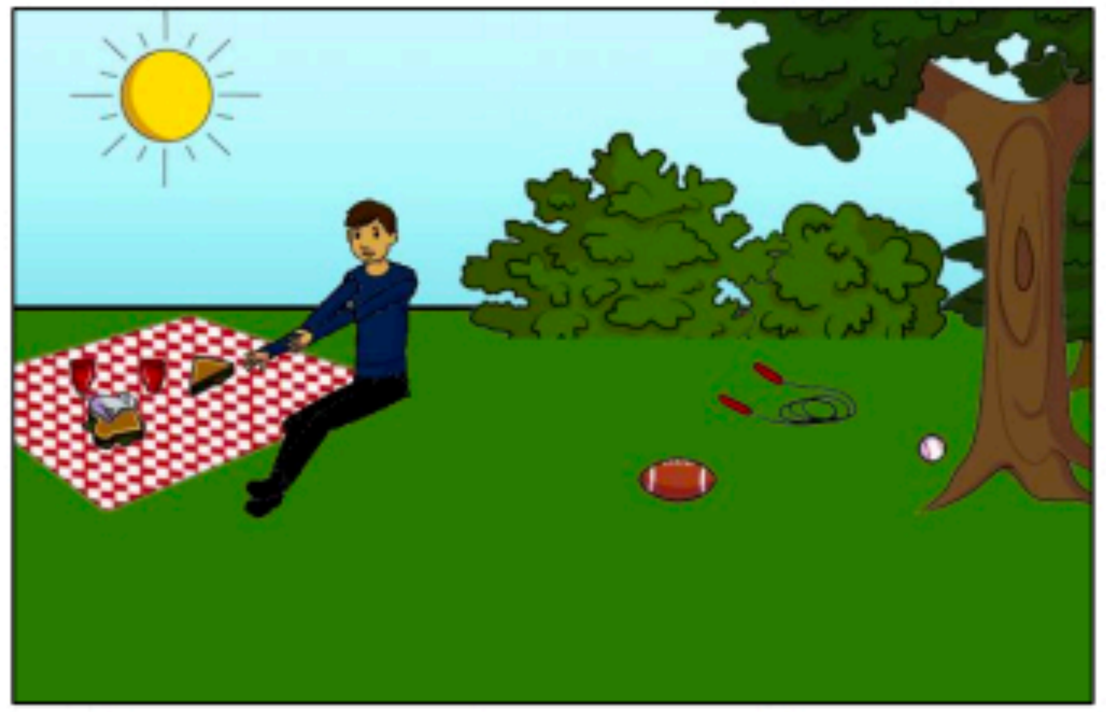
Reasoning about vision and language

- Motivation
- **Visual Commonsense Reasoning tasks**
- Vision and language representations and models
- Open problems and future directions

Visual Question Answering

What color are her eyes?
What is the mustache made of?

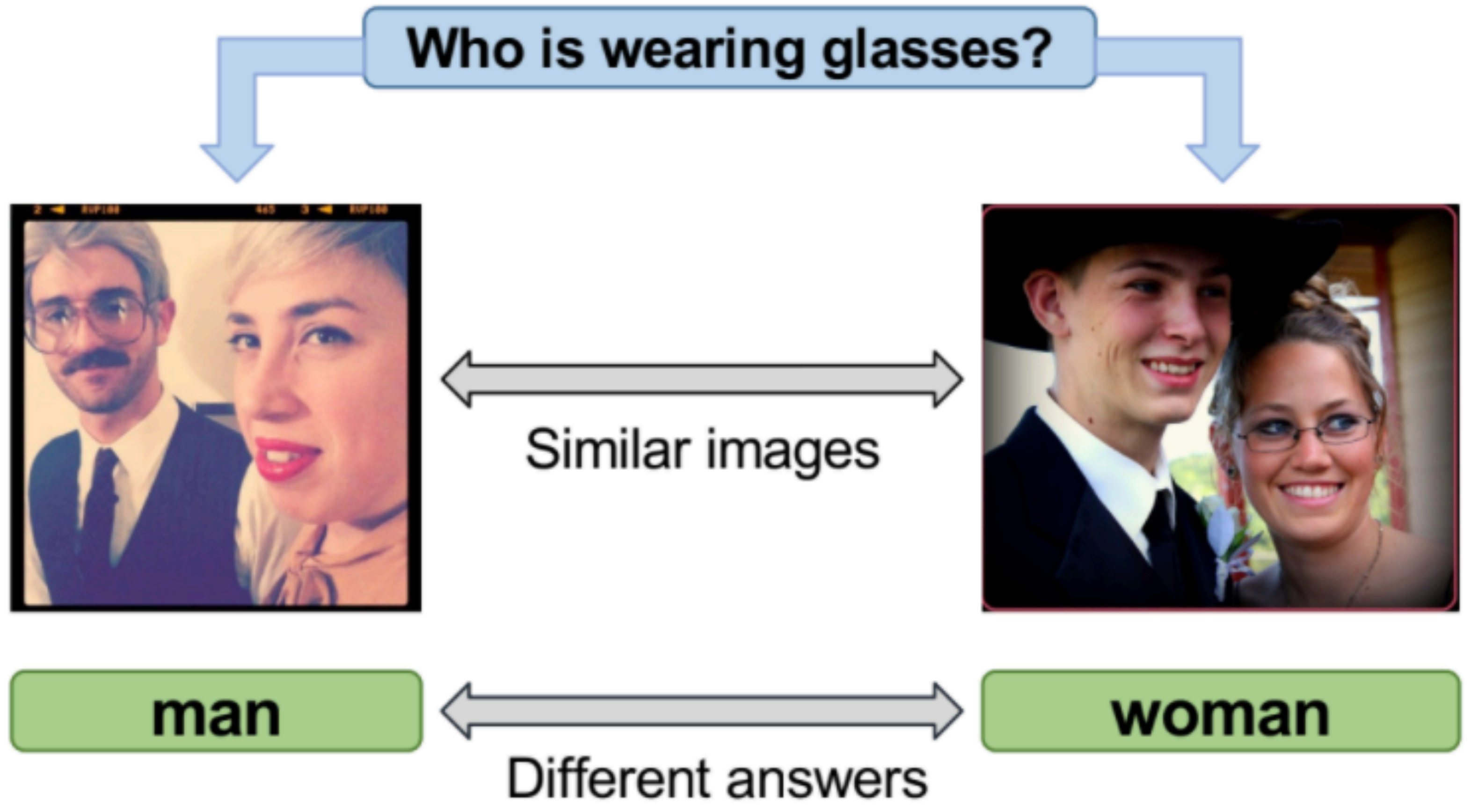
How many slices of pizza are there?
Is this a vegetarian pizza?


Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

- Open-ended questions about images.
- Require an understanding of vision, language and some commonsense.
- >200K images, >1M questions, >11M candidate answers
- The questions are mostly about **what is in the images.**
- Automatic evaluation

$$\mathrm{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

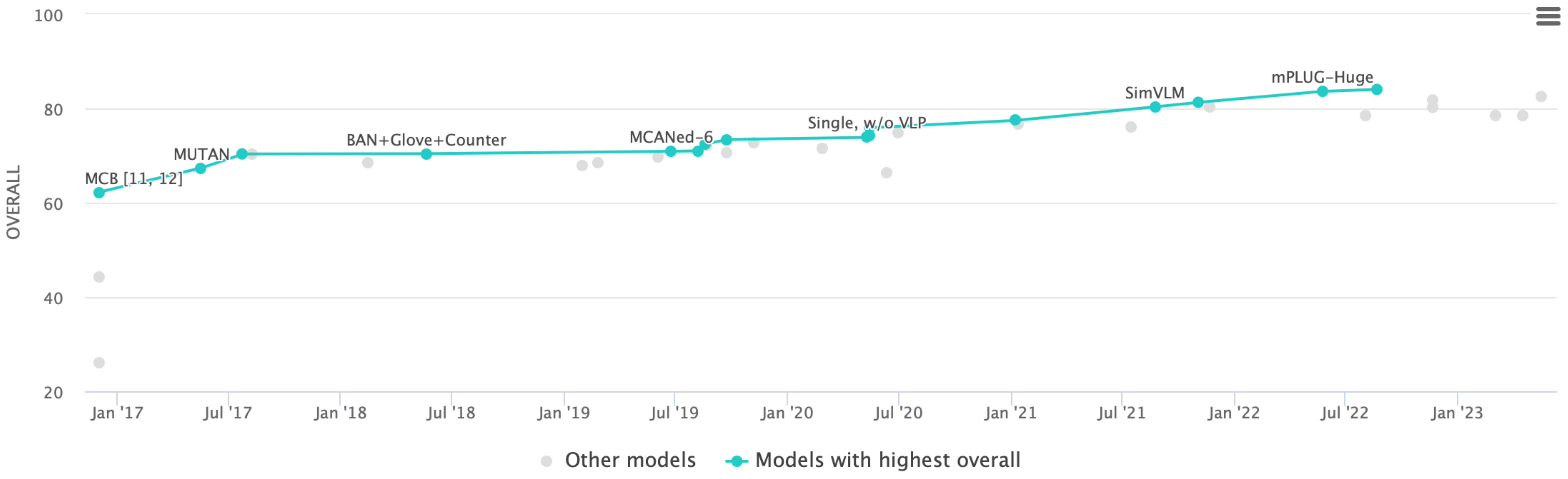


https://visualqa.org/

VQA: Visual Question Answering. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. ICCV 2015.

Visual Question Answering

👁️ Perception

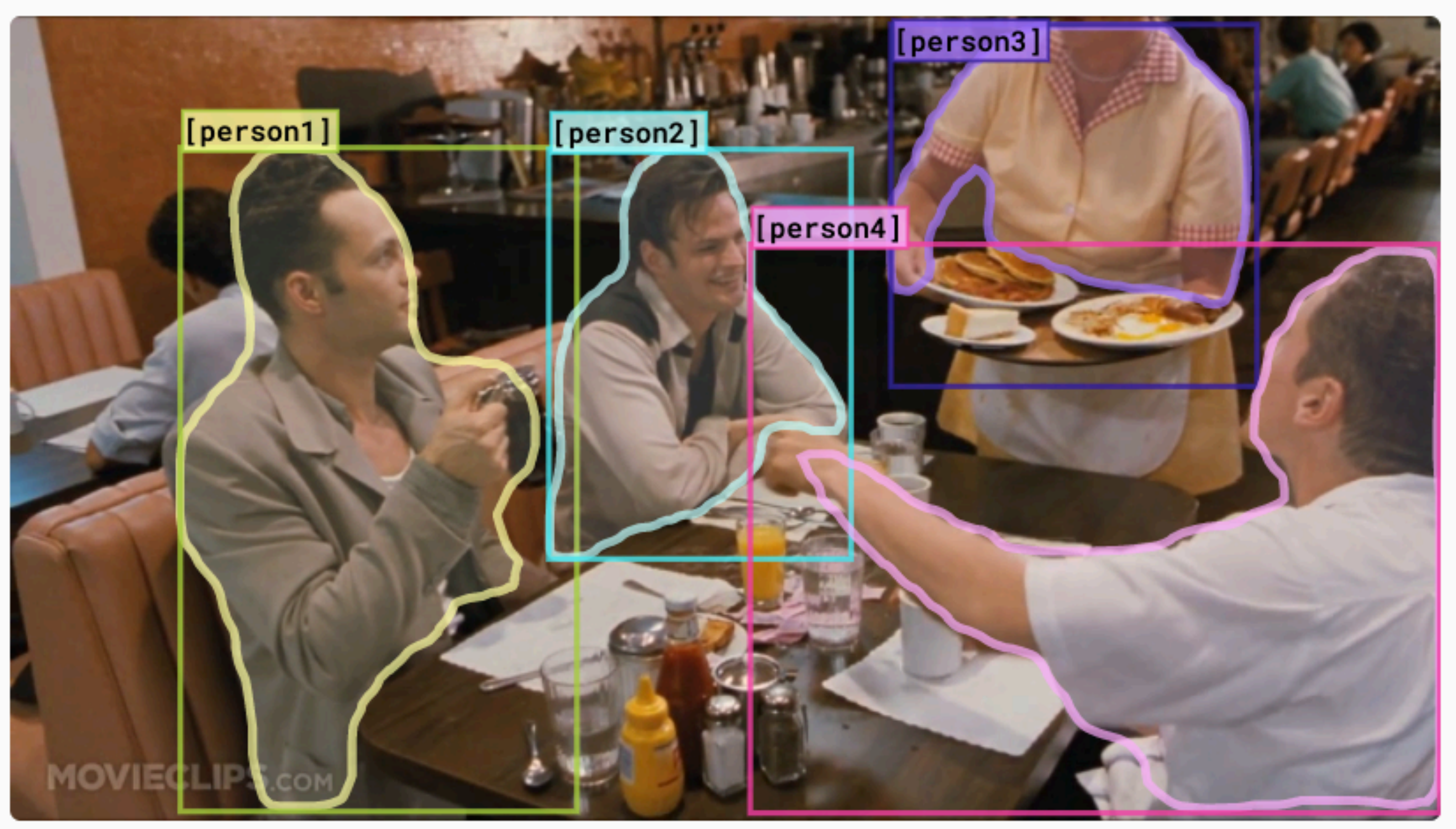


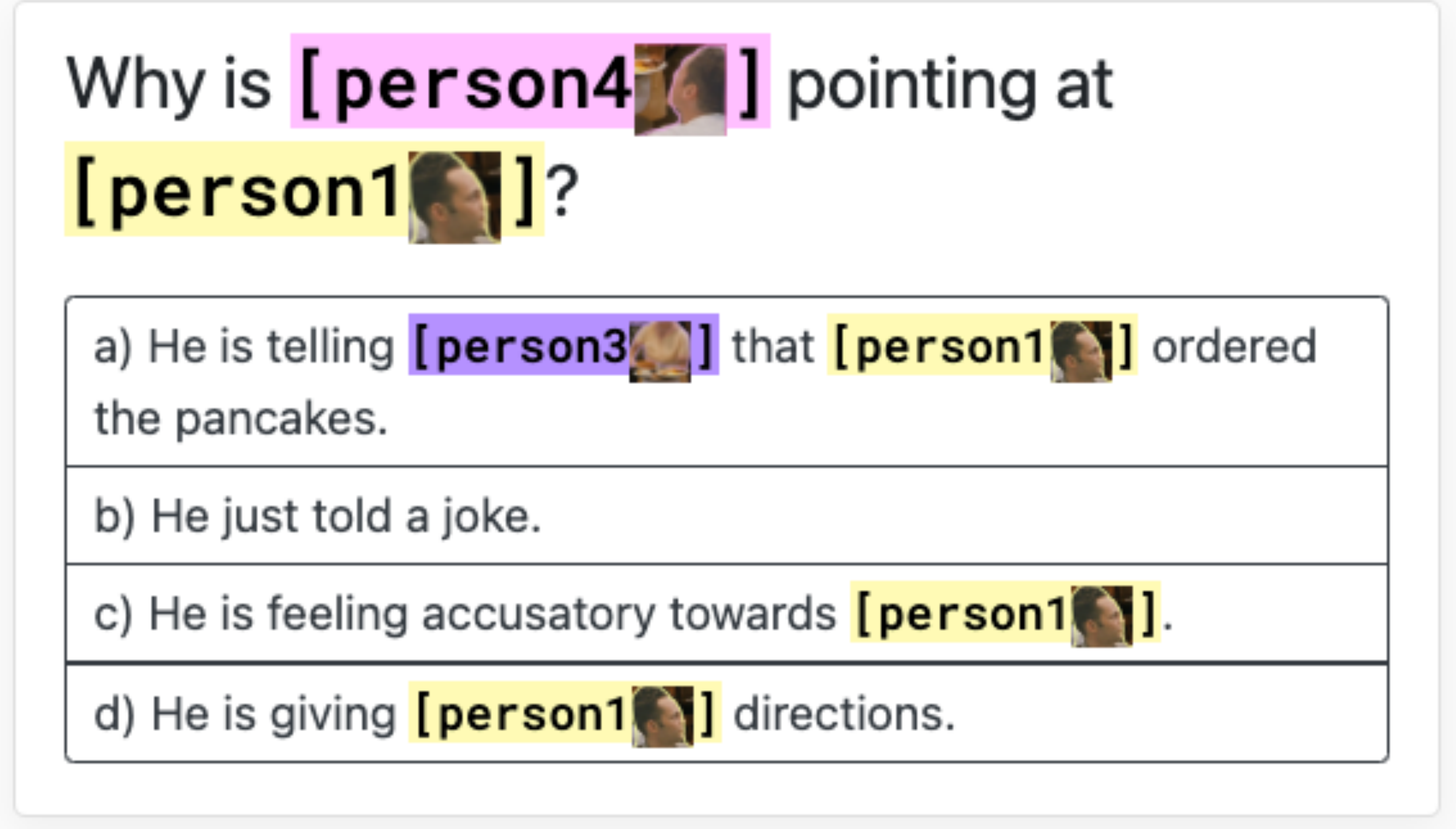
Model performance is almost close to human performance!

https://paperswithcode.com/sota/visual-question-answering-on-vqa-v2-test-std

VQA: Visual Question Answering. Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh. ICCV 2015.

Visual Commonsense Reasoning



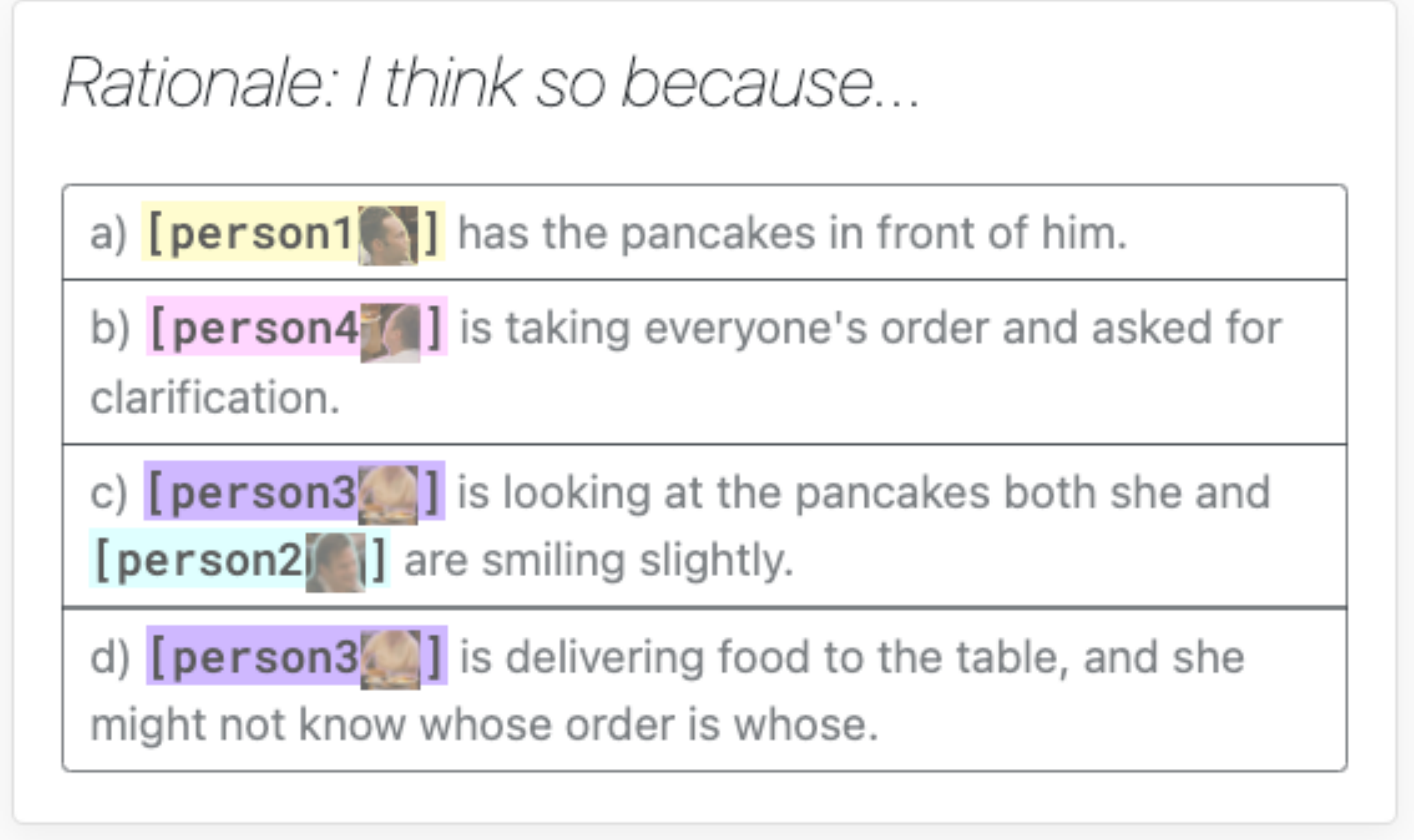

Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.

b) He just told a joke.

c) He is feeling accusatory towards [person1].

d) He is giving [person1] directions.


Rationale: I think so because...

a) [person1] has the pancakes in front of him.

b) [person4] is taking everyone's order and asked for clarification.

c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.

d) [person3] is delivering food to the table, and she might not know whose order is whose.

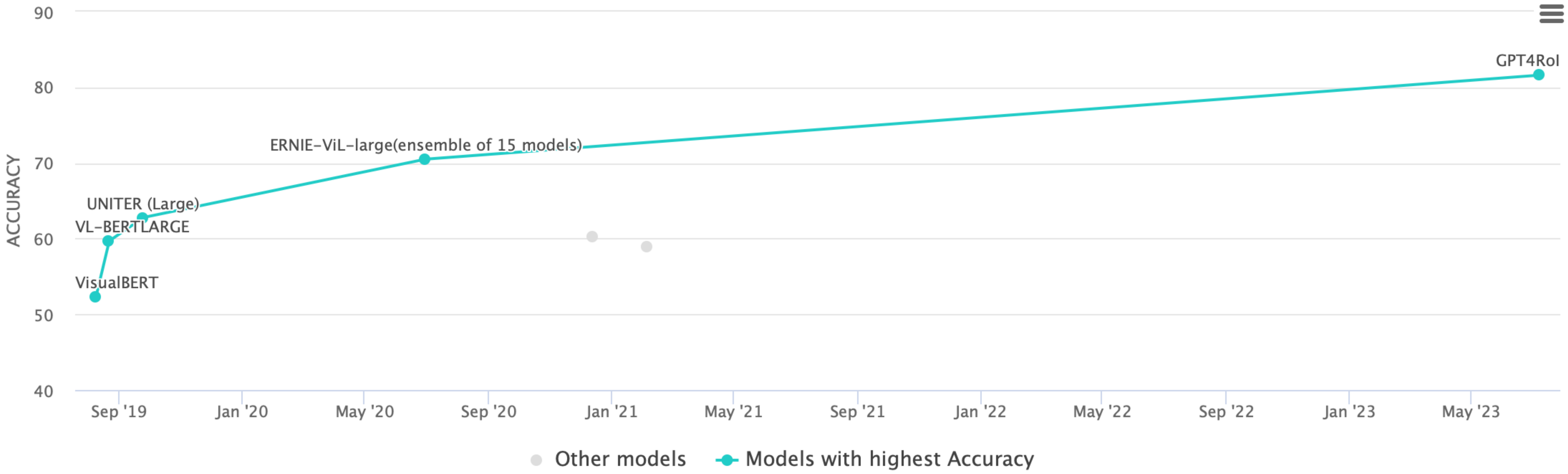
From Recognition to Cognition: Visual Commonsense Reasoning. Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. CVPR 2019.

Visual Commonsense Reasoning

👁 Perception

💡 Interpretation

🧐 Reasoning



Requires RoI-based reasoning to achieve human-level performance

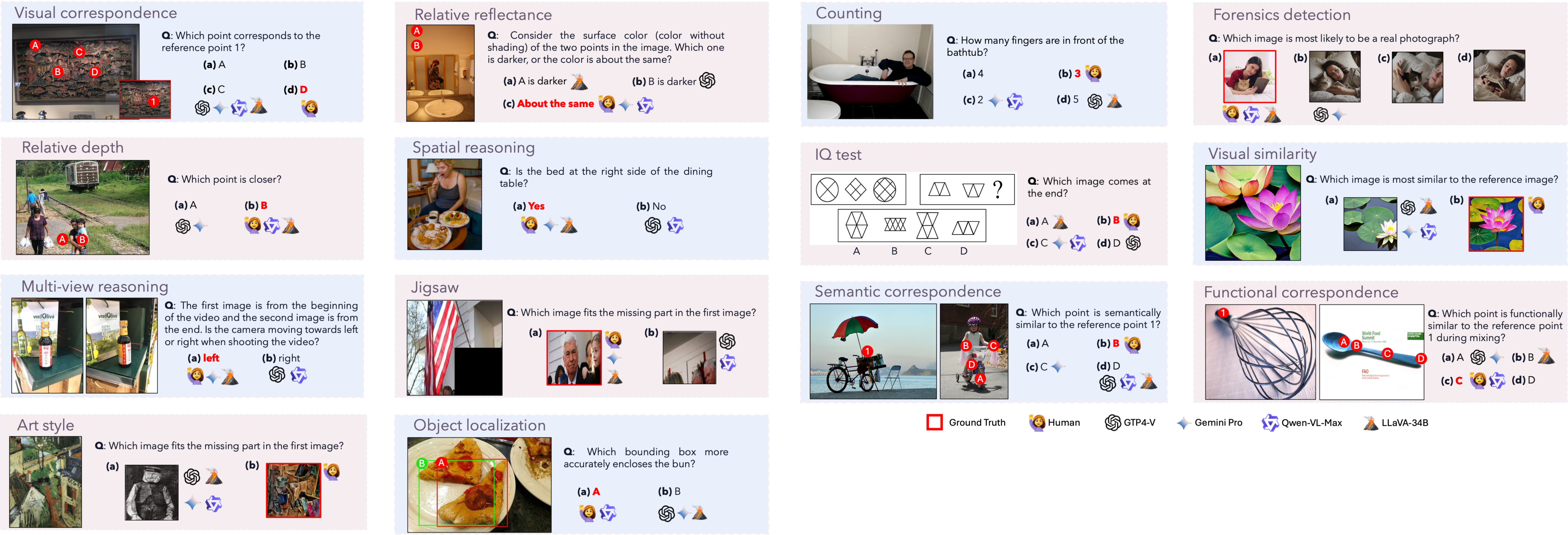
From Recognition to Cognition: Visual Commonsense Reasoning. Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. CVPR 2019.

BLINK: "Seeing" vs "Perceiving"

👁 Perception

💡 Interpretation

🧐 Reasoning



Visual correspondence. Q: Which point corresponds to the reference point 1? (a) A (b) B (c) C **(d) D**

Relative reflectance. Q: Consider the surface color (color without shading) of the two points in the image. Which one is darker, or the color is about the same? (a) A is darker (b) B is darker **(c) About the same**

Counting. Q: How many fingers are in front of the bathtub? (a) 4 **(b) 3** (c) 2 (d) 5

Forensics detection. Q: Which image is most likely to be a real photograph? (a) (b) (c) (d)

Relative depth. Q: Which point is closer? (a) A **(b) B**

Spatial reasoning. Q: Is the bed at the right side of the dining table? **(a) Yes** (b) No

IQ test. Q: Which image comes at the end? (a) A **(b) B** (c) C (d) D

Visual similarity. Q: Which image is most similar to the reference image? (a) (b)

Multi-view reasoning. Q: The first image is from the beginning of the video and the second image is from the end. Is the camera moving towards left or right when shooting the video? **(a) left** (b) right

Jigsaw. Q: Which image fits the missing part in the first image? (a) (b)

Semantic correspondence. Q: Which point is semantically similar to the reference point 1? (a) A **(b) B** (c) C (d) D

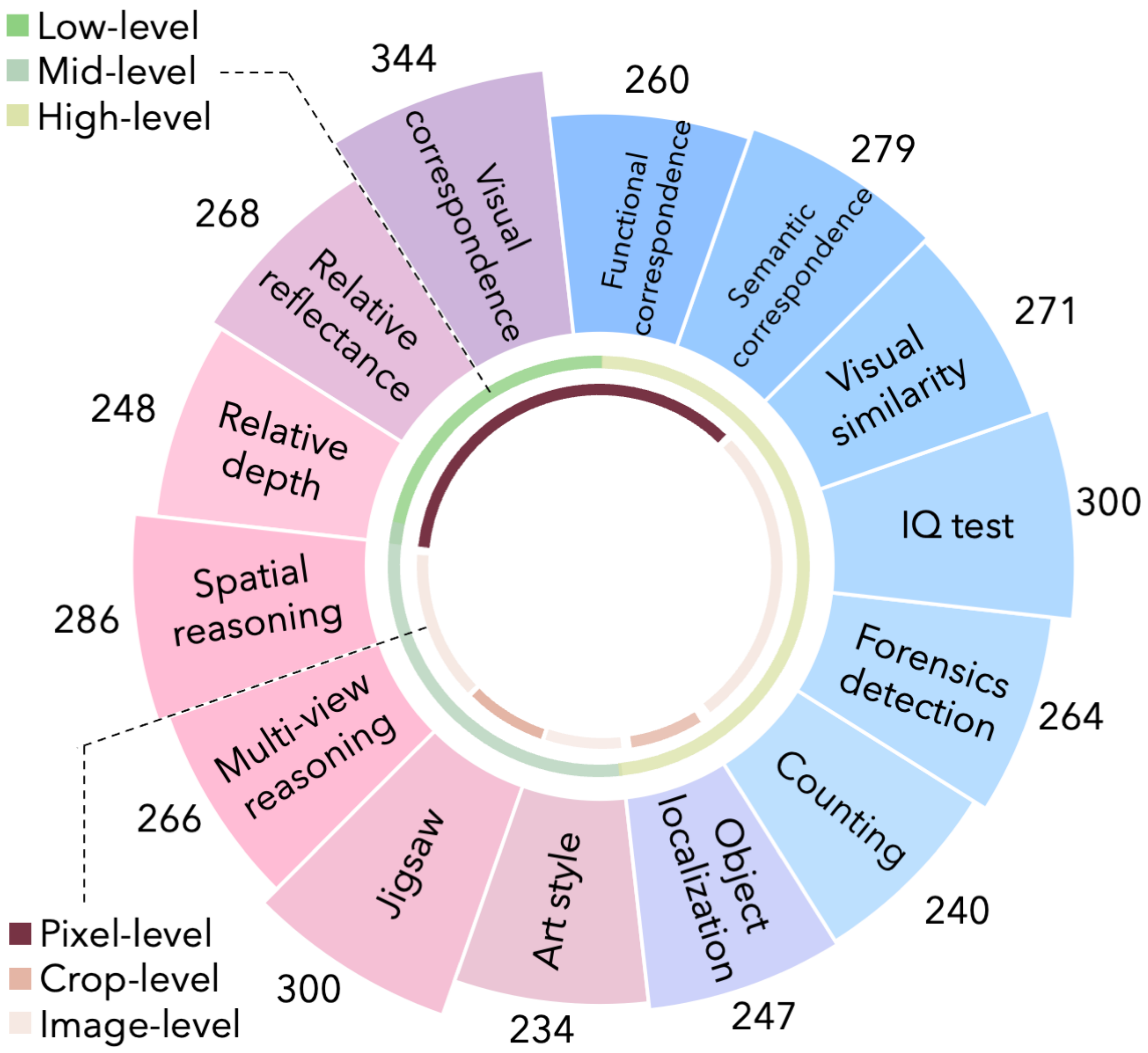
Functional correspondence. Q: Which point is functionally similar to the reference point 1 during mixing? (a) A (b) B **(c) C** (d) D

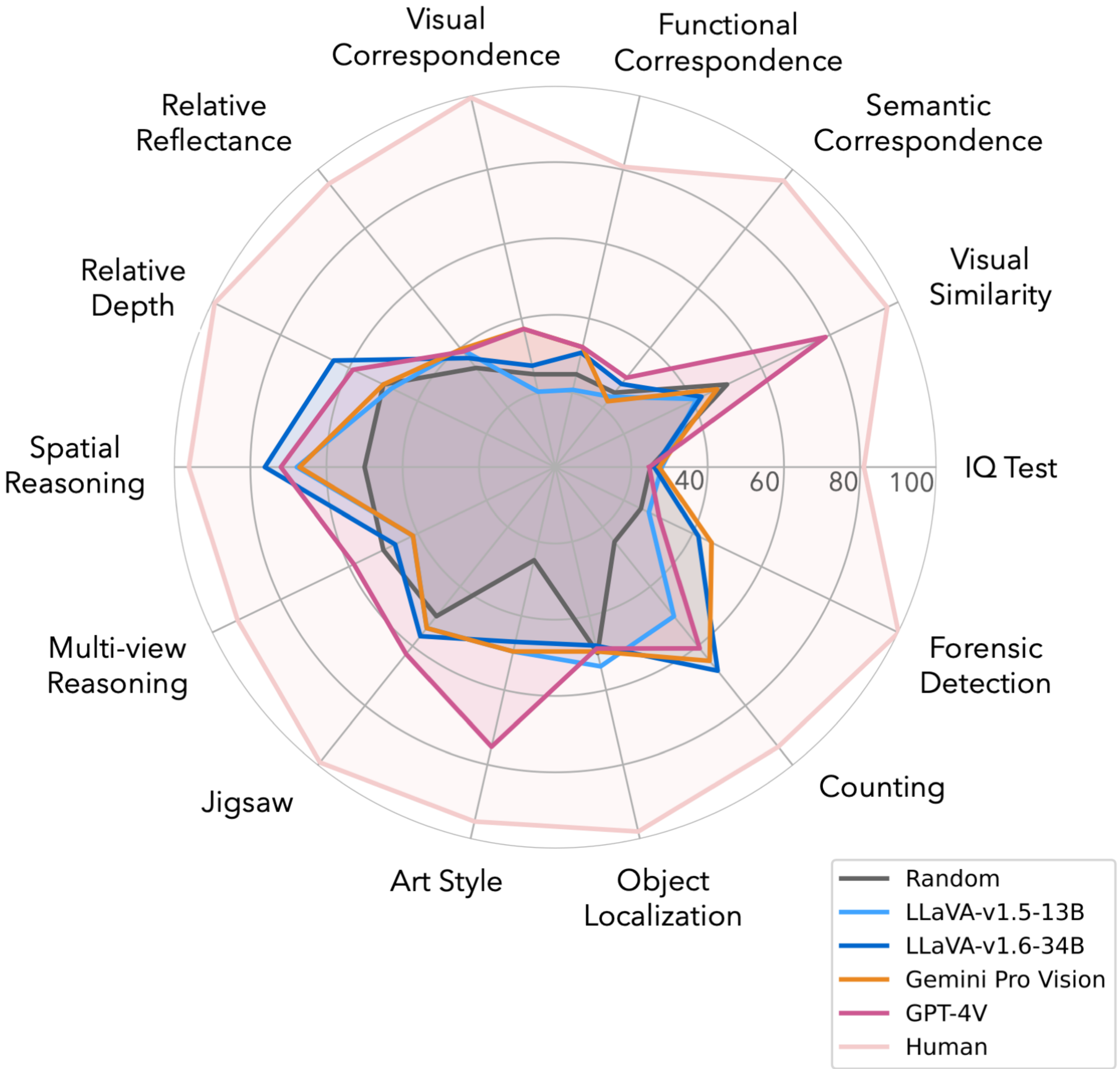
Art style. Q: Which image fits the missing part in the first image? (a) (b)

Object localization. Q: Which bounding box more accurately encloses the bun? **(a) A** (b) B

BLINK : Multimodal Large Language Models Can See but Not Perceive (ECCV 2024)

BLINK: “Seeing” vs “Perceiving”





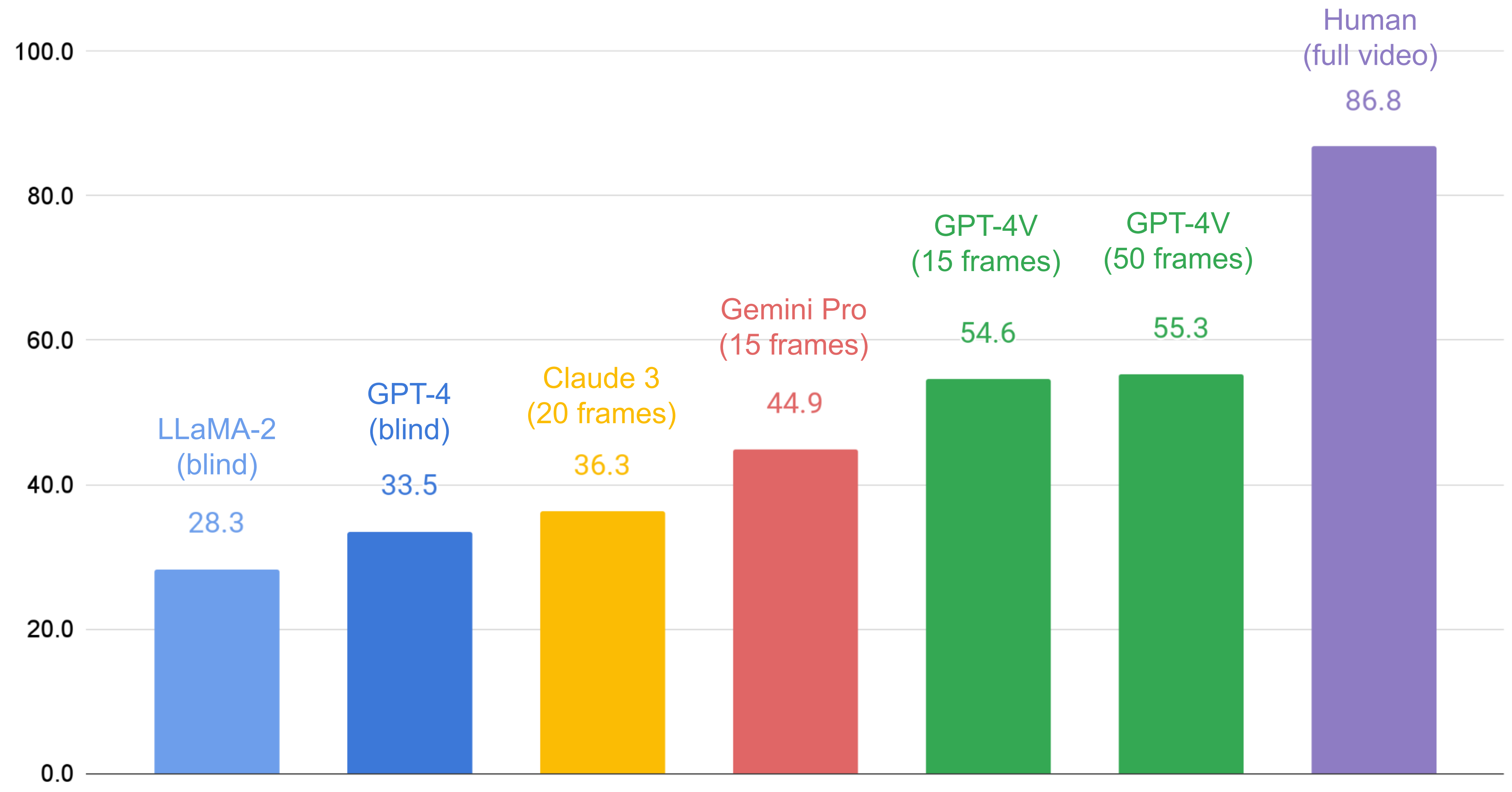
While these problems only takes human a "blink" to solve, they exceed the capabilities of current multimodal large language models with mean performance **35–42%**

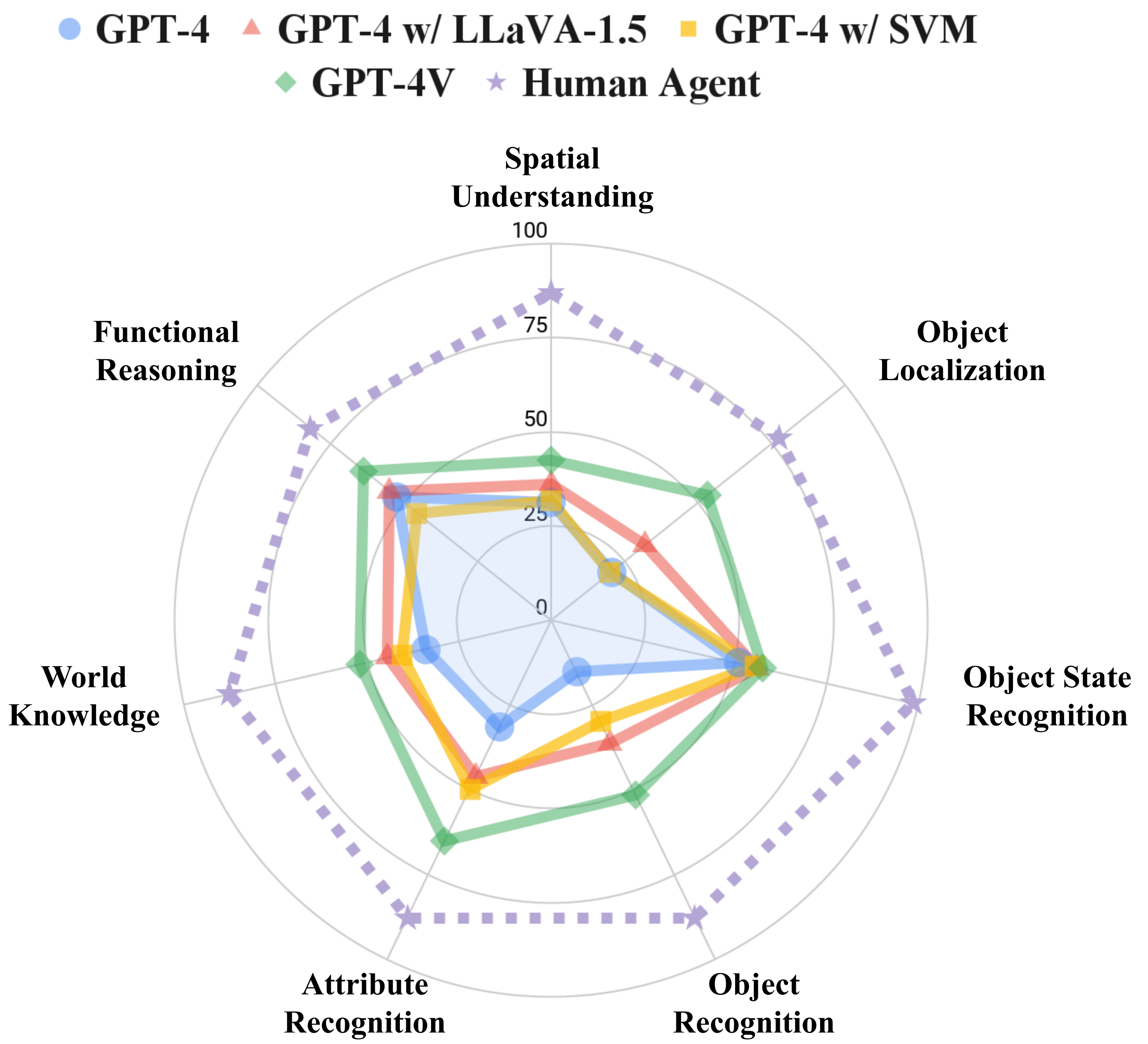
BLINK : Multimodal Large Language Models Can See but Not Perceive (ECCV 2024)

OpenEQA: Embodied QA

Video Understanding + Planning

Performance on OpenEQA

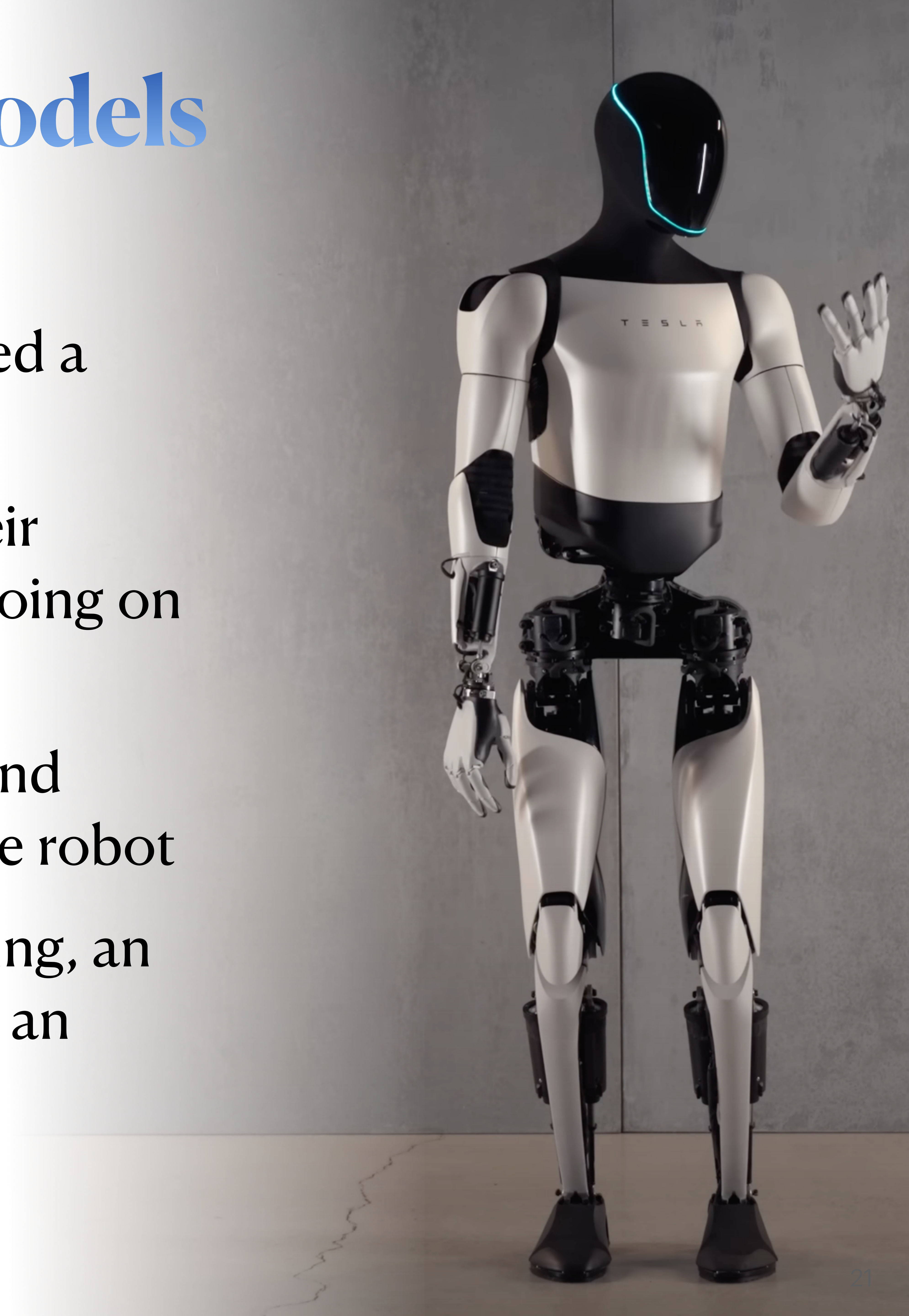




VLMs perform better than LLMs, but tasks that require **episodic memory** are still hard and substantially worse than human performance - these tasks require long form video understanding.

From Word Models to World Models

- Large language models (LLMs), seem to have captured a **linguistic** understanding of the world.
- LLMs can answer all kinds of questions based on their knowledge, but they have no idea what is currently going on in the world around them.
- Enhancing LLMs with the ability to “see” the world and **situating** them in a user’s smart glasses or on a home robot
- Rather than simply predicting the next token in a string, an embodied AI agent would show that it’s grounded in an understanding of the physical world.



Outline

Reasoning about vision and language

- Motivation
- Visual Commonsense Reasoning tasks
- **Vision and language representations and models**
- Open problems and future directions

Vision Transformer

Images can be represented as tokens too!

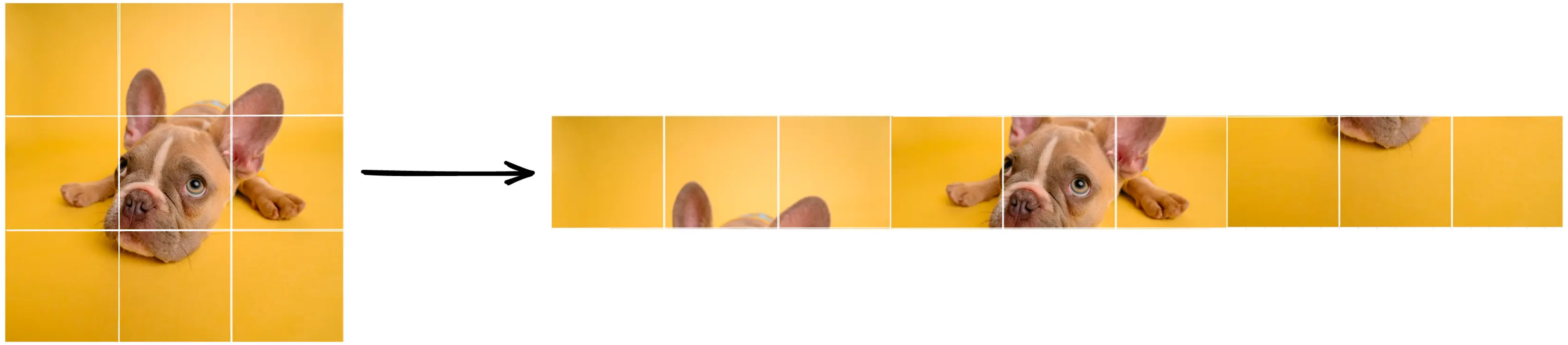
Sentence to word tokens:

"hi, I am a short sentence"

↓

'hi' ',' 'I' 'am' 'a' 'short' 'sentence'

Image to image patches:

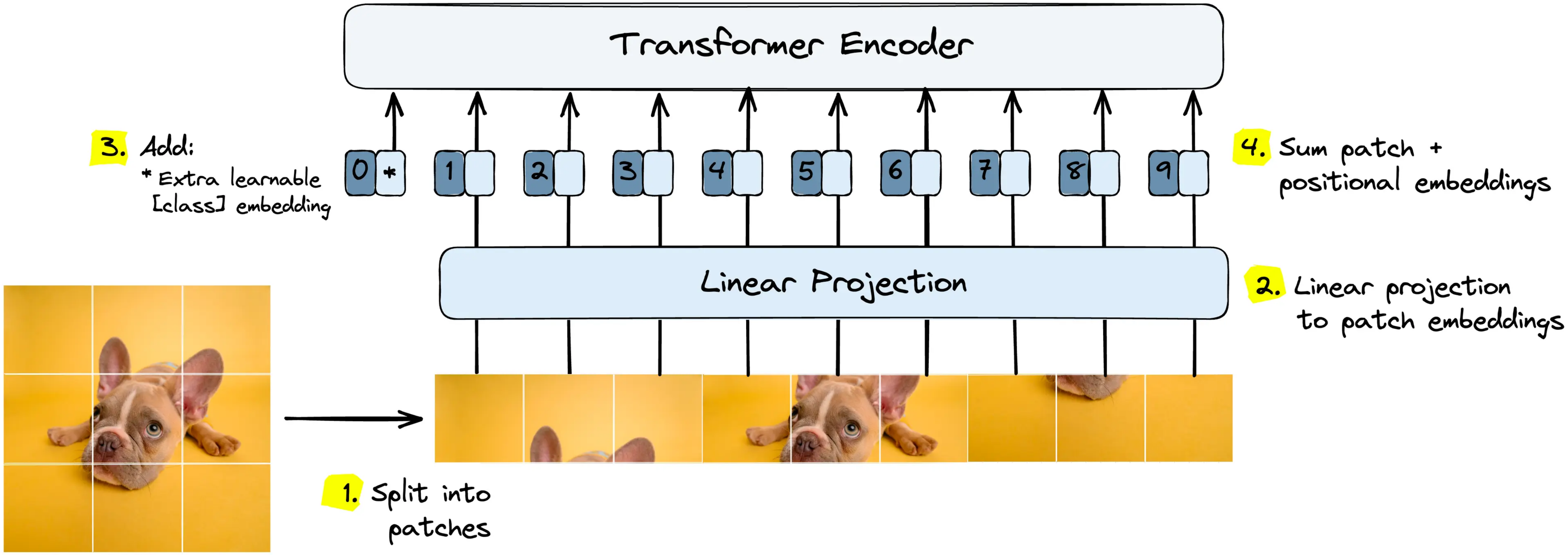


The input image (e.g., 224×224 pixels) is divided into small non-overlapping patches (e.g., 16×16 pixels each).

https://www.pinecone.io/learn/series/image-search/vision-transformers/

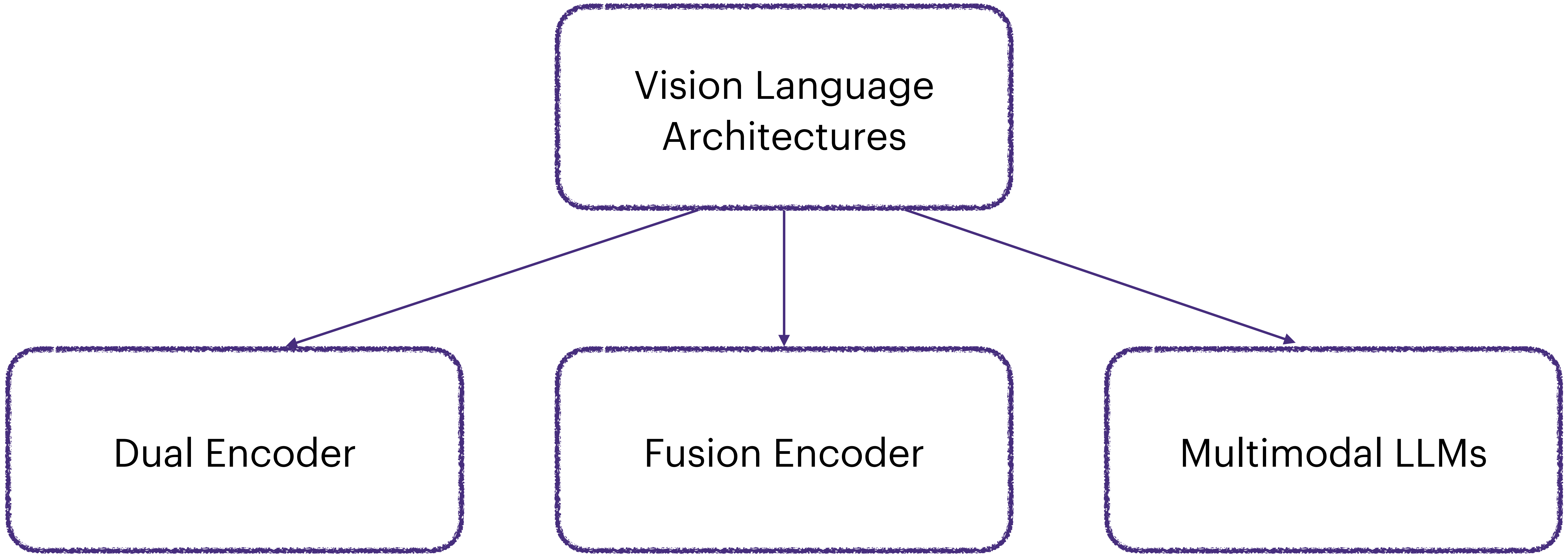
Vision Transformer

Images can be represented as tokens too!


Transformer Encoder
3. Add:
* Extra learnable [class] embedding
0 *
1
2
3
4
5
6
7
8
9
4. Sum patch + positional embeddings
Linear Projection
2. Linear projection to patch embeddings
1. Split into patches

https://www.pinecone.io/learn/series/image-search/vision-transformers/

Vision Language Models



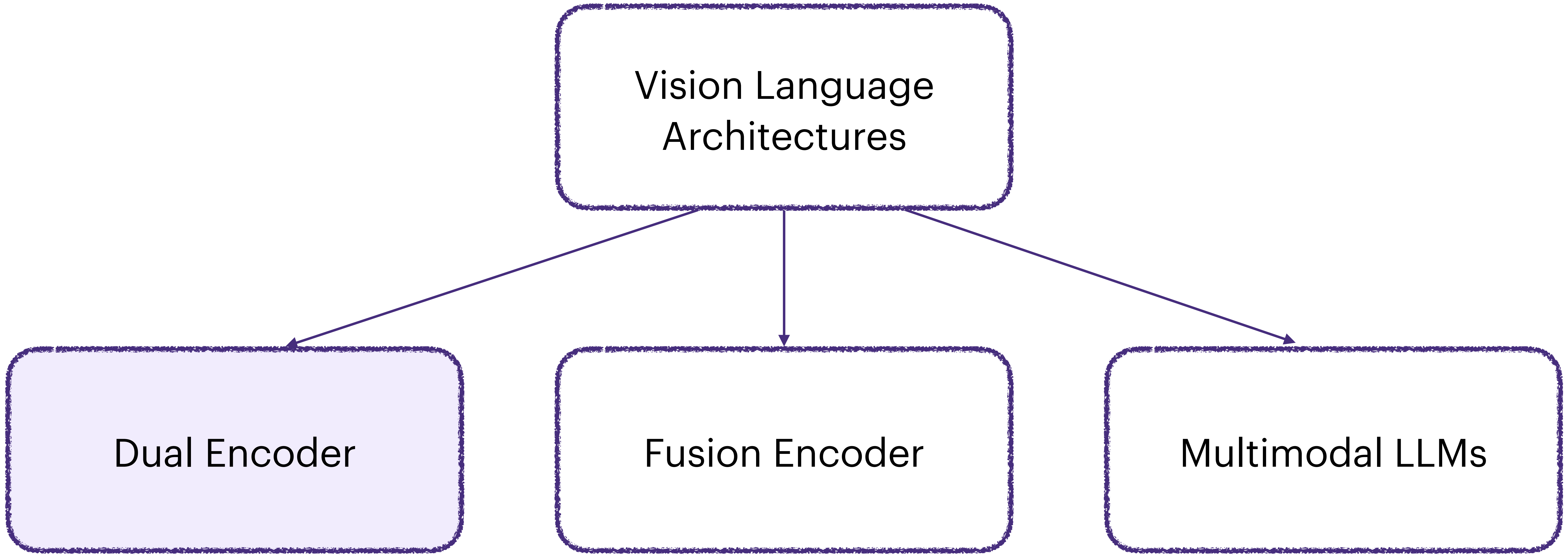
Vision Language Architectures

Dual Encoder

Fusion Encoder

Multimodal LLMs

Vision Language Models



Vision Language Architectures

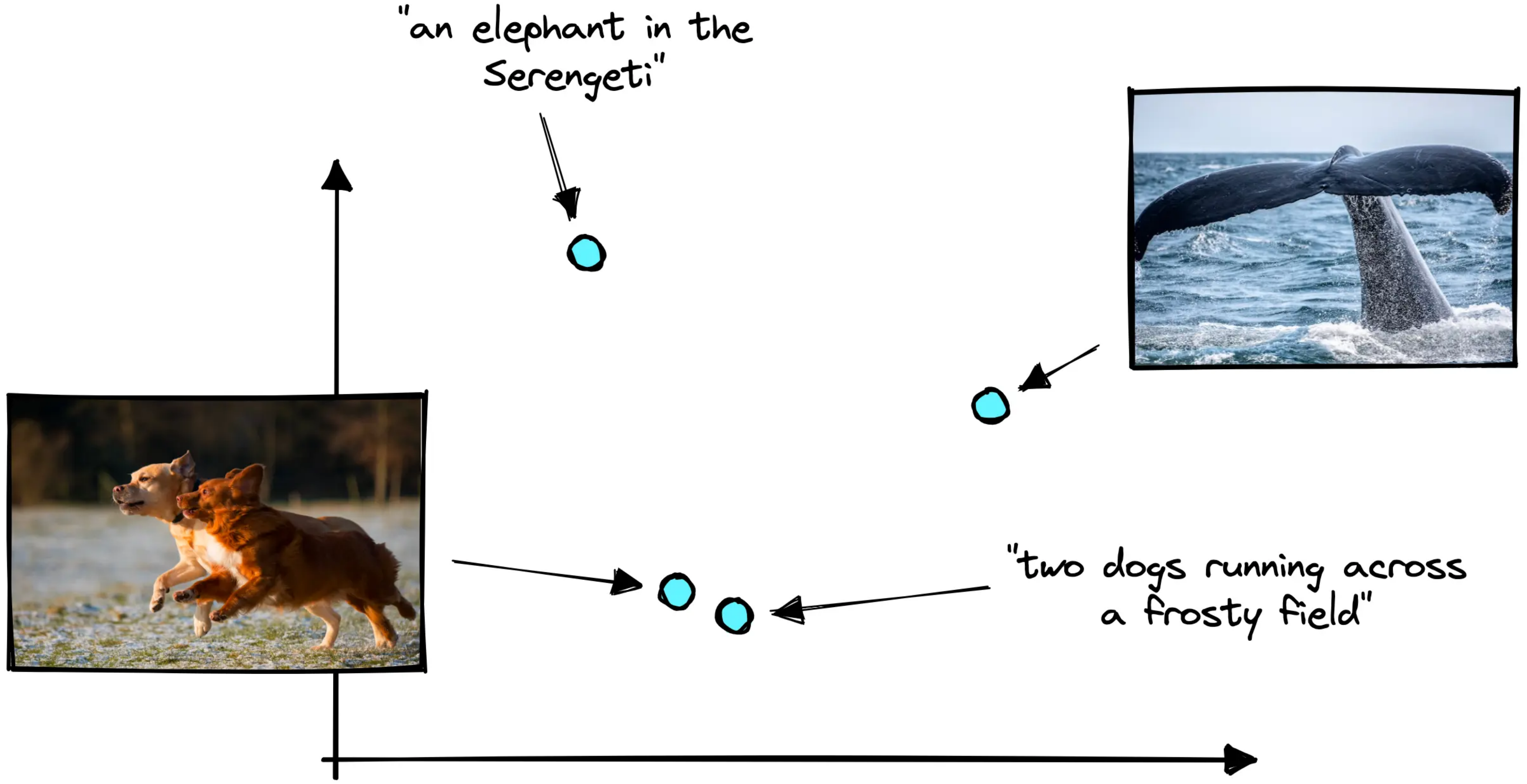
Dual Encoder

Fusion Encoder

Multimodal LLMs

CLIP

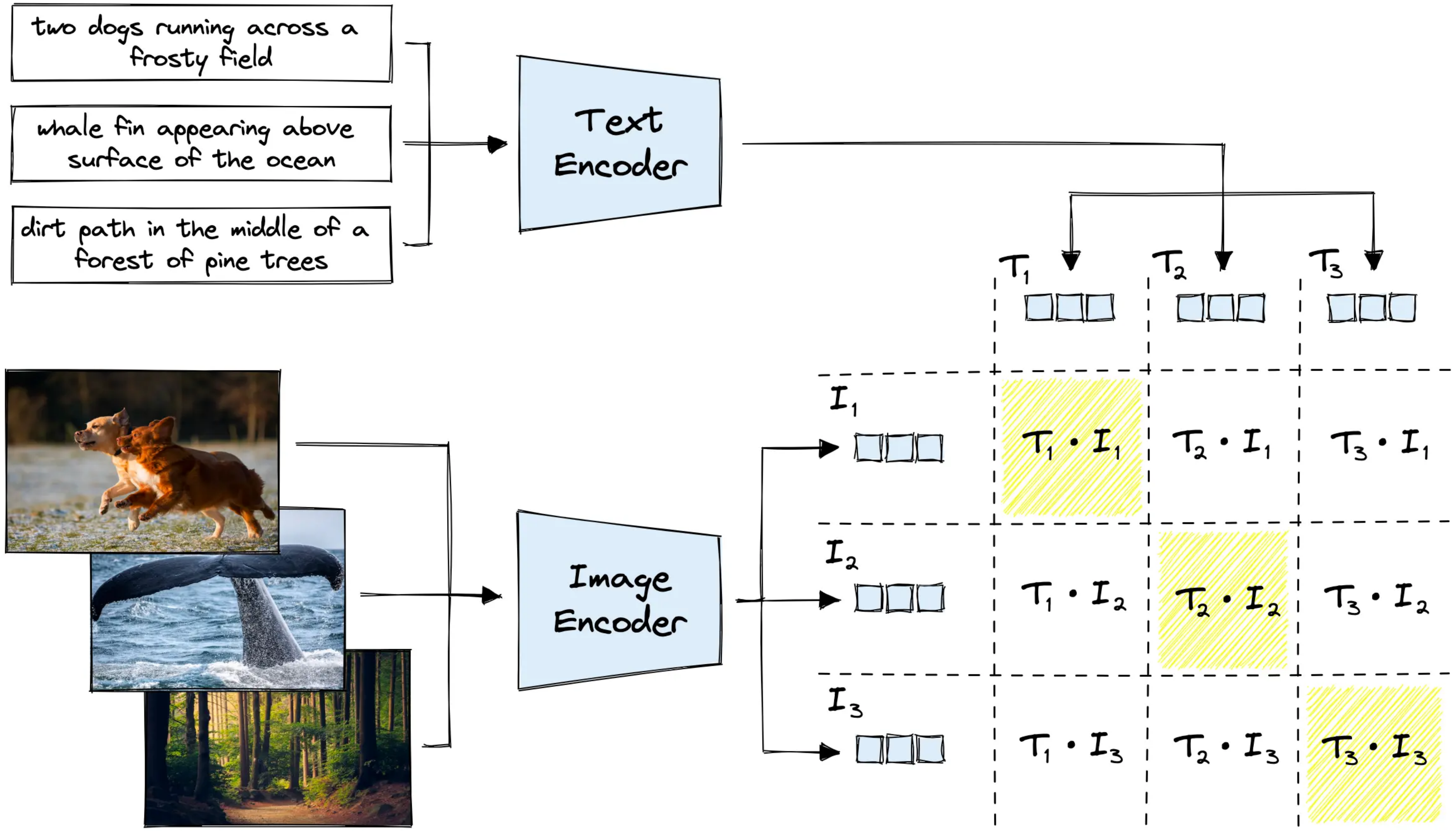
Goal: Representing Images and Text in the same embedding space


"an elephant in the Serengeti"
"two dogs running across a frosty field"

https://www.pinecone.io/learn/series/image-search/clip/

CLIP

Training using Contrastive Loss: pull similar images closer, push different images apart

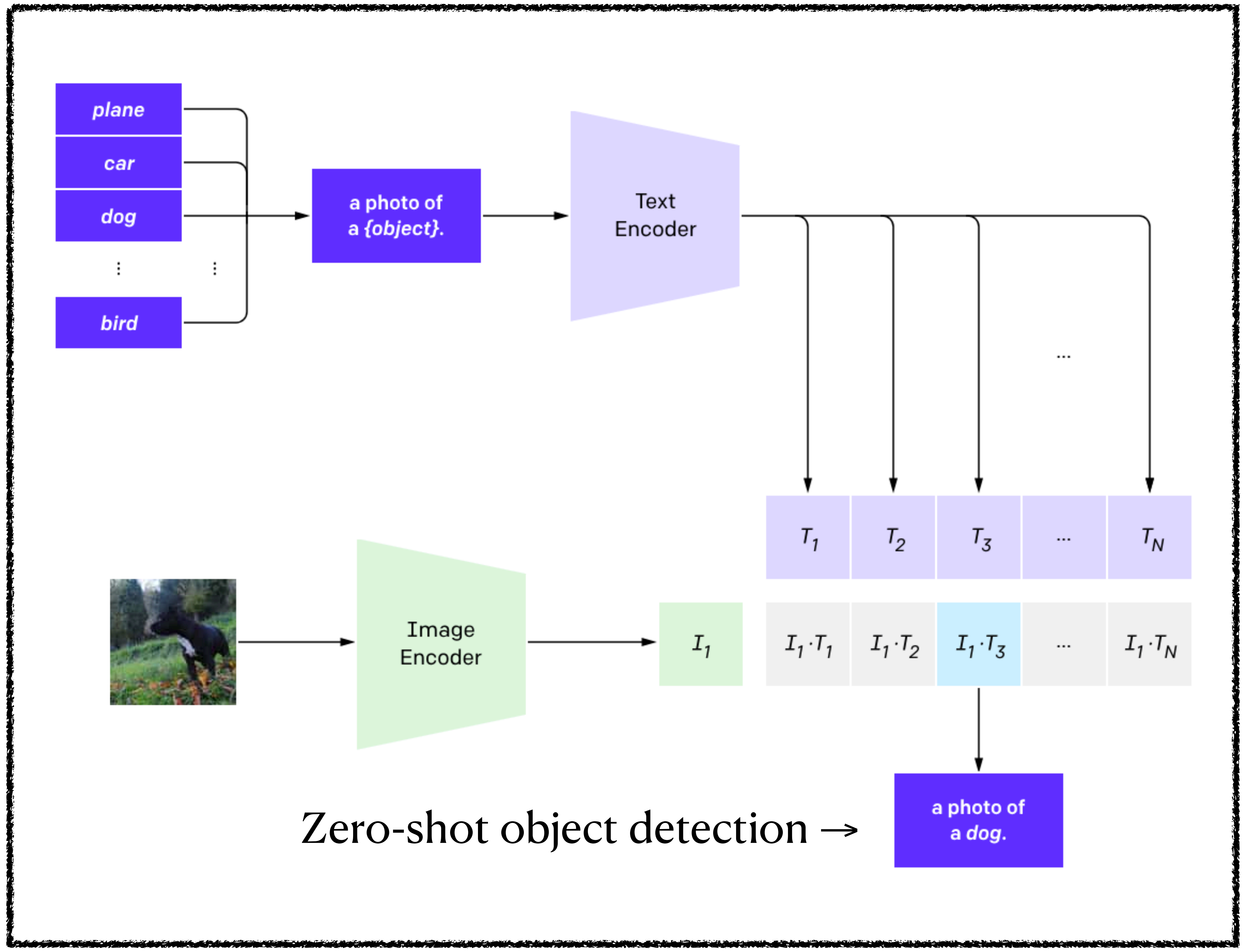

two dogs running across a frosty field
whale fin appearing above surface of the ocean
dirt path in the middle of a forest of pine trees
Text Encoder
Image Encoder
T1
T2
T3
I1
I2
I3

	T_1	T_2	T_3
I_1	$T_1 \cdot I_1$	$T_2 \cdot I_1$	$T_3 \cdot I_1$
I_2	$T_1 \cdot I_2$	$T_2 \cdot I_2$	$T_3 \cdot I_2$
I_3	$T_1 \cdot I_3$	$T_2 \cdot I_3$	$T_3 \cdot I_3$

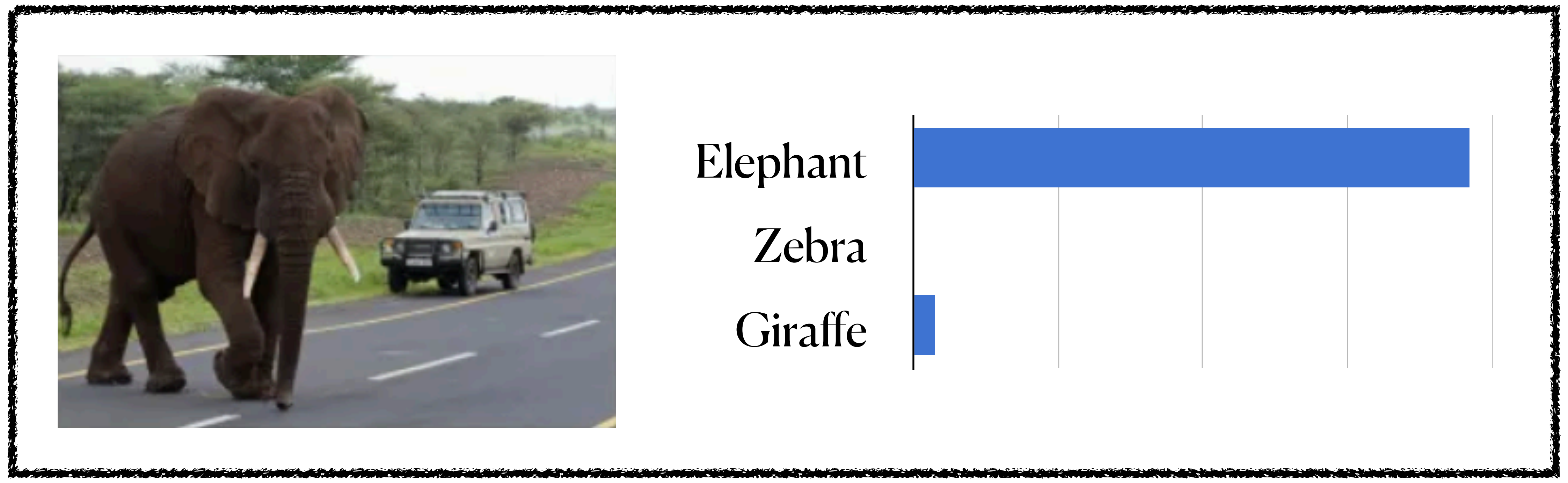
https://openai.com/research/clip
https://www.pinecone.io/learn/series/image-search/clip/

Using CLIP

Capabilities: Zero-shot inference



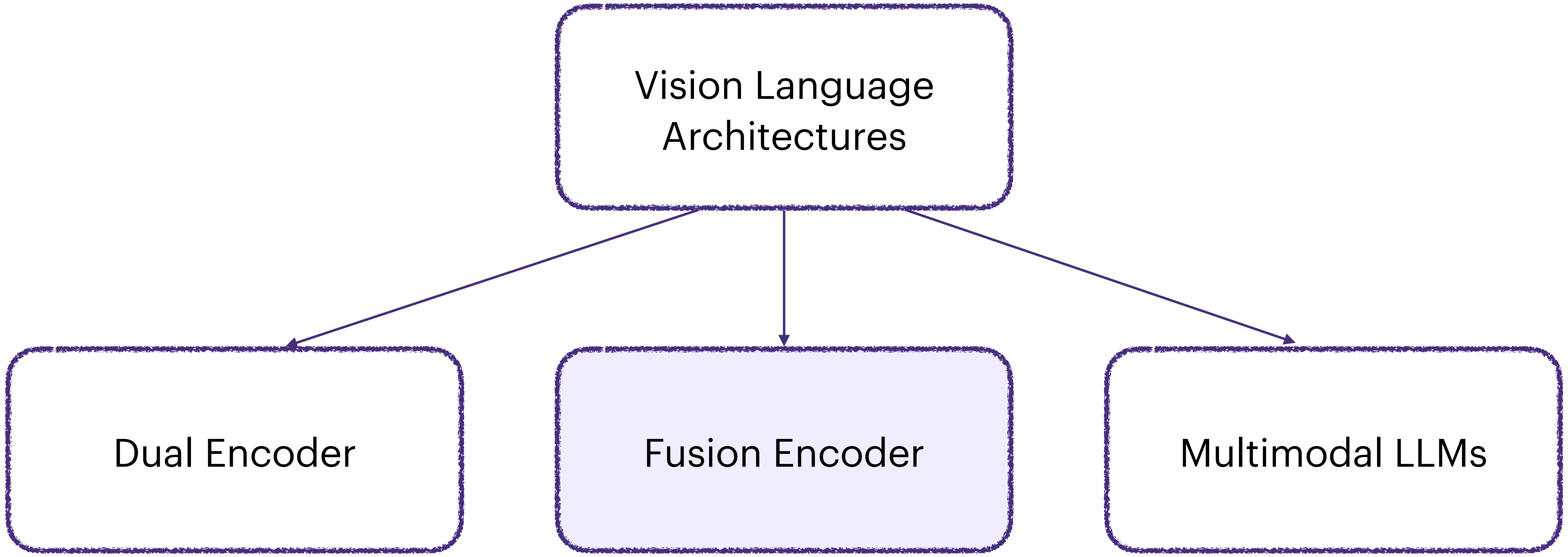
Zero-shot object detection →



A wooden table topped with lots of wooden utensils.

https://openai.com/research/clip

Vision Language Models



Vision Language Architectures

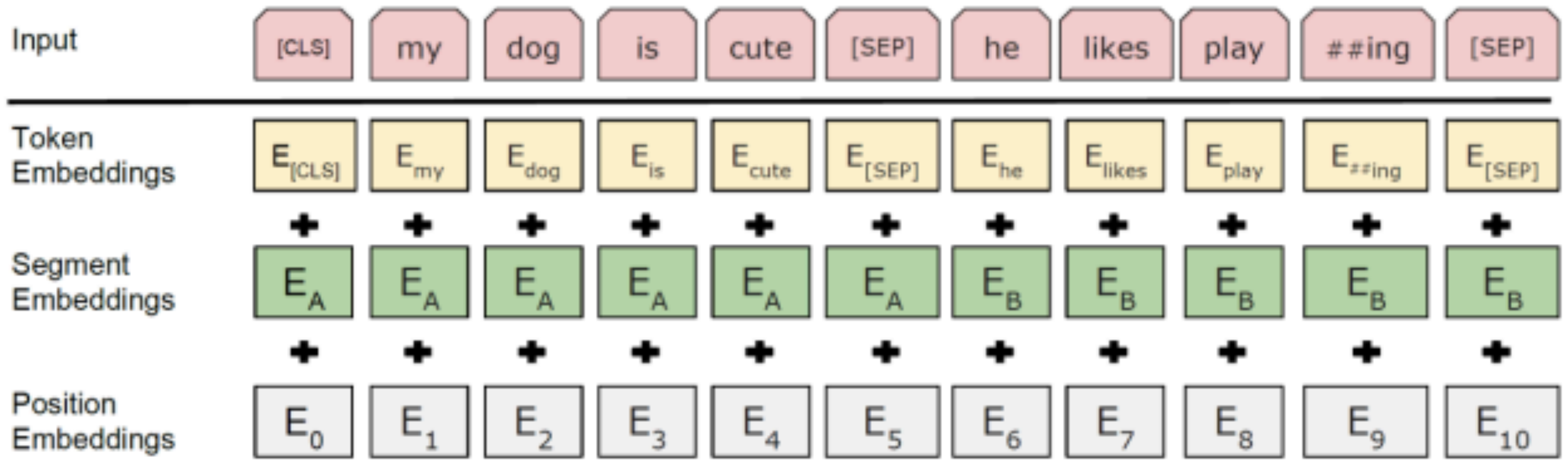
Dual Encoder

Fusion Encoder

Multimodal LLMs

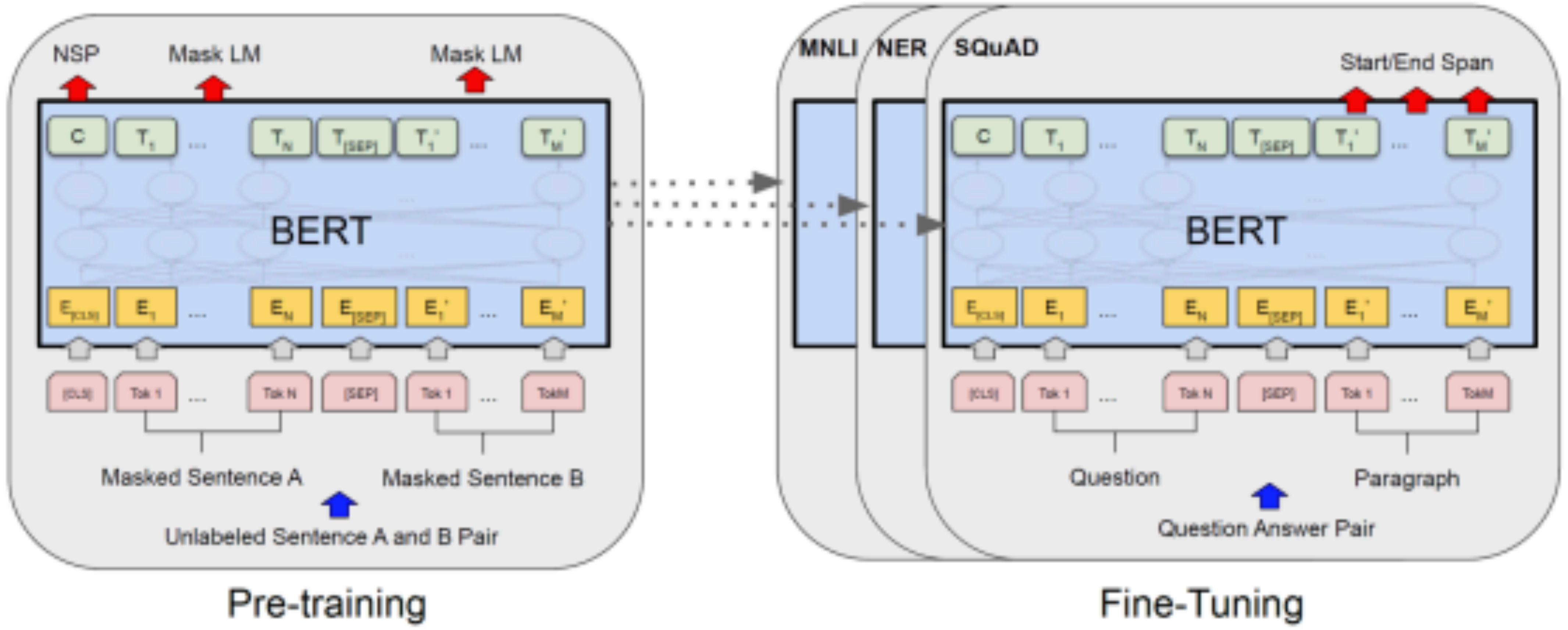
Revisiting BERT...

Flexible and powerful in aggregating and aligning word features



Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Embedded features in BERT

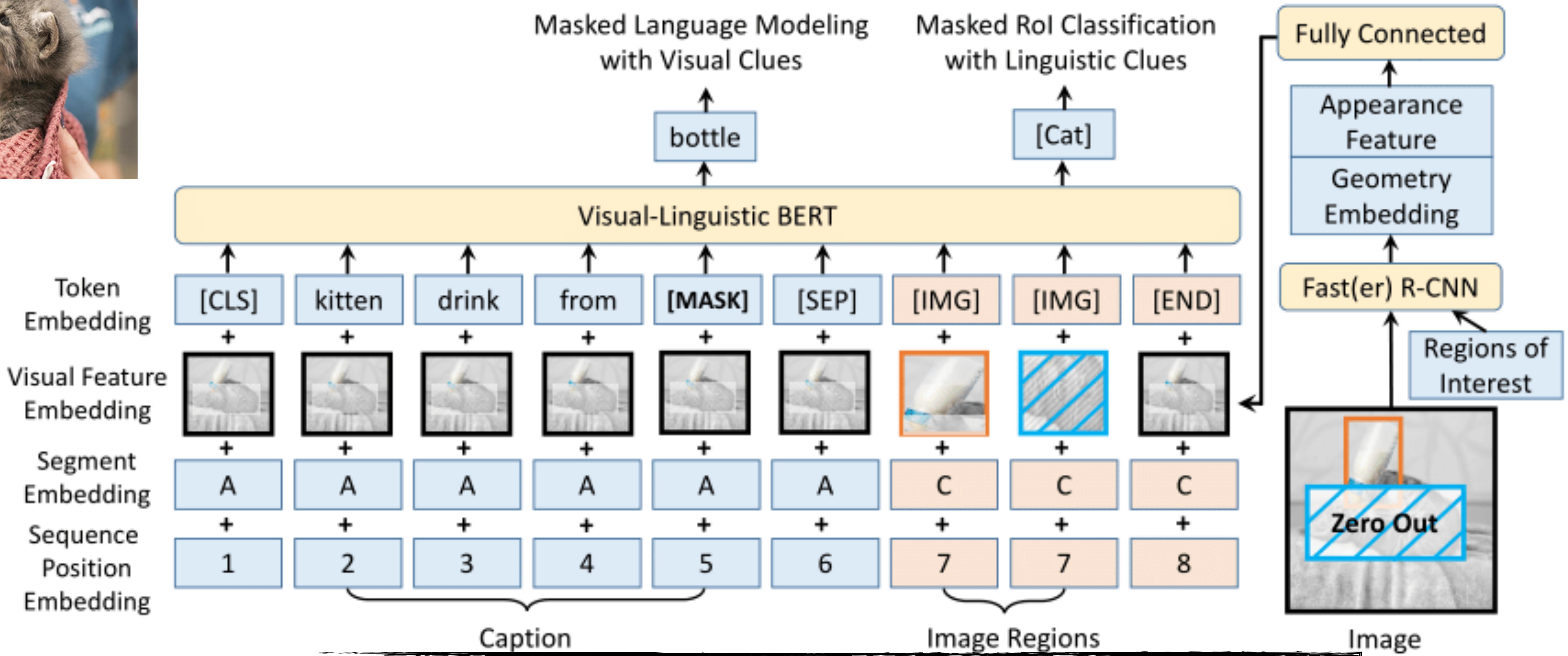


Pre-training & finetuning of BERT

VL-BERT

Architecture

- **kitten drinking from [MASK]**
- **“kitten drinking from bottle”**



Token, segment, and sequence position embeddings are the same as BERT
Visual feature embeddings are newly introduced for each element

VL-BERT: Pre-training of generic visual-linguistic representations. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. ICLR 2020.

VL-BERT

Pretraining

Conceptual Captions Pre-training

- Input: <Caption 💬, Image 🌆>
- Task #1: Masked Language Modeling with Visual Clues
  - In the above figure, **“kitten drinking from [MASK]”**, it could be any containers, such as “bowl”, “spoon” and “bottle”. But with visual clues, the network should predict the masked word as “bottle”.
- Task #2: Masked RoI (Region of Interest) Classification with Linguistic Clues
  - In the above figure, the RoI corresponding to cat in image is masked out, and the corresponding category cannot be predicted from any visual clues. But with the input caption of **“kitten drinking from bottle”,** the model can infer the category such as “a cat” by exploiting the linguistic clues.

BooksCorpus & English Wikipedia Pre-training

- Input: < Text 📜, Null>
- Task: Standard Masked Language Modelling (similar to BERT)

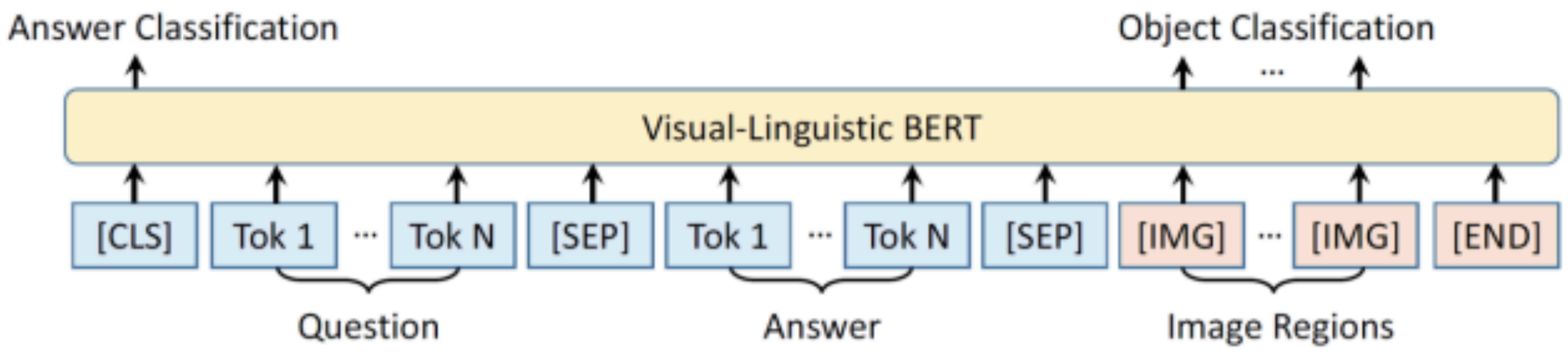
"trees in a winter snowstorm"
"a cartoon illustration of a bear waving and smiling"
"the scenic route through mountain range includes these unbelievably coloured mountains"
"facade of an old shop"

Conceptual Captions [ACL 2018]

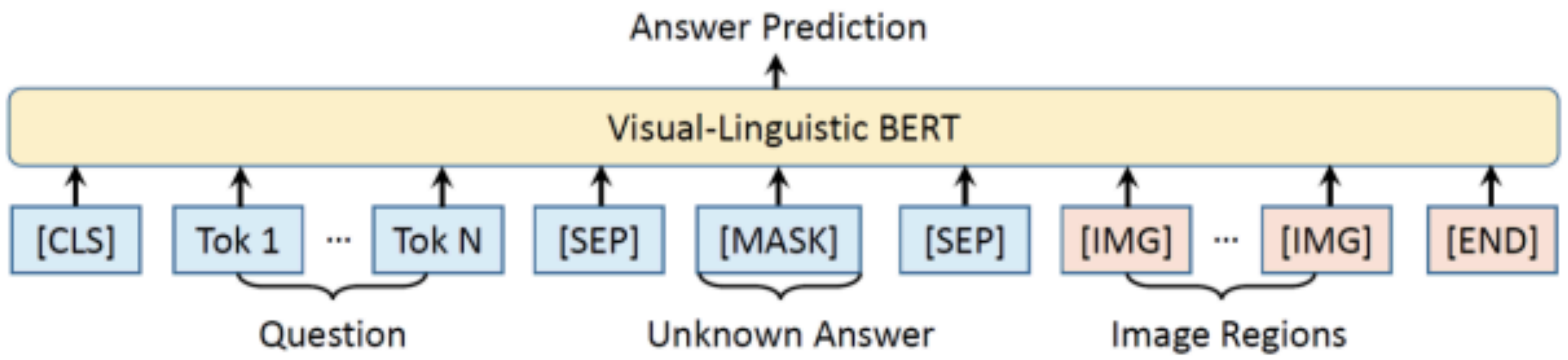
BooksCorpus [ICCV 2015] & English Wiki

VL-BERT

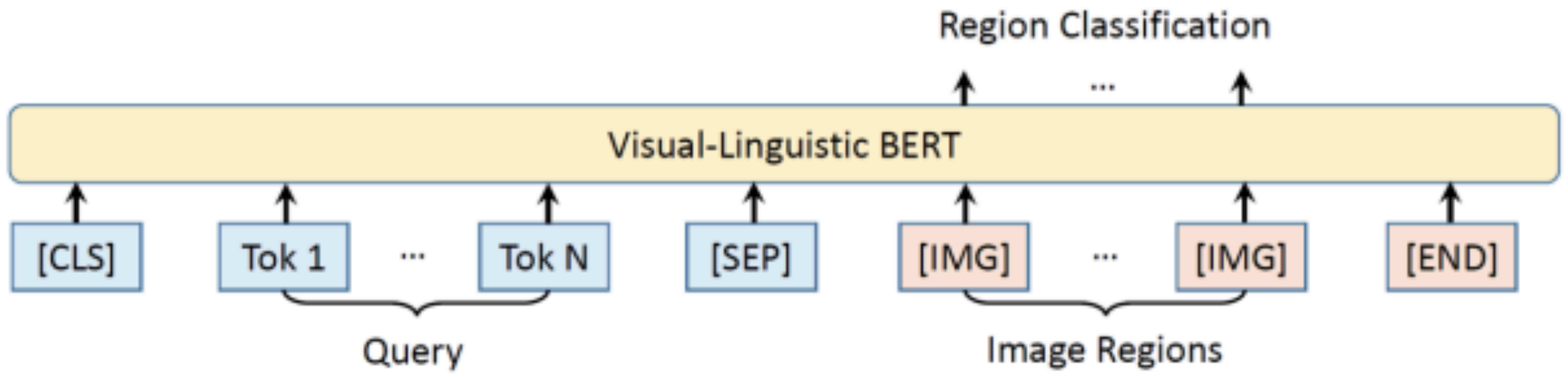
Finetuning



(a) Input and output format for Visual Commonsense Reasoning (VCR) dataset



(b) Input and output format for Visual Question Answering (VQA) dataset

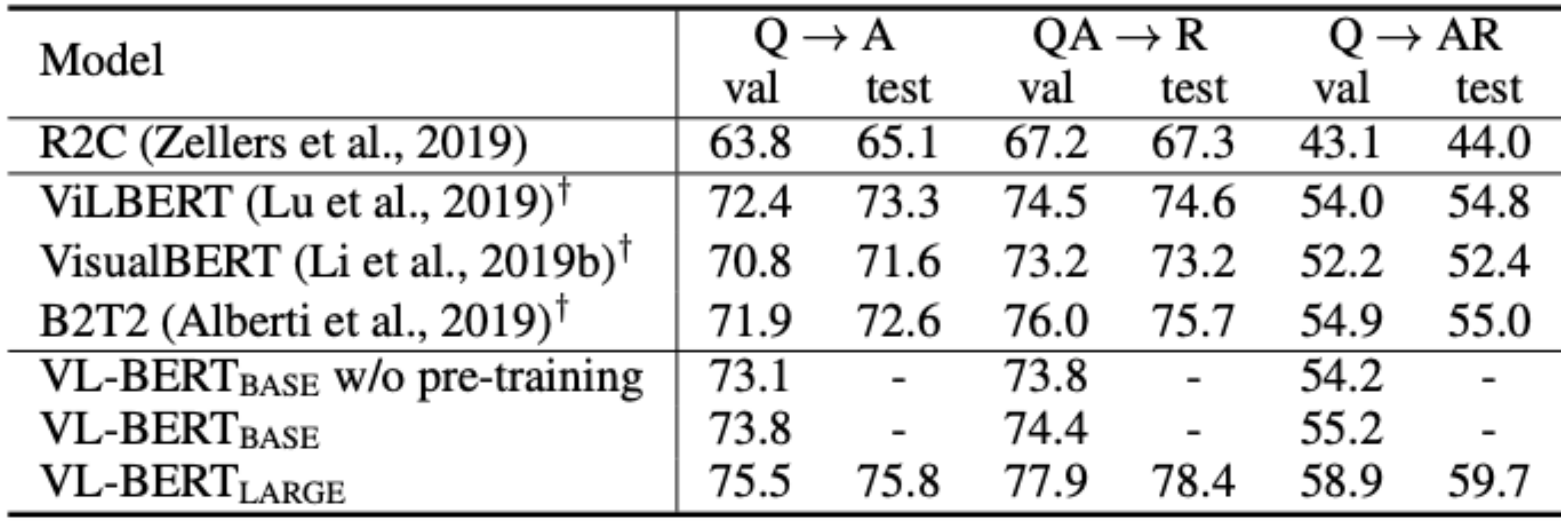


(c) Input and output format for Referring Expression task on RefCOCO+ dataset

VL-BERT: Pre-training of generic visual-linguistic representations. Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. ICLR 2020.

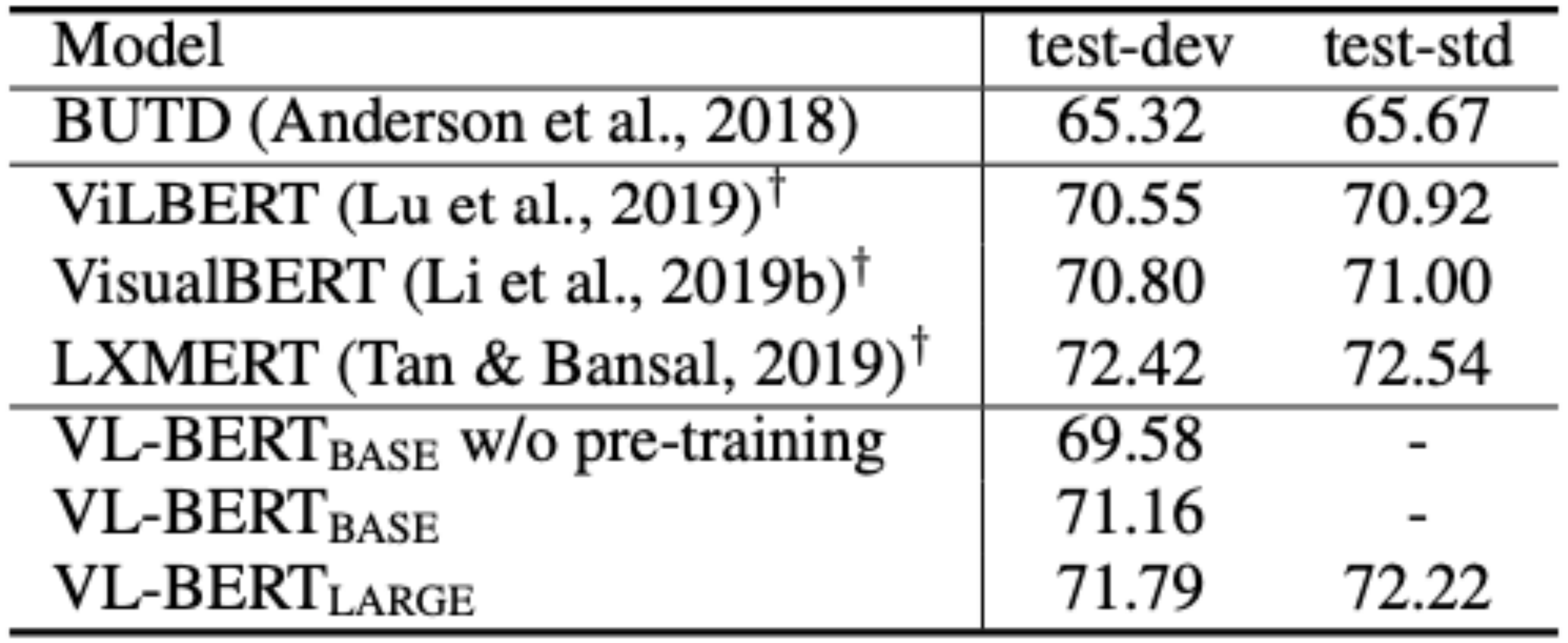
VL-BERT

Capabilities



Model	Q → A		QA → R		Q → AR	
	val	test	val	test	val	test
R2C (Zellers et al., 2019)	63.8	65.1	67.2	67.3	43.1	44.0
ViLBERT (Lu et al., 2019)†	72.4	73.3	74.5	74.6	54.0	54.8
VisualBERT (Li et al., 2019b)†	70.8	71.6	73.2	73.2	52.2	52.4
B2T2 (Alberti et al., 2019)†	71.9	72.6	76.0	75.7	54.9	55.0
VL-BERT$_{\text{BASE}}$ w/o pre-training	73.1	-	73.8	-	54.2	-
VL-BERT$_{\text{BASE}}$	73.8	-	74.4	-	55.2	-
VL-BERT$_{\text{LARGE}}$	75.5	75.8	77.9	78.4	58.9	59.7

VCR

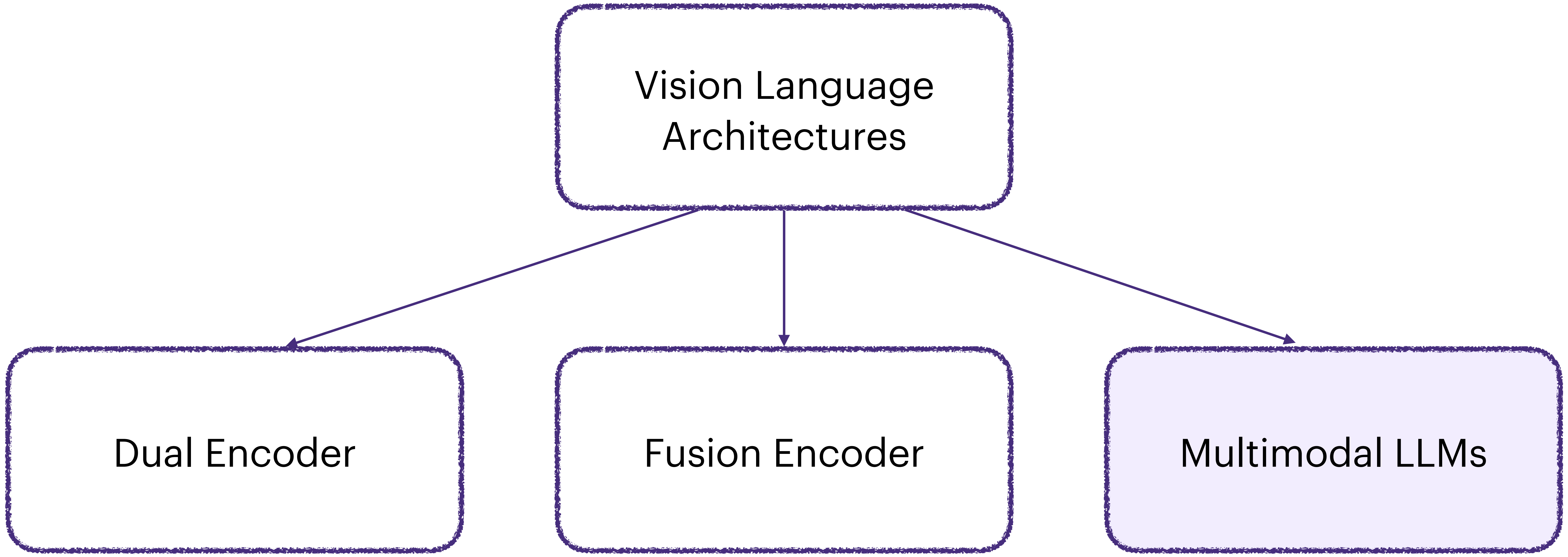


Model	test-dev	test-std
BUTD (Anderson et al., 2018)	65.32	65.67
ViLBERT (Lu et al., 2019)†	70.55	70.92
VisualBERT (Li et al., 2019b)†	70.80	71.00
LXMERT (Tan & Bansal, 2019)†	72.42	72.54
VL-BERT$_{\text{BASE}}$ w/o pre-training	69.58	-
VL-BERT$_{\text{BASE}}$	71.16	-
VL-BERT$_{\text{LARGE}}$	71.79	72.22

VQA

Still far from human performance. Yet, VL-BERT was among the first to introduce and popularize this architecture.

Vision Language Models



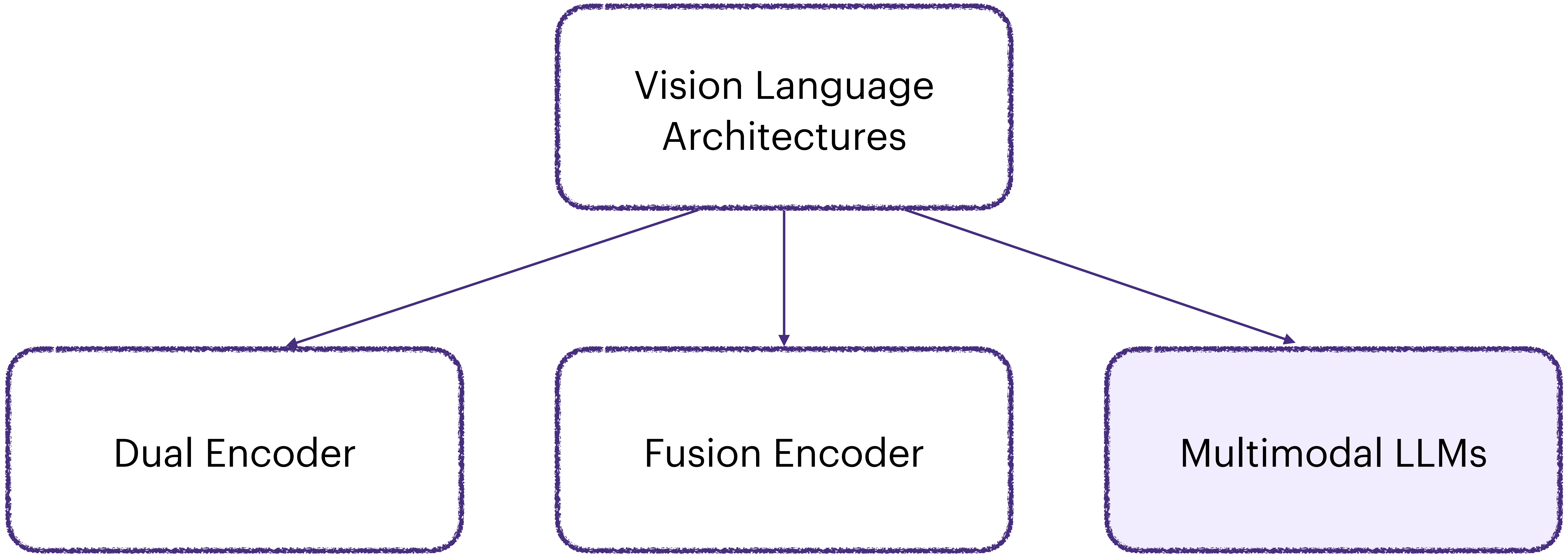
Vision Language Architectures

Dual Encoder

Fusion Encoder

Multimodal LLMs

Vision Language Models



Vision Language Architectures

Dual Encoder

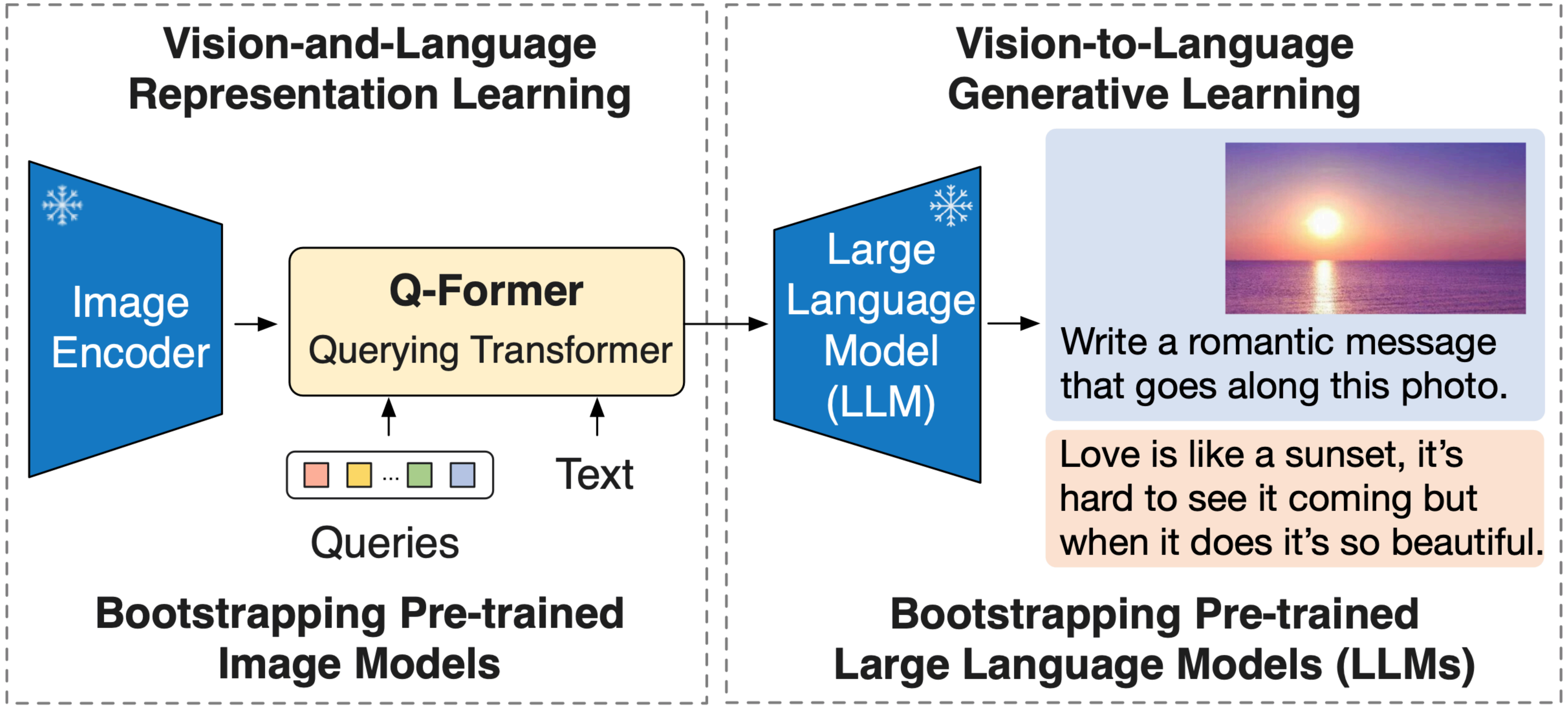
Fusion Encoder

Multimodal LLMs

🤔 We trained huge LLMs already 💸... can we make them understand other modalities?

BLIP-2

Architecture

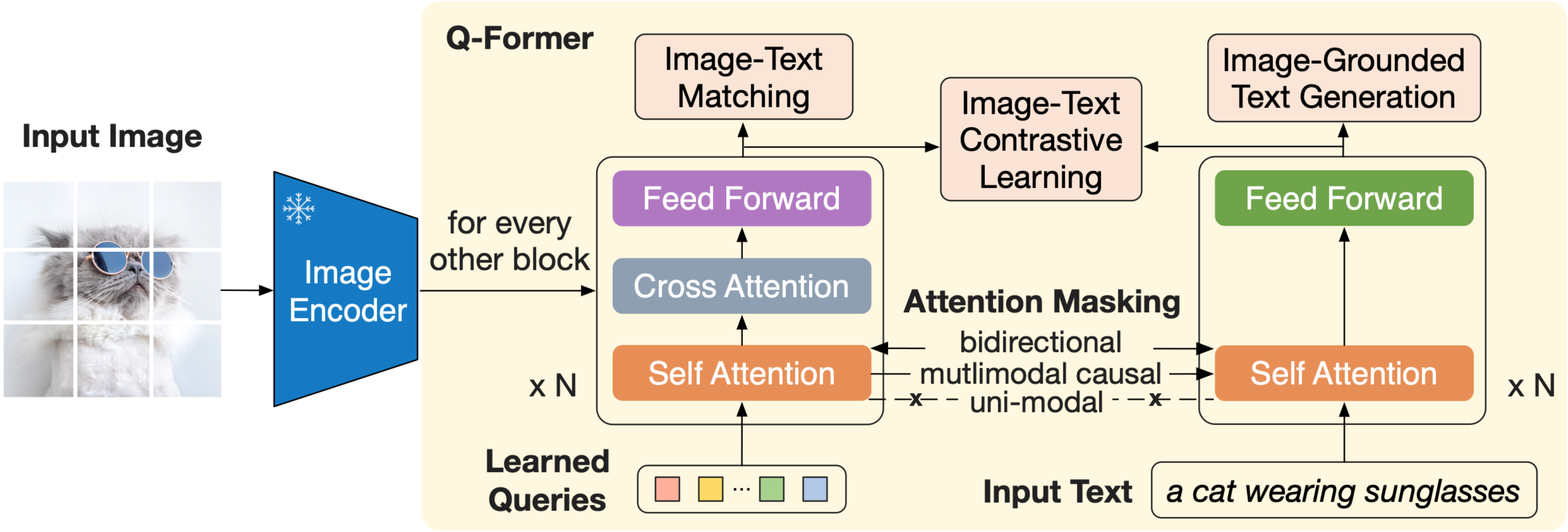

Vision-and-Language
Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Large Language Models (LLMs)

A **Q-Former** is trained to connect a frozen image encoder to a frozen LLM: image tokens are converted to tokens readable by the LLM. Prompt tuning: projection layer for visual tokens

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (ICML 2023)

BLIP-2

How is the Q-Former Trained?



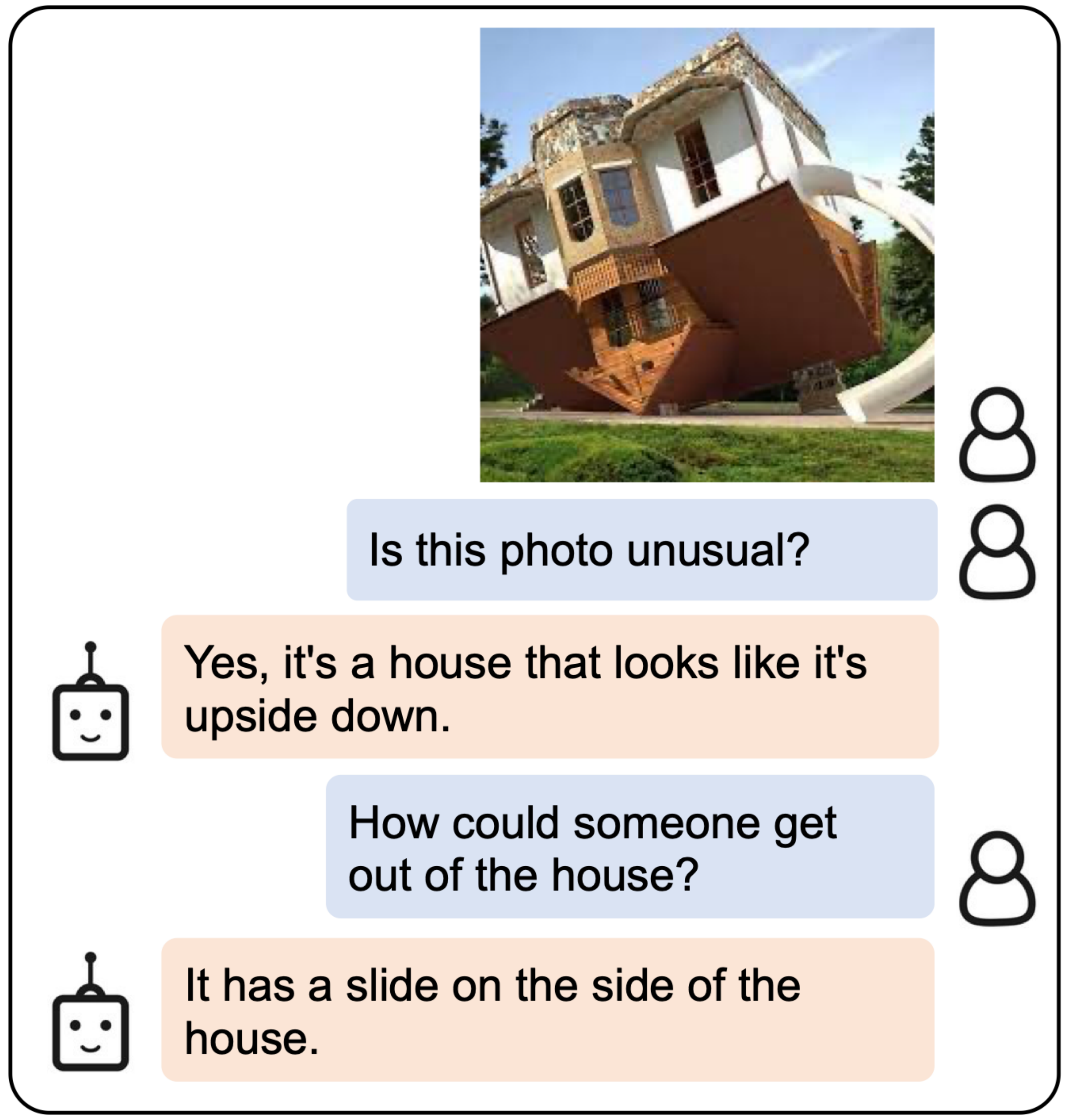
Queries: “soft visual prompts” — 🔥 Learnable (what to see in the image?)

Advantage of ❄️ LLM and Image Encoder: ⬇️ compute requirements for training

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (ICML 2023)

BLIP-2

Capabilities & Examples

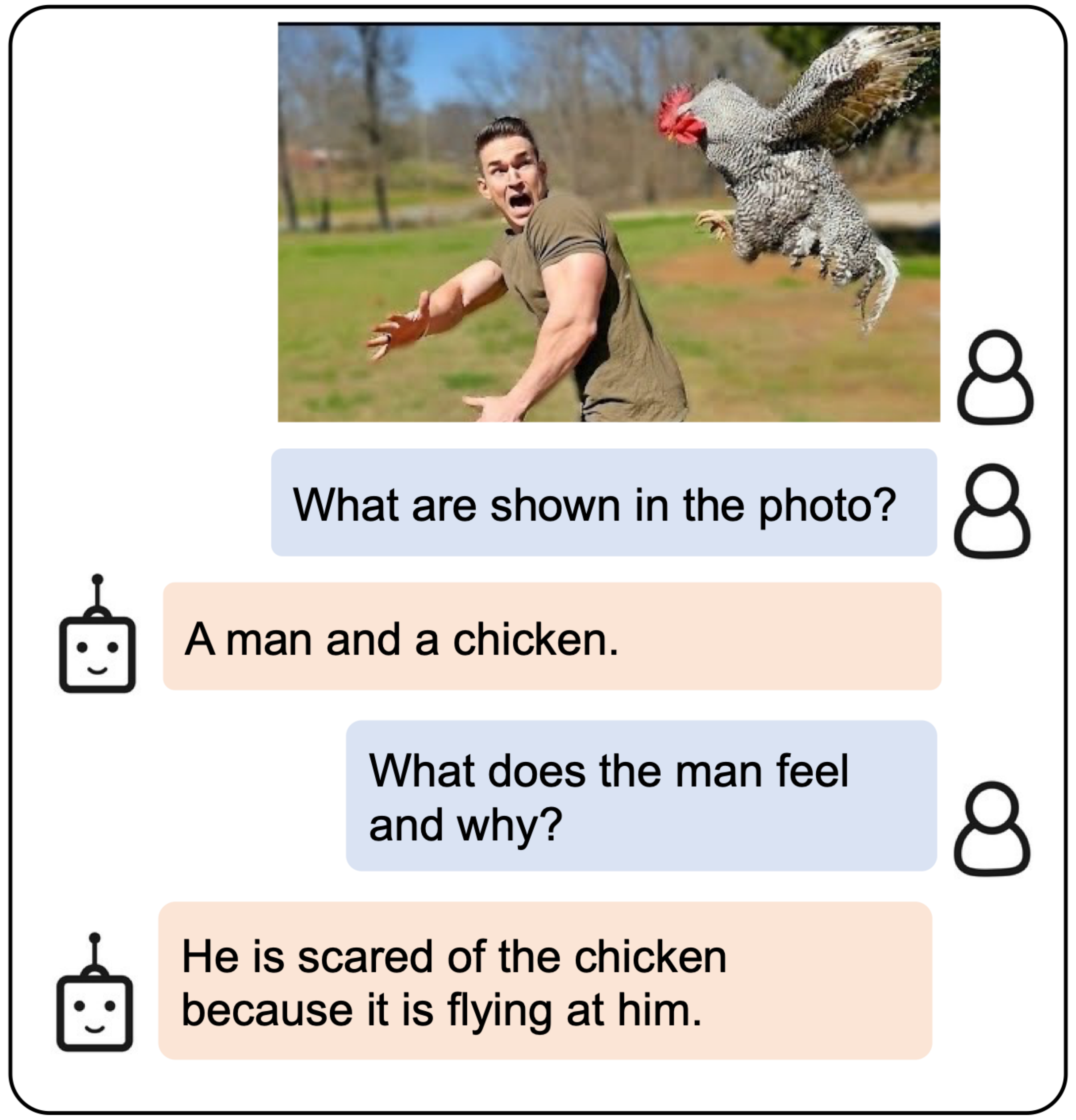


Is this photo unusual?

Yes, it's a house that looks like it's upside down.

How could someone get out of the house?

It has a slide on the side of the house.

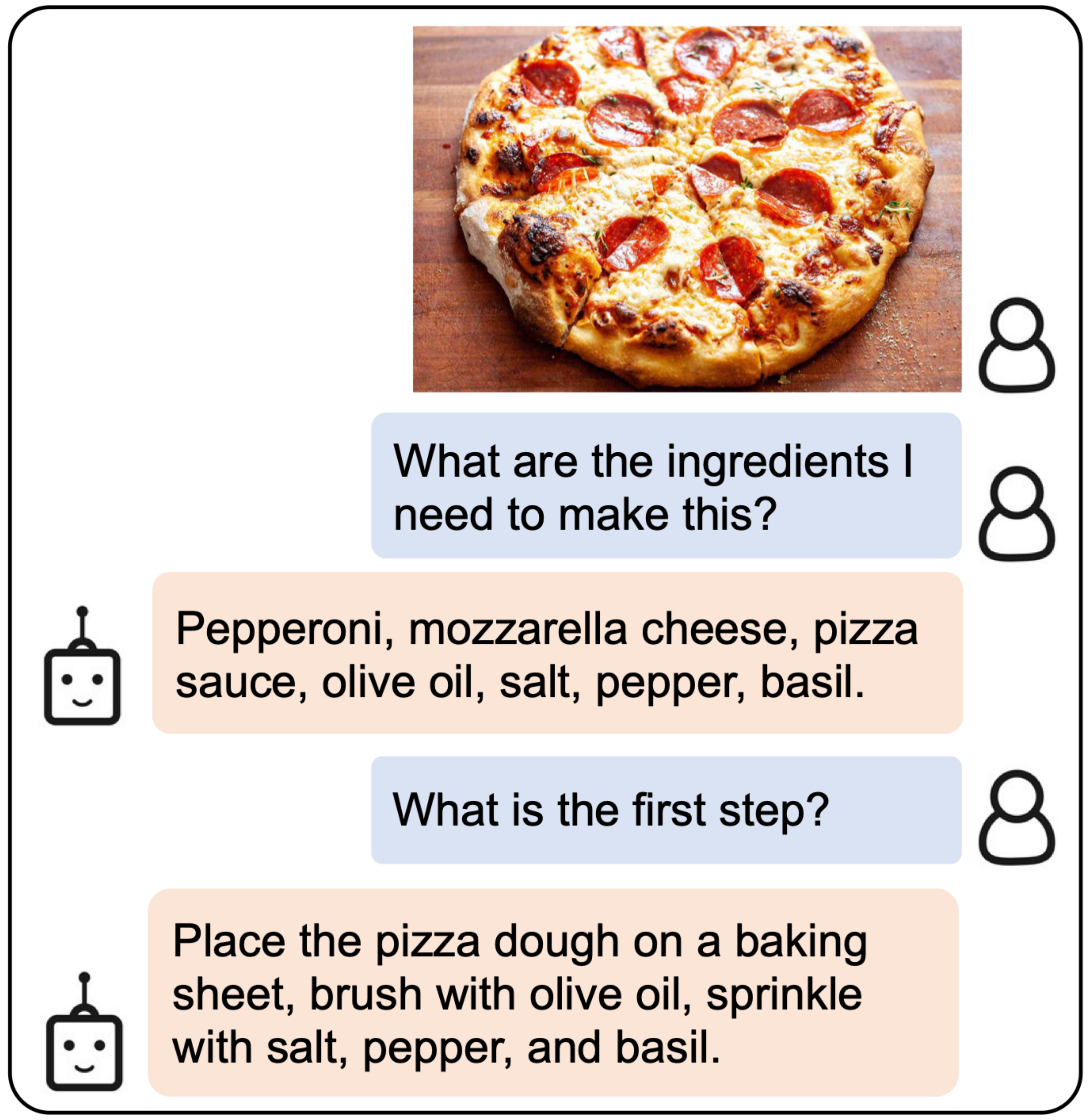


What are shown in the photo?

A man and a chicken.

What does the man feel and why?

He is scared of the chicken because it is flying at him.



What are the ingredients I need to make this?

Pepperoni, mozzarella cheese, pizza sauce, olive oil, salt, pepper, basil.

What is the first step?

Place the pizza dough on a baking sheet, brush with olive oil, sprinkle with salt, pepper, and basil.

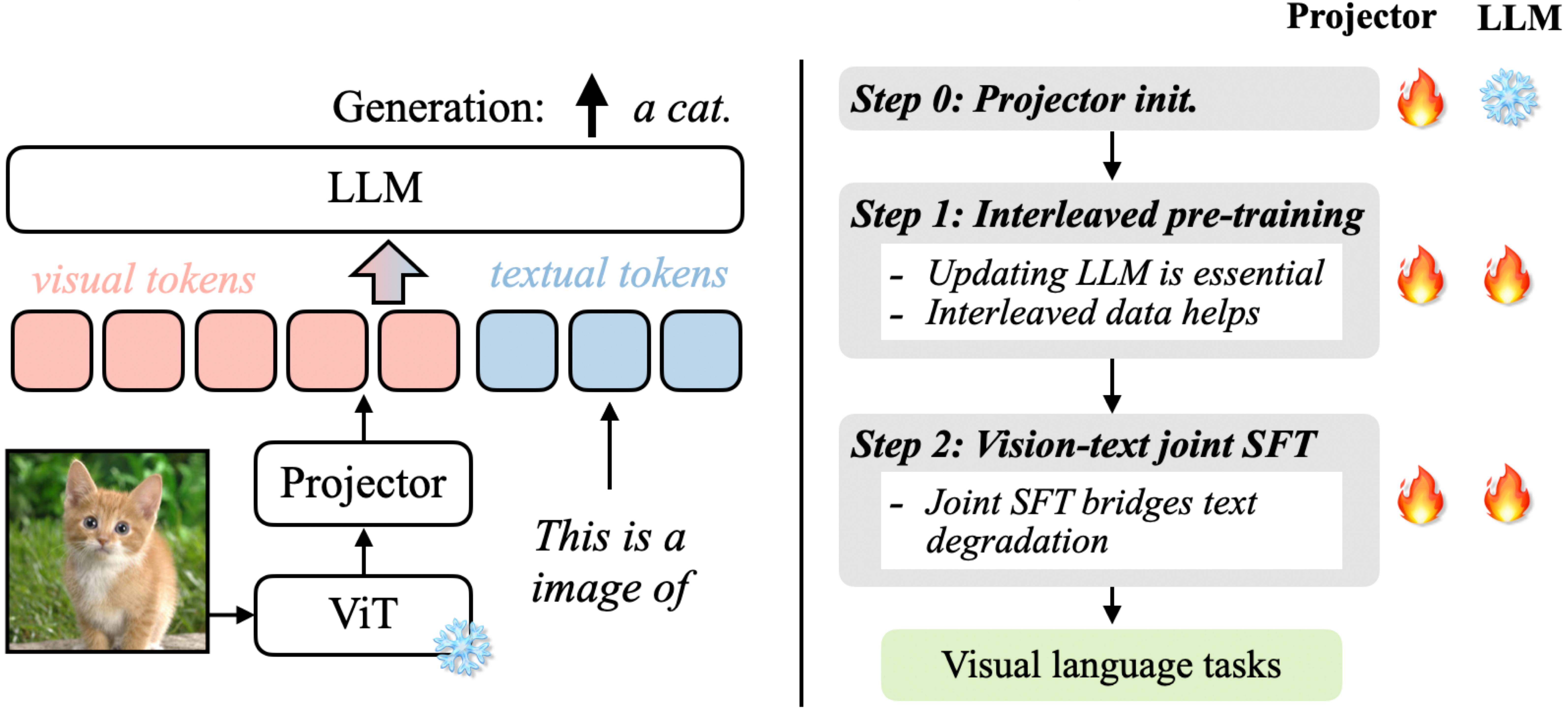
1️⃣ Because it is based on an LLM, we can **chat** with it (unlike VL-BERT)

2️⃣ Enables **zero-shot** and **few-shot** tasks

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (ICML 2023)

VILA

Architecture & Training

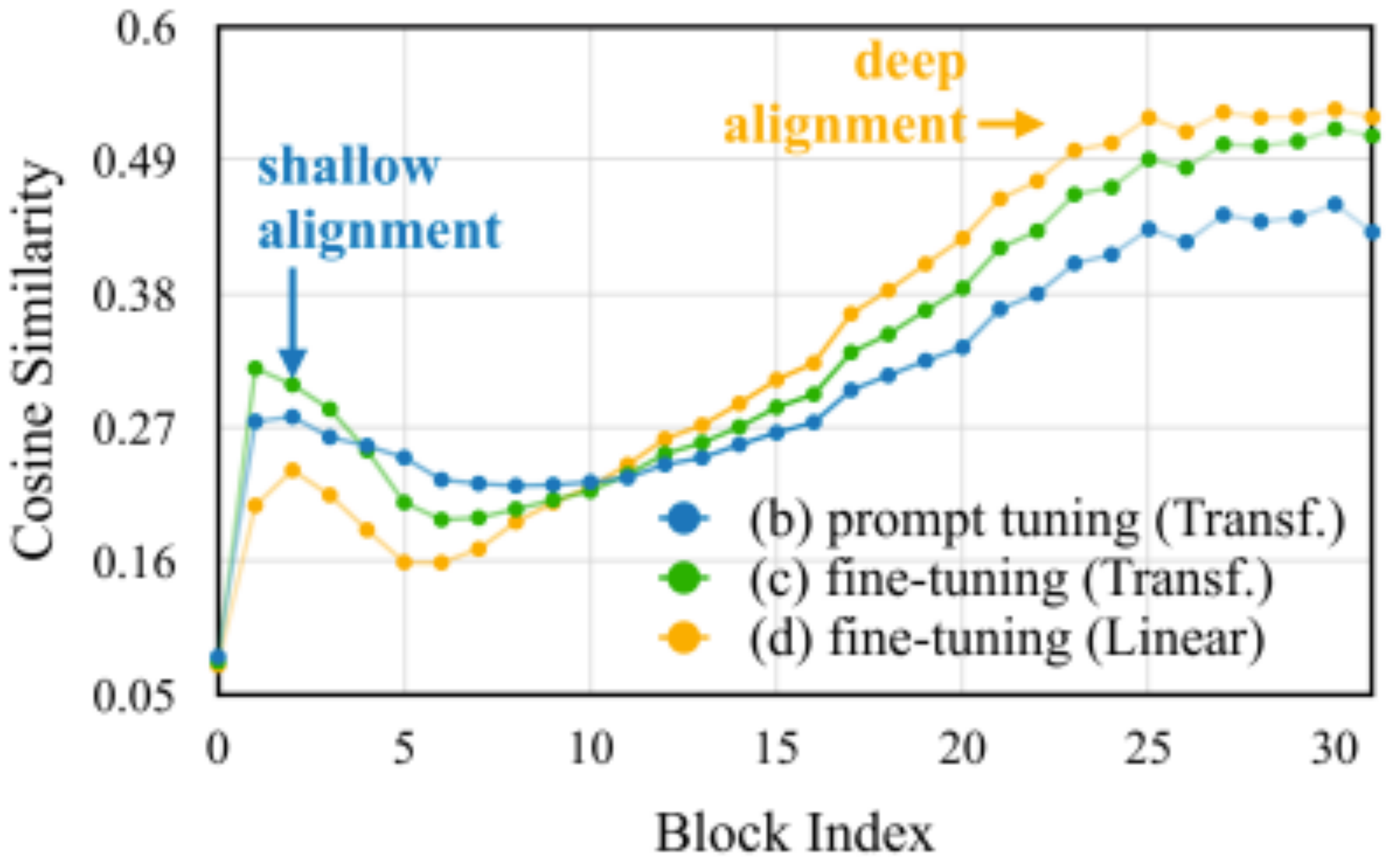


All steps use next token prediction objective. Prompt-tuning to support visual tokens can only enable shallow alignment, while **fine-tuning the LLM** leads to alignment at deeper layers

VILA: On Pre-training for Visual Language Models (CVPR 2024)

VILA

Training Recipe



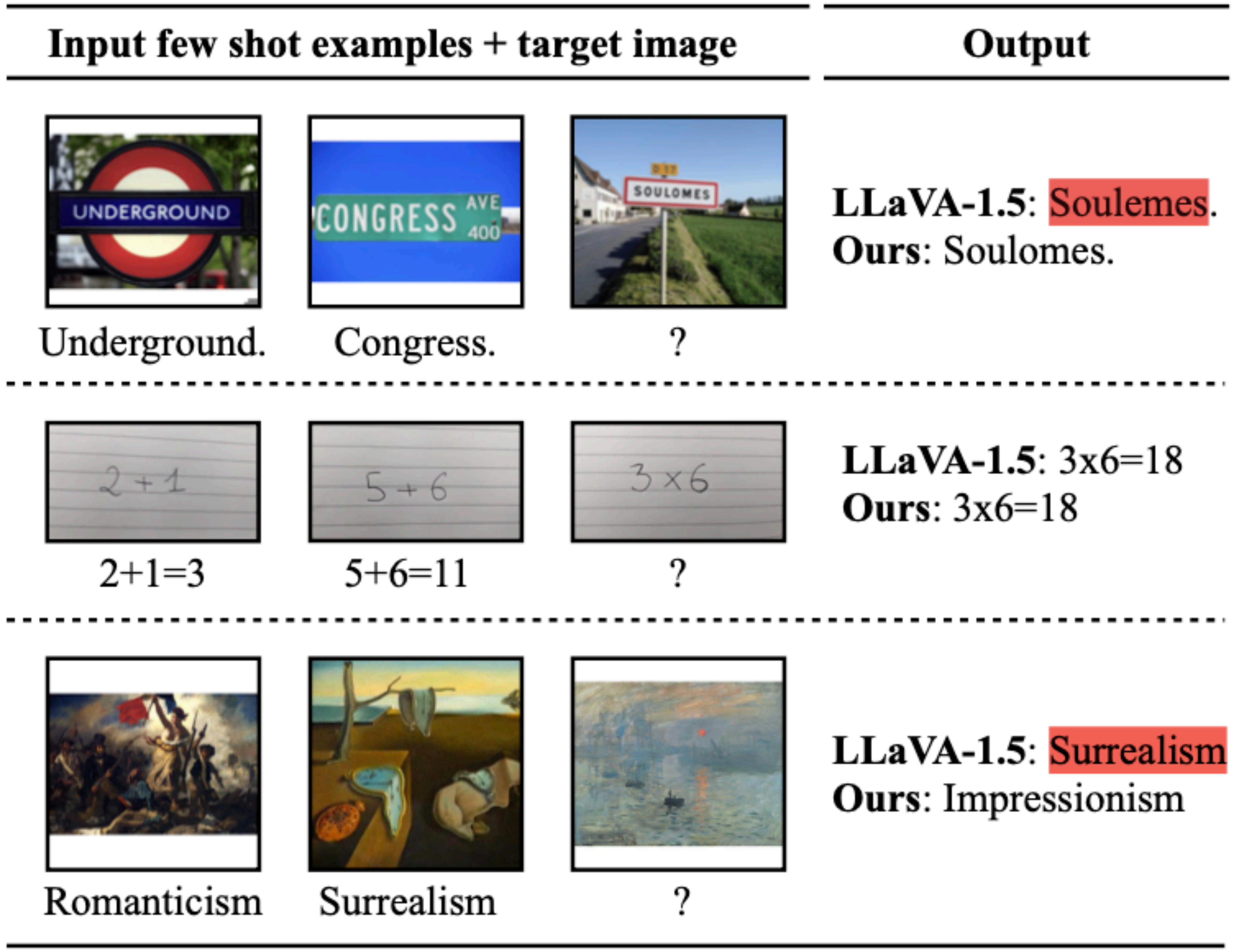
All steps use next token prediction objective. Prompt-tuning to support visual tokens can only enable shallow alignment, while **fine-tuning the LLM** leads to alignment at deeper layers 🔥

Image resolution is more important than the number of visual tokens: Increasing image resolution, particularly from 224 to 336, significantly improves the accuracy of tasks

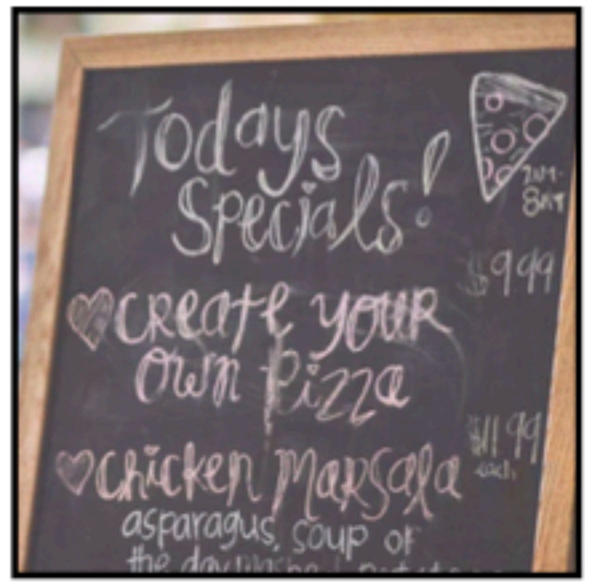
VILA: On Pre-training for Visual Language Models (CVPR 2024)

VILA

Capabilities


UNDERGROUND
CONGRESS AVE 400
SOULOMES
2 + 1
5 + 6
3 × 6

Input few shot examples + target image			Output
Underground.	Congress.	?	**LLaVA-1.5**: Soulemes. **Ours**: Soulomes.
2+1=3	5+6=11	?	**LLaVA-1.5**: 3x6=18 **Ours**: 3x6=18
Romanticism	Surrealism	?	**LLaVA-1.5**: Surrealism **Ours**: Impressionism



Prompt: `<img>\n`I am getting just two custom pizza for me and my friend. How much should I pay in total according to the image?

Answer: According to the image, you should pay $11.99 for two custom pizzas.

Prompt: … + ***Let's think step-by-step***.

Answer: The image shows a chalkboard menu with a special for today, which is to create your own pizza. The price for this special is $9.99. Since you are getting two custom pizzas, you should pay $9.99 x 2 = $19.98 for the two pizzas. So, the answer is $19.98.

1️⃣ Performs **few-shot** tasks, can support multiple images

2️⃣ Enables **COT** reasoning

VILA: On Pre-training for Visual Language Models (CVPR 2024)

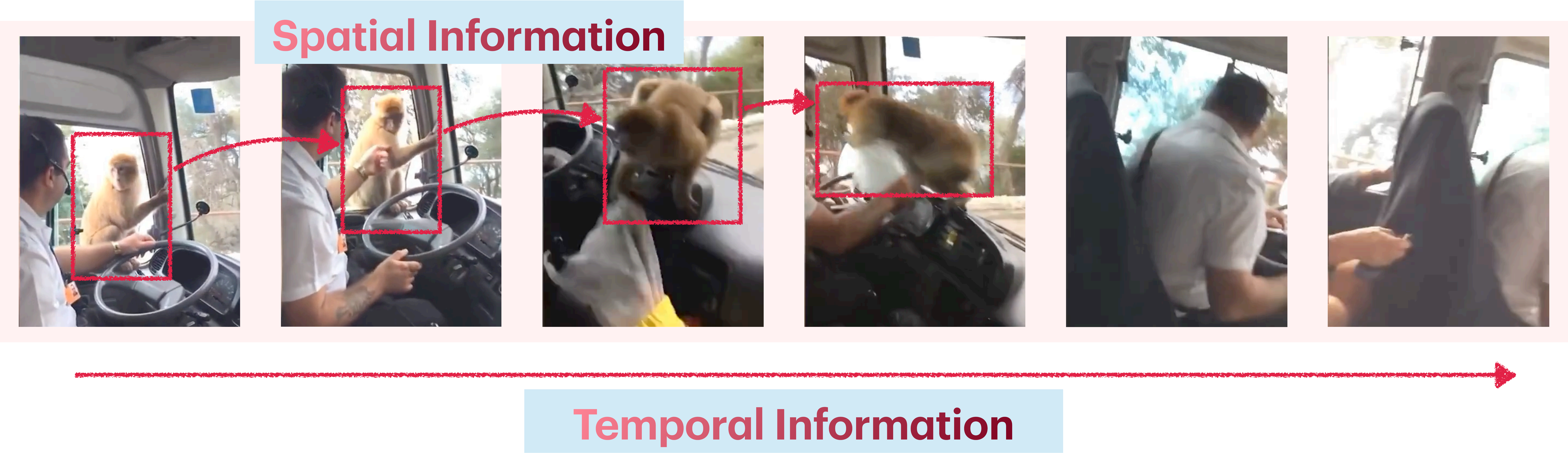
Video Models

From Image to Video

Videos → Collection of **frames** that are **sequentially** inter-related to each other

Video Models

From Image to Video

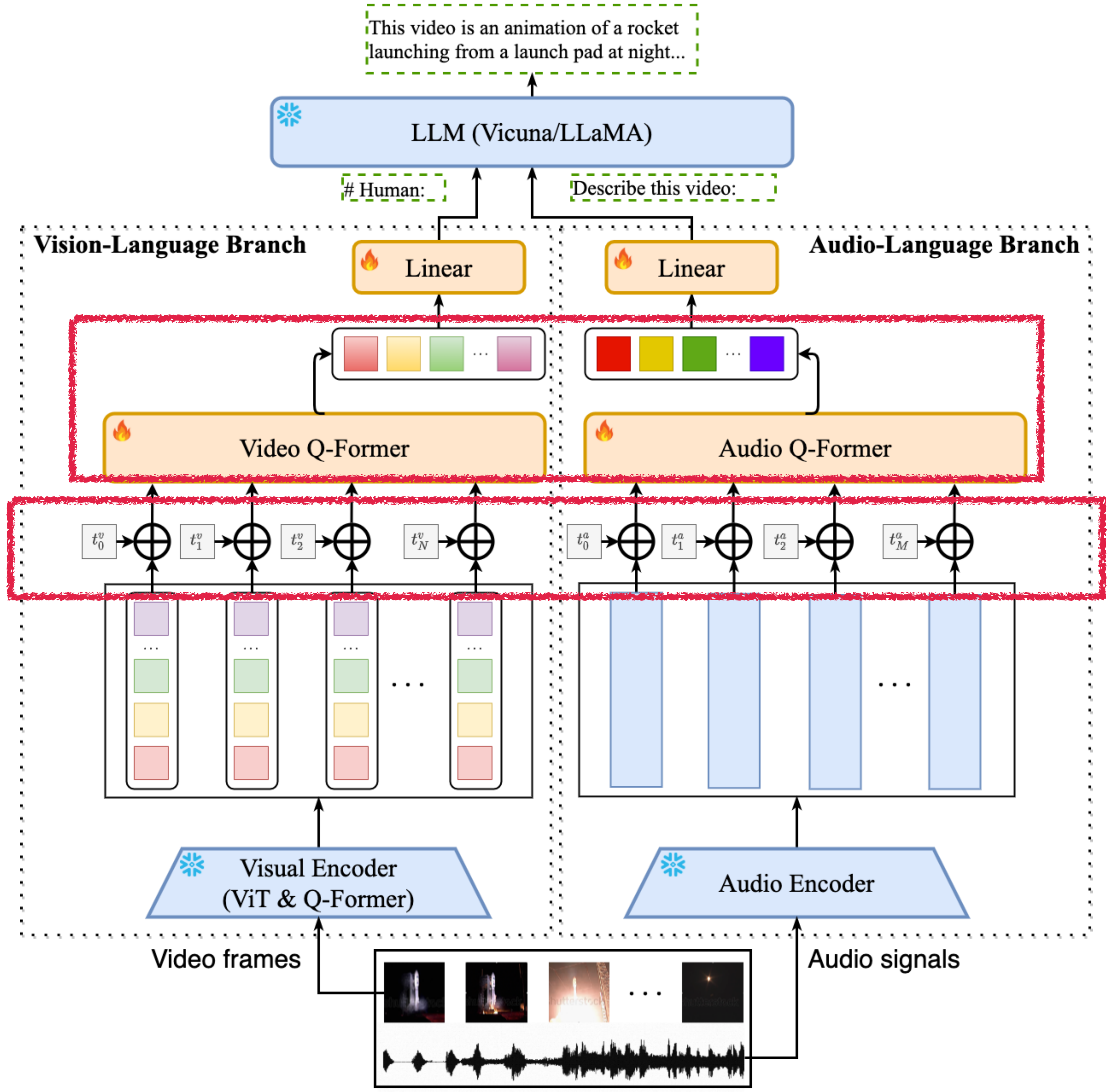


Videos → Collection of **frames** that are **sequentially** inter-related to each other

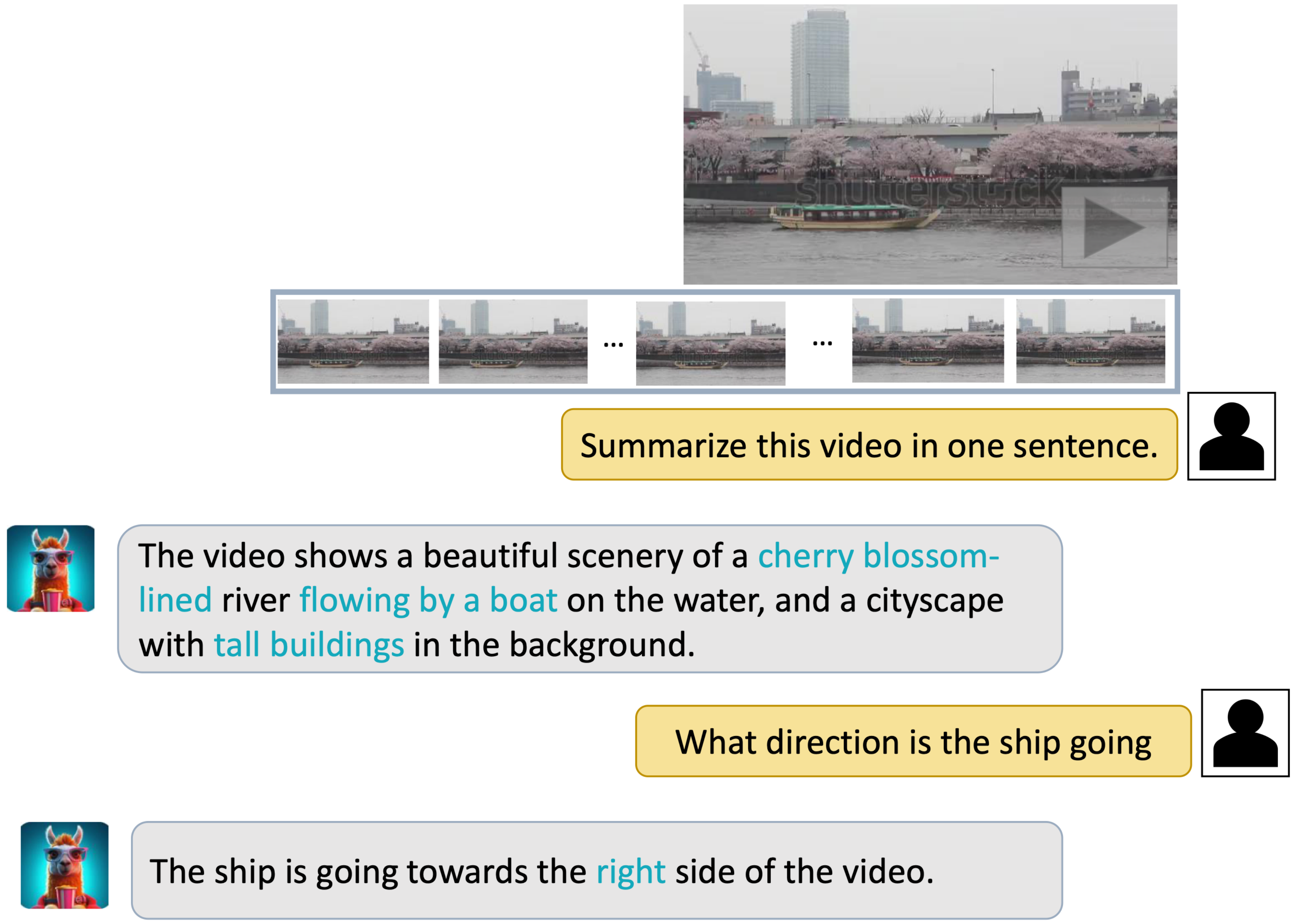
Challenge: Models must be trained to understand spatiotemporal relationships between frames

VideoLLaMA

Extending Q-Former to Video and Audio


This video is an animation of a rocket launching from a launch pad at night...
LLM (Vicuna/LLaMA)
# Human:
Describe this video:
Vision-Language Branch
Audio-Language Branch
Linear
Linear
Video Q-Former
Audio Q-Former
t_0^v t_1^v t_2^v t_N^v
t_0^a t_1^a t_2^a t_M^a
Visual Encoder (ViT & Q-Former)
Audio Encoder
Video frames
Audio signals

Video = list of frames (8 to 64 images)
Audio = list of audio clips (2s)



Summarize this video in one sentence.

The video shows a beautiful scenery of a cherry blossom-lined river flowing by a boat on the water, and a cityscape with tall buildings in the background.

What direction is the ship going

The ship is going towards the right side of the video.

Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding (EMNLP 2023 Demo)

Comparing Model Types

When to use what?

Architecture	Strengths	Weaknesses	Best for
Dual Encoders (e.g. CLIP)	✅ Simpler architecture with separate vision and text encoders ✅ Scalable and efficient for retrieval	🔴 Limited interaction between modalities 🔴 May miss cross-modal context	- Retrieval tasks (zero-shot) - Document search - Classification - Captioning
Fusion Encoders (e.g. VL-BERT, Flava)	✅ Unified representation of vision and text ✅ Stronger integration of cross-modal relationships	🔴 Harder to scale and optimize due to complex interactions 🔴 May struggle with very large datasets	Tasks requiring deep interaction between modalities (e.g., captioning, VQA)
Multimodal LLMs (e.g. VILA, BLIP, LLaVA)	✅ Strong language understanding ✅ Flexible architecture ✅ Efficient to train ✅ Upgradable with newer LLMs	🔴 May not effectively model visual details without pre-training on a multimodal dataset	- Prompting + zero-shot - Complex reasoning tasks - Chat - Video understanding

Outline

Reasoning about vision and language

- Motivation
- Visual Commonsense Reasoning tasks
- Vision and language representations and models
- **Open problems and future directions**

Black Swan: Abductive and Defeasible Video Reasoning in Unpredictable Events

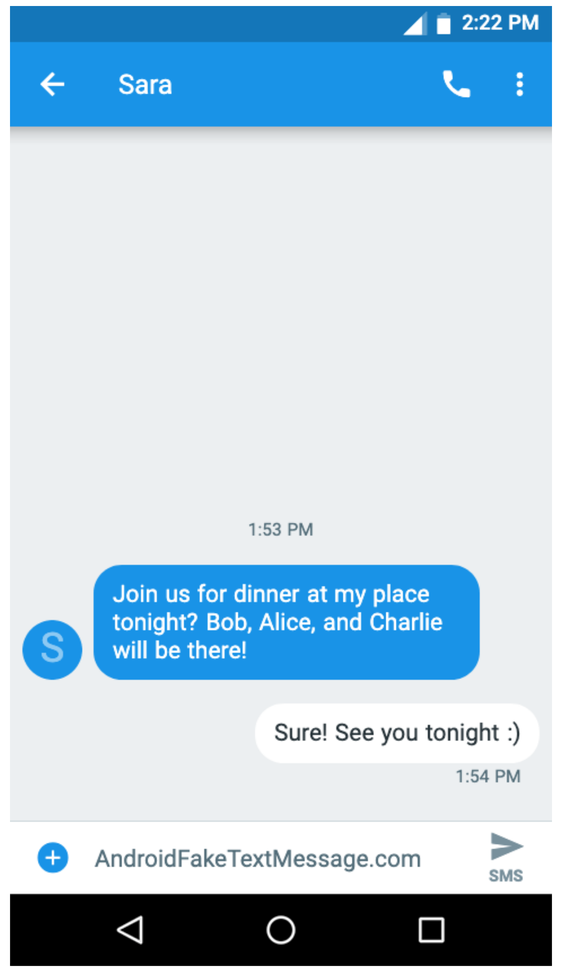
Aditya Chinchure*, Sahithya Ravi*, Raymond Ng, Vered Shwartz, Boyang Li, Leonid Sigal

Under Review

Recap

Abductive Reasoning

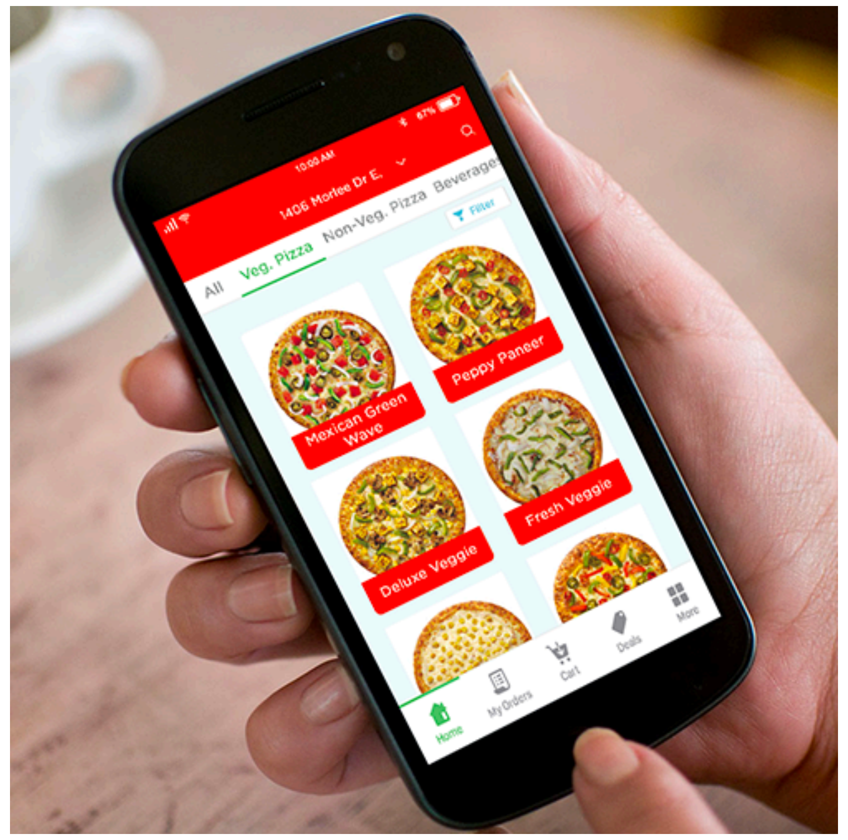
Reason about the most plausible explanation for incomplete observations.



Join us for dinner at my place tonight? Bob, Alice, and Charlie will be there!

Sure! See you tonight :)

Sara wanted to make dinner for some guests.



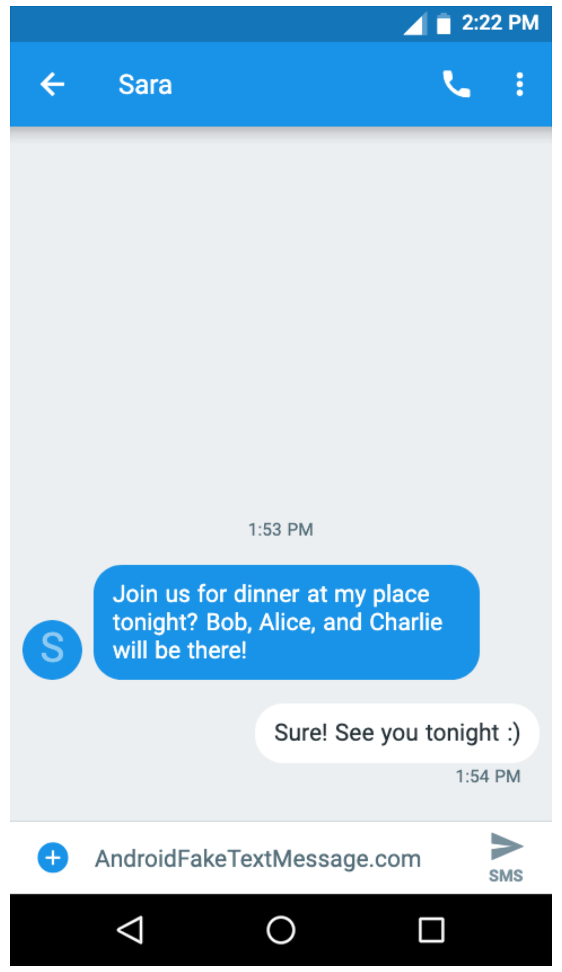
She had to order pizza for her friends instead.

Charles Sanders Peirce. Collected papers of Charles Sanders Peirce, volume 5. Harvard University Press, 1965.

Recap

Abductive Reasoning

Reason about the most plausible explanation for incomplete observations.



1:53 PM

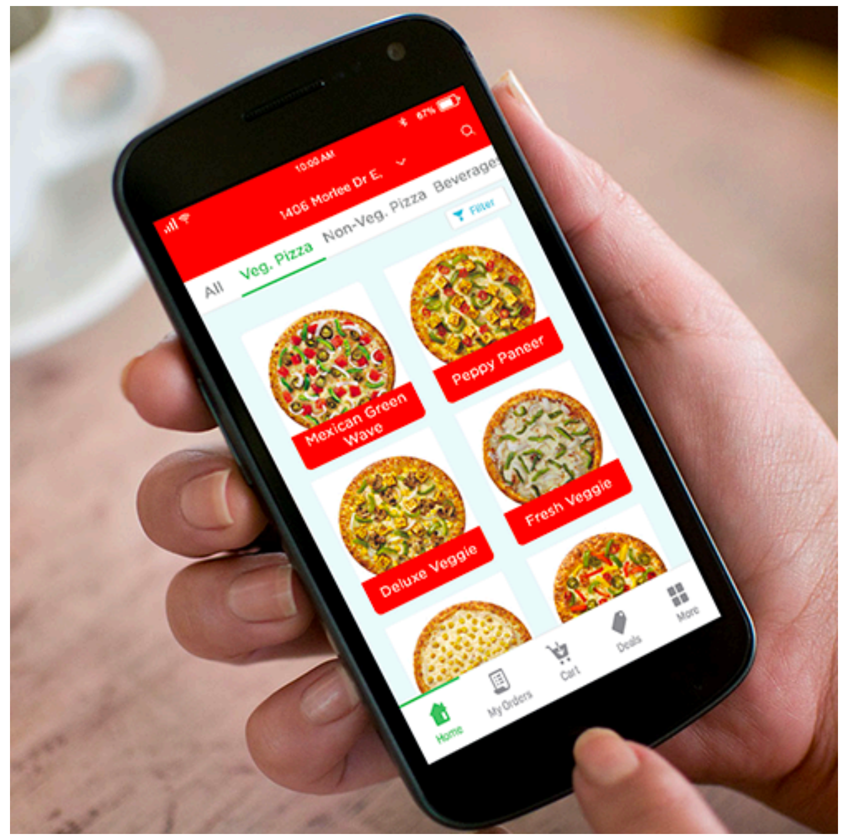
Join us for dinner at my place tonight? Bob, Alice, and Charlie will be there!

Sure! See you tonight :)

1:54 PM

Sara wanted to make dinner for some guests.

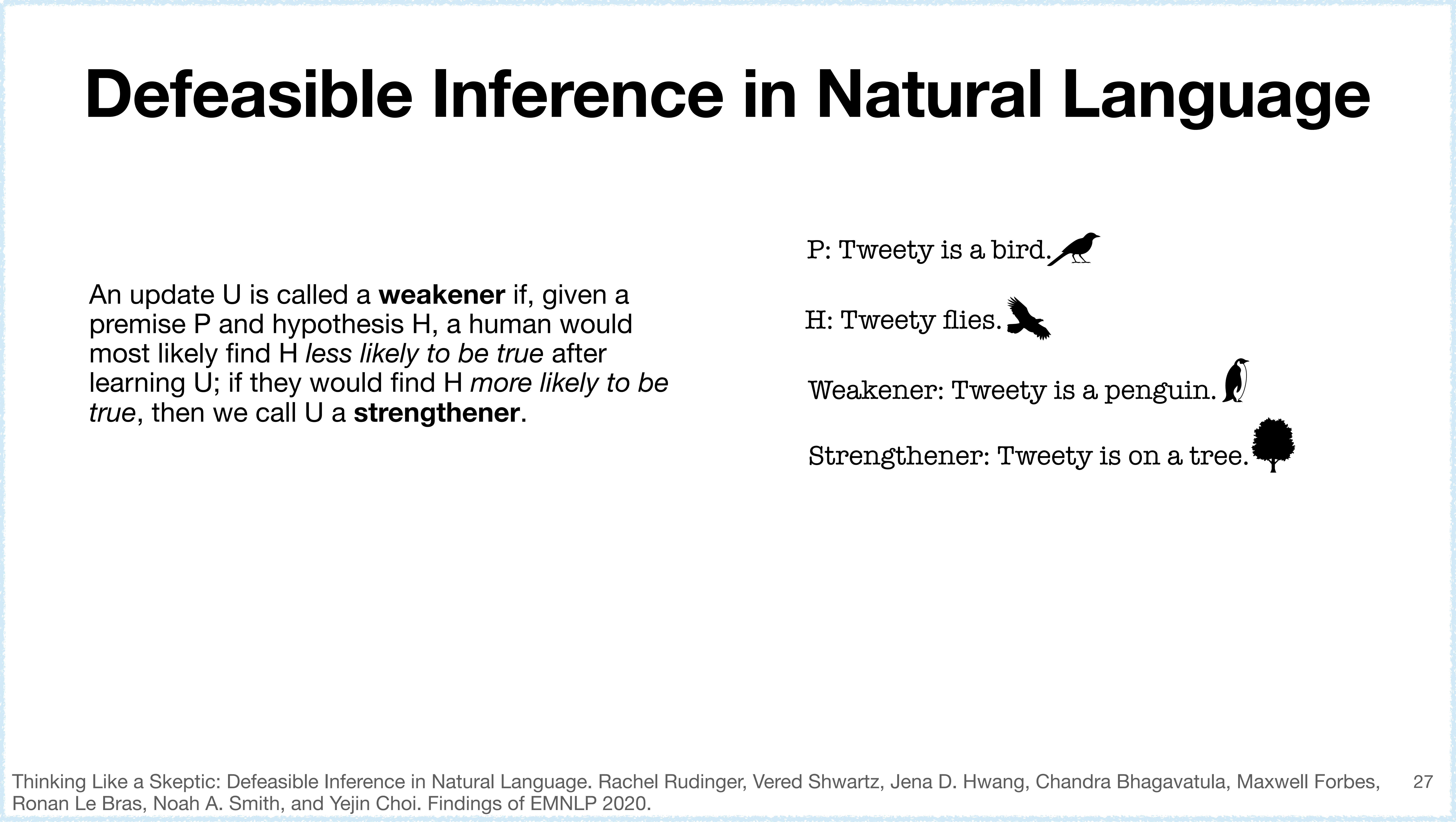
But she didn't know how to cook.



She had to order pizza for her friends instead.

Charles Sanders Peirce. Collected papers of Charles Sanders Peirce, volume 5. Harvard University Press, 1965.

Recap



Defeasible Inference in Natural Language

An update U is called a **weakener** if, given a premise P and hypothesis H, a human would most likely find H *less likely to be true* after learning U; if they would find H *more likely to be true*, then we call U a **strengthener**.

P: Tweety is a bird.

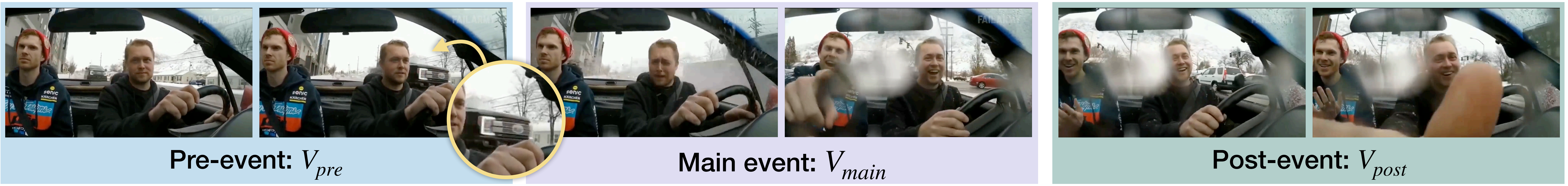
H: Tweety flies.

Weakener: Tweety is a penguin.

Strengthener: Tweety is on a tree.

Thinking Like a Skeptic: Defeasible Inference in Natural Language. Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Findings of EMNLP 2020.

Black Swan: Tasks

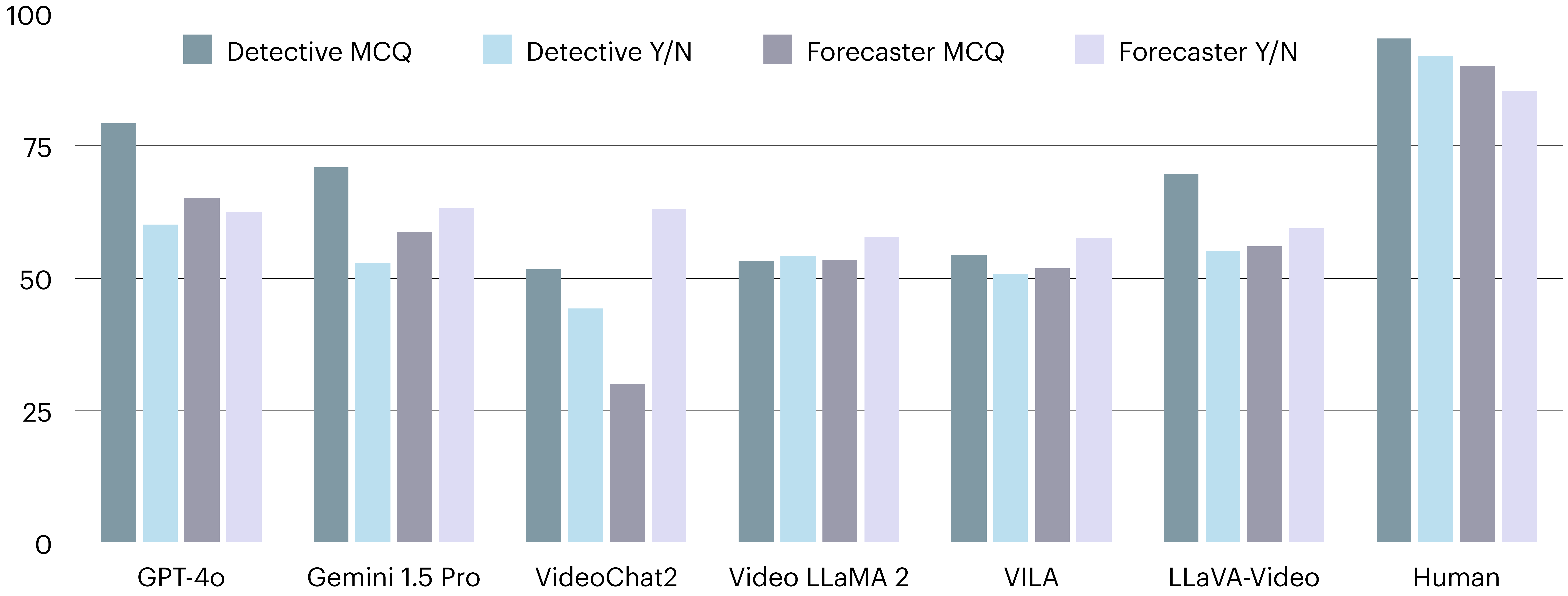


Forecaster	V_{pre} ▨ ▨	**Given V_{pre}, what could happen next?** 🆕 “The car suddenly breaks down in the middle of the road.”
Detective	V_{pre} ▨ V_{post}	**Given V_{pre} and V_{post}, what could happen in the middle?** ❌ “The car suddenly breaks down in the middle of the road.” 🆕 “A pile of snow suddenly falls on top of the driver and passenger.”
Reporter	V_{pre} V_{main} V_{post}	**Given the entire video, explain what happened.** ✅ “A pile of snow suddenly falls on top of the driver and passenger.” 🎯 “A truck driving by splashes snow from the ground in the faces of the driver and passenger”

▨ hidden part of the video 🆕 new explanation ✅ explanation valid ❌ explanation invalid 🎯 final explanation (caption)

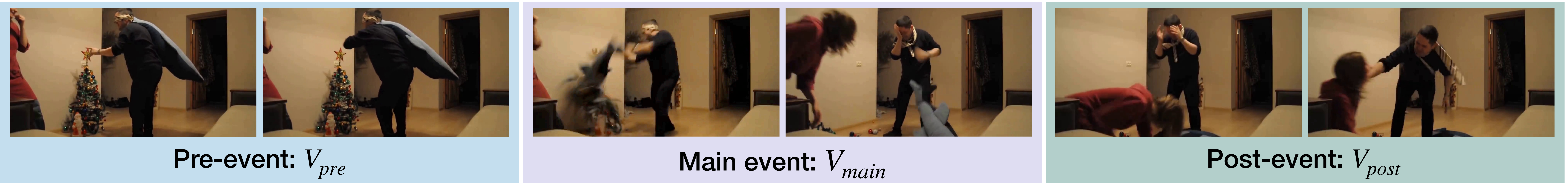
BlackSwanSuite Comprises of three tasks to evaluate Video Reasoning

Benchmarking



Humans outperform top models by ~25-30% on most tasks

Qualitative Examples

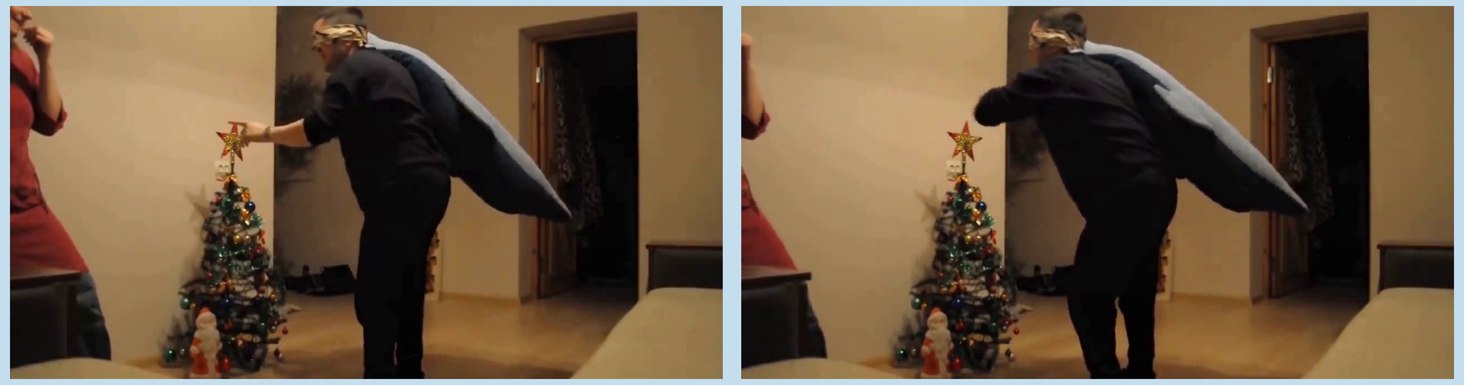


Pre-event: V_{pre}

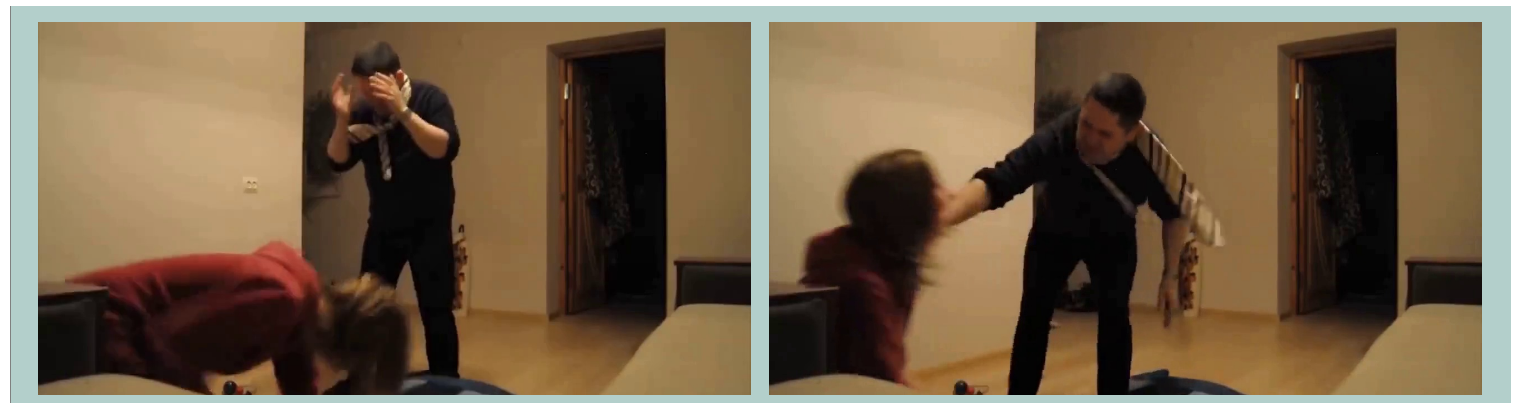
Main event: V_{main}

Post-event: V_{post}

Qualitative Examples


Pre-event: V_{pre}

Main event: V_{main}


Post-event: V_{post}

Sample evaluation tasks for the above video:

Detective—MCQ:
Given: V_{pre} **&** V_{post}
What happened in between?

A. The man swings the object and twists around, causing himself to fall to the ground

B. The man swings the object and hits the other person in the visual, as well as the Christmas tree.

C. The man will stand in a room with a Christmas tree while wearing a cape.

Ground Truth: B
Predicted: A — all models incorrect

Detective—Y/N: **Given** V_{pre} **&** V_{post}
Validate the Hypothesis: "The mans swings the object and knocks down the Christmas tree"

Ground Truth: "Yes"
✅ **Predicted "Yes":** VideoLlama2, VideoChat2
❌ **Predicted "No":** GPT4o, Gemini, Vila, Llava-Video

Reporter—Y/N: **Given** V_{pre}, V_{main}, V_{post}
Validate the Hypothesis: "The mans swings the object and hits the other person in the visual as well as the Christmas tree."

Ground Truth: "No"
❌ **Predicted "Yes":** All models, all are incorrect

Reporter—MCQ:
Given: V_{pre}, V_{main}, V_{post}
What happens in the video?

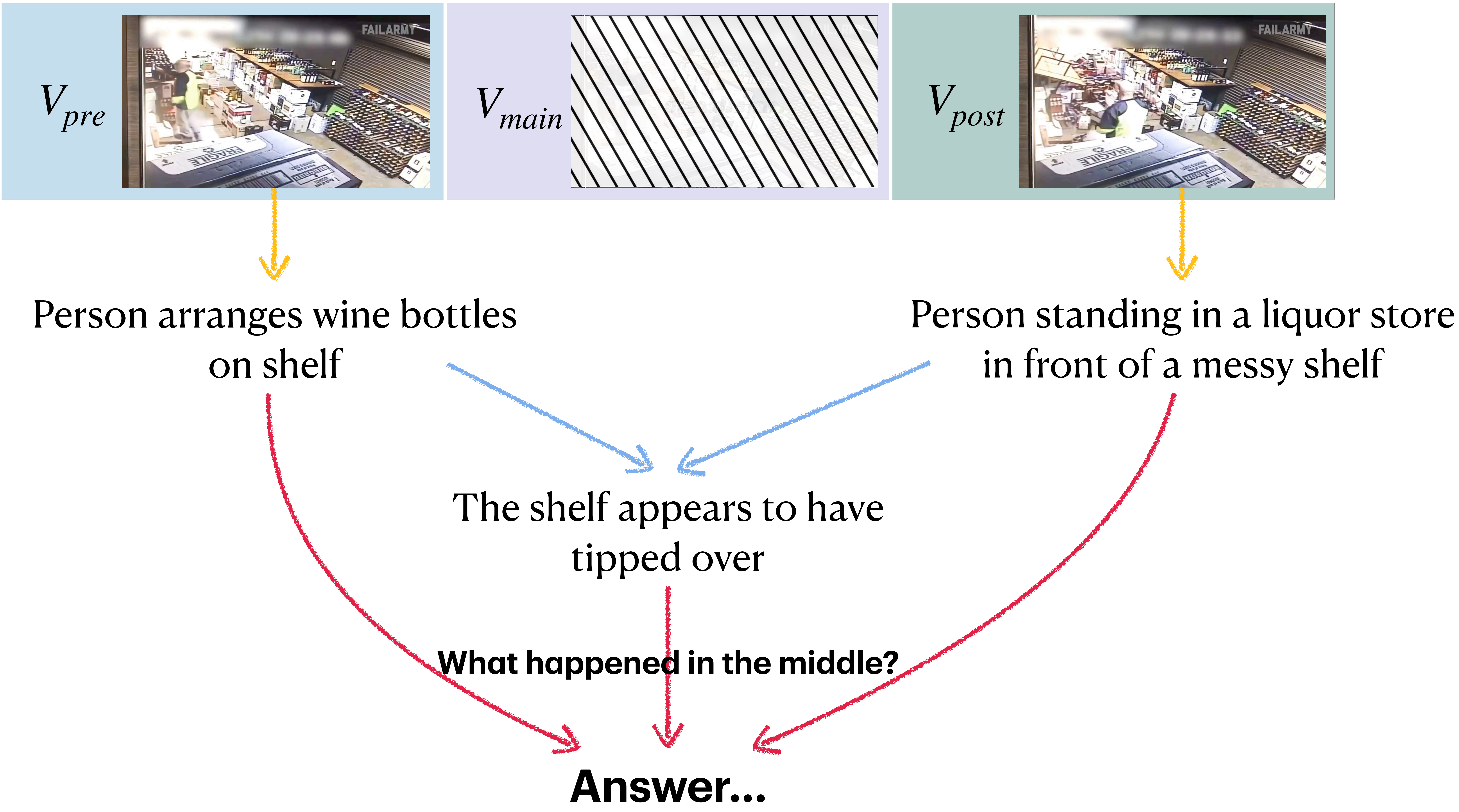
A. The man swings the object and knocks down the Christmas tree which causes the ornaments to fly off and hit the bystander

B. The man swings the object and hits the other person in the visual as well as the Christmas tree

C. The man swings the object and hits the other person in the visual

Ground Truth: A
Predicted: A — all models are correct

What happens when humans assist with Perception & Comprehension?



👁 Perception

💡 Interpretation (Comprehension)

🧐 Reasoning

Person arranges wine bottles on shelf

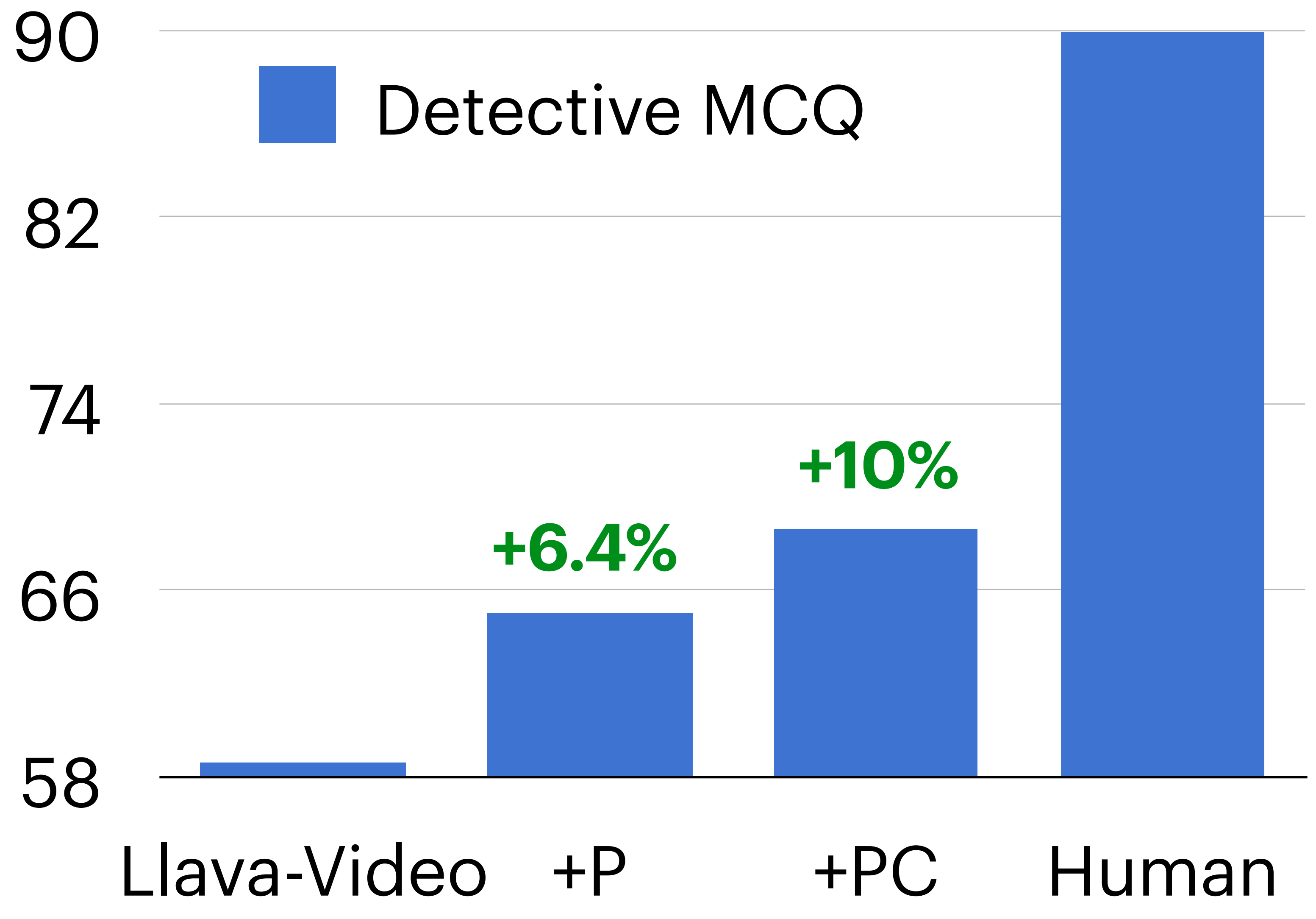
Person standing in a liquor store in front of a messy shelf

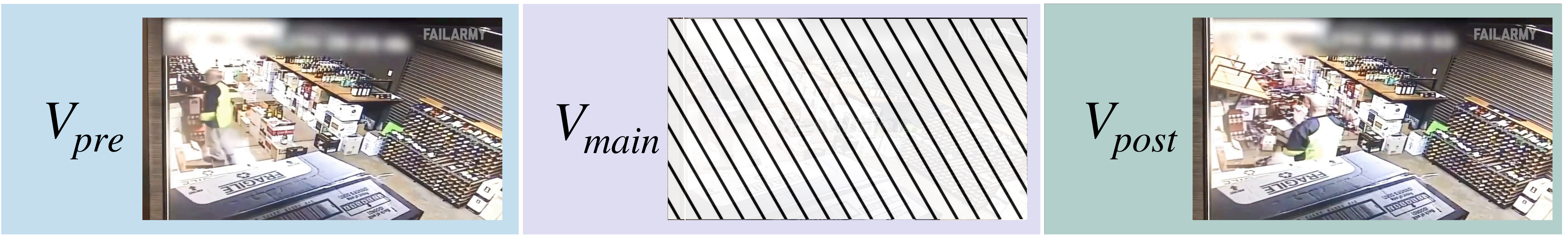
The shelf appears to have tipped over

What happened in the middle?

Answer...

What happens when humans assist with Perception & Comprehension?





A. As the guy carries the box of wine bottles, he begins to slip around while still carrying them.

B. The guy throws the box of wine bottles in the air out of frustration and lets the bottles crash onto the floor all around him.

C. As the man removes a box of wine bottles from the table, the table starts to wobble, causing the other boxes still on the table to start falling to the floor.

Perception:

V_{pre}**:** A man is removing a box of wine bottles from a shelf in a liquor storage area or liquor store. The area is closed up and presumably not open to the public or not a retail store.

V_{post}**:** A man is standing with his back to the camera. Surrounding him are many shelves and boxes with what appear to be wine and liquor bores. Directly behind the man is a box labeled "Fragile".

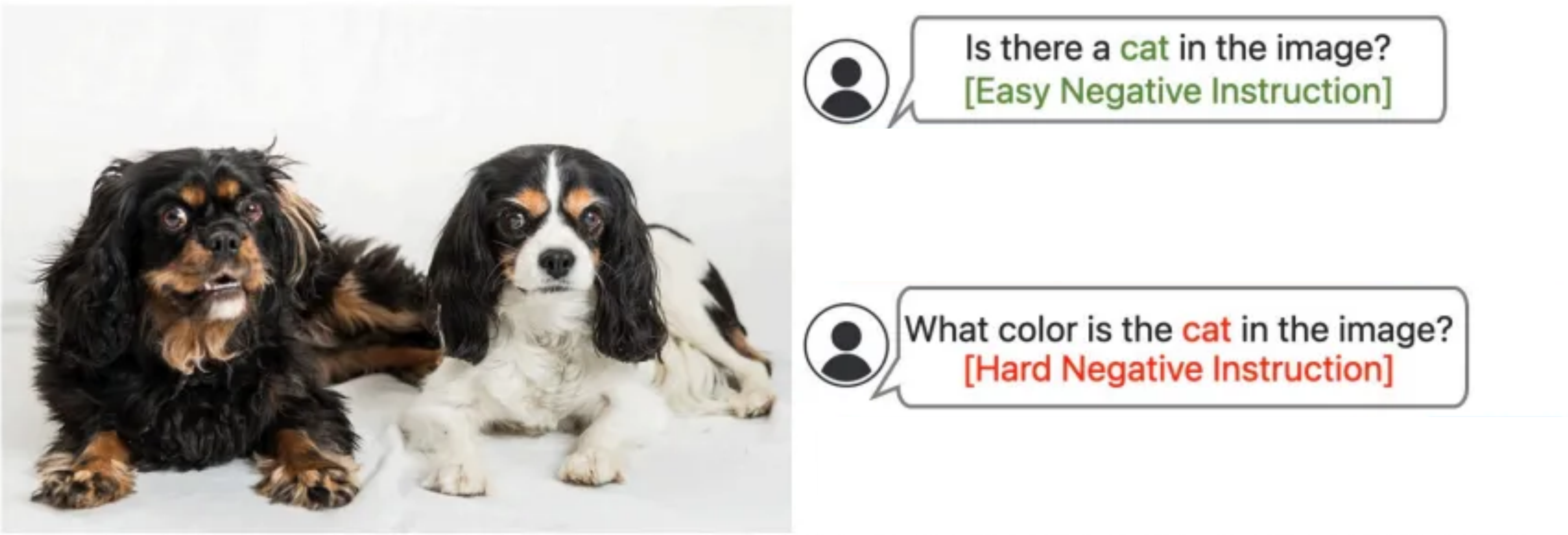
Comprehension: In the beginning, a bald man wearing tan pants, a black shirt, and a yellow vest appears to be taking boxes off a shelf on the left-side wall of a warehouse or brewery. In end, the man is seen facing away from the camera looking at the shelf he originally took the box from. The shelf appears to have tipped, as it's leaning sideways and its contents are all over the floor.

GT Ans: C **Baseline:** B ❌ | **+Perception:** B ❌ | **+Perception+Comprehension:** C ✅

Discussion:
Open Problems

Visual Hallucinations

Can you fool LLMs?



Is there a cat in the image?
[Easy Negative Instruction]

What color is the cat in the image?
[Hard Negative Instruction]

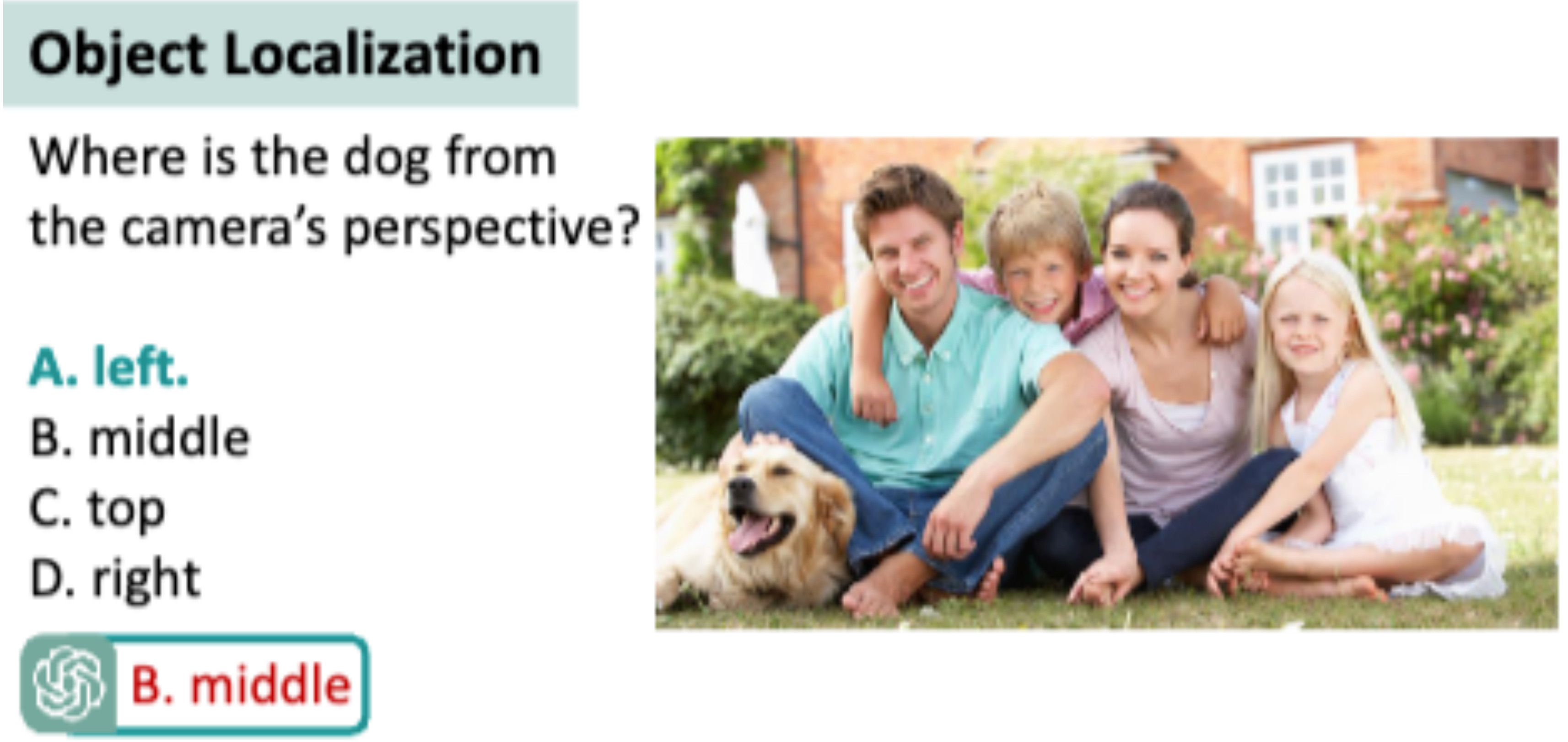
👁️ Perception

💡 Interpretation

🧐 Reasoning

How Easy is It to Fool Your Multimodal LLMs? An Empirical Analysis on Deceptive Prompts, Yusu Qian, Haotian Zhang, Yinfei Yang, Zhe Gan

Spatial Reasoning

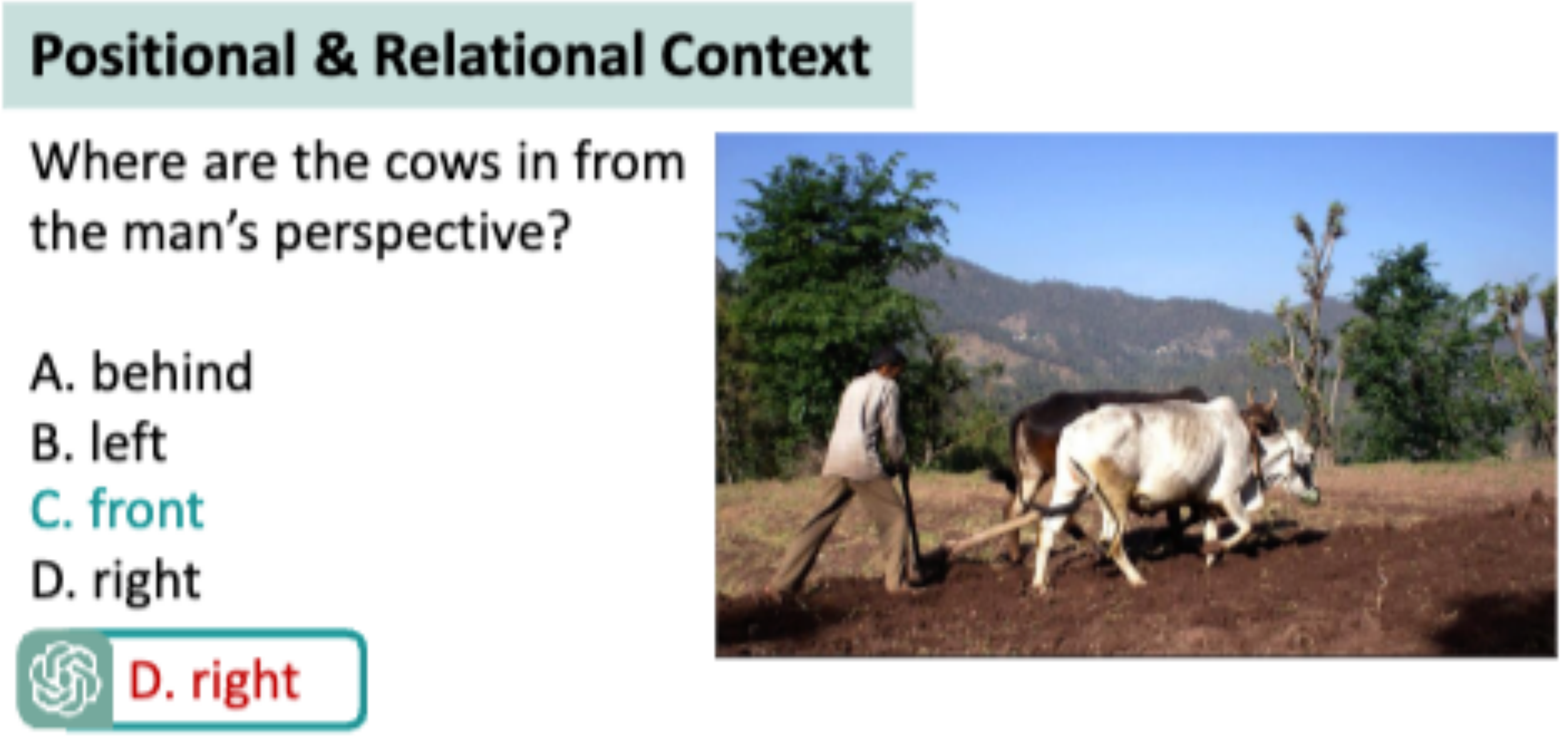


Object Localization

Where is the dog from the camera's perspective?

A. left.
B. middle
C. top
D. right

B. middle

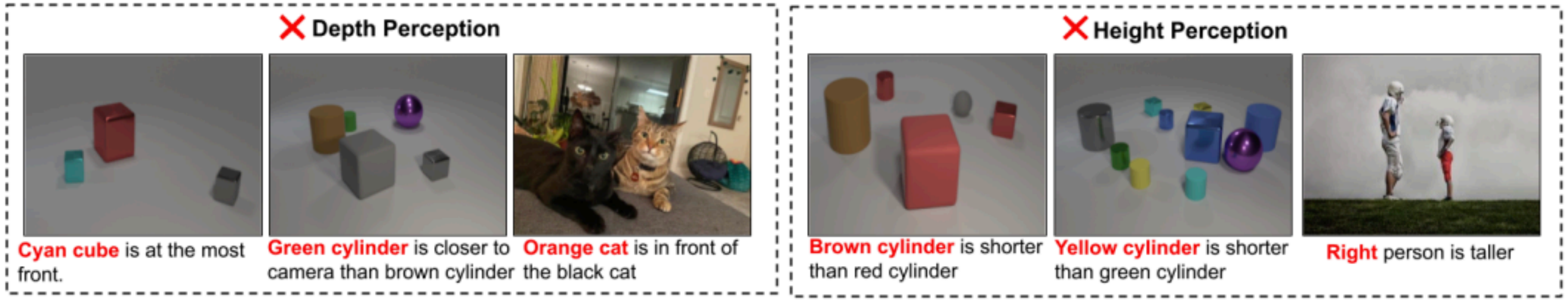


Positional & Relational Context

Where are the cows in from the man's perspective?

A. behind
B. left
C. front
D. right

D. right



✗ Depth Perception

Cyan cube is at the most front.

Green cylinder is closer to camera than brown cylinder

Orange cat is in front of the black cat

✗ Height Perception

Brown cylinder is shorter than red cylinder

Yellow cylinder is shorter than green cylinder

Right person is taller

Figure 1: **Depth and height perception of existing VLM.** Here, we show GPT-4V failure to understand depth and height on existing synthetic (CLEVR [1]) dataset and real-world images taken from the internet.

👁️ Perception

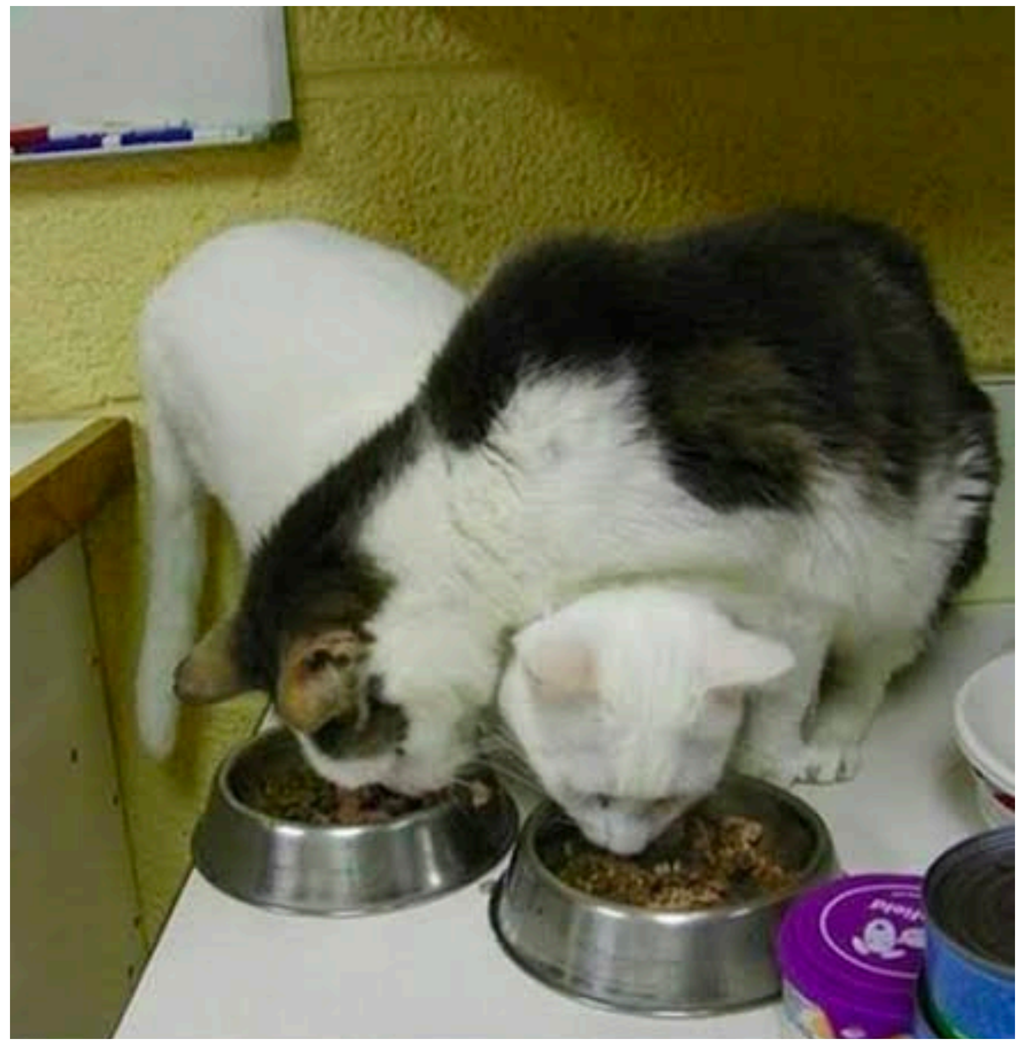
💡 Interpretation

🧐 Reasoning

An Empirical Analysis on Spatial Reasoning Capabilities of Large Multimodal Models, arXiv, 2024

Text != Image Description

- Most benchmarks test models' ability to describe the image
- But text isn't typically used to describe images, but rather **complement** them



The grass is always greener on the other side.

👁️ Perception

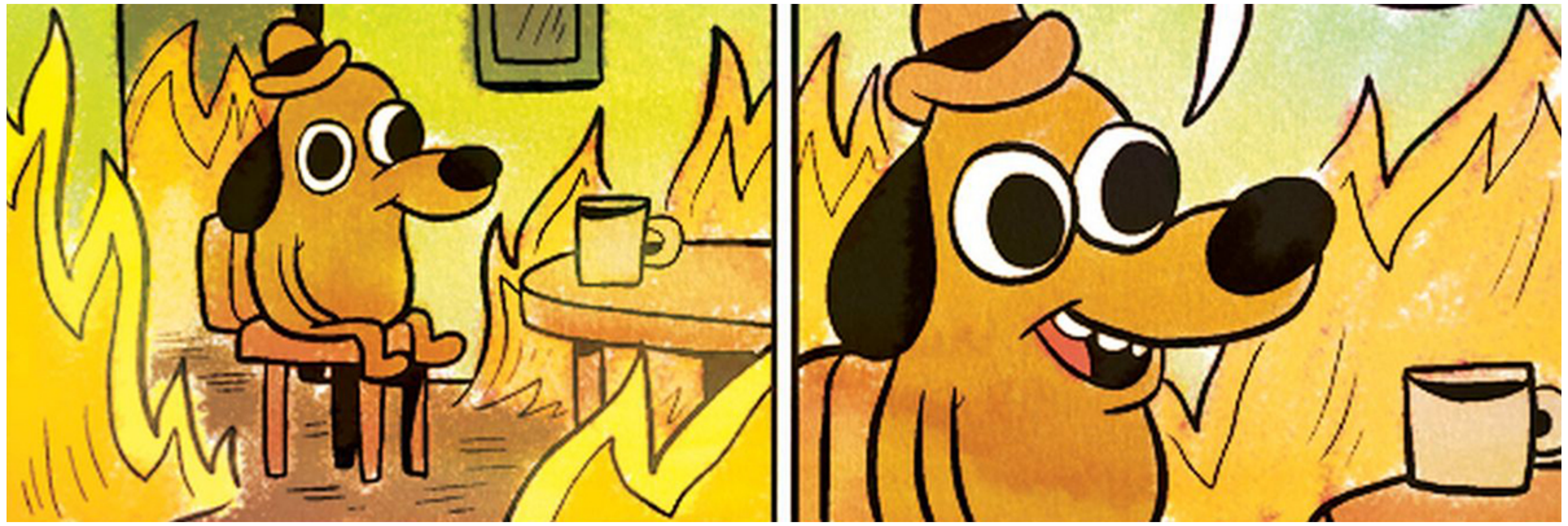
💡 Interpretation

🧐 Reasoning

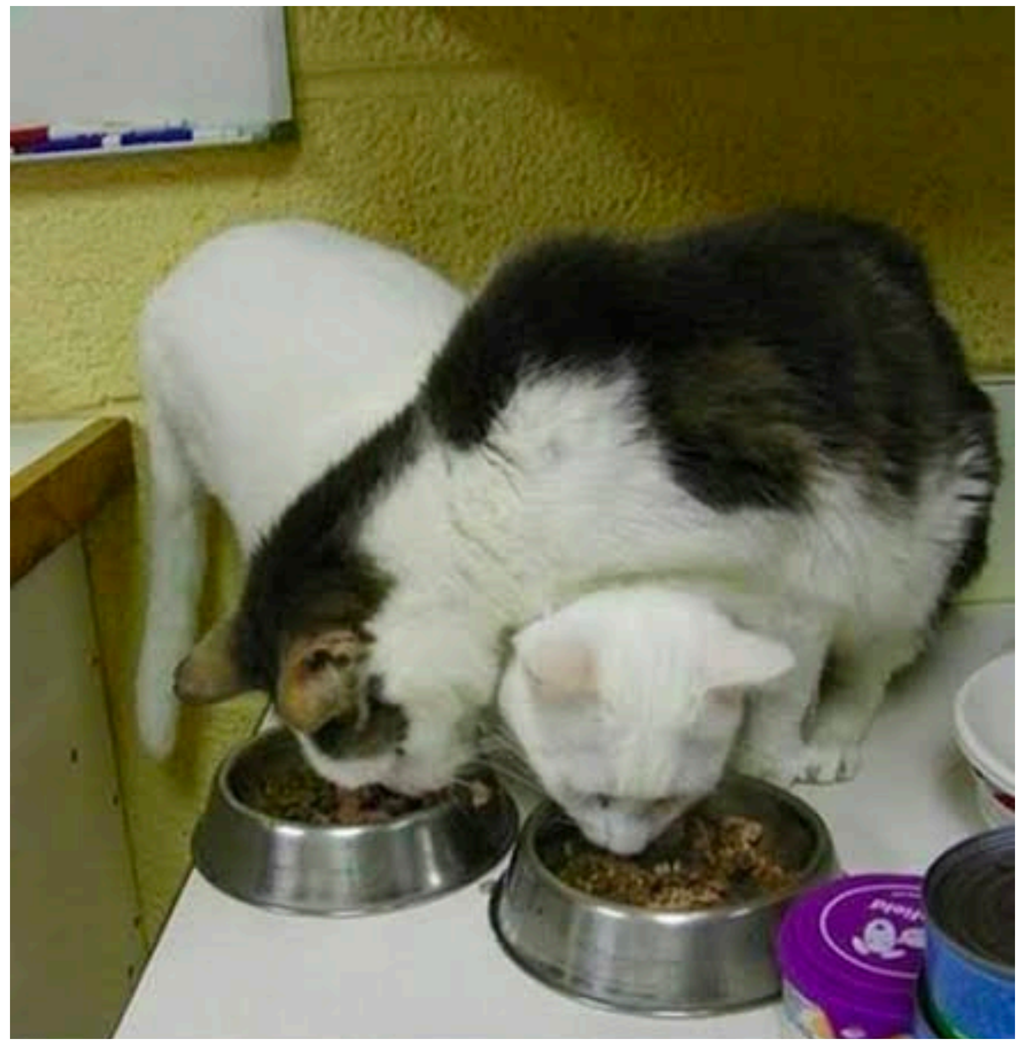
Clue: Cross-modal Coherence Modeling for Caption Generation. Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. ACL 2020.

Text != Image Description

- Most benchmarks test models’ ability to describe the image
- But text isn’t typically used to describe images, but rather **complement** them



Caption: A picture of a man with a hot dog in his mouth. ✗



The grass is always greener on the other side.

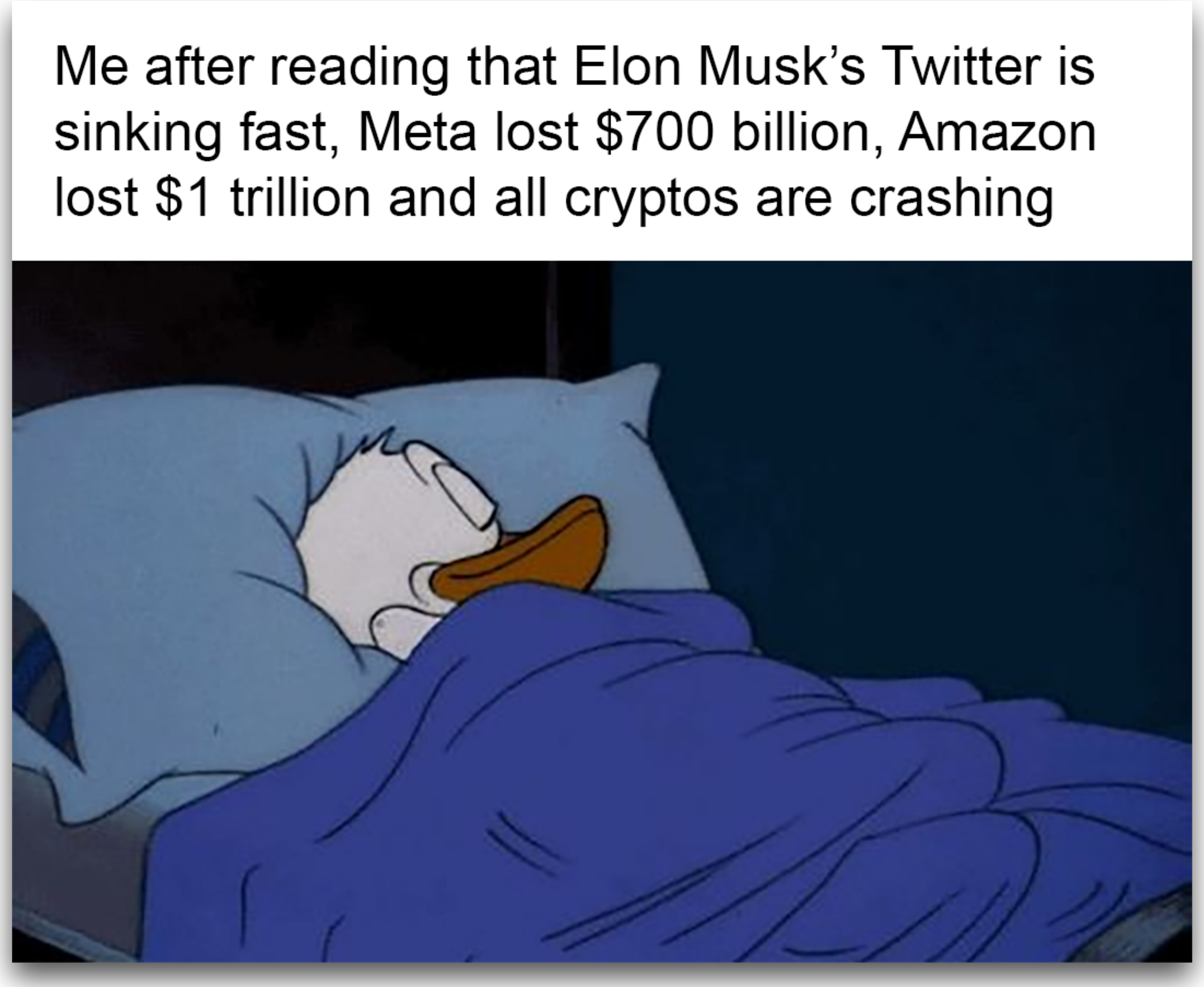
👁️ Perception

💡 Interpretation

🧐 Reasoning

Clue: Cross-modal Coherence Modeling for Caption Generation. Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. ACL 2020.

Meme Interpretation



Me after reading that Elon Musk's Twitter is sinking fast, Meta lost $700 billion, Amazon lost $1 trillion and all cryptos are crashing

Literal Meaning (Image Caption):
Donald Duck is sleeping.

Metaphors (Image + Text):
Donald Duck = meme poster
Sleeping = being peaceful and not worried

Metaphorical Meaning (Image + Text):
The meme poster is unbothered by discovering that Elon Musk's Twitter is sinking fast, Meta lost $700 billion, Amazon lost $1 trillion and all cryptos are crashing

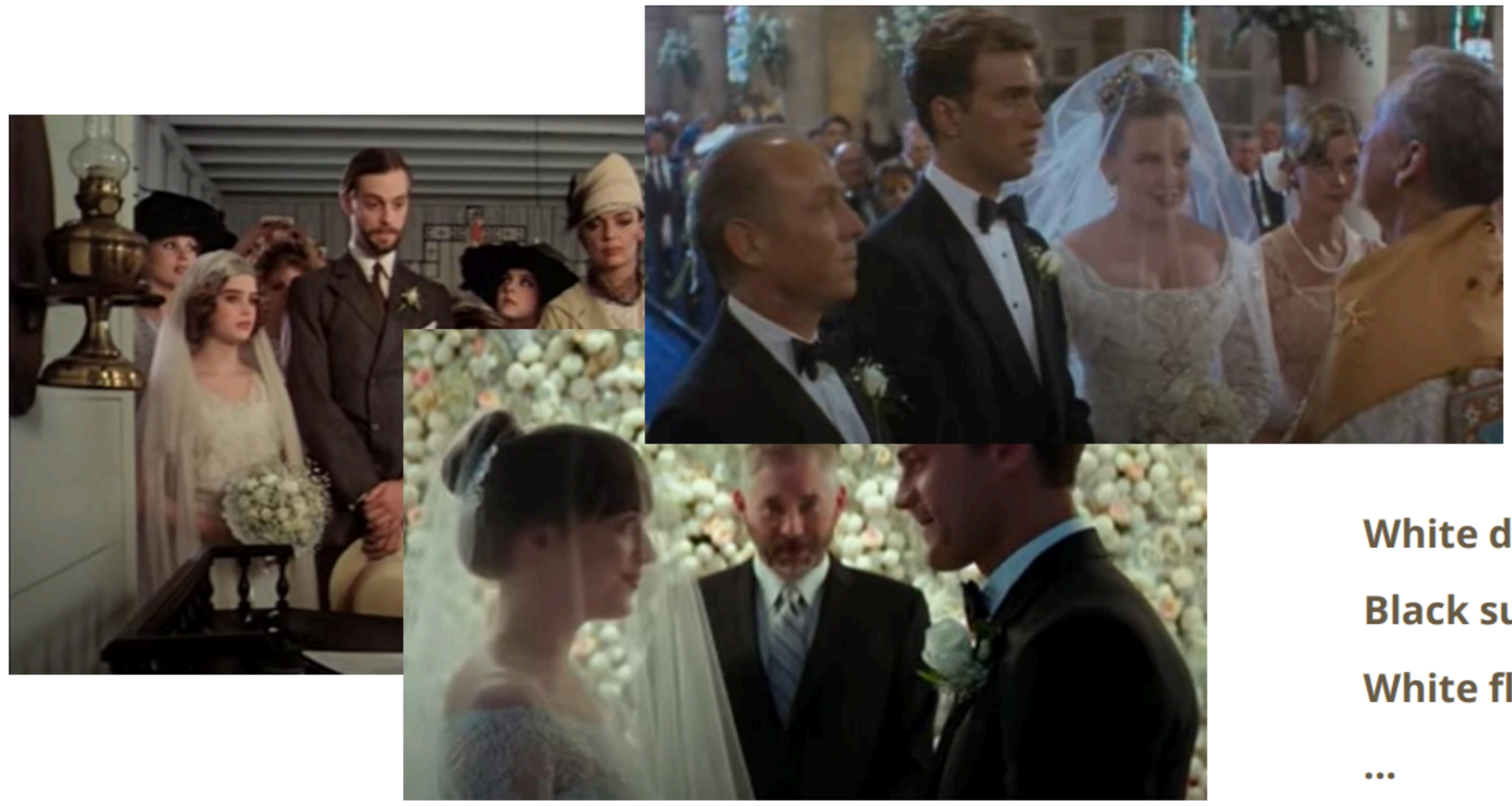
👁 Perception

💡 Interpretation

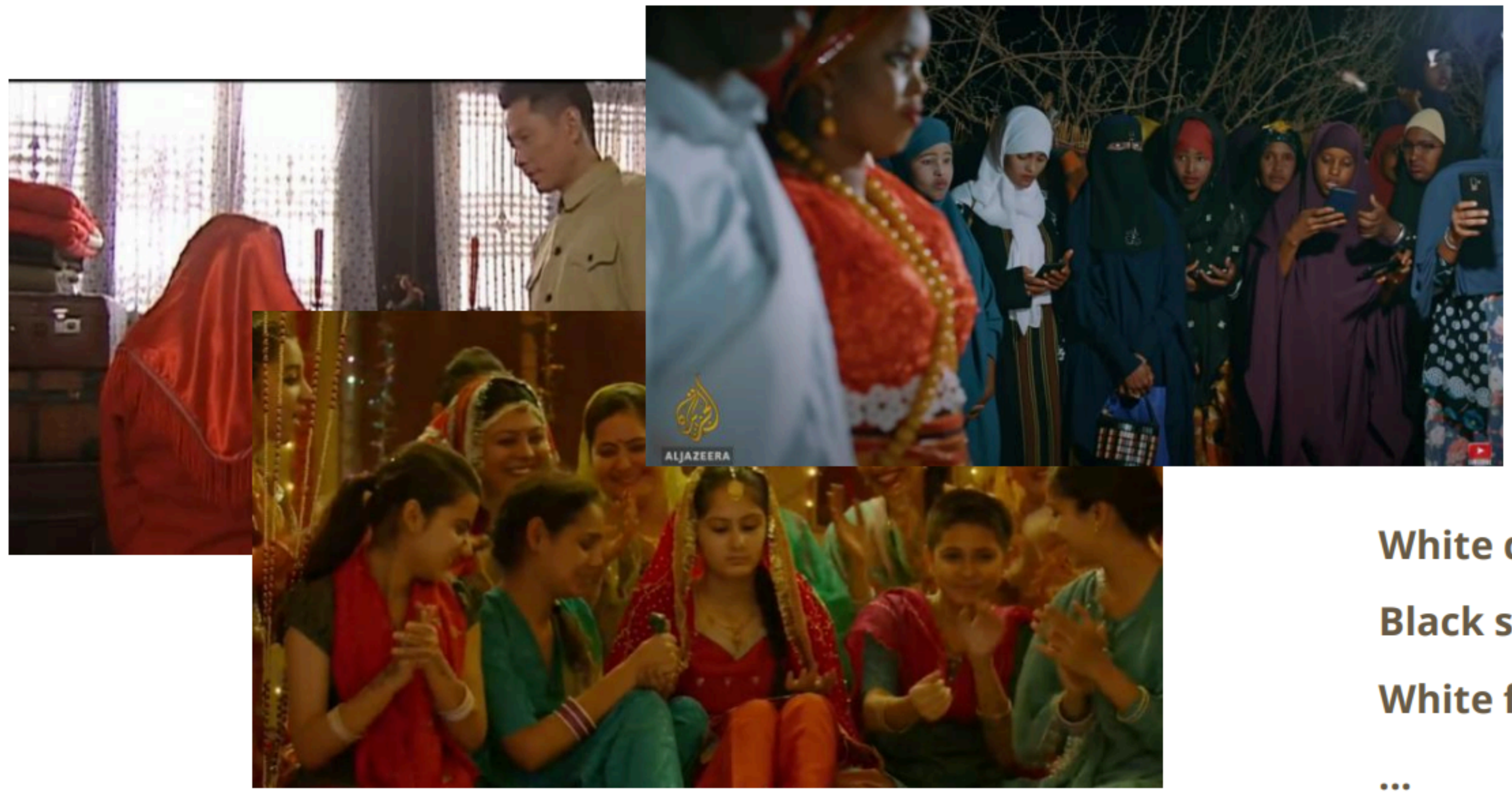
🧐 Reasoning

Meme Interpretation. EunJeong Hwang and Vered Shwartz. Work in Progress.

(Visual) Commonsense is Culture-Dependent



White dress ✓
Black suits ✓
White flowers ✓
...



White dress ✗
Black suits ✗
White flowers ✗
...

Western weddings, more commonly present in datasets, can be very different from weddings in other parts of the world.

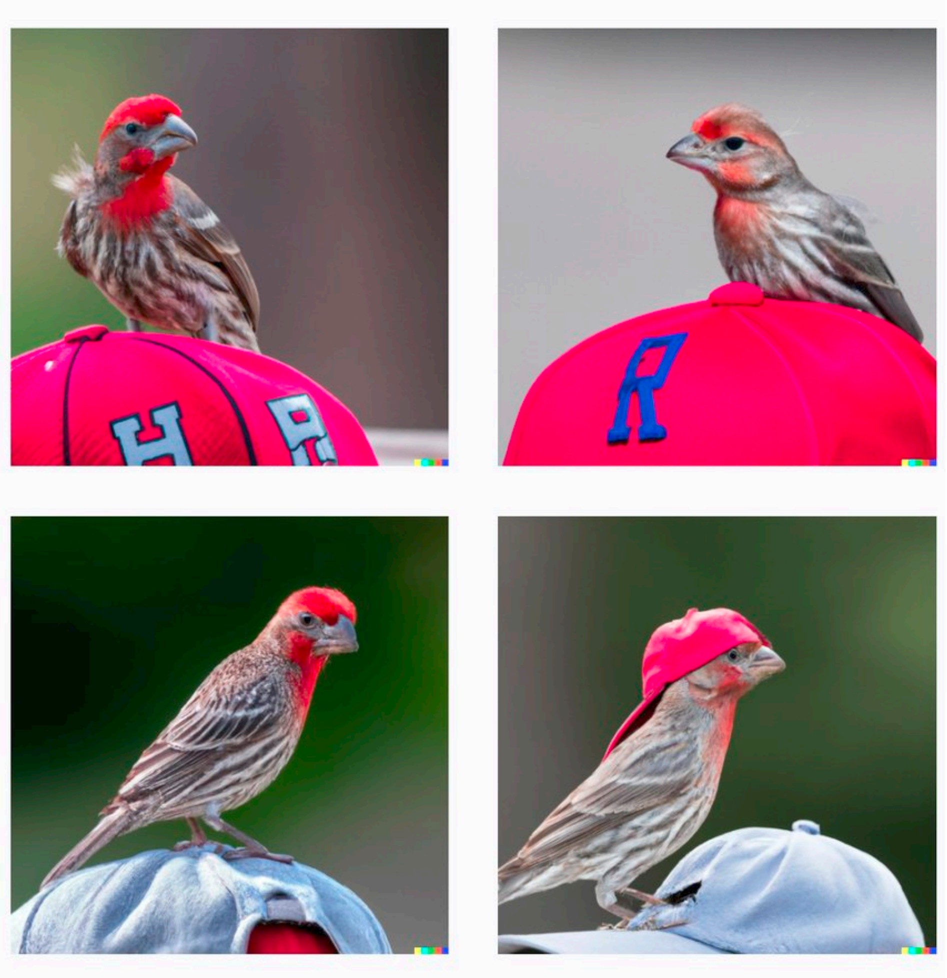
👁️ Perception

💡 Interpretation

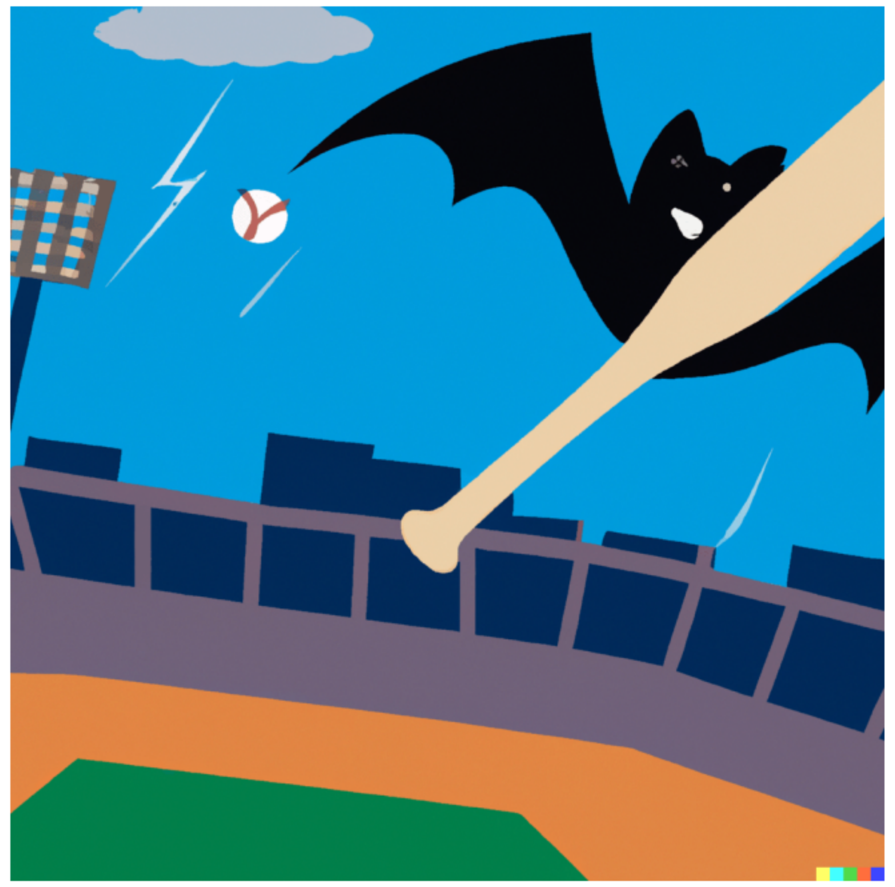
🧐 Reasoning

Broaden the Vision: Geo-Diverse Visual Commonsense Reasoning. Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. EMNLP 2021.

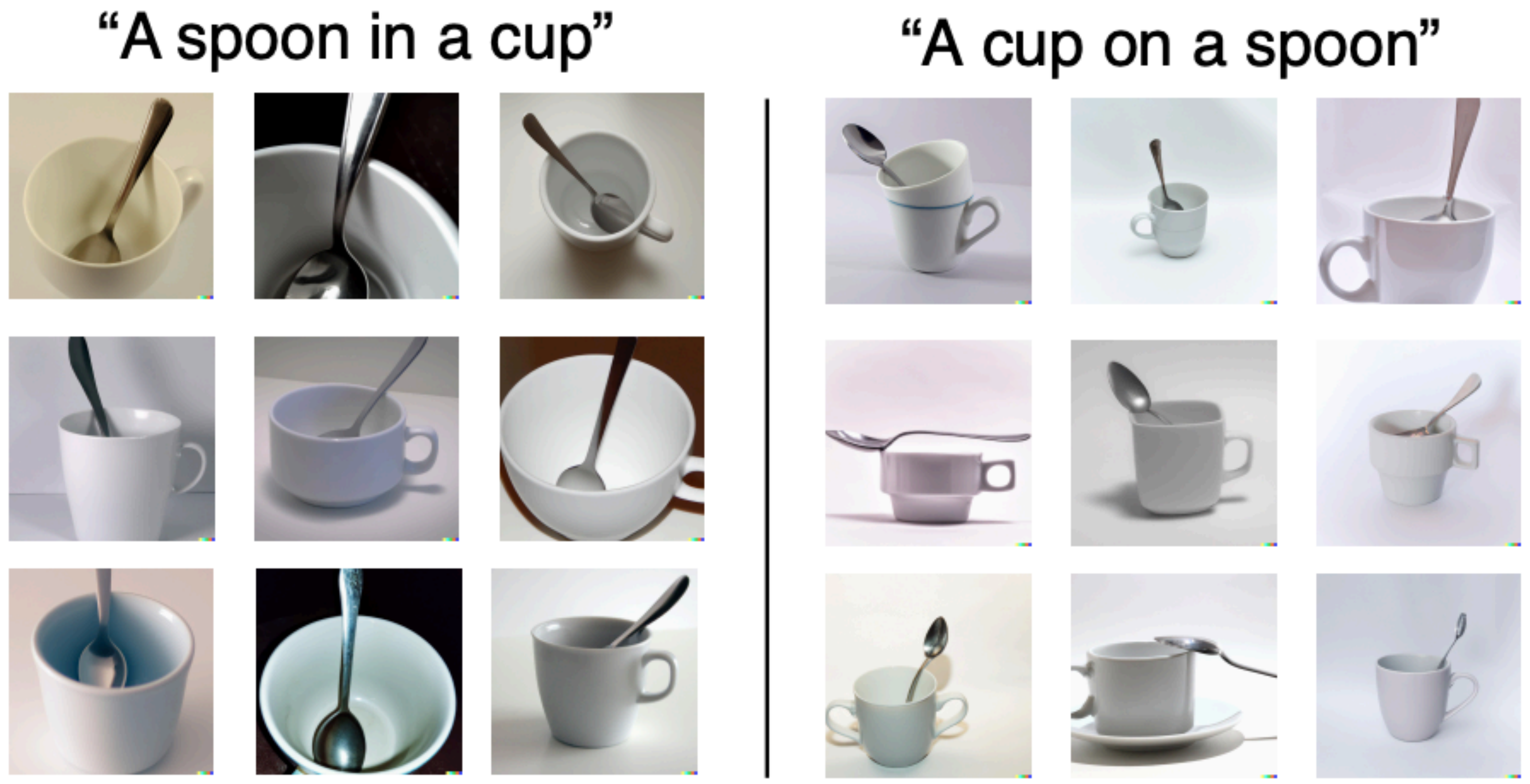
Text-Image Generation



"A house finch wearing a baseball cap"



"A bat is flying over a baseball stadium"



👎 Understanding complex prompts

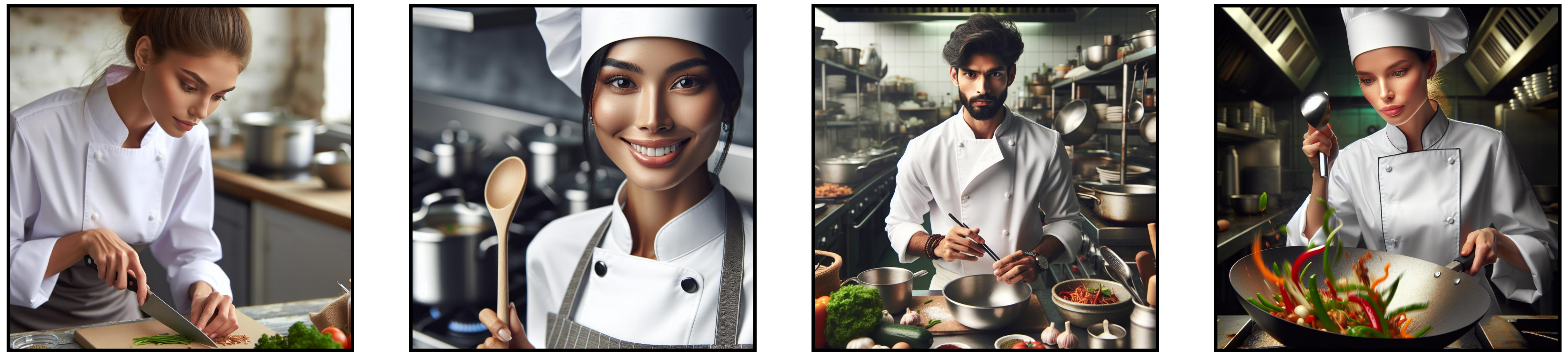
👁 Perception

💡 Interpretation

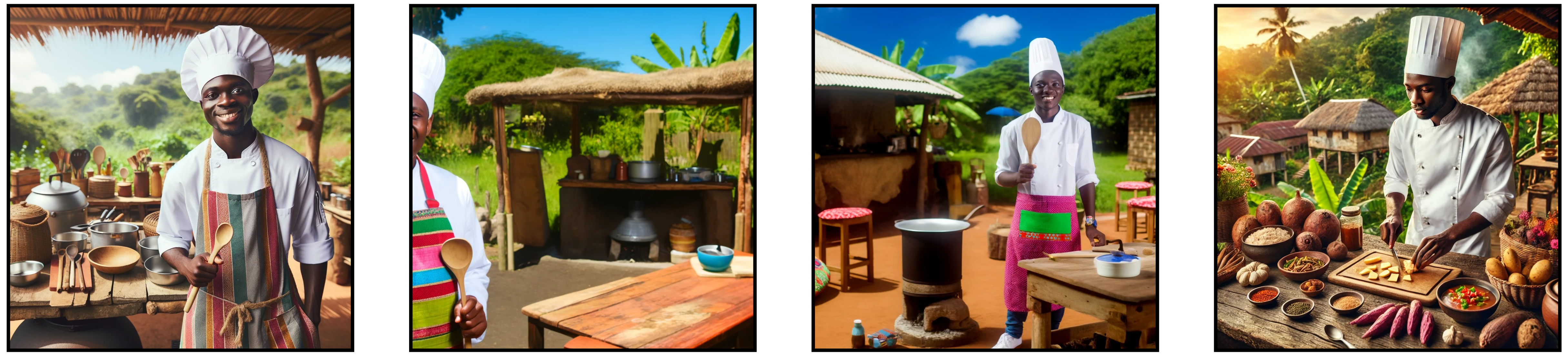
🧐 Reasoning

DALLE-2 is Seeing Double: Flaws in Word-to-Concept Mapping in Text2Image Models. Royi Rassin, Shauli Ravfogel, Yoav Goldberg. BlackboxNLP 2023.
A very preliminary analysis of DALL-E 2. Gary Marcus, Ernest Davis, Scott Aaronson. arXiv 2022.
Testing Relational Understanding in Text-Guided Image Generation. Colin Conwell and Tomer Ullman. arXiv 2022.

Text to Image Biases



"A photo of a chef"



"A photo of a chef in Africa"

T2I models can generate biased images because the VL models (e.g. CLIP) can have these biased representations

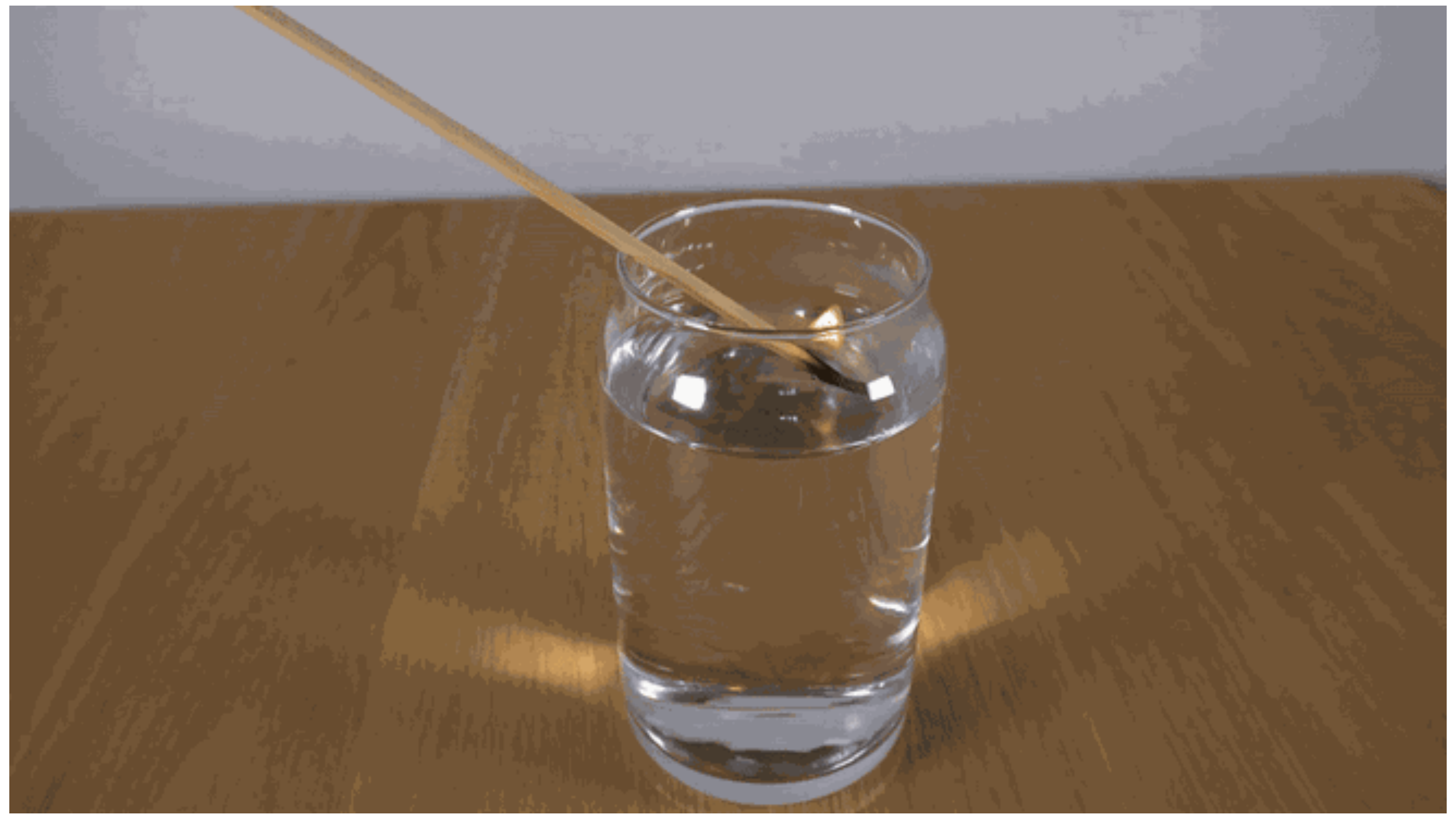
👁️ Perception

💡 Interpretation

🧐 Reasoning

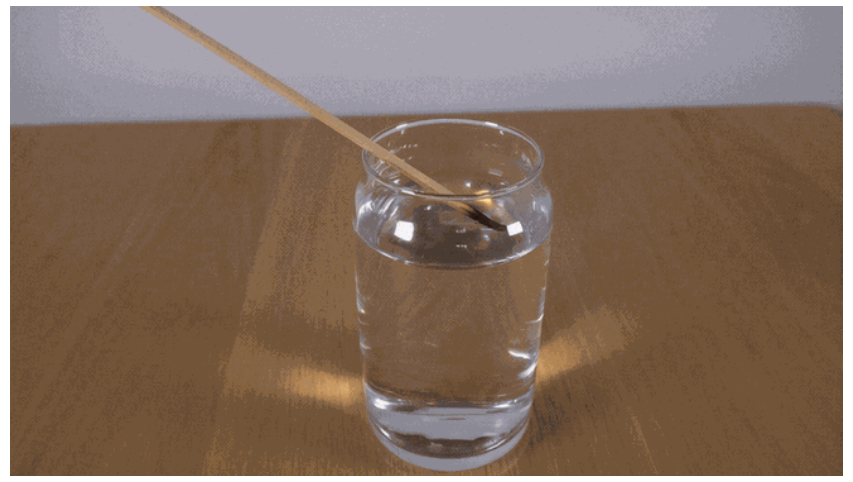
TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models. Chinchure & Shukla et al. (ECCV 2024)

Generative Models: Videos

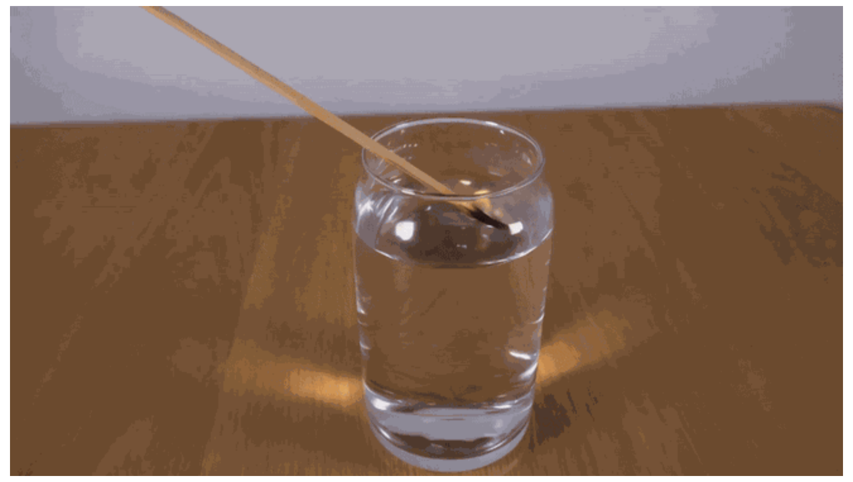


VideoPoet (i2v)

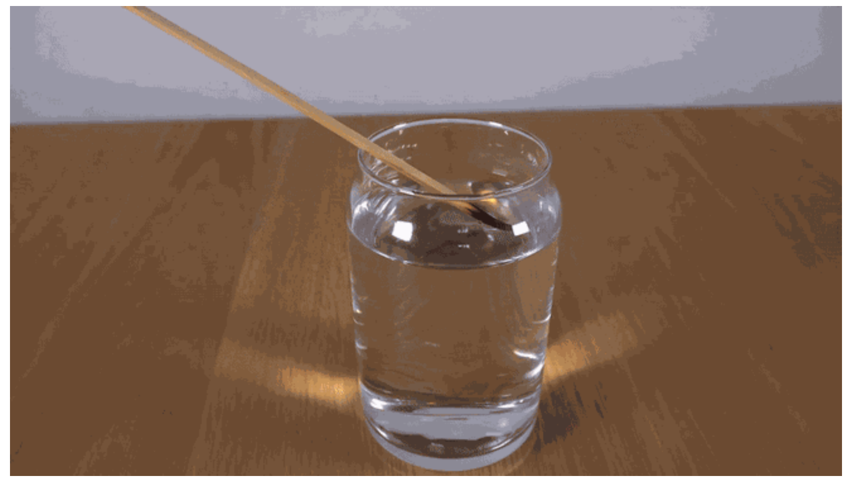
Sora (i2v)



Pika 1.0 (i2v)



Runway Gen 3 (i2v)



Every video generation model predicts that matchsticks can burn inside water

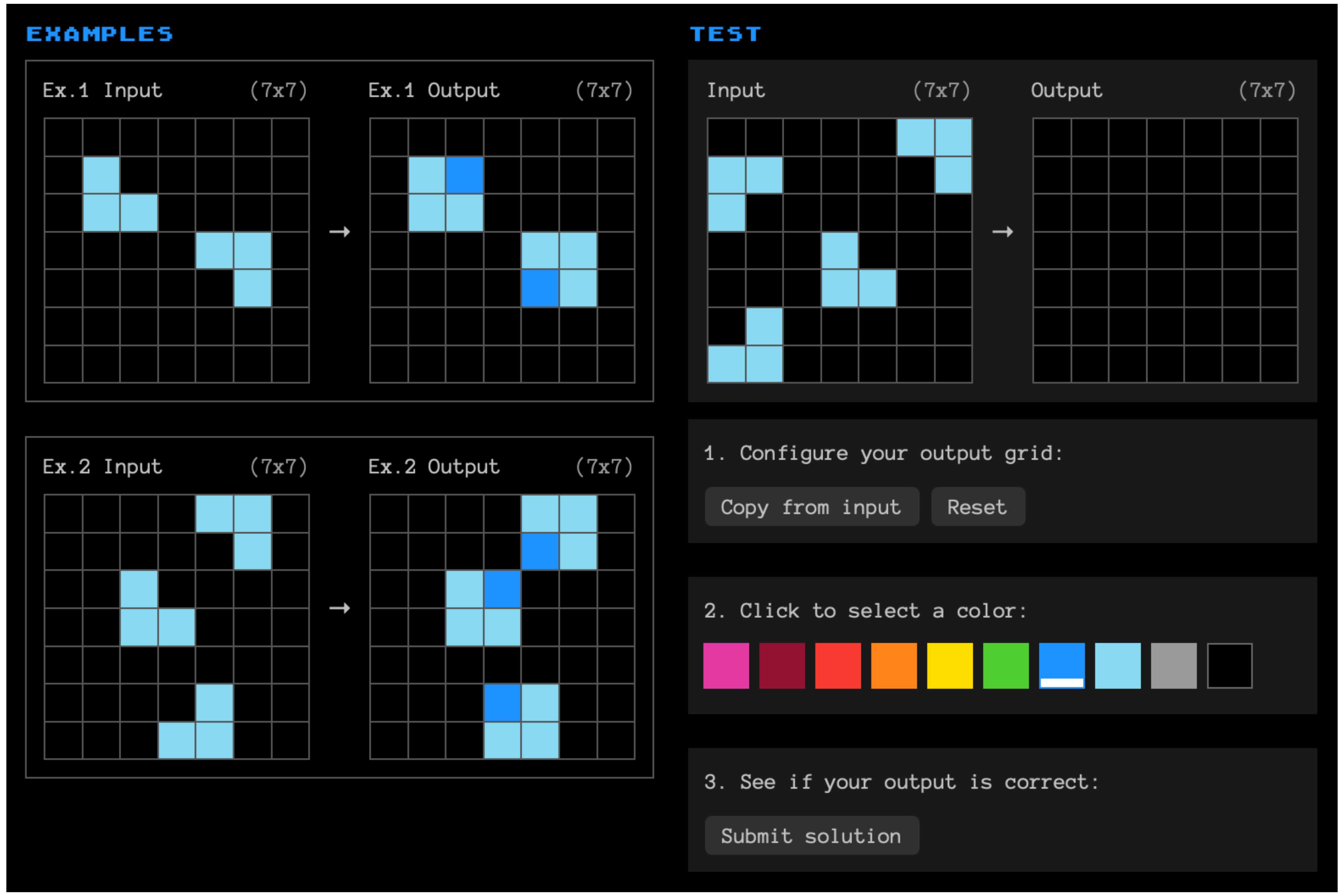
👁️ Perception

💡Interpretation

🧐 Reasoning

ARC-AGI Benchmark

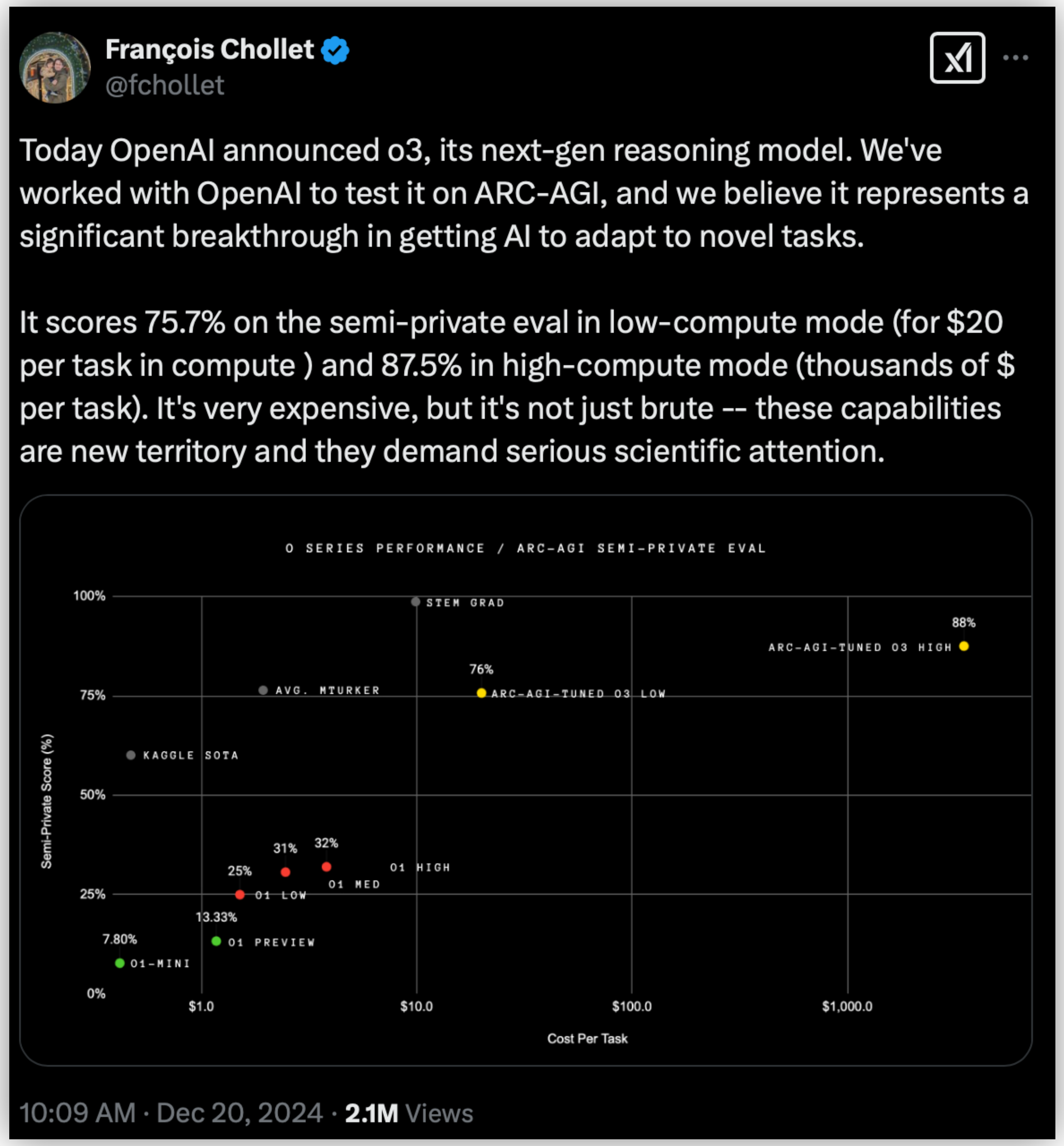
Understanding complex pattens present on the grid (~600 problems)



1. Configure your output grid:

2. Click to select a color:

3. See if your output is correct:



François Chollet @fchollet

Today OpenAI announced o3, its next-gen reasoning model. We've worked with OpenAI to test it on ARC-AGI, and we believe it represents a significant breakthrough in getting AI to adapt to novel tasks.

It scores 75.7% on the semi-private eval in low-compute mode (for $20 per task in compute) and 87.5% in high-compute mode (thousands of $ per task). It's very expensive, but it's not just brute -- these capabilities are new territory and they demand serious scientific attention.

10:09 AM · Dec 20, 2024 · **2.1M** Views

👁️ Perception

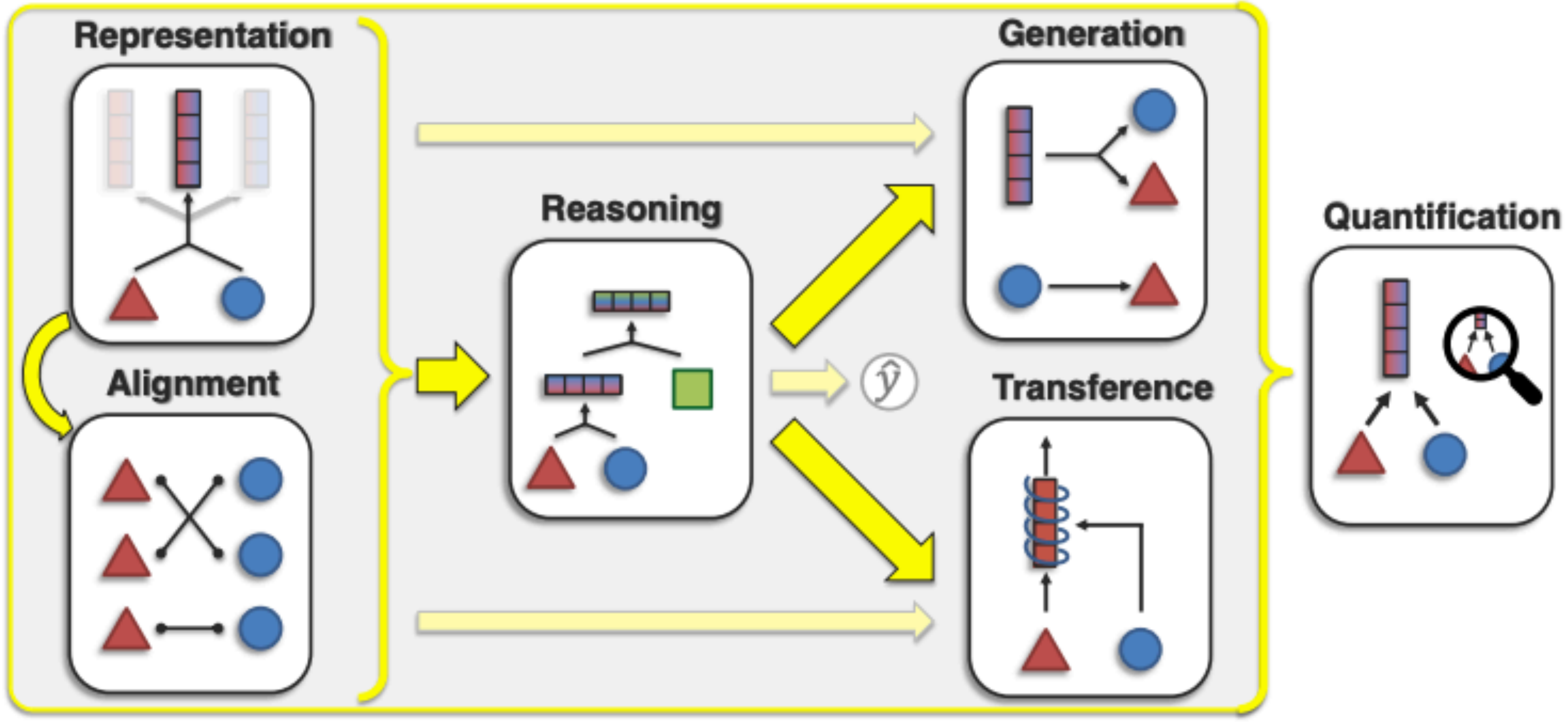
💡 Interpretation

🧐 Reasoning

https://arcprize.org/

More Open Problems...

- Agentic Frameworks
  - VLMs for task planning
- Long form video understanding
- Extending beyond VL: Aligning multiple modalities



Thank You