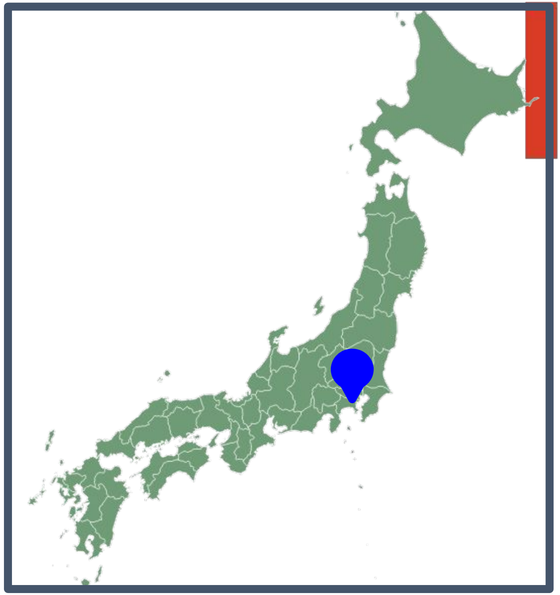
Developing multimodal language models in real-world scenarios

Syrielle Montariol

Geographic mobility

université PARIS-SACLAY

Inria



EPFL

Jožef Stefan Institute

NII
National Institute of Informatics

Thematic mobility

Inria

Modeling lexical semantic variation across time

Semantic variation across languages

Detecting harmful content across languages

Detecting harmful content in different modalities

EPFL

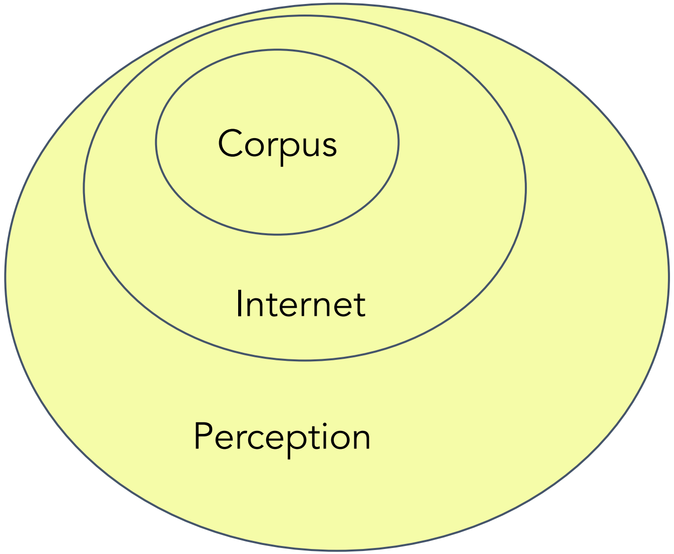
Models that perceive and reason across different modalities

université PARIS-SACLAY

How does AI learn language?

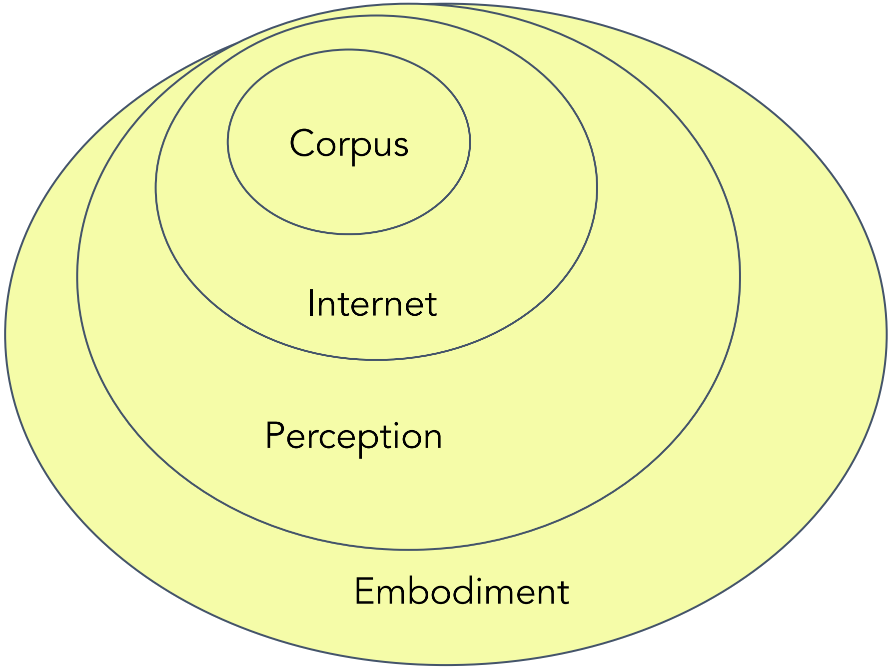
You can't learn language from books.

Can you perform a task using perceptual input?



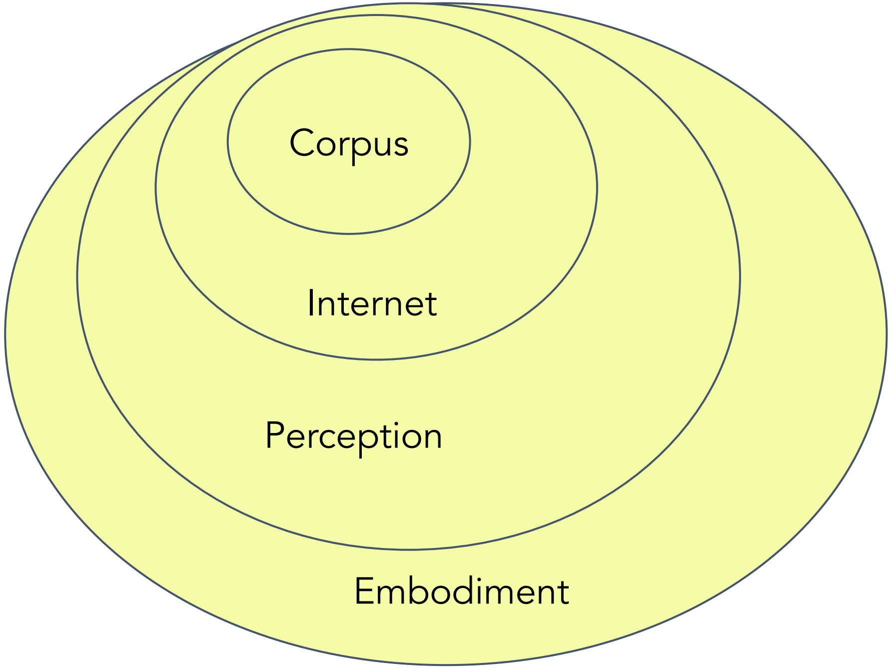
You can't learn language from the television.

Can you interact with your environment?



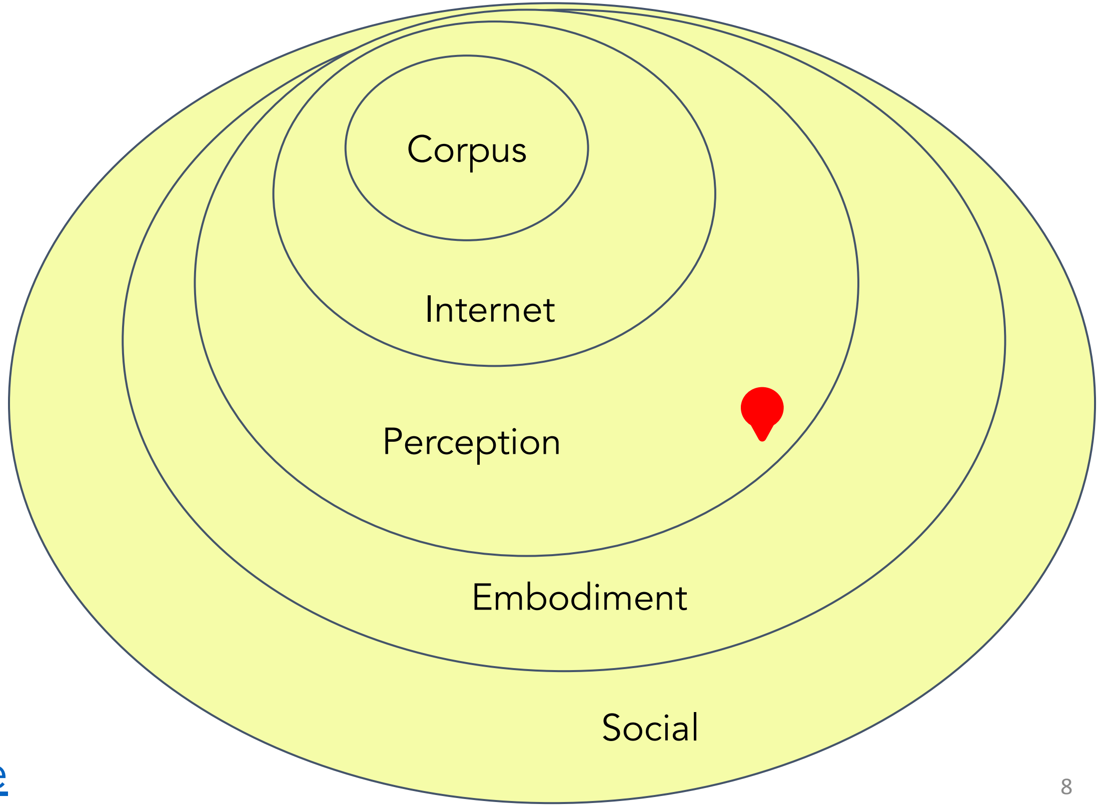
You can't learn language by yourself.

Can you cooperate with humans to achieve your goal?



You can't learn language...

- From the radio (Internet)
- From a television
- By yourself



[Experience Grounds Language]

Why Multimodality?

- **Human learning & experience** is multimodal
- Multimodal data is **richer than text**
- More data is usually better, and the amount of available textual data is limited

In the literature, they are called…

MLLMs

LMMs

LVLMs

VLMs

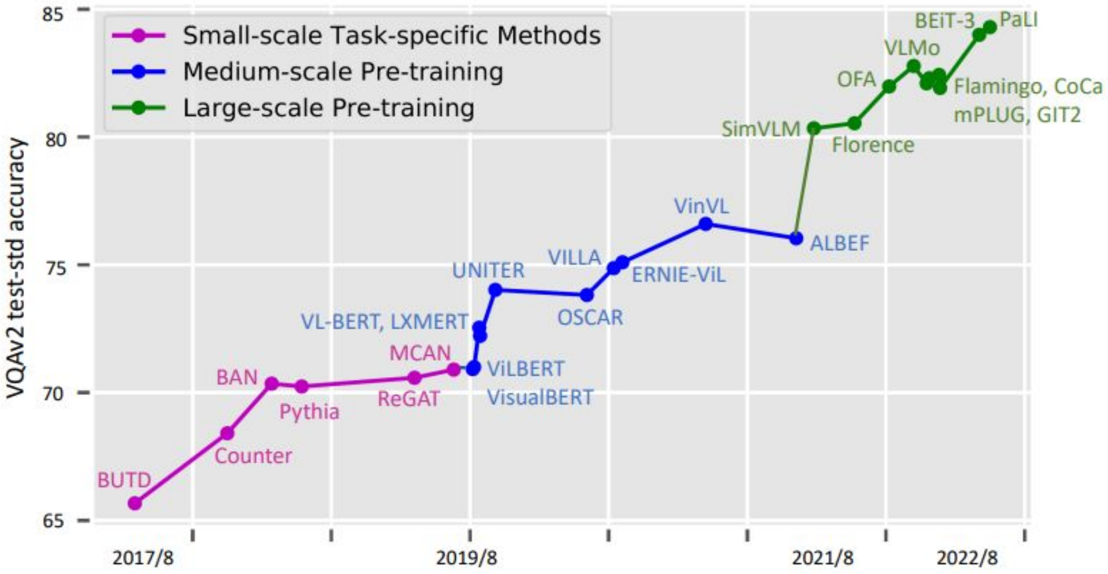
VLLMs

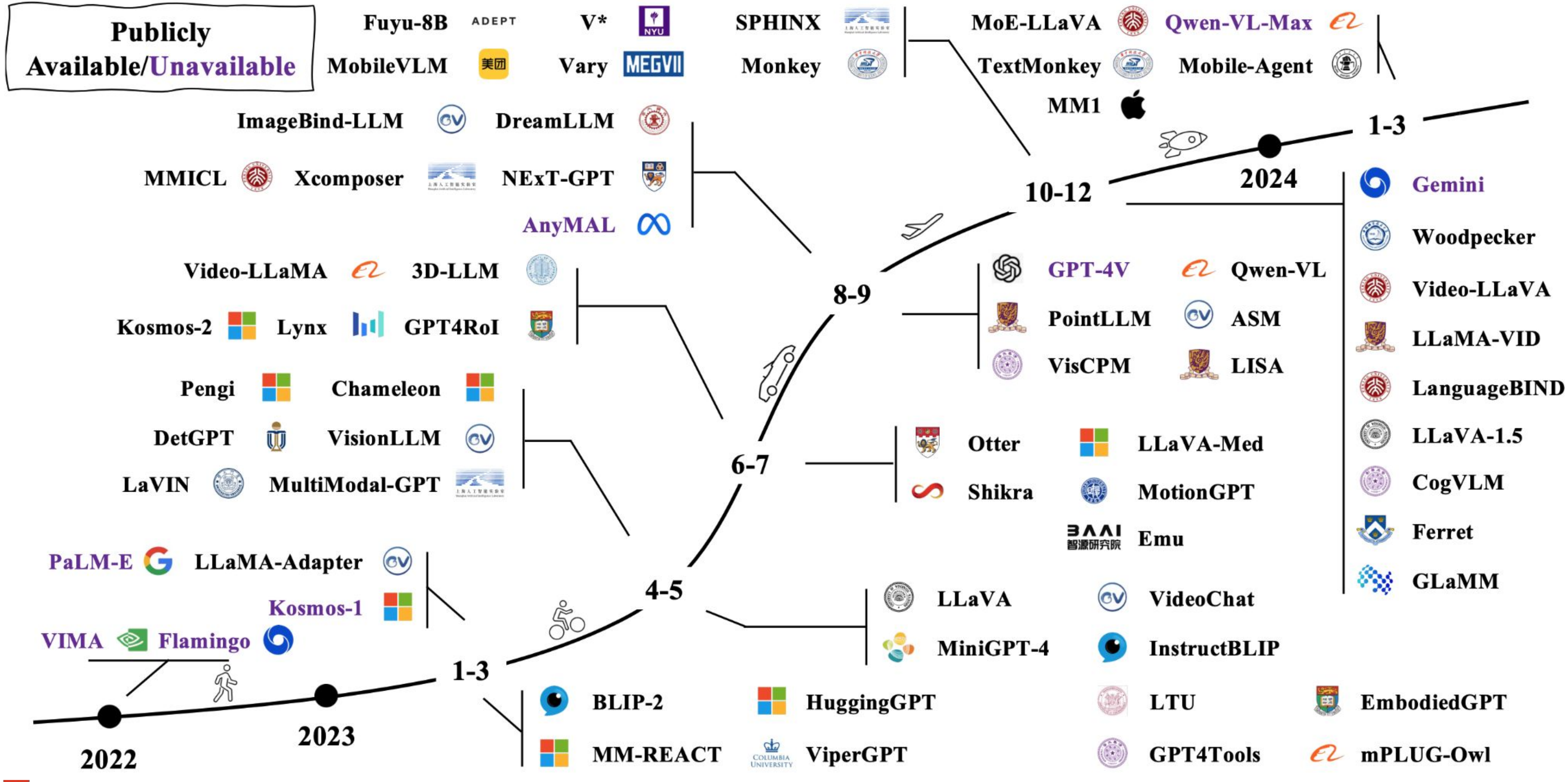
Evolution of Vision and Language Models (VLMs)

1. **Small-scale, task-specific methods**: ResNet & FasterRCNN, Glove & Word2Vec

2. **Medium-scale pre-training**, since 2019 (up to 340M parameters with BERT-Large): Inspired by BERT, transformers-based multi-modal fusion.

3. **Large-scale pre-training**, since 2021: starting with CLIP, then adapting pre-trained LLMs.





https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models

What are the limits of VLMs' perception and reasoning, and how to improve it?

How to train a VLM?

Open medical foundation models adapted for clinical practice

Meditron

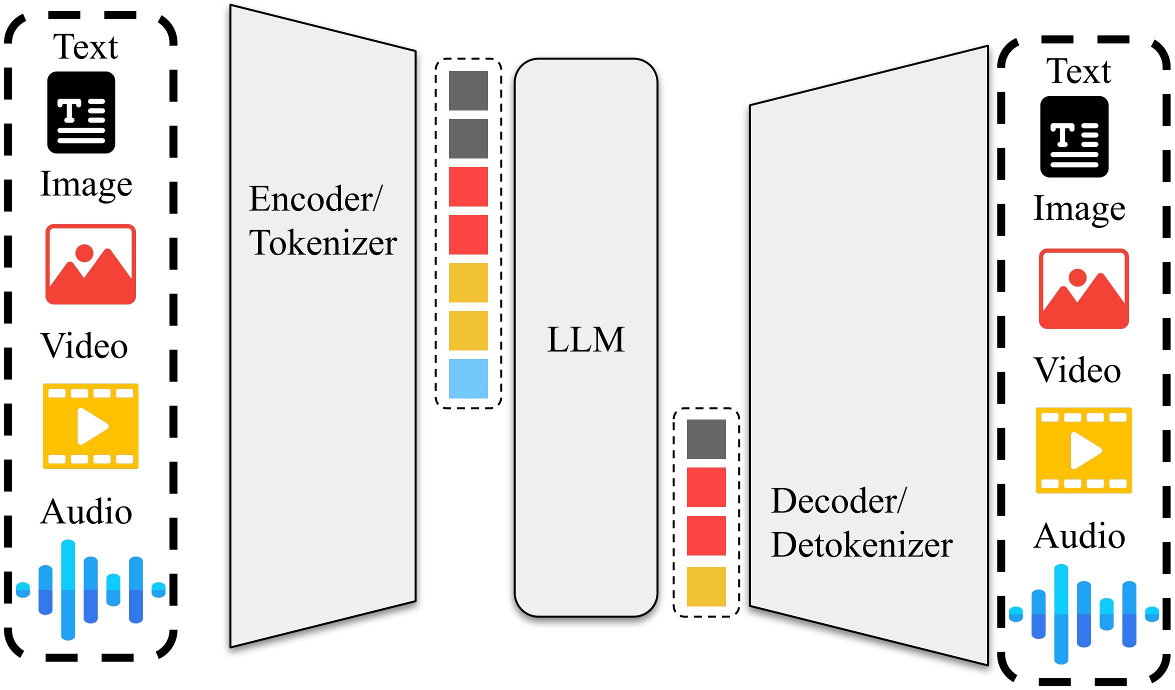
What are key VLM failures?

CAVE: Detecting and Explaining Commonsense Anomalies in Visual Environments

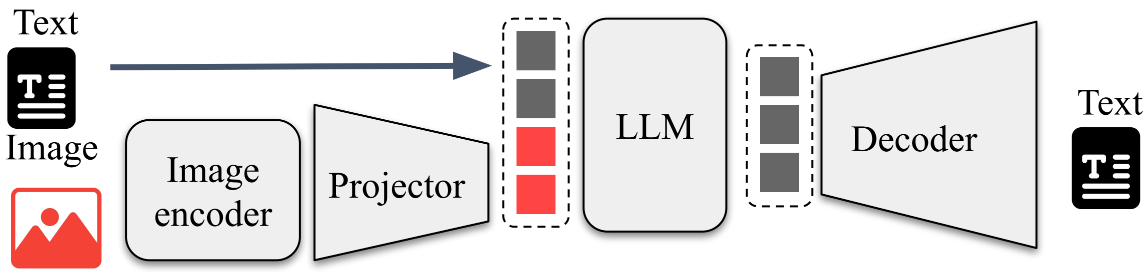
How to improve VLM reasoning?

Visual Chain-of-Thought Reasoning in Real-World Driving Scenarios

What's a Large Multimodal Model?



What's a VLM?



Let's make one!

Meditron

EPFL

idiap RESEARCH INSTITUTE

Yale University School of Medicine

Open Assistant

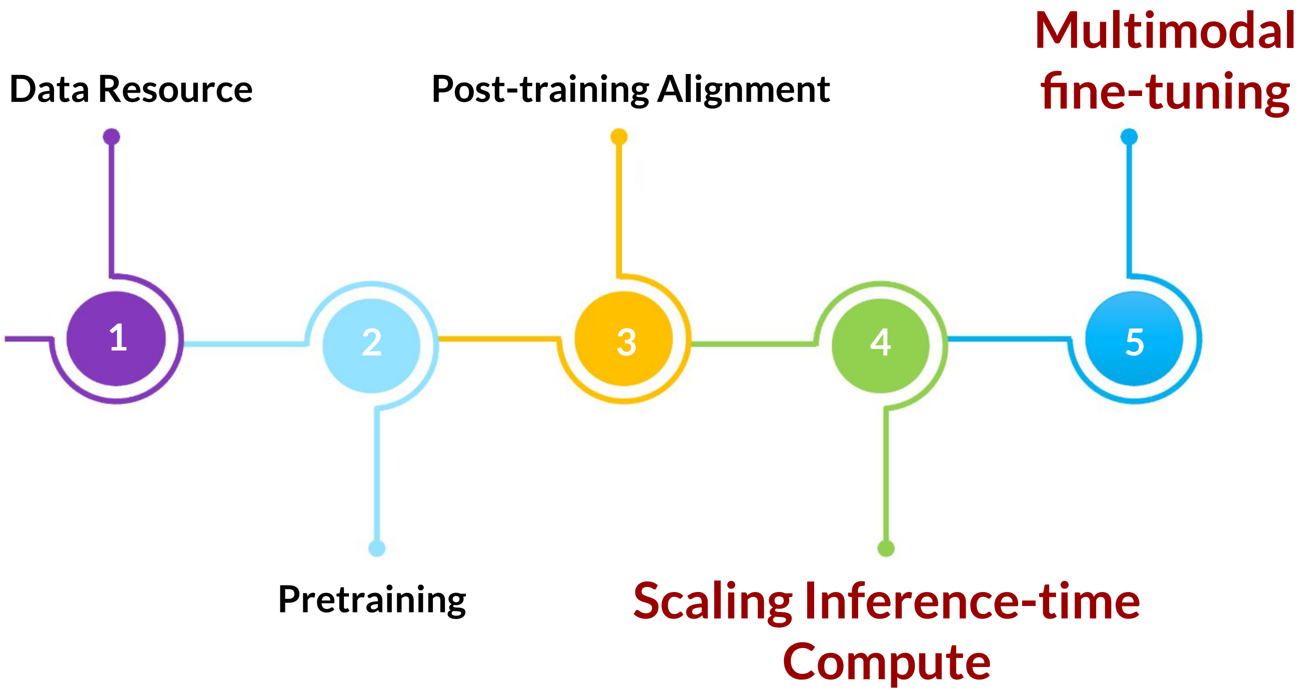
COMITE INTERNATIONAL GENEVE ICRC

CHUV Centre hospitalier universitaire vaudois

INSELSPITAL UNIVERSITÄTSSPITAL BERN HÔPITAL UNIVERSITAIRE DE BERNE

Open Medical Foundation Models Adapted for Clinical Practice

Development Outline



Slides credit: Zeming Chen

Chain-of-Thought Reasoning

Model Input

Question: A 65-year old man presents with gradually worsening rigidity of his arms and legs and slowness in performing tasks. He says he has also noticed hand tremors, which increase at rest and decrease with focused movements. On examination, the patient does not swing his arms while walking and has a shortened, shuffling gait. An antiviral drug is prescribed which alleviates the patient's symptoms. Which of the following drugs was most likely prescribed to this patient?

Options: (A) Amantadine (B) Ribavirin (C) Acyclovir (D) Levodopa (E) Zidovudine
Let's think step by step:

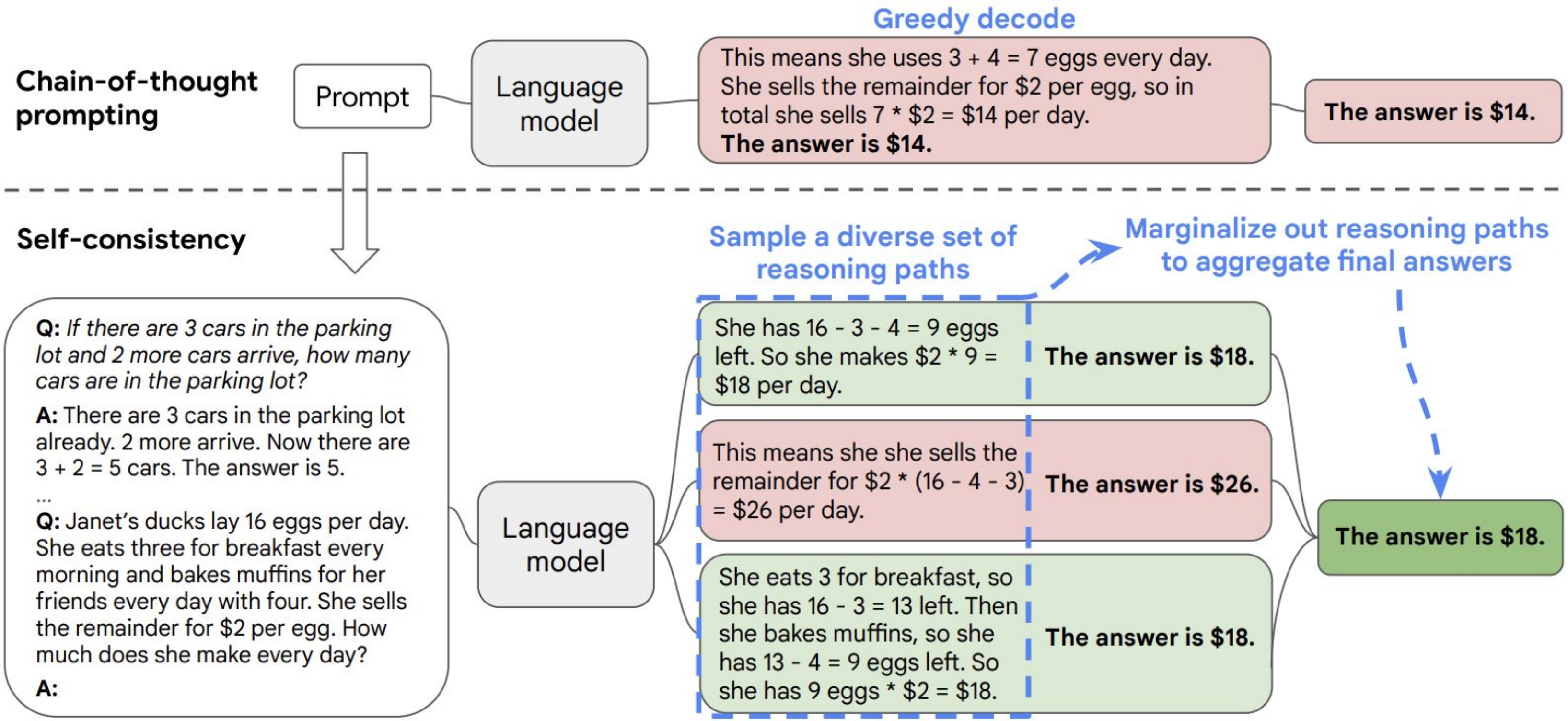
Step-by-step Thought Process

Model Output

Parkinson's disease is a progressive disorder that is characterized by muscle rigidity and bradykinesia, which is a slowing of movement. Levodopa is used to alleviate symptoms of Parkinson's disease. However, it is not an antiviral drug. Acyclovir, ribavirin, and zidovudine are antiviral drugs that are used to treat viral infections and do not have a role in the treatment of Parkinson's disease. Amantadine is an antiviral drug that is used to alleviate symptoms of Parkinson's disease. The most likely drug prescribed for this patient is **(A) Amantadine**.

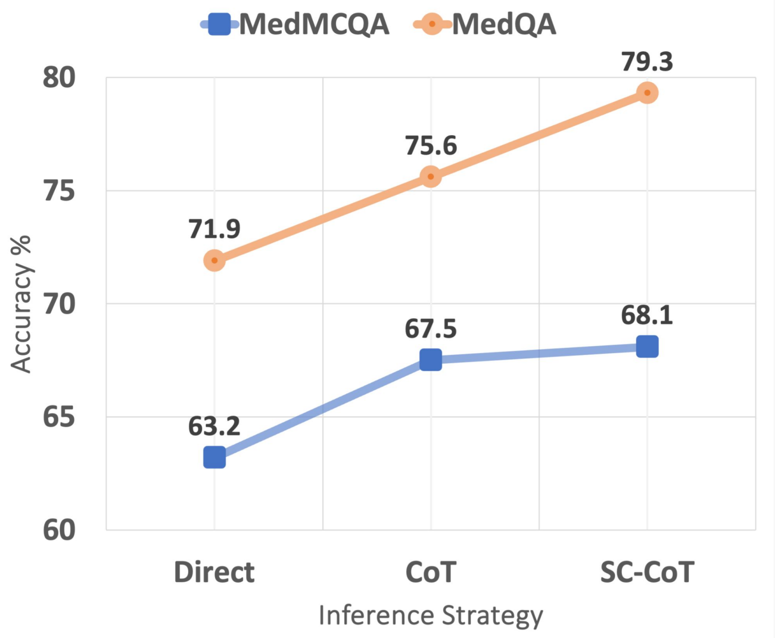
Slides credit: Zeming Chen

Self-Consistency



Wang, Xuezhi, et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." ICLR 2023.

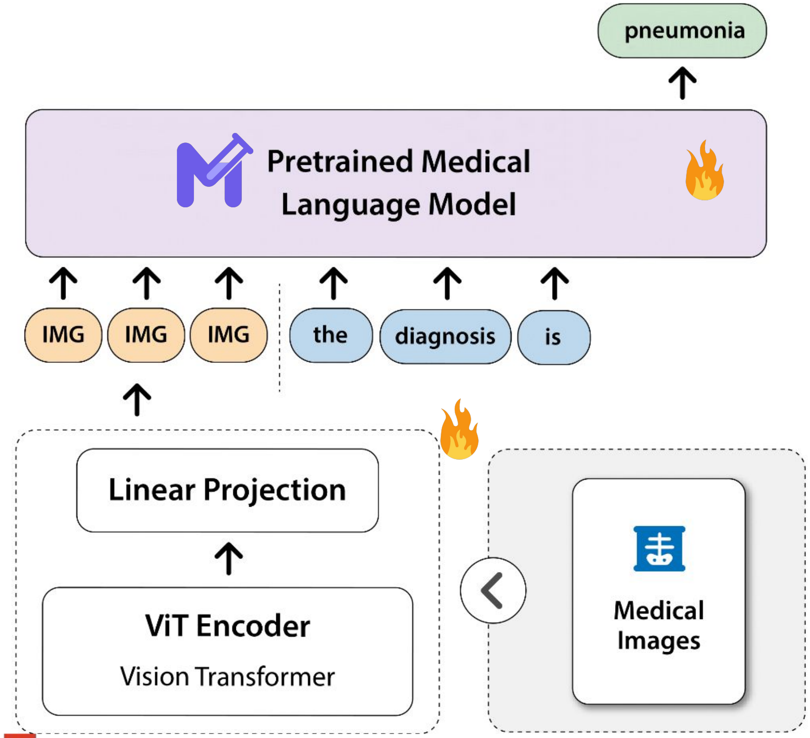
Performance Gain

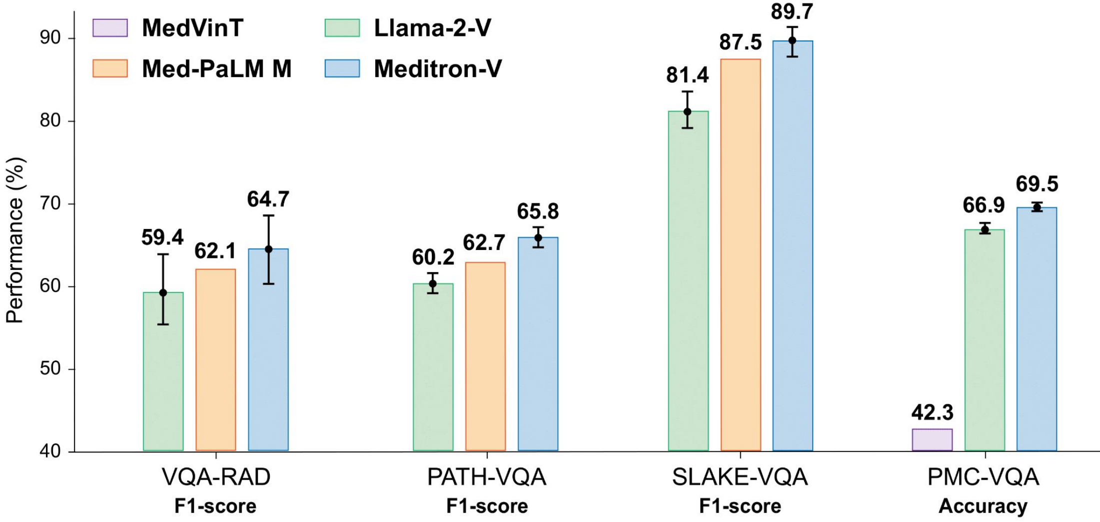


Benefits of scaling inference-time compute:
Meditron's performance continues to improve when more advanced inference strategies are applied.

Slides credit: Zeming Chen

Multimodal Fine-tuning for Multimodal Reasoning





Meditron-V (7B) outperforms both open-access models and the commercial SOTA, Med-PaLM M (562B) on all benchmarks

Slides credit: Zeming Chen

What about the general domain?

COCO Captioning Dataset

The man at bat readies to swing at the pitch while the umpire looks on.

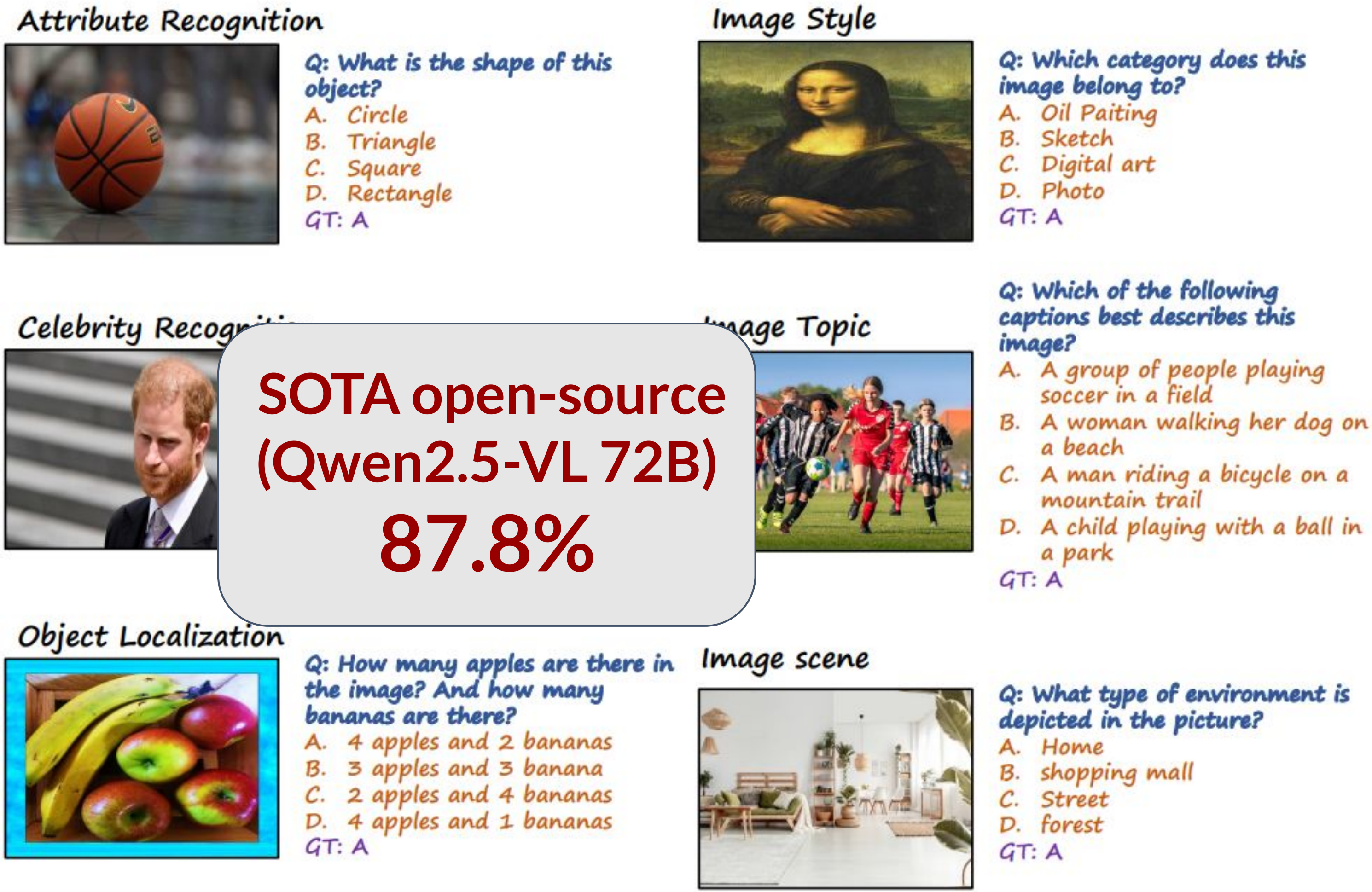
A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

MMBench

3k single-choice questions over 20 different skills



SOTA open-source (Qwen2.5-VL 72B)
87.8%

Liu et al. *MMBench: Is Your Multi-modal Model an All-around Player?*. ECCV 2024

Everything seems great, right?

Let's play a game!

- There is a chair missing on the second row from the back.
- There are three consecutive red chairs in the front row

✗ o1: No anomaly
✗ GPT-4o: The seating arrangement is missing a seat in the bottom row, creating an empty space.
✗ LlavaOneVision-72B: No anomaly
✗ InternVL-78B: No anomaly
✗ QwenVL-72B: The seat at the bottom right corner appears to be partially cut off by the edge of the image.

The button for floor number 2 is missing.

✗ o1: The braille labels for 3 and 5 appear identical.
✗ GPT-4o: The button for floor 1 is labeled with a star instead of a number.
✗ LlavaOneVision-72B: The close button is red, which is unusual as it is typically found in a different color.
✗ InternVL-78B: The number 1 button is missing its label.
✗ QwenVL-72B: The button labeled ‘1’ has a star symbol next to it, which is not present on other numbered buttons.

The poster says forty carrots but there are only twenty eight illustrations of carrots.

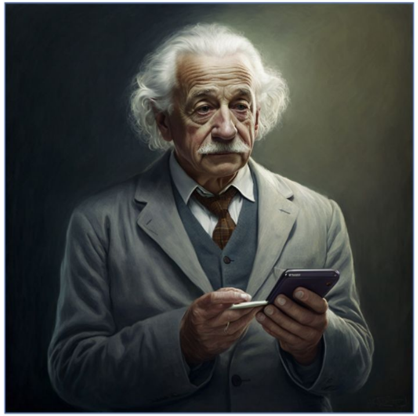
- ✗ o1: No anomaly
- ✗ GPT-4o: No anomaly
- ✗ LlavaOneVision-72B: No anomaly
- ✗ InternVL-78B: No anomaly
- ✗ QwenVL-72B: No anomaly

What’s going on?

CAVE: Detecting and Explaining Commonsense Anomalies in Visual Environments

Rishika Bhagwatkar*, **Syrielle Montariol***, Angelika Romanou, Irina Rish, Antoine Bosselut

How to evaluate VLM’s anomaly detection ability ?



What makes this image weird?

Einstein’s death (1955) was before the modern smartphone was invented (2007).

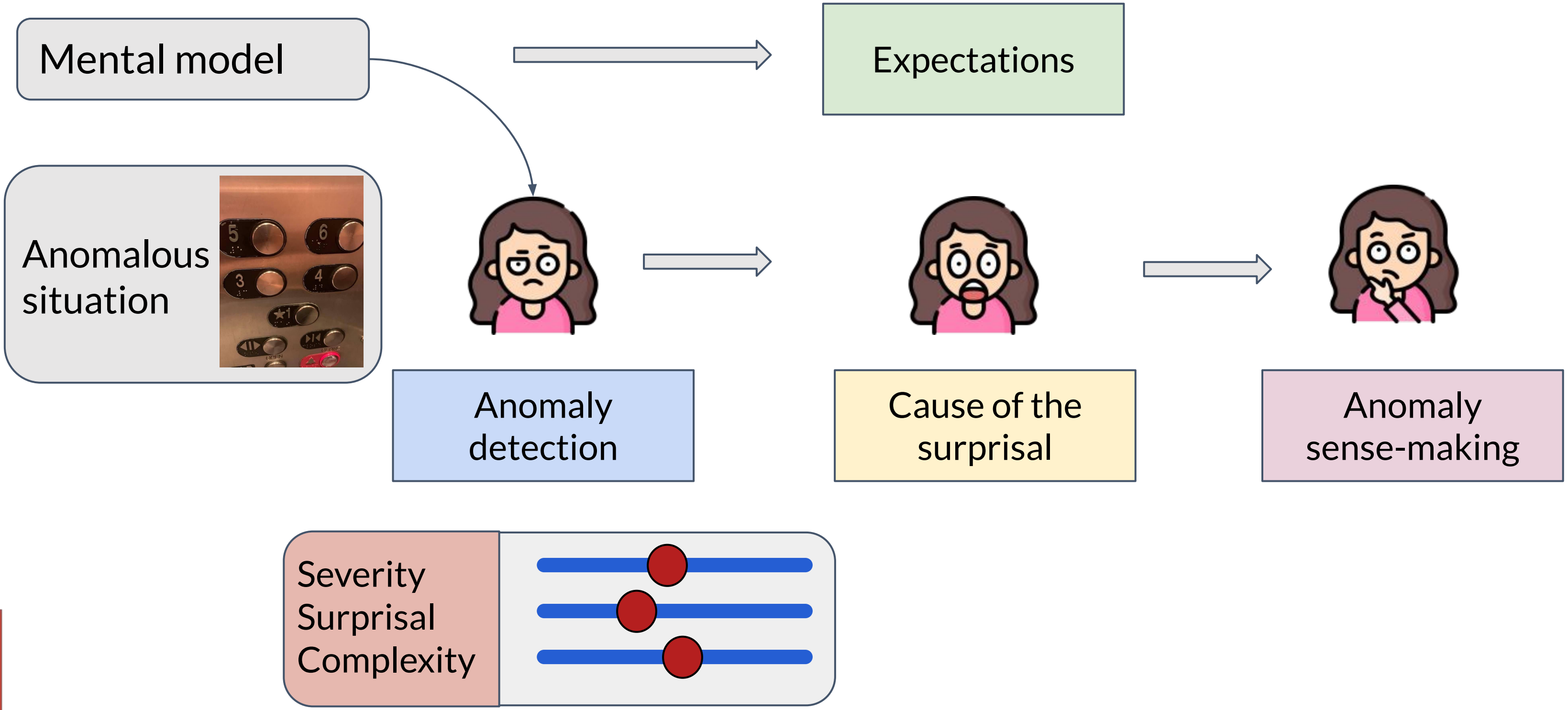
A candle needs a constant supply of oxygen to burn, which does not exist in a sealed bottle.

Question: What is wrong with this image?
Ground truth: It is raining inside the building

Bitton-Guetta, Nitzan, et al. "Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images." *ICCV* 2023.

Taesiri, Mohammad Reza, et al. "GlitchBench: Can large multimodal models detect video game glitches?" *CVPR* 2024.

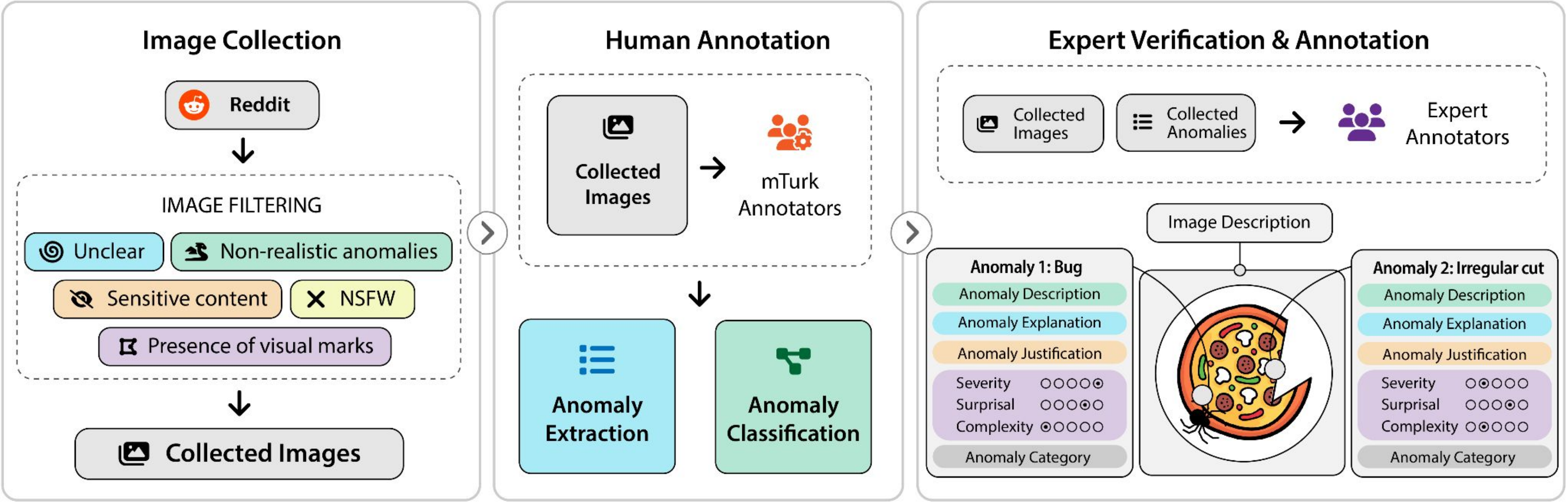
How do humans detect and understand anomalies?



Let’s make a dataset like this:

1. Natural images, not centered around the anomaly
2. Realistic situations and large diversity of commonsense anomalies
3. Open answers and fine-grained annotations

Dataset creation process



Example

Anomaly Description

There is a chair missing on the second row from the back

Correct Version Description

All the rows have the same number of chairs

Anomaly Explanation

The seating capacity of the place is reduced by one

Anomaly Justification

The chair was damaged but there was no replacement one.

Entity absence

Severity
Surprisal
Complexity

Anomaly Description

There are three consecutive red chairs in the front row

Correct Version Description

The alternating pattern of red and white is the same everywhere

Anomaly Explanation

The color inconsistency disrupts the aesthetics

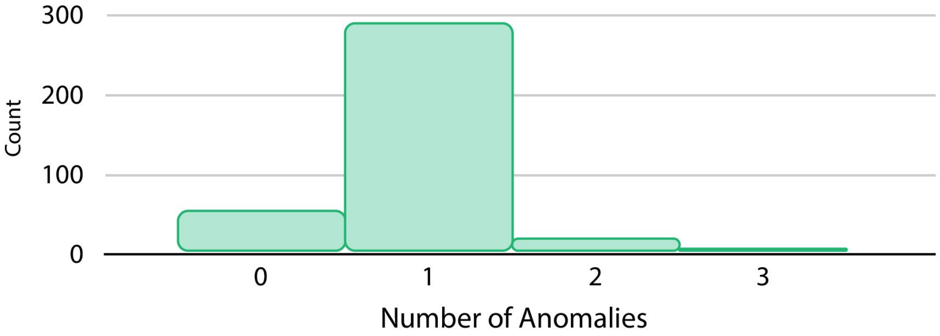
Anomaly Justification

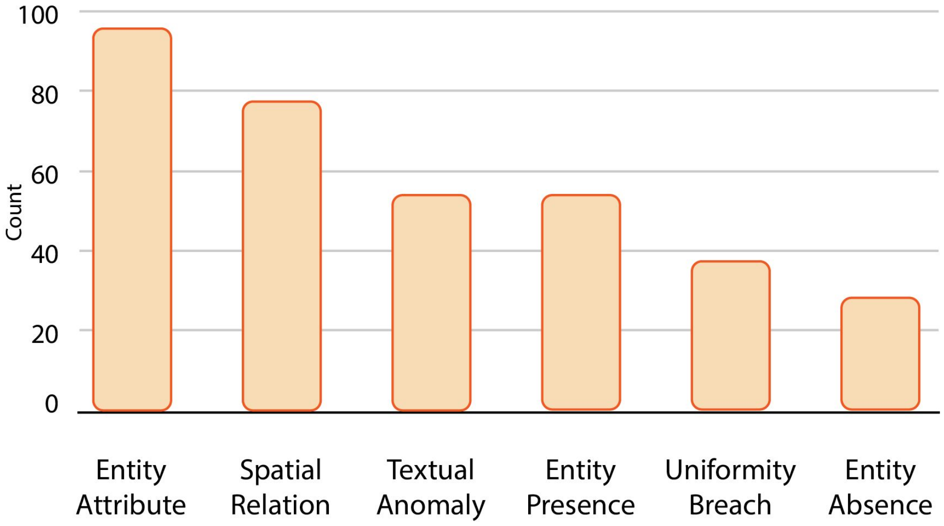
A chair was broken and the only replacement one was red

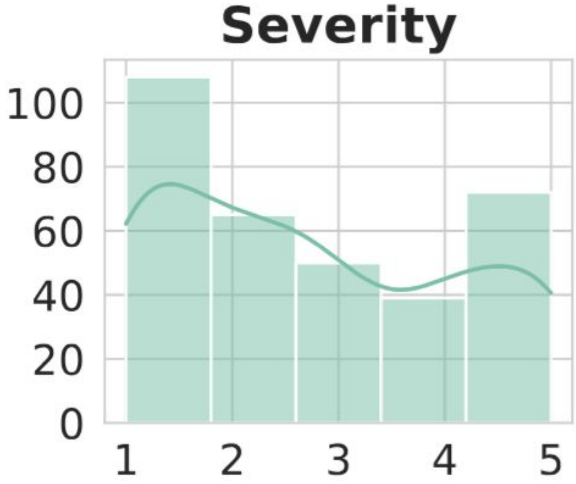
Uniformity breach

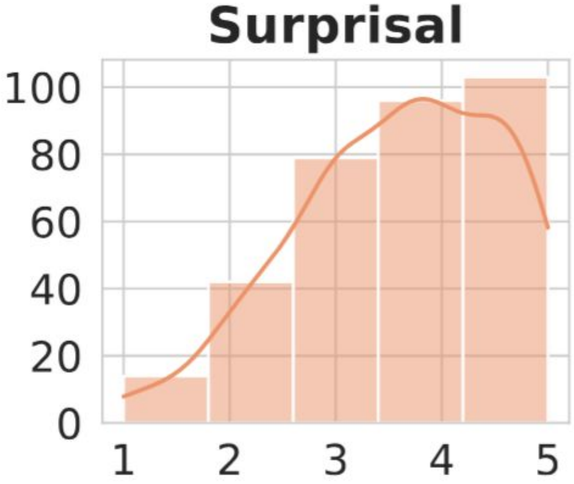
Severity
Surprisal
Complexity

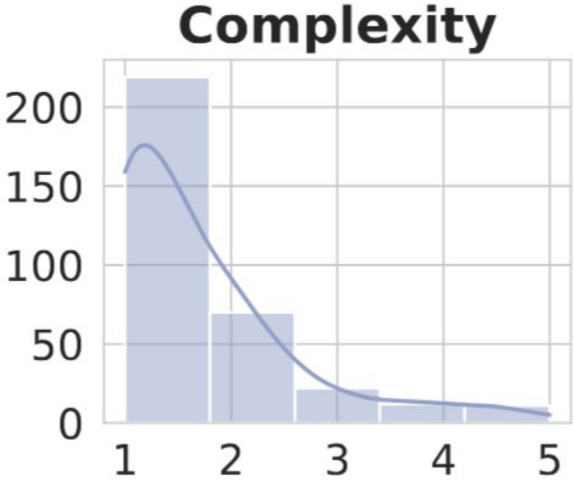
Some statistics









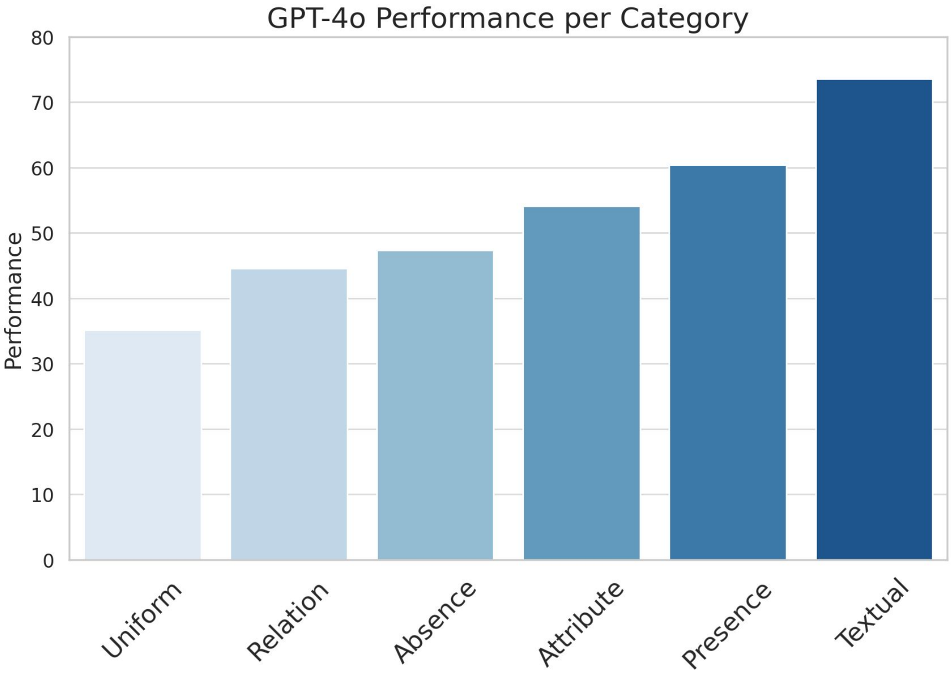


Anomaly Description task

anomalies correctly detected by the model | anomalies hallucinated by the model | anomalies missed by the model

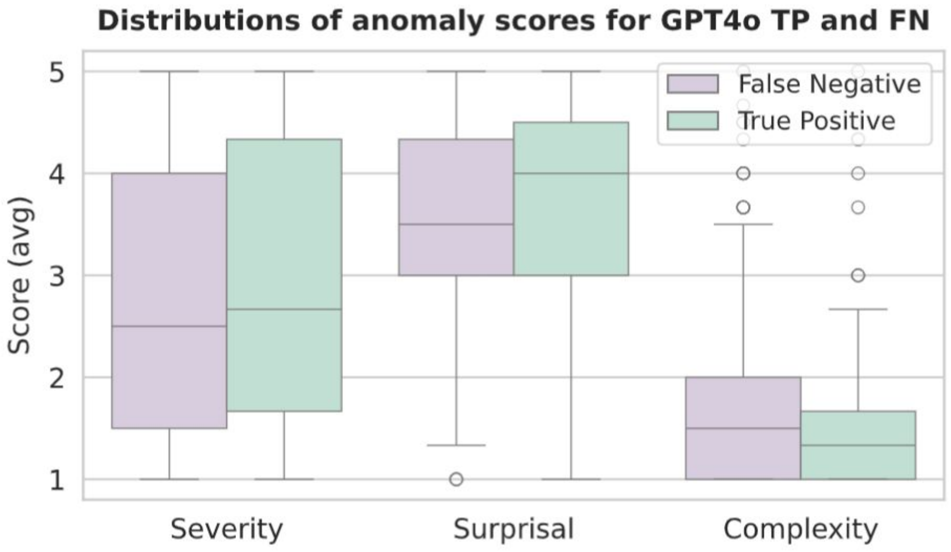
Model	TP	FP	FN	Precision	Recall	F1
Llama3.2 90B	105	407	227	20.5	31.6	24.9
LlavaOV 72B	77	**155**	255	33.2	23.2	27.3
InternVL2.5 38B	110	210	222	34.4	33.1	33.7
QwenVL2.5 72B	**126**	247	**206**	33.8	**38**	35.7
InternVL2.5 78B	112	167	220	**40.1**	33.7	**36.7**
o1	131	257	201	33.8	39.5	36.4
Claude 3.5	133	**128**	199	**50.9**	40.1	44.9
GPT-4o	**166**	197	**166**	45.7	**50**	**47.8**

What’s hard /easy for the VLM?



- Textual anomalies are easy
- Spatial reasoning problems are hard

What’s hard /easy for the VLM?



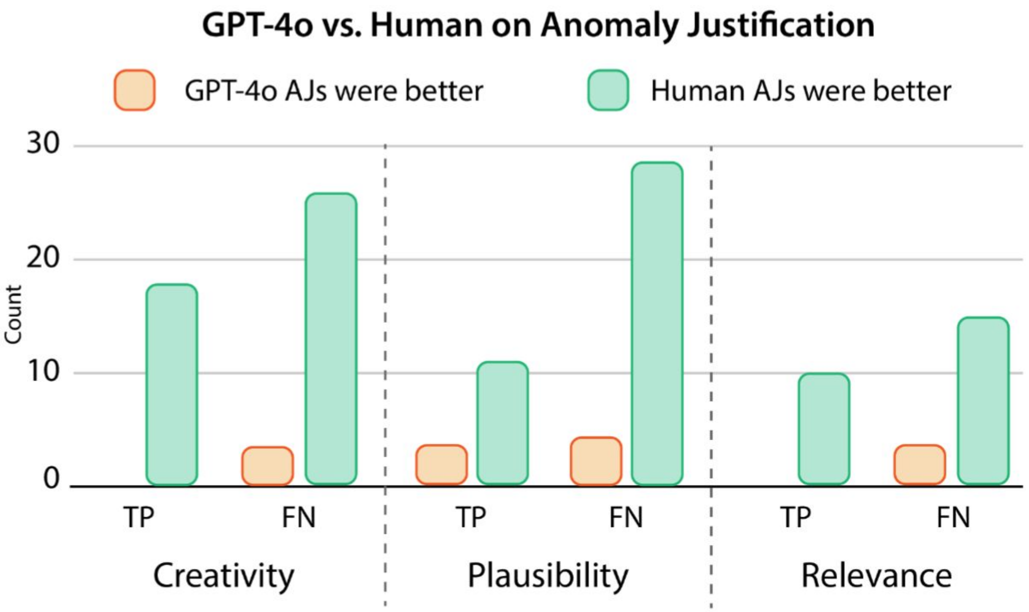
- Low severity and high complexity samples are harder
- High severity and high surprisal are easier

Given the anomaly description, can the model explain the anomalies ?

Subset	GPT-4o	InternVL2.5 78B
TP on AD	78%	74%
FN on AD	78%	66%

(Manual annotation on 50 TP and 50 FN)

Given the anomaly description and explanation, can the model justify the anomalies ?



→ Lower plausibility, creativity and relevant

→ Especially for FN

Lack of creativity

Anomaly Description

The logo for the restaurant is off-center and sticks out of the side of the building

Human Justification

There used to be a second sign on the right, but it fell before the photo was taken

GPT4o Justification

The logo may have been improperly installed due to lack of precise measurement or miscommunication during the installation process

Reasoning error / low plausibility

Anomaly Description

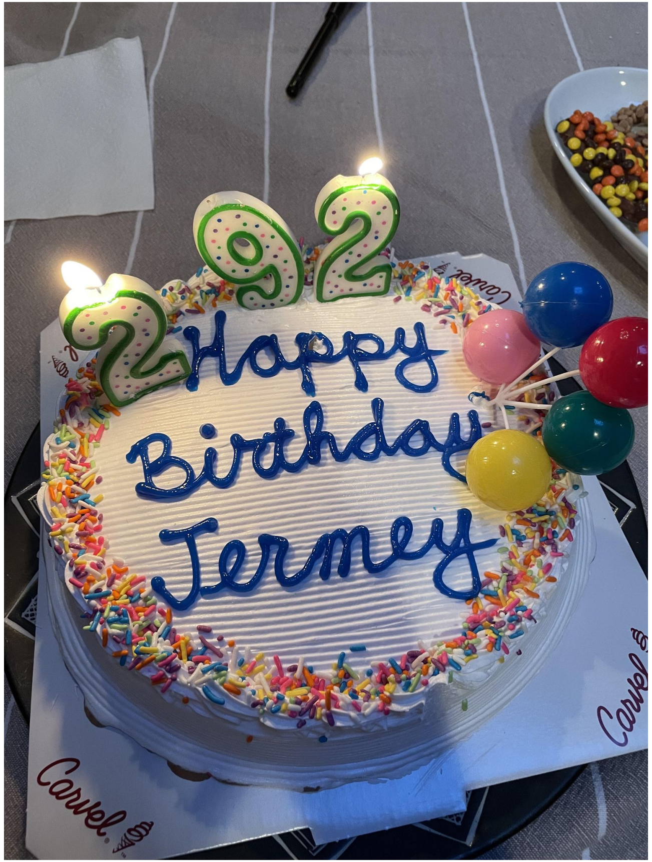
The candles depict the age of “Jermey” as 292 years old

Human Justification

The candles have been placed as a joke and exaggerating Jermey’s age

GPT4o Justification

The candles might have been arranged incorrectly, resulting in the age being displayed as 292 instead of the intended order, such as 29 or 92.



Perception error

Anomaly Description

There is trash on the floor next to the trash can

Human Justification

Users threw the trash in the bin and missed it

GPT4o Justification

The trash can is full or was not opened properly, causing overflow and resulting in trash being left outside

How to increase a model's visual reasoning and perception ability?

→ Inference-time scaling: Visual chain-of-thought reasoning

What datasets to train / evaluate visual reasoning?

MMBench

Physical Property Reasoning

Q: The object shown in this figure:
A. Is the hardest naturally occurring substance on Earth.
B. Conducts electricity well at room temperature.
C. Is typically found in igneous rocks like basalt and granite.
D. Has a low melting point compared to other minerals.
GT: A

Function Reasoning

Q: What's the function of the demonstrated object?
A. Cut vegetables
B. stir
C. Water purification
D. Boiling water
GT: B

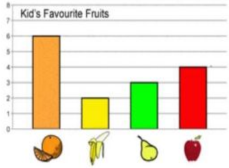
Identity Reasoning

Q: What's the profession of the people in this picture?
A. Librarian
B. radio host
C. gardener
D. lawyer
GT: C

Future Prediction

Q: What will happen next?
A. this person is gonna cry
B. this person is gonna laugh
C. this person is gonna get mad
D. both A,B, and C
GT: A

Structuralized Image-text Understanding



Q: According to this image, which fruit did the most kids like?
A. Orange
B. Banana
C. Pear
D. Apple
GT: A

Nature Relation

Q: In nature, what's the relationship between these two creatures?
A. Predatory relationships
B. Competitive relationships
C. Parasitic relationships
D. Symbiotic relationship
GT: B

Physical Relation

Q: Who is closer to the football in the image, the player in the black jersey or the player in the green jersey?
A. The player in the black jersey
B. The player in the green jersey
C. They are equally close
D. It cannot be determined
GT: A

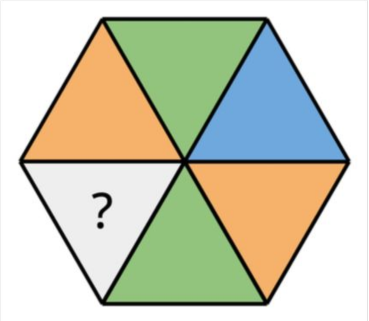
Liu et al. *MMBench: Is Your Multi-modal Model an All-around Player?* ECCV 2024

GQA

A1. Is the tray on top of the table black or light brown? light brown
A2. Are the napkin and the cup the same color? yes
A3. Is the small table both oval and wooden? yes
A4. Is there any fruit to the left of the tray the cup is on top of? yes
A5. Are there any cups to the left of the tray on top of the table? no
B1. What is the brown animal sitting inside of? box
B2. What is the large container made of? cardboard
B3. What animal is in the box? bear
B4. Is there a bag to the right of the green door? no
B5. Is there a box inside the plastic bag? no

PuzzleVQA

Model Input



Question: What is the missing color of the part denoted with a question mark?
Options: (A) purple (B) green (C) blue (D) yellow
Answer: Let's describe the image first and think step by step.

Hudson et al. "Gqa: A new dataset for real-world visual reasoning and compositional question answering." CVPR 2019.

Ghosal et al. Are Language Models Puzzle Prodigies? Algorithmic Puzzles Unveil Serious Challenges in Multimodal Reasoning. arXiV preprint 2024.

The problem with visual reasoning benchmarks

- Synthetic or schematic visuals
- LLM-generated explanations
- Simplified / unrealistic questions

→ What's a domain with **realistic problems and images** involving **complex visual reasoning**, where we can get **expert-written explanations**?

Question: Can I overtake the vehicle in front of me?

Possible answers:
(A) Yes (B) No

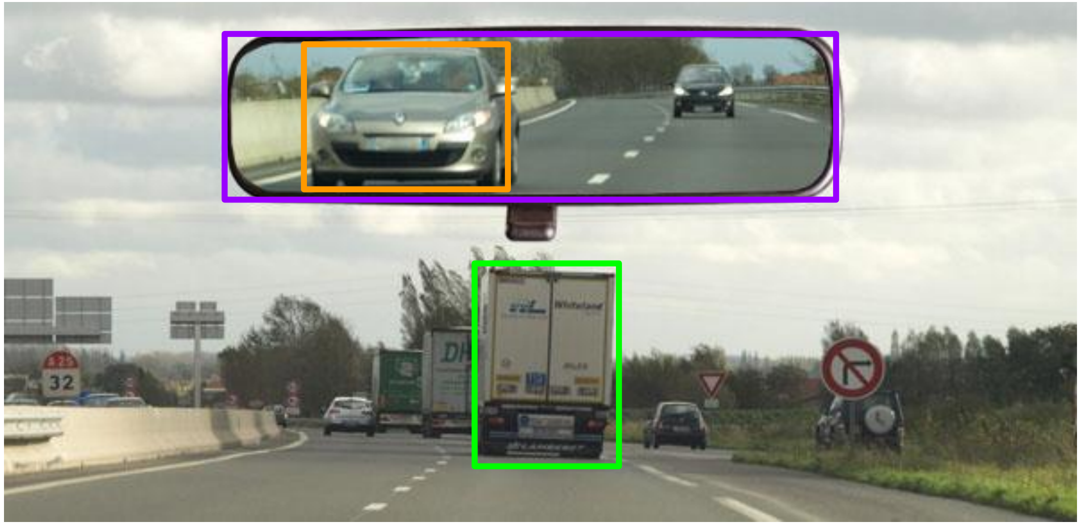
Answer: B
Explanation: On this two-way road, no one is coming the other way. I don't see anyone in my rearview mirror either. However, next to the shadow of my vehicle, I see another shadow. Another vehicle is therefore in my blind spot. I decide not to overtake.

Visual Chain-of-Thought Reasoning in Real-World Driving Scenarios

Syrielle Montariol*, Charles Corbière*, Simon Roburin*, Antoine Bosselut, Alexandre Alahi

How would a human reason about an image?

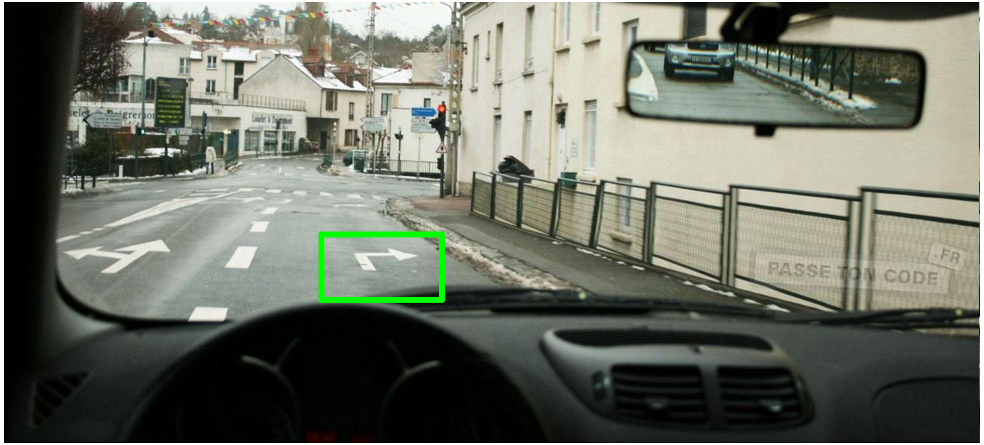
```
Can I overtake the truck?
A: Yes, B: No
```



1) To overtake the truck in front of me, I need to go to the left lane.
2) There's a rear view mirror showing the scene behind me.
3) A car is about to overtake me. It is in the left lane and will be soon next to me.
4) It would be dangerous to go to the left lane now.

How should a VLM reason?

```
To go straight, can I stay in this lane?
A: Yes, B: No
```



```
The road markings indicate that the right lane
can only be used for turning right. Therefore,
I must position myself in the left lane to go
straight or to the left.
Answer: B
```

1) Identify **relevant entities** in the image: directional arrow
2) Extract its **attribute**: it points to the right.
3) Extract its **location**: it is in my lane
4) Conclude

Grounding tasks: linking image regions with textual descriptions

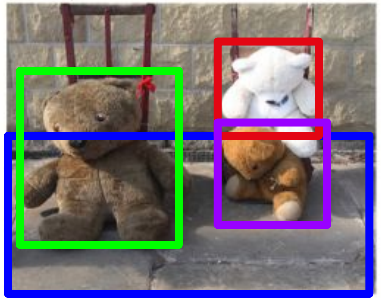
- Referring Expression Comprehension (REC)
  Prompt: *Where is the white teddy bear ?*

  → *[x1, x2, y1, y2]*

- Grounded Captioning

  Prompt: Describe the image.
  → *A dark brown [x1, x2, y1, y2], a light brown [x1, x2, y1, y2] and a white teddy bear[x1, x2, y1, y2] on a sidewalk [x1, x2, y1, y2].*

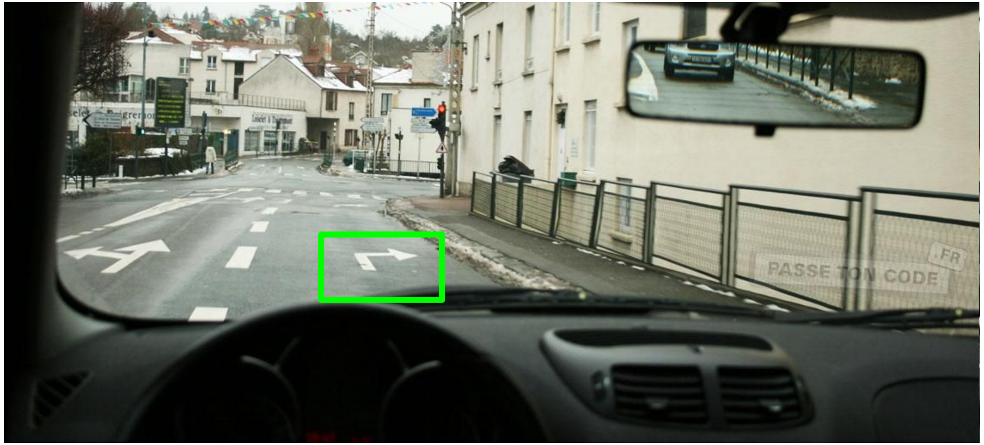
Output:



Let's use it to guide the VLM through these steps!

To go straight, can I stay in this lane?

A: Yes, B: No



The relevant entity is: **directional arrow.**

The relevant entity is: **directional arrow [ww,xx,yy,zz].**

The relevant entity is: **directional arrow**  **.**

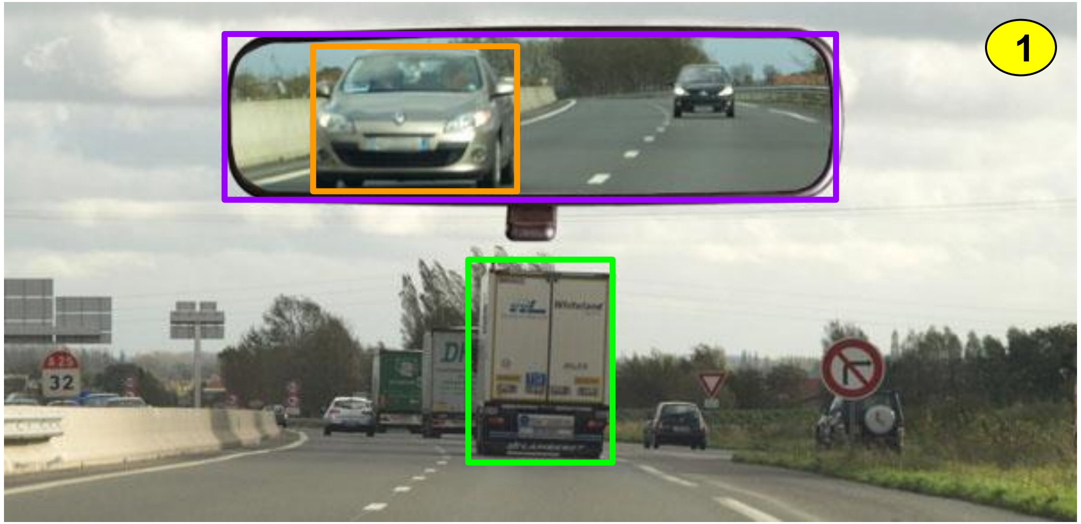
The **road markings** indicate that the right lane can only be used for turning right. Therefore, I must position myself in the left lane to go straight or to the left.

Answer: B

Let’s make a dataset that includes all these elements!

1. Image
2. Question + options
3. List of relevant entities + coordinates
4. Interleaved explanation

Can I overtake the truck?
A: Yes, B: No (2)



A vehicle is about to overtake you. You cannot overtake.

vehicle
rear-view mirror
truck (3)

(4) A **vehicle** is about to overtake you, as seen in the **rear-view mirror**. You cannot overtake the **truck**.

Dataset creation process

Data collection

- **Scraping** from 15 online platforms with free practice tests
- **Cleaning** and format standardisation
- **Filtering** of questions where the image isn't necessary to answer

Entities annotation

- **Pre-annotation** using GPT4o + GroundingDino + rule-based refinement
- **Manual annotation** with CVAT: 5 657 entities, spanning 256 unique labels

Interleaved explanations

- Interleave entities inside the expert explanation, using GPT4o + self-correction + rule-based refinement

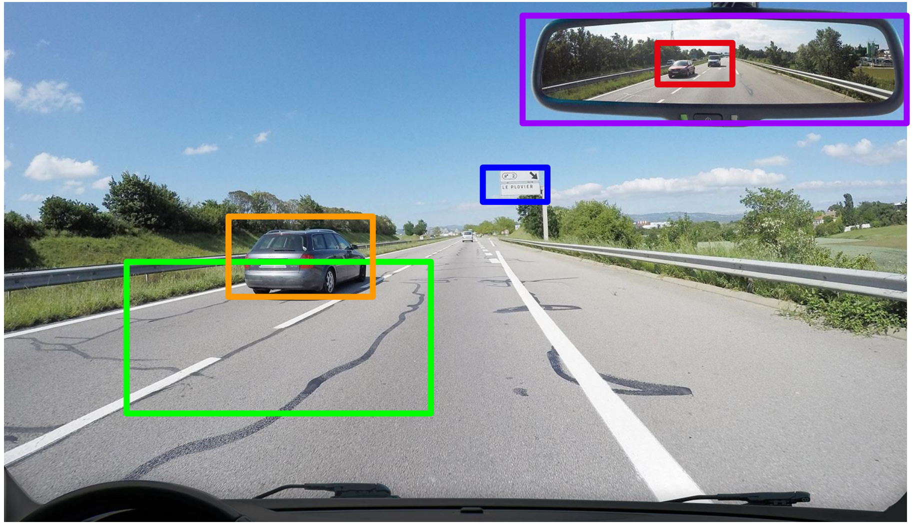
Total: 3,931 samples

DrivingVQA example

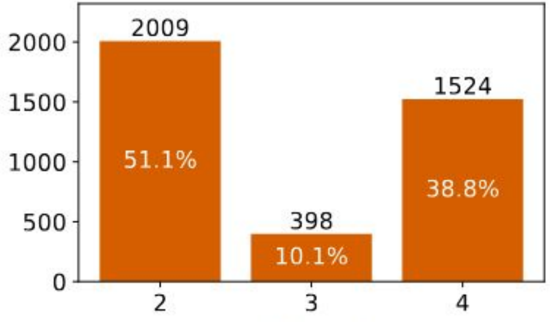
The vehicle in front can still take the next exit.
A: Yes, B: No
I can follow this vehicle to overtake.
C: Yes, D: No

The marking in the middle of the two lanes of traffic only prohibits vehicles in the left lane from merging to take the exit. However, vehicles are approaching from behind, so I must give up overtaking.

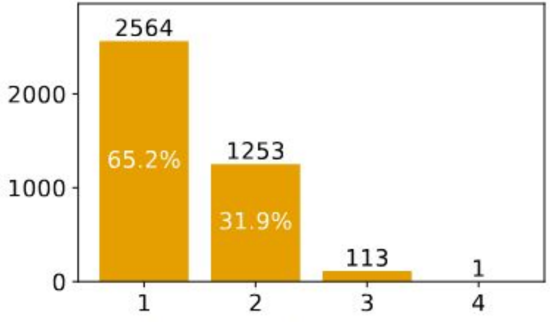
The **dashed line [317.36, 423.1, 394.94, 188.64]** in the middle of the two lanes of traffic only prohibits **vehicles in the left lane [394.01, 371.19, 211.56, 119.92]** from merging to take the **exit indicated by the exit sign [819.58, 272.31, 75.71, 51.58]**. However, **vehicles [1101.09, 77.8, 93.61, 49.79]** are approaching from behind, as seen in the **rear-view mirror [873.99, 9.21, 623.74, 196.48]**, so I must give up overtaking.
Answers: B, D



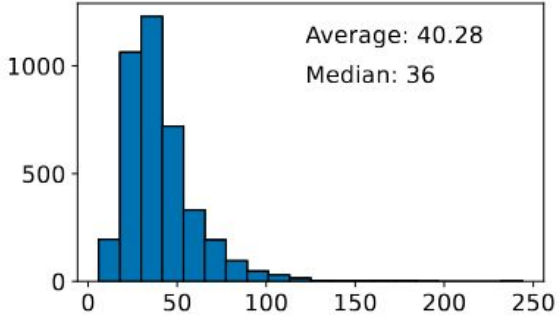
Dataset statistics



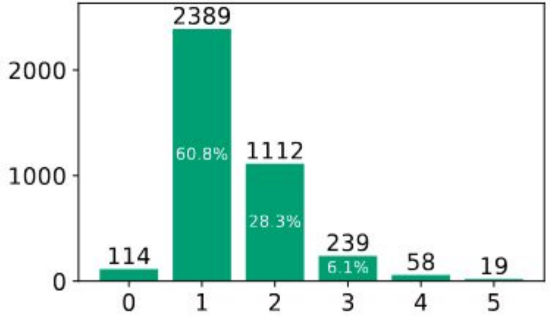
(a) Number of possible answers.



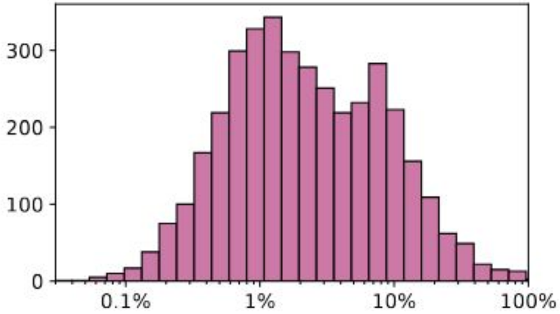
(b) Number of correct answers.



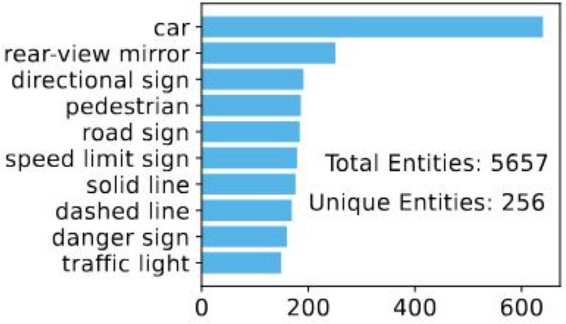
(c) Number of words in explanations.



(d) Number of relevant entities.



(e) Image coverage by entities (%), log-scale.



(f) Top 10 most frequent entities.

Research questions

1. How do VLMs perform **out-of-the-box** in real-world scenarios involving complex visual reasoning?

2. Which **type of entity-related information** (names, coordinates, visual content) most effectively improves VLMs' performance?

3. Can VLMs learn **how to predict the relevant entities** to generate visual-chain-of-thought reasoning?

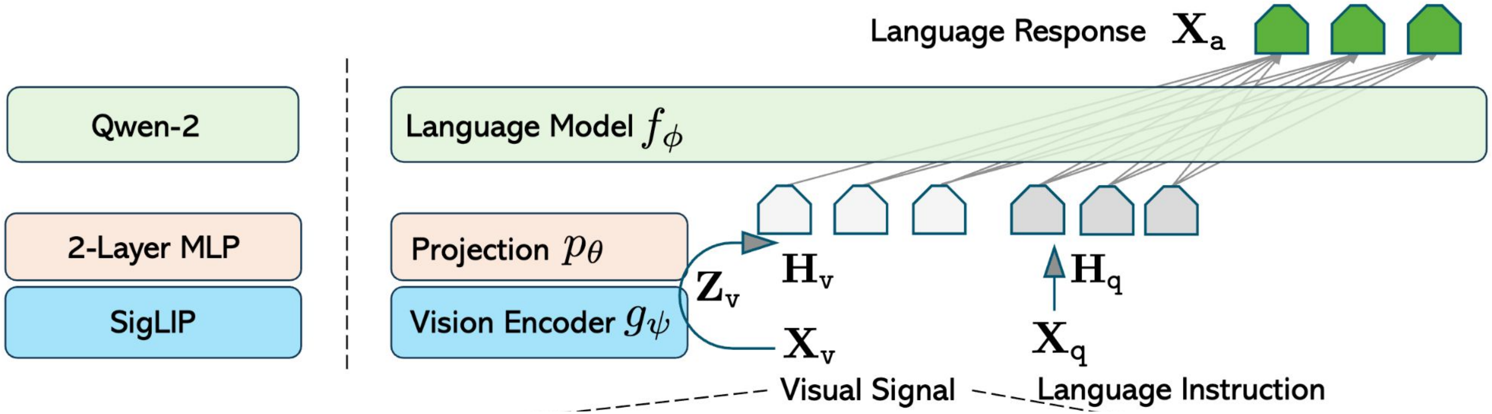
4. Can we achieve the same performance with silver entities & coordinates instead of gold?

Experimental setup

Evaluation metric:
→ Exam score: all correct answers must be selected

Model fine-tuning:
→ LLaVA-OneVision-7B

Train-test split: 80-20%



Bo et al. *LLaVA-OneVision: Easy Visual Tas kTransfer.* arXiV preprint 2024.

Conversation format

Q = Question

P = Possible answers

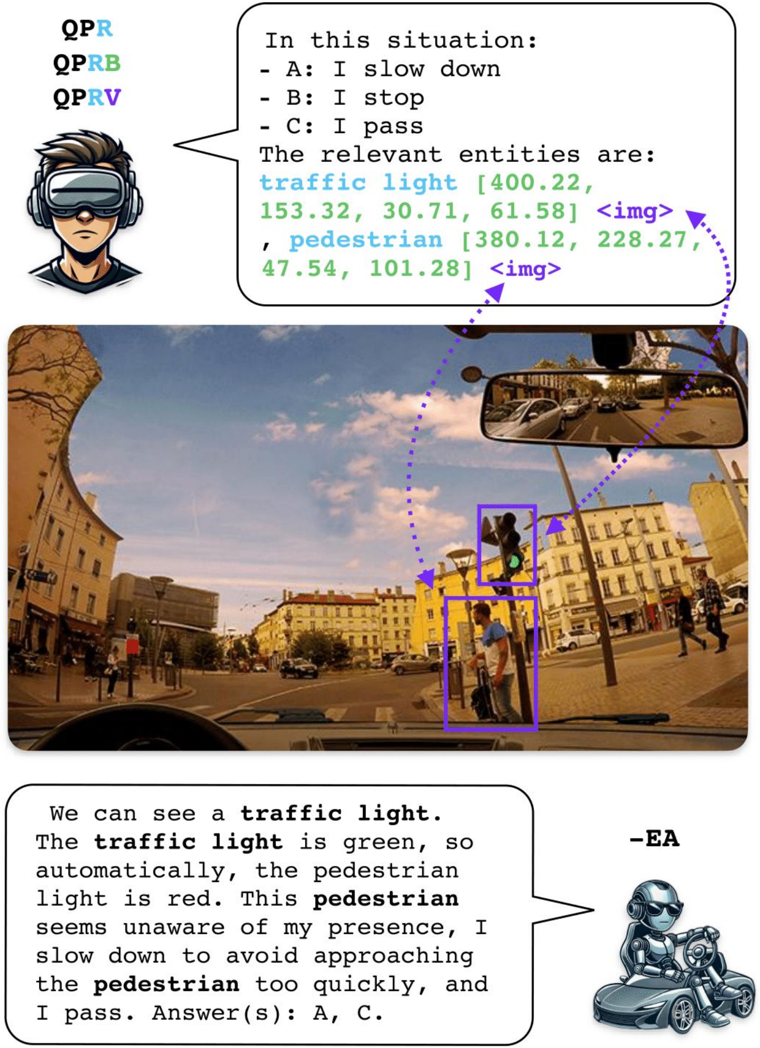
R = Relevant entities

B = Bounding box coordinates

V = Visual crops

E = Explanation

A = Answer



R1: Zero-shot performance

Model	Conv. Format	Exam Score (%)
Random baseline		19.8
LLaVA-OV-0.5B	QP-A	29.0
LLaVA-OV-7B	QP-A	33.5
LLaVA-OV-72B	QP-A	45.5
GPT-4o	QP-A ***w/o image***	33.1
	QP-A	59.1
	QP-EA	**60.7**
Human baseline		79.6

R2: Oracle setting: entity format

Conv. Format	Explanation	Entity Name	Entity BBox	Entity visual patch	Exam Score (%)
QP-A					53.7 (± 0.6)
QP-EA	✔				55.5 (± 1.8)
QPR-EA	✔	✔			58.7 (± 0.8)
QPR-EA†	✔	✔		image crops	59.0 (± 0.5)
QPRB-EA	✔	✔	✔		59.1 (± 1.2)
QPRV-EA	✔	✔		✔	**60.4** (± 1.2)

- Incorporating the **list of relevant entities** into the prompt (QPR-EA) → **+3.2 pts** over QP-EA
- Limited use of **localization information** from coordinates (QPRB-EA)
- **Prompting with visual crops**, (QPRV-EA) achieves the **best exam score** (60.4%)

R2: Oracle performance

To go straight, can I stay in this lane?

A: Yes, B: No

QPR-EA:

Input: The relevant entity is: directional arrow.

Output: If I want to go straight, the road markings indicate that I am correctly positioned, as I can either continue straight or turn right. Therefore, I can stay in this lane.

Answer: A.

QPRV-EA:

Input: The relevant entity is: directional arrow

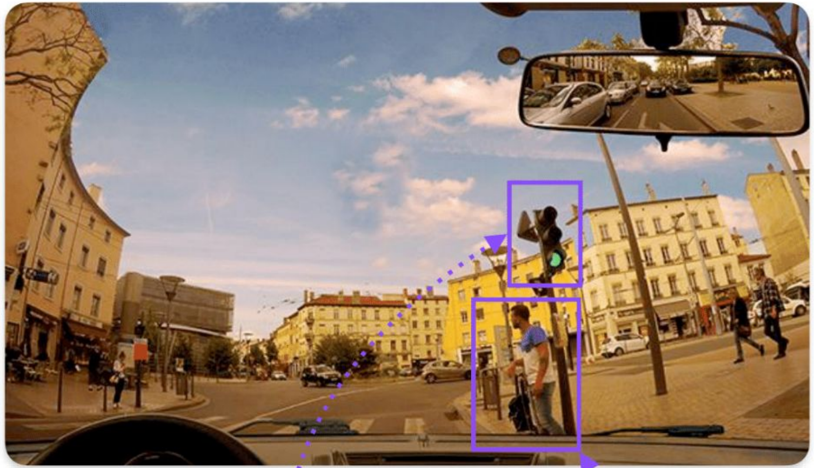
Output: The lane I am in is intended for users wishing to turn right. This is indicated by the directional arrow on the road markings. Therefore, I cannot stay in this lane if my intention is to go straight.

Answer: B.

QP-RB-RV-EA: Two-steps conversation format

QP

Unless explicitly stated otherwise, assume you are driving in France. List all relevant entities from the scene that are necessary to answer the following question, such as road signs, markings, signals or other vehicles in the image.
In this situation:
- A: I slow down
- B: I stop
- C: I pass



-R
-RB

The relevant entities for this problem are: traffic light [395.22, 140.14, 25.71], pedestrian [380.51, 227.18, 46.82, 100.75]

-RV

Their corresponding image patches are: traffic light <image>, pedestrian <image>. Then select all correct answers to the following question from the available options. Detail your reasoning step by step based on these entities and relevant driving rules. Provide the letters corresponding to you answer in the form A'Answer(s):<letters>.

-EA

I can see a green **traffic light**. The **pedestrian** must wait to pass. But pedestrians behaviour are often unpredictable.
Answer(s): A, C.

R3: Entity-predictive methods

Conv. Format	Explanation	Entity Name	Entity BBox	Entity visual patch	Exam Score (%)
QP-A					53.7 (± 0.6)
QP-EA	✔				55.5 (±1.8)
QP-REA	✔	✔			56.6 (± 0.7)
QP-RBEA	✔	✔	✔		56.2 (± 1.6)
QP-RB-RV-EA	✔	✔		✔	**57.5** (± 1.2)

- All **entity-predictive methods** outperform CoT prompting baseline (QP-EA)
- **Reasoning by leveraging visual crops** (QP-RB-RV-EA) ⟶ **+2.0 pts** over QP-EA

R4: Automatically-extracted entities

GPT-4o + GroundingDINO to extract and localize relevant entities (road signs, markings, traffic control, vehicles, …)
→Keeping top-5 by confidence scores

Conv. Format	Entities	Exam Score (%)
QP-EA	Human-annotated	55.5 (±1.8)
QP-RBEA	Automatically-extracted	54.8 (± 0.6)
	Human-annotated	**56.2** (±1.5)
QP-RB-RV-EA	Automatically-extracted	56.3 (± 2.2)
	Human-annotated	**57.5** (± 1.2)
QPRV-EA	Automatically-extracted	58.9 (± 1.4)
	Human-annotated	**60.4** (± 1.2)

- Replacing human-annotated entities with automatically extracted ones leads to a performance drop
- But **visual patches based methods** still outperform baseline!

Liu et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection.* ECCV 2024.

Next step: Scaling up inference-time compute with mixed-modality generation

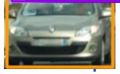
Can I overtake the truck?

A: Yes, B: No

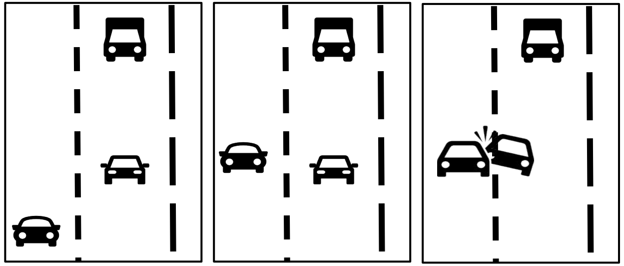
Grounded reasoning

The **rear-view mirror [ww, xx, yy, zz]** shows that a **vehicle [ww, xx, yy, zz]** is about to overtake you. It will prevent you from moving to the left lane to overtake the **truck [ww, xx, yy, zz]**. Answer: B

Reasoning with visualization

The rear-view mirror  shows that a vehicle  is about to overtake you. It will prevent you from moving to the left lane to overtake the truck.  Answer: B

Reasoning with world simulation



The rear-view mirror shows that a vehicle is about to overtake you. It will prevent you from moving to the left lane to overtake the truck. Answer: B

Chain-of-thought reasoning

The rear-view mirror shows that a vehicle is about to overtake you. It will prevent you from moving to the left lane to overtake the truck. Answer: B

What are the limits of SOTA VLMs' perception and reasoning, and how to improve it?

Meditron

- Scaling inference-time compute unlocks complex reasoning
- How an LLM can be fine-tuned into a VLM

- VLMs are very bad at detecting visual anomalies
- It mostly results from perception errors

- VLMs can be trained to reason using visual crops
- It can be done with only silver annotations

Thank you for your attention!

- Alignment between modalities
- Unsupervised entity extraction and linking
- Remote sensing, in particular for disaster assessment
- Textual causal, temporal and commonsense reasoning

syrielle.montariol@epfl.ch