Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 9.2: Reasoning – Inference Time

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Administrative Stuff

Upcoming Course Assignments

Midterm project report (due this Sunday 10/25 at 8pm ET)

- For each N-1 research idea, implement a first (baseline) model
  - The emphasis is on the result discussion (to update research ideas)
- Read carefully the instructions!!

Midterm project presentations (next week)

- In-class presentations: Tuesday 10/27 and Thursday 10/29
  - One presentation per team, maximum of N+1 minutes
  - Focus on the updated research ideas (from report)
- Attendance required for both sessions
  - Give feedback to at least 6 teams per session
- Upload your slides on shared Google Drive
  - Same laptop used for all presentations

Attendance using Poll Everywhere

1. Scan QR Code
   or
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

❌ *Do not try to sign-in*

3. Click ”Share location”

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 9.2: Reasoning – Inference Time

Louis-Philippe Morency

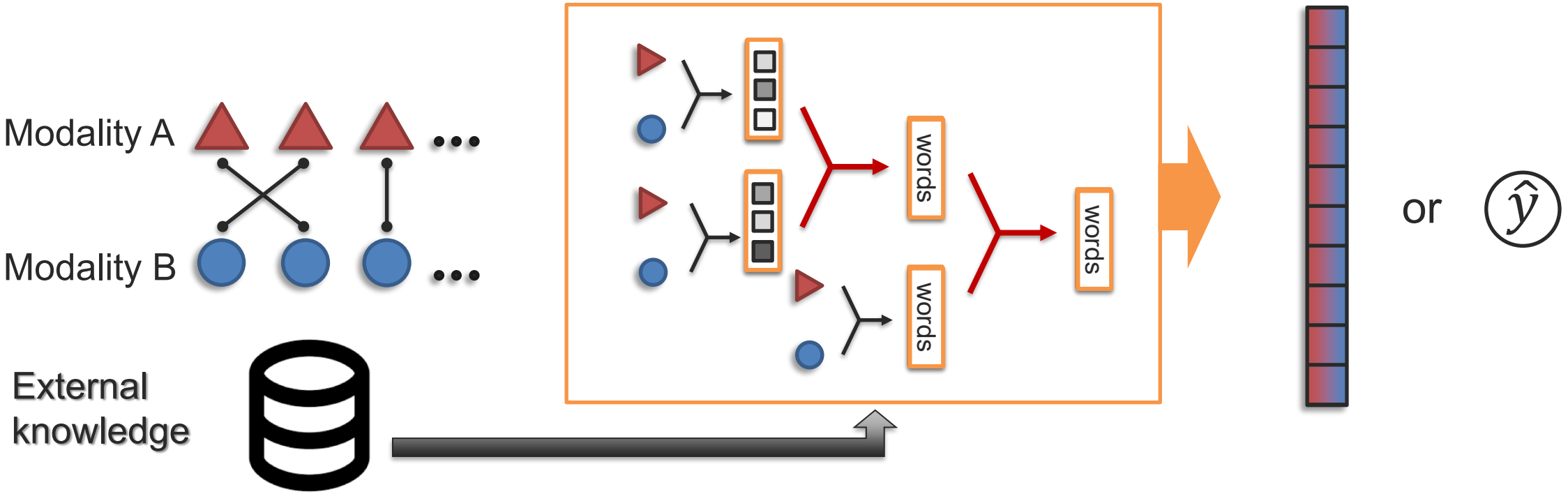
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

- Inference Design
  - Logical inference, causal inference
- Reasoning Traces
  - Chain-of-thought prompting and few-shot learning
  - Test-time compute scaling, and its multimodal ceiling
- Two strategies for improving multimodal reasoning
  - Improving perception
    - RLVR, GRPO, rule-based rewards
  - Active perception
    - Visual search, frame selection, designed vs learned policies

Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

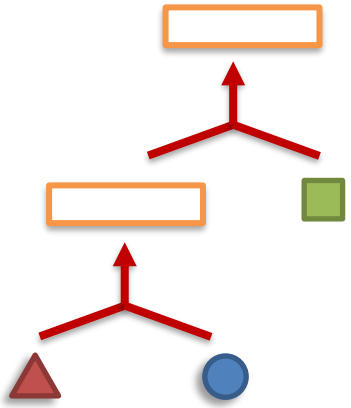


Reasoning

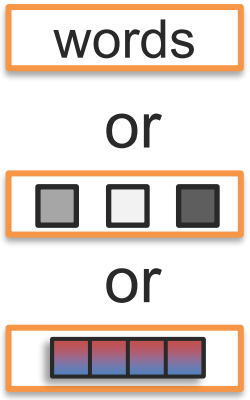
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

Sub-challenges:

Structure Modeling

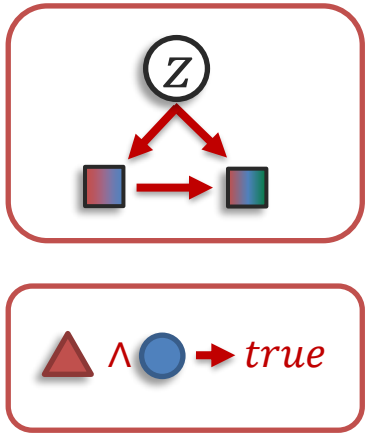


Intermediate concepts



External Knowledge

Inference Paradigm



Inference Design

Logical Inference

Inference through logical operators in question

Is there beer AND is there a WINE GLASS?

Is the man NOT wearing shoes AND is there beer?

Is there beer? *Is the man wearing shoes?*

Adversarial antonyms ✗

Logical connectives ✗

Basic premises ✓

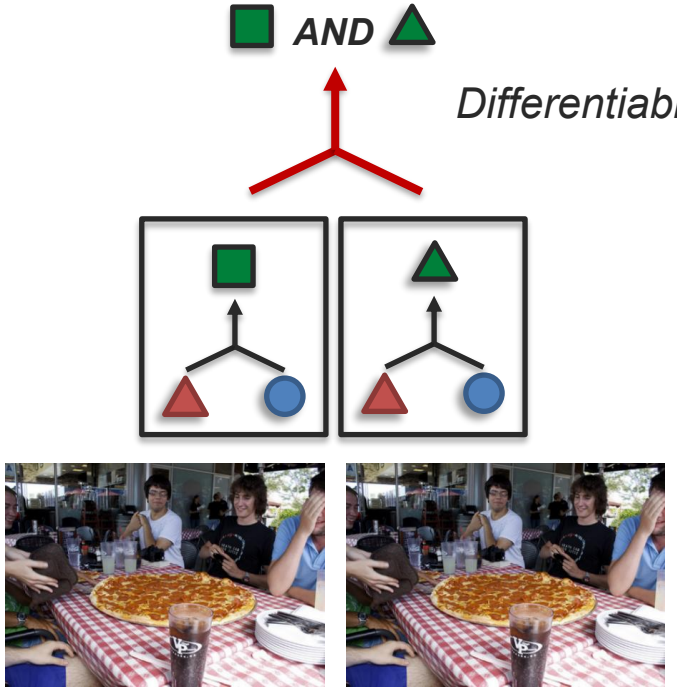
Existing models struggle to capture logical connectives.
How can we make them more logical?

[Gokhale et al., VQA-LOL: Visual Question Answering Under the Lens of Logic. ECCV 2020]

Logical Inference

Inference through logical operators in question

*Are they in a restaurant **AND** are they all boys?*



Are they in a restaurant?

Are they all boys?

*Differentiable **AND** composition operator!*

Also applies to other logic connectives: *AND, OR, NOT*

[Gokhale et al., VQA-LOL: Visual Question Answering Under the Lens of Logic. ECCV 2020]

Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

Covariant VQA

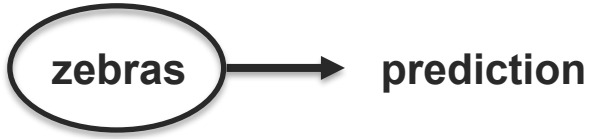
Target object in question

Q: How many zebras are there in the picture?

A: 2

Baselines: **2**

i.e., treatment variable



BUT: correlation or causation?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

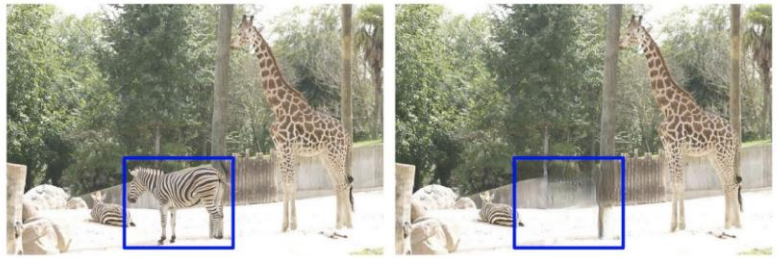
Causal VQA: does my multimodal model capture causation or correlation?

Covariant VQA

Target object in question

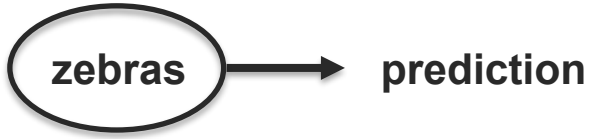
Q: How many zebras are there in the picture?

A: 2 *zebra removed* A: 1



Baselines: **2** **2**

i.e., treatment variable



Interventional conditional: $p(y|do(zebras = 1))$

Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

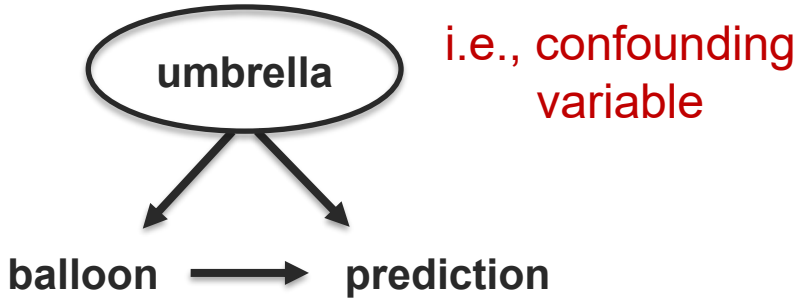
Invariant VQA

Target irrelevant object

Q: What color is the balloon?

A: red

Baselines: **pink**



i.e., confounding variable

Is my model picking up irrelevant objects?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Causal Inference

Causal VQA: does my multimodal model capture causation or correlation?

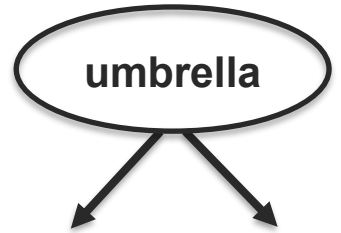
Invariant VQA

Target irrelevant object

Q: What color is the balloon?

A: red *umbrellas removed*; A: red

Baselines: **pink** **red**



i.e., confounding variable

Interventional conditional: $p(y|do(no\ umbrella))$

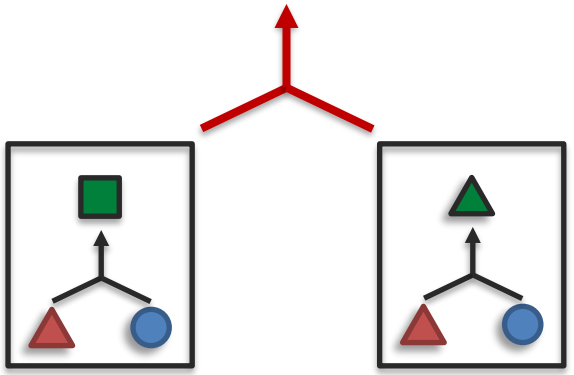
Existing models struggle to adapt to targeted causal interventions.
How can we make them more robust to spurious correlations?

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

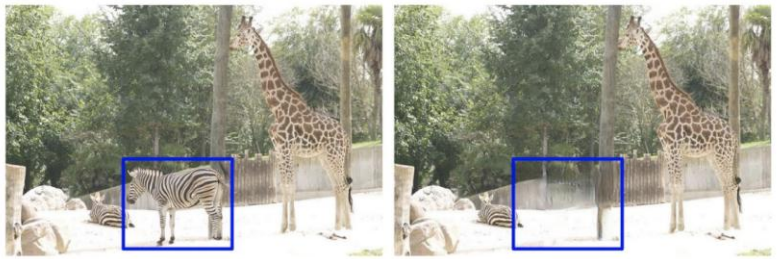
Causal Inference

Causal inference via data augmentation

Covariance
(targeted changes to answer)

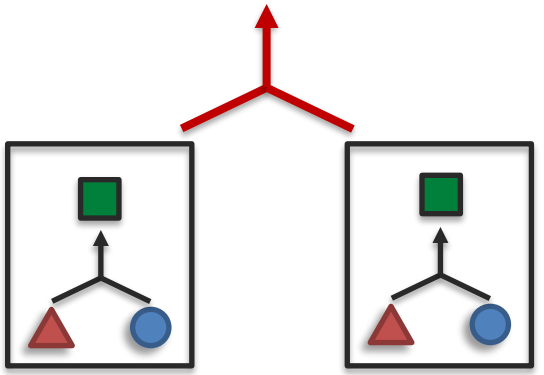


Q: How many zebras are there in the picture?
A: 2 *zebra removed* A: 1



With **relevant** object | Without **relevant** object

Invariance
(answer stays the same)



Q: What color is the balloon?
A: red *umbrellas removed*; A: red

With **irrelevant** object | Without **irrelevant** object

[Agarwal et al., Towards Causal VQA: Revealing & Reducing Spurious Correlations by Invariant & Covariant Semantic Editing. CVPR 2020]

Reasoning Traces

Chain-of-Thought and Reasoning Traces

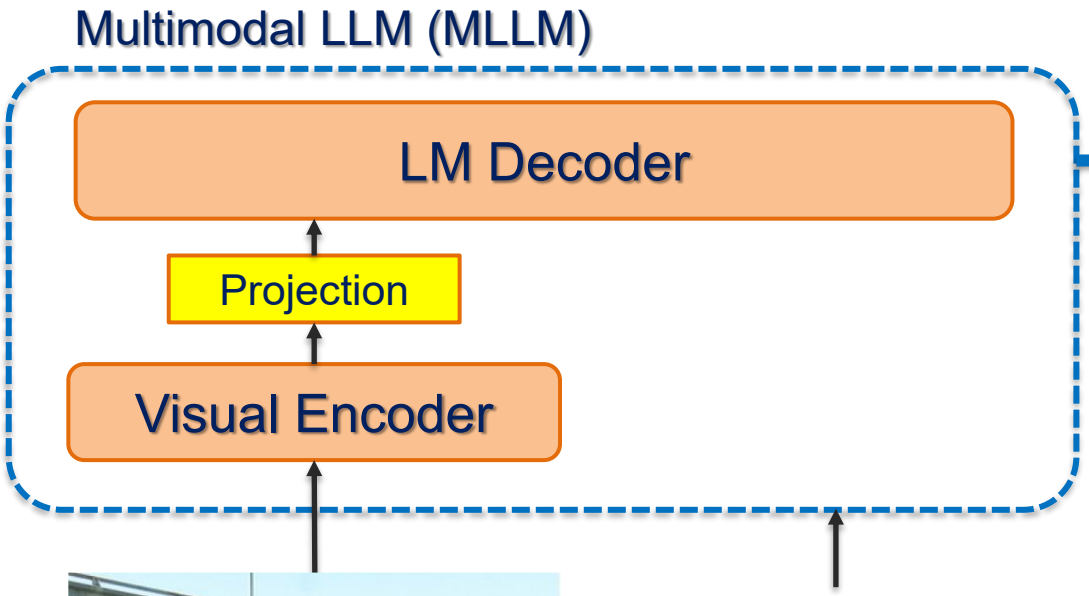
Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?
A: ***Let's think step by step.***

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

<think> There is … **</think>**
<answer> The answer is … **</answer>**

Kojima et al., 2022

MLLM Reasoning Traces

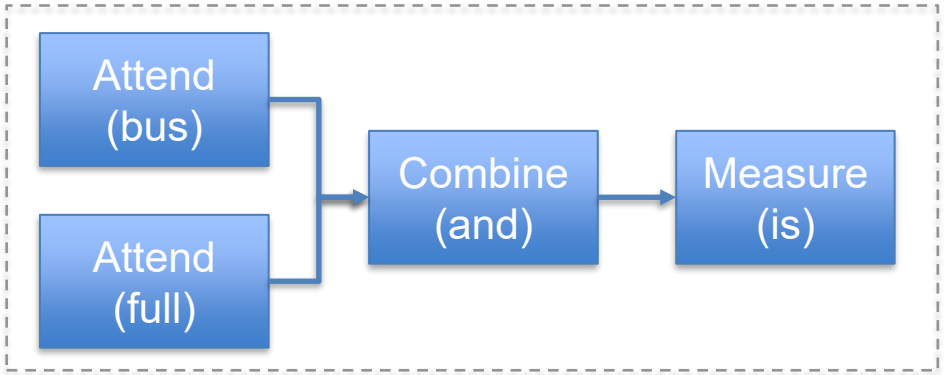


"Is the bus full of passengers?

Let's think step by step."

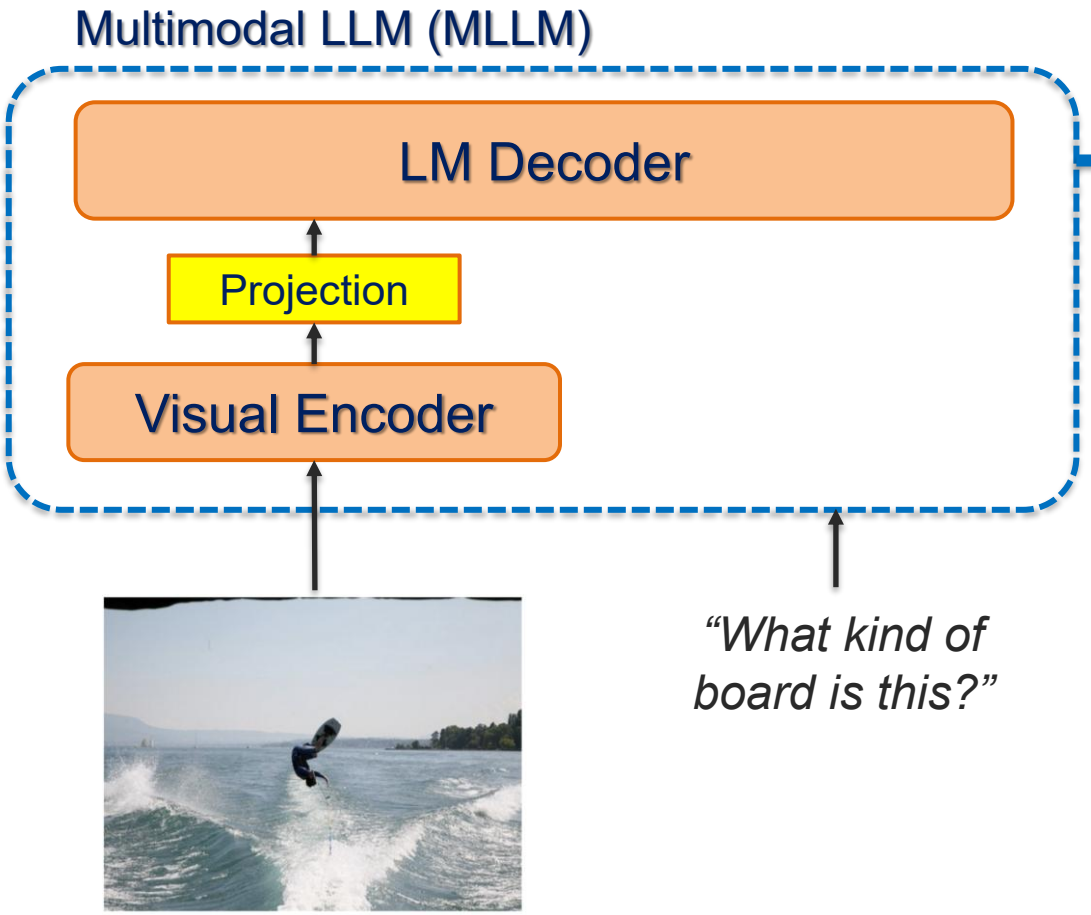
<think>
First, find the bus.
Now look at the passengers inside it.
Is it full?
</think>
<answer> Yes **</answer>**

Computation layout



Andreas et al., 2016

MLLM Reasoning Traces and Few-Shot Learning

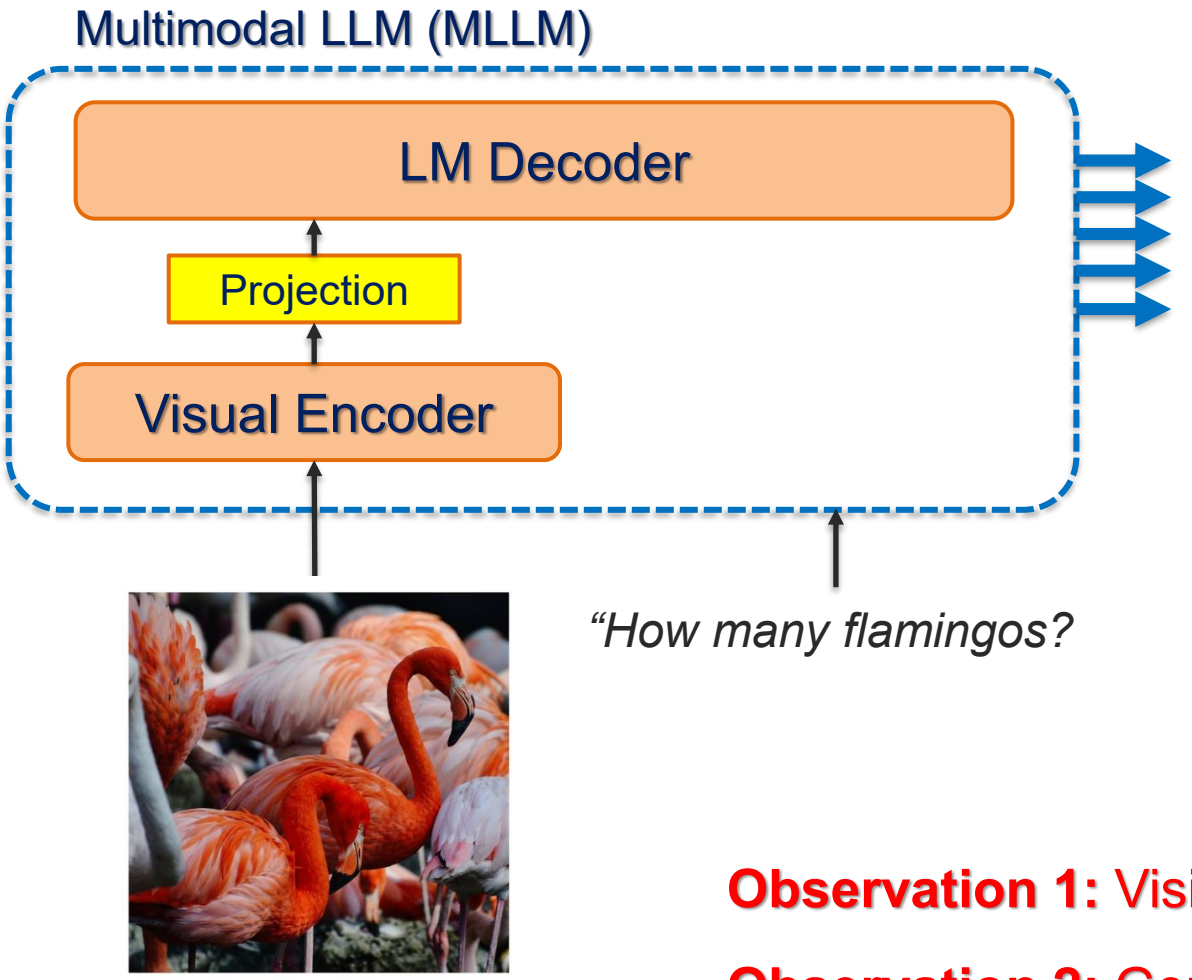


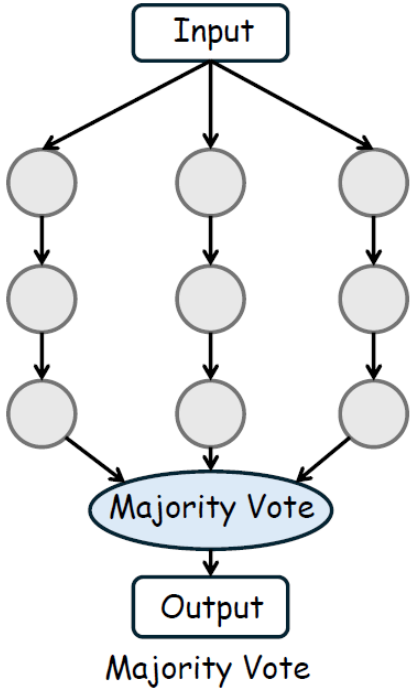
<think>
First, what is the person doing?
They are being pulled behind a boat on water.
-> 3a decomposition
The board is short and wide, with bindings for both feet.
-> 3b intermediate concepts (in text)
Boards used behind boats with foot bindings are wakeboards.
-> 3c knowledge, recalled not retrieved
</think>
<answer> A wakeboard **</answer>**

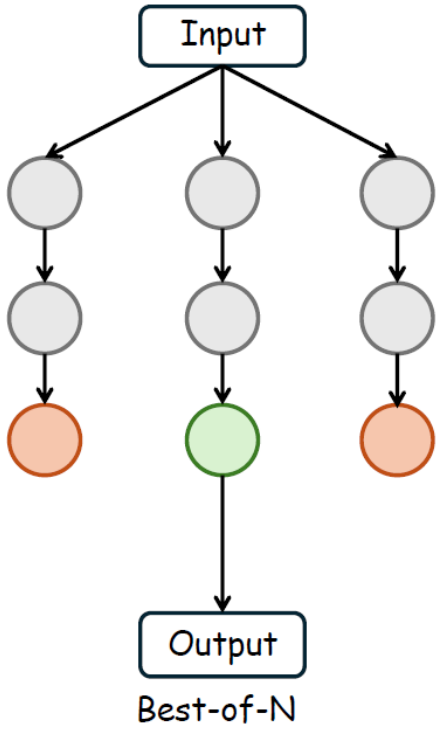
Traces are flexible, but nothing guarantees they're right

Andreas et al., 2016

Multiple CoT Reasoning Traces (Test-time Scaling)







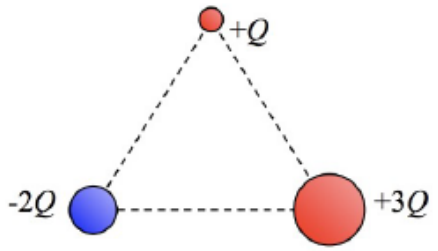
Observation 1: Vision encoder runs once, deterministically.

Observation 2: Generator-verifier gap can be small (hard to verify)

Wan et al., 2026

EMMA: Can MLLMs Reason in Multimodality?

Multimodal reasoning question in EMMA



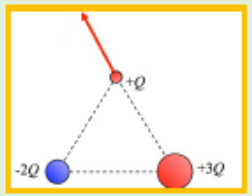
Question: Three point charges, of charge $+Q$, $-2Q$, and $+3Q$, are placed equidistant as shown. Which vector best describes the net direction of the electric force acting on the $+Q$ charge?

a. b. c. d. e.

Human reasoning

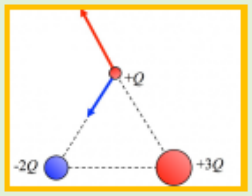
The direction of the electric force due to $-2Q$ and $+3Q$ can be visualized with a quick free-body sketch of the electric forces:

1



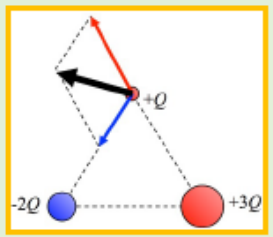
repel

2



attract

3



Combining the forces:
∵ |+3Q|>|-2Q|
∴ repel > attract

b.

Model reasoning (GPT-4o)

1. **Forces Due to the Charges:**
   - The $-2Q$ charge attracts the $+Q$ charge. (Force points downward-left).
   - The $+3Q$ charge repels the $+Q$ charge. (Force points downward-right).
2. **Force Magnitudes (Coulomb's Law):**
   - Force due to $-2Q$: $F_{-2Q} = \frac{2kQ^2}{r^2}$.
   - Force due to $+3Q$: $F_{+3Q} = \frac{3kQ^2}{r^2}$.

Error: While the model understands that like charges repel, it incorrectly identifies the force from +3Q to +Q as downward-right ↘, when it should be upward-left ↖.

Hao et al., Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark, 2025

EMMA: Can MLLMs Reason in Multimodality?

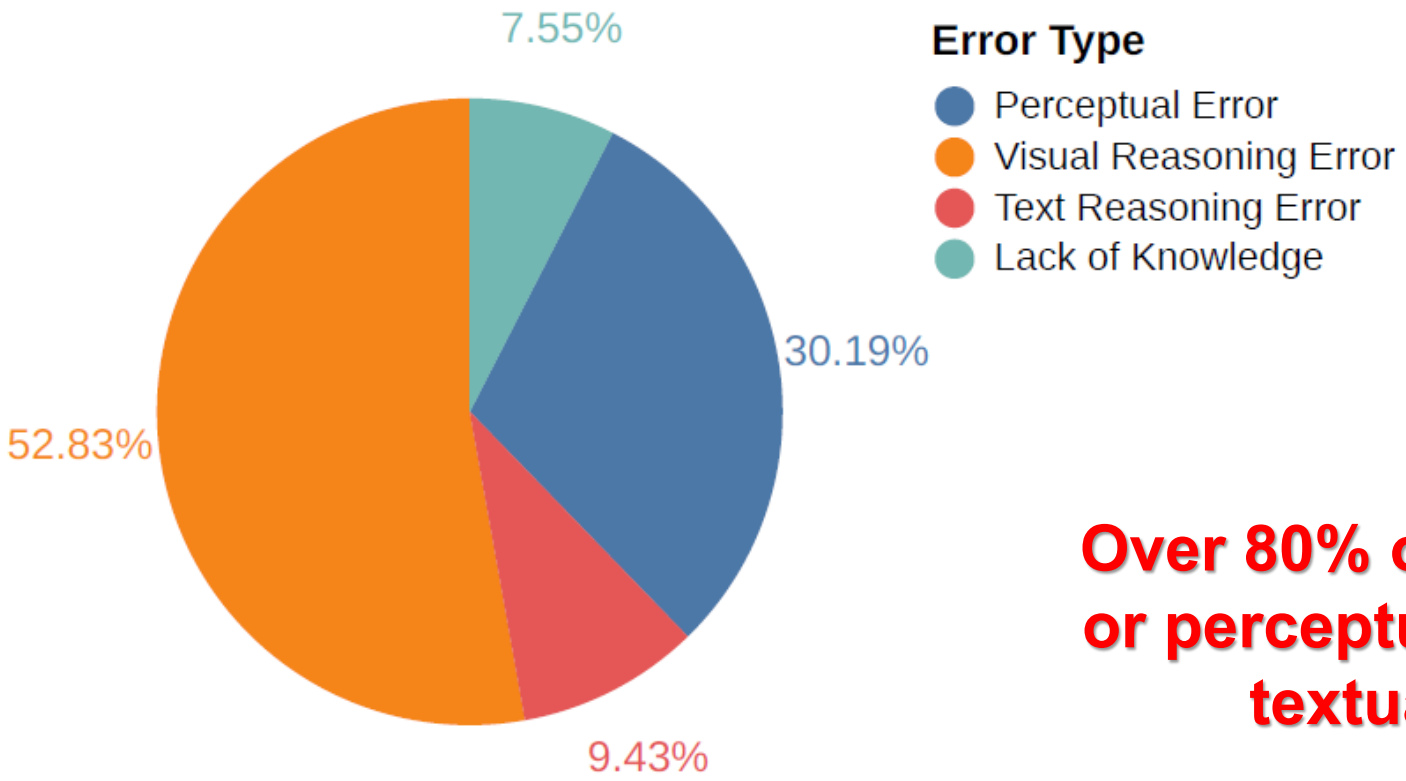
Model	Method	N=1	N=2	N=4	N=8	N=16
GPT-4o	Majority Voting	36.00	–	37.25	36.25	38.25
	BoN (self)		35.50	35.75	36.75	–
	BoN		40.75	36.25	36.5	–
	Tournament		40.75	39.25	**41.25**	35.25
	Pass@N		45.00	53.25	65.75	74.00
Gemini 2.0 Flash	Majority Voting	36.25	–	37.75	39.25	39.75
	BoN (self)		38.25	36.50	36.00	–
	BoN		36.75	37.00	40.25	–
	Tournament		36.75	37.25	**40.75**	38.75
	Pass@N		45.25	56.25	64.50	75.00

Observation 1: majority voting is *mostly flat*. Most samples inherit the same misperception and converge on the same wrong answer.

Observation 2: The gap between Pass@N and the other rows is the verifier gap. The correct answer is in the set of traces, but not selected.

Hao et al., Can MLLMs Reason in Multimodality? EMMA: An Enhanced MultiModal ReAsoning Benchmark, 2025

EMMA: Can MLLMs Reason in Multimodality?



Over 80% of errors are visual or perceptual; under 10% are textual reasoning.

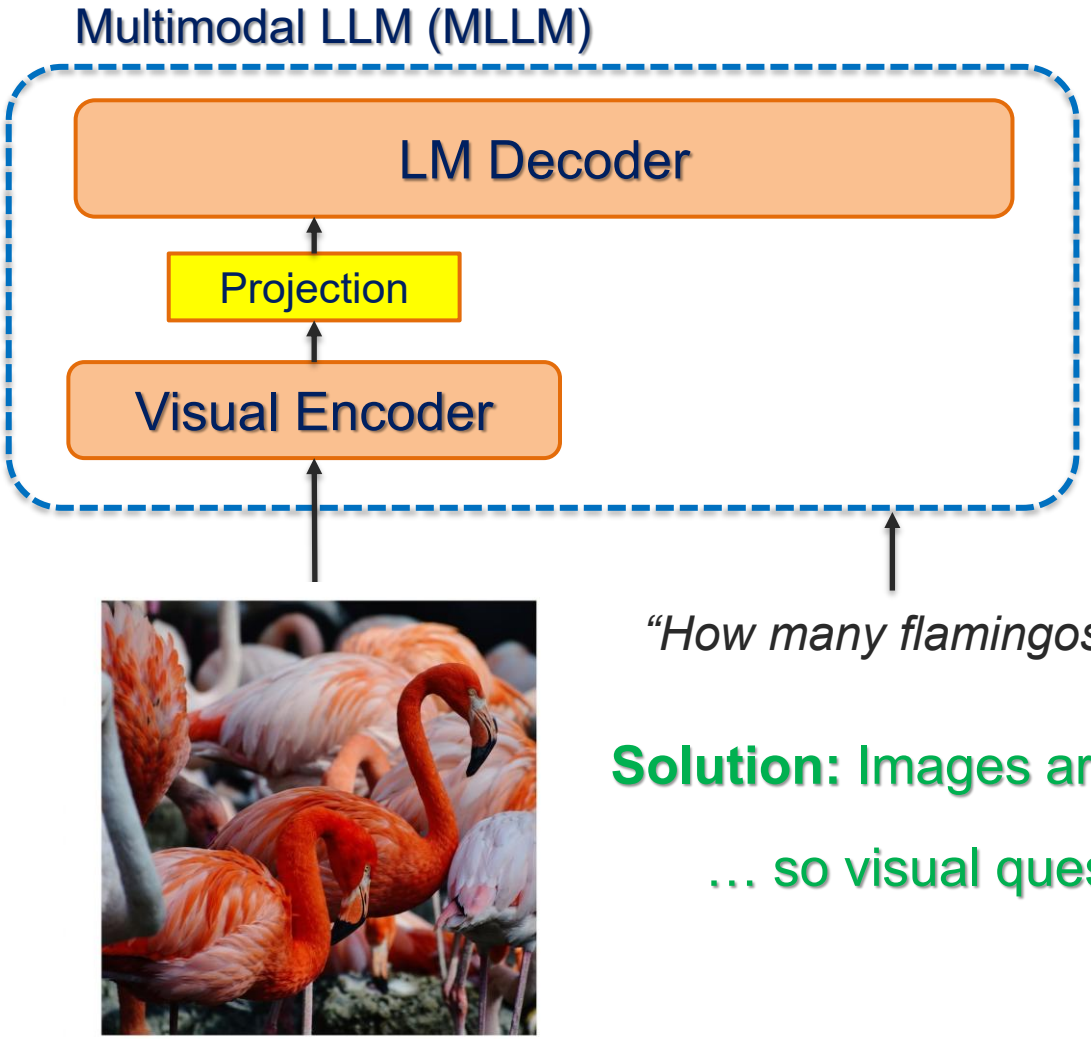
Two Strategies to Improve Multimodal Reasoning

Improving Visual perception

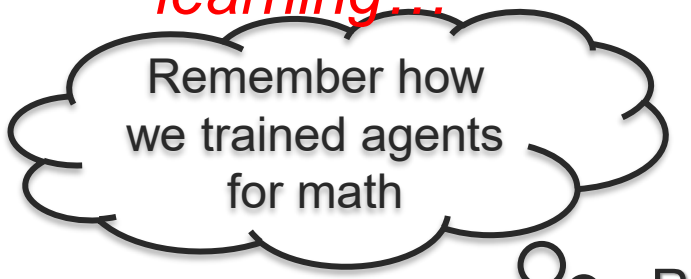
Active Perception

Improving Visual Perception in Multimodal Reasoning

Improving Visual Perception



Challenge: If we *reward the answer, then maybe only the LM decoder ends up learning…*

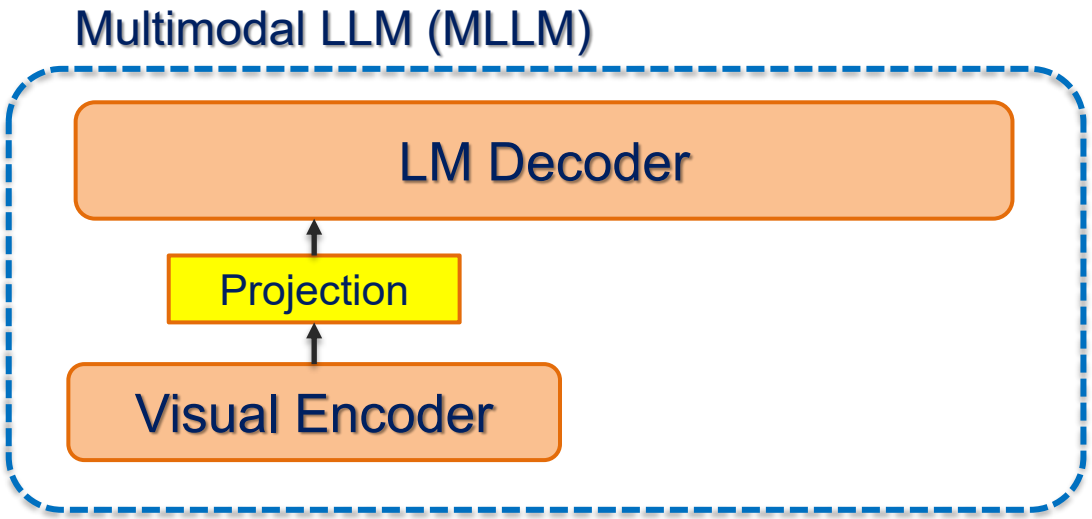


Rule-based reward functions

Solution: Images are grounded in the real world…

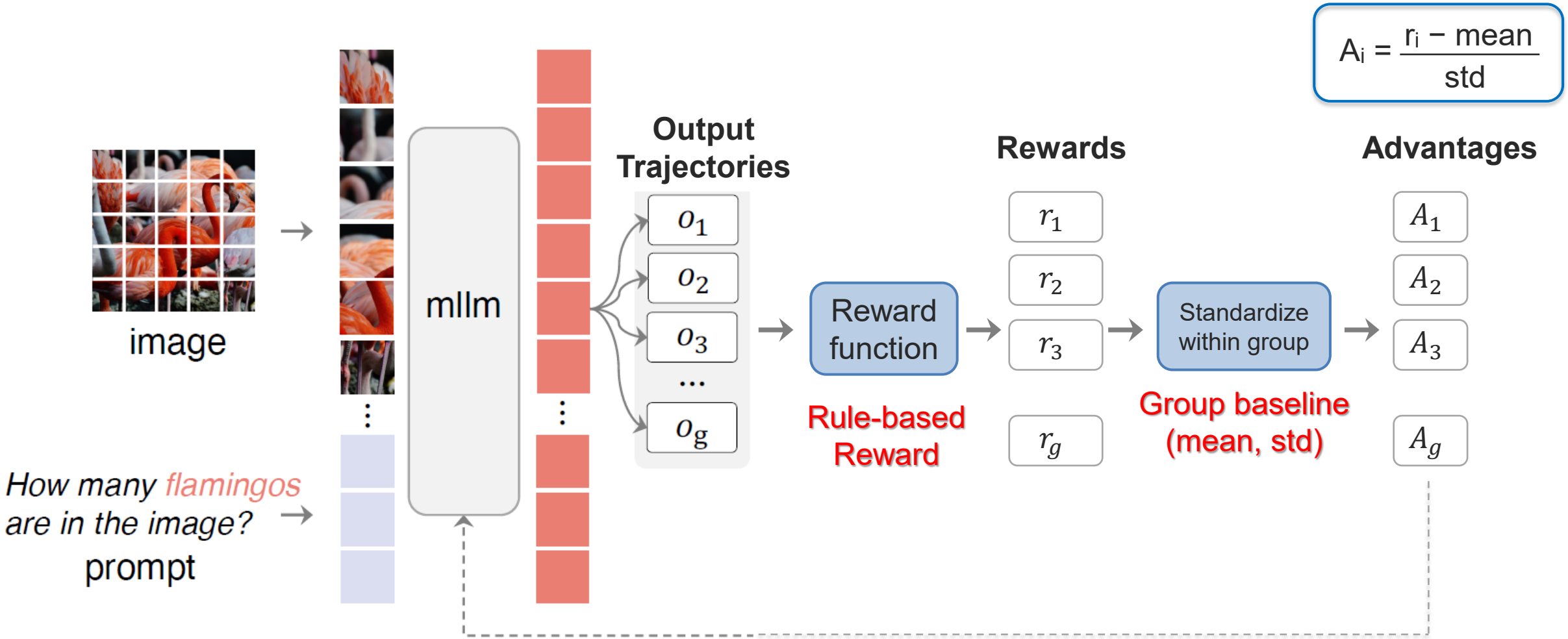
… so visual questions often have a concrete ground truth

RL Vocabulary with MLLM



RL term	In an MLLM	Example: counting
Policy π_θ	the MLLM itself	Model answering
Action a_t	next token	“8”
State s_t	prompt + image + tokens so far	
Trajectory	the full response	reasoning + answer (“8”)
Reward r	score at the end	digit match → 0/1

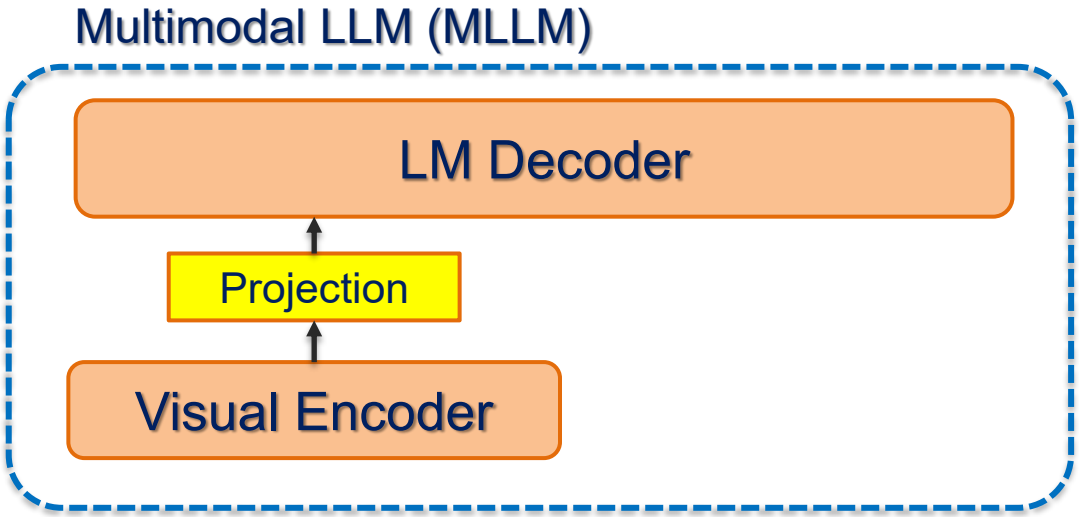
GRPO with Multimodal LLM



NOTE: if all trajectories get the same reward, then std → 0, no gradient

Shao et al., *DeepSeekMath*, 2024

RL with Verifiable Rewards (RLVR)



Verifiable (rule-based) rewards

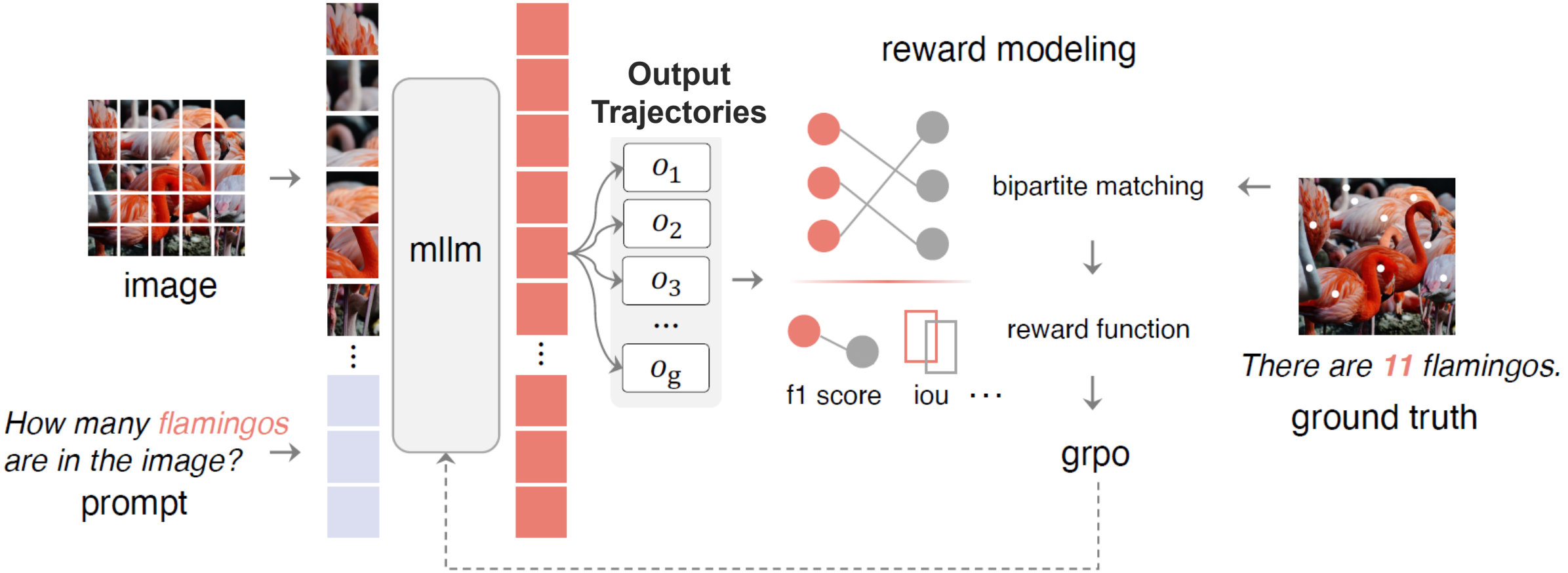
Domain	Verifier
Counting	Digit match
OCR	Edit distance
Grounding	Box overlap
Detection	Match boxes, then overlap

In contrast to RLVR…

RLHF: Learns a reward model based on preference annotations. Needed when nothing can verify the answer

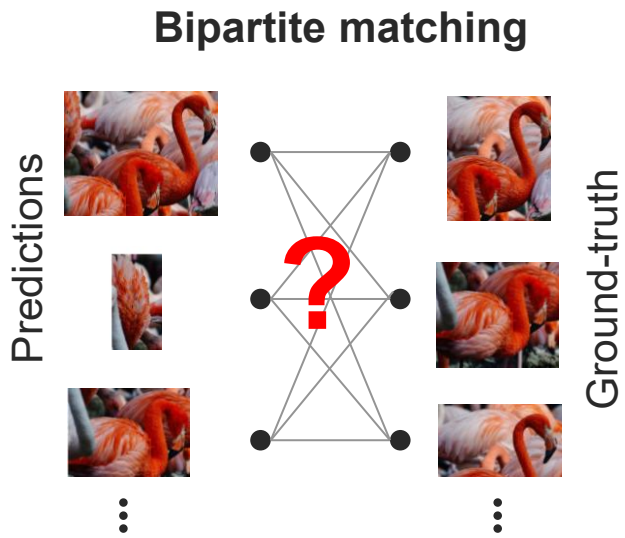
Lambert et al., Tülu 3
DeepSeek-AI, DeepSeek-R1

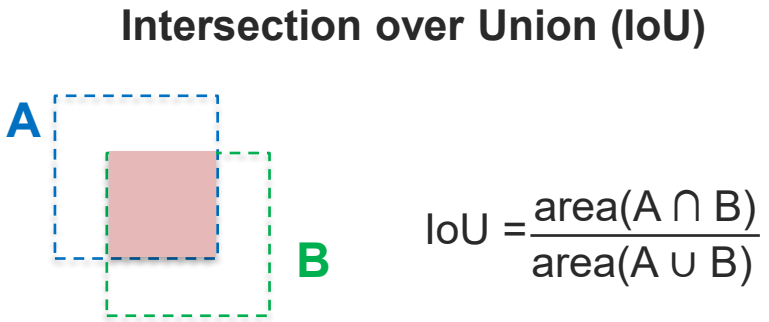
Perception-R1: GRPO + RLVR for MLLM

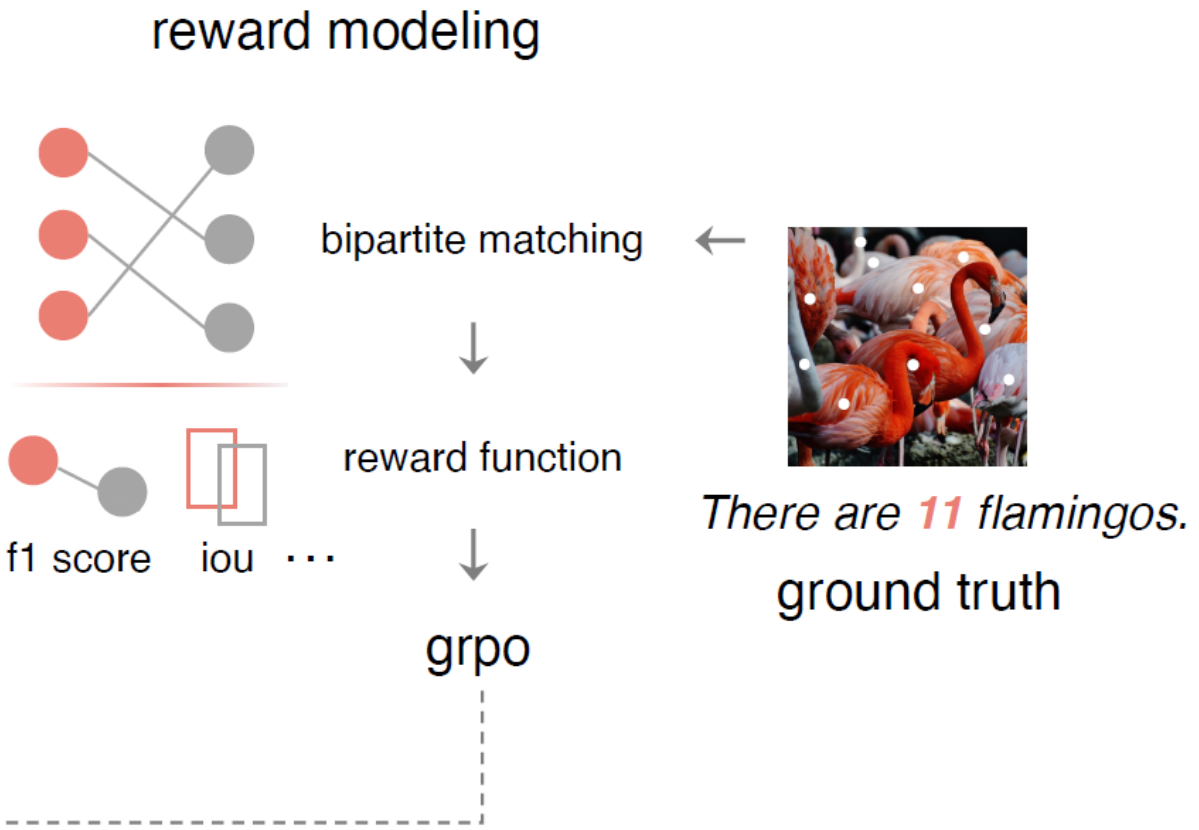


Yu et al., Perception-R1: Pioneering Perception Policy with Reinforcement Learning, 2025

Perception-R1: Matching Prediction Boxes







Yu et al., Perception-R1: Pioneering Perception Policy with Reinforcement Learning, 2025

Perception-R1: A Surprising Result…

case	Visual Grounding			OCR	Visual Counting		Detection
	RefCOCO	RefCOCO+	RefCOCOg	PageOCR	$\text{Pixmo}_{\text{val}}$	$\text{Pixmo}_{\text{test}}$	COCO2017
Perception-R1	89.1	81.7	85.7	**98.4**	78.1	75.6	31.9
w/o reward matching	-	-	-	-	77.1	75.4	23.5
w/o RL	86.8	77.1	83.3	94.4	60.2	50.5	16.1
w thinking	75.1	67.9	71.3	77.3	74.9	72.8	25.7
w/o thinking	**89.1**	**81.7**	**85.7**	95.7	**78.1**	**75.6**	28.1

Surprising result: Explicit thinking processes prove non-essential for contemporary visual perception.

Yu et al., Perception-R1: Pioneering Perception Policy with Reinforcement Learning, 2025

Active Perception

Active Perception

"They Did Not Expect Him", painting by Ilya Repin, 1884

Yarbus, Eye Movements and Vision, Plenum Press, 1967

Where to look?

What to look at first?

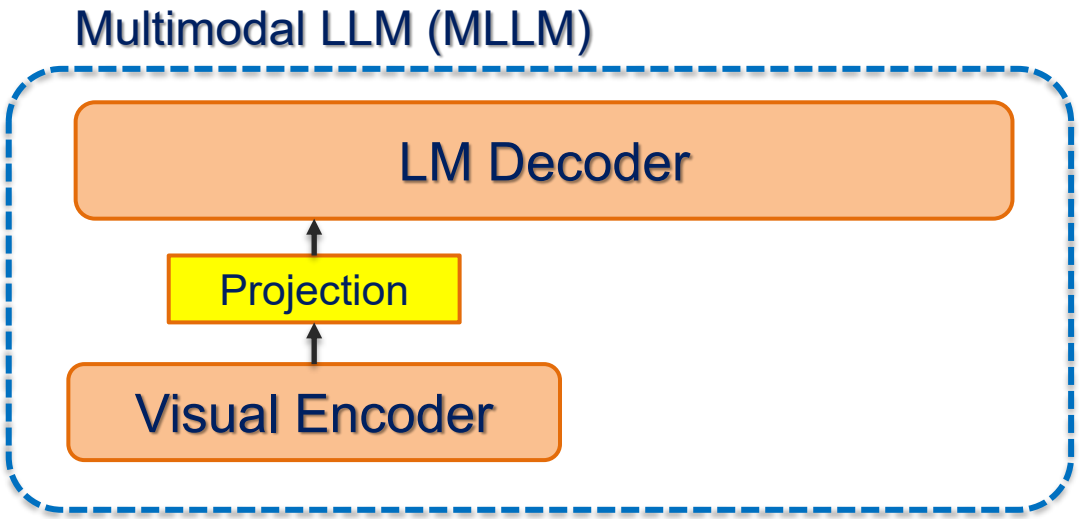
What to remember?

Also important in robotics…

Active observer: one "engaged in some kind of activity whose purpose is to control the geometric parameters of the sensory apparatus."

— Aloimonos, Weiss & Bandyopadhyay, *Active Vision*, IJCV 1988

Active Perception with MLLMs



Challenge: MLLM's visual evidence comes from a one-shot, query-independent encoder

Solution: Allow the MLLMs to be active in its perception of the world (active agent)

Goal	Action	Domain
Where	crop, zoom	Spatial (e.g., visual search)
When	frame selection	Temporal (e.g., long videos)
What if	generate alternative hypotheses	Epistemic (e.g., causal inference)

Guided Visual Search

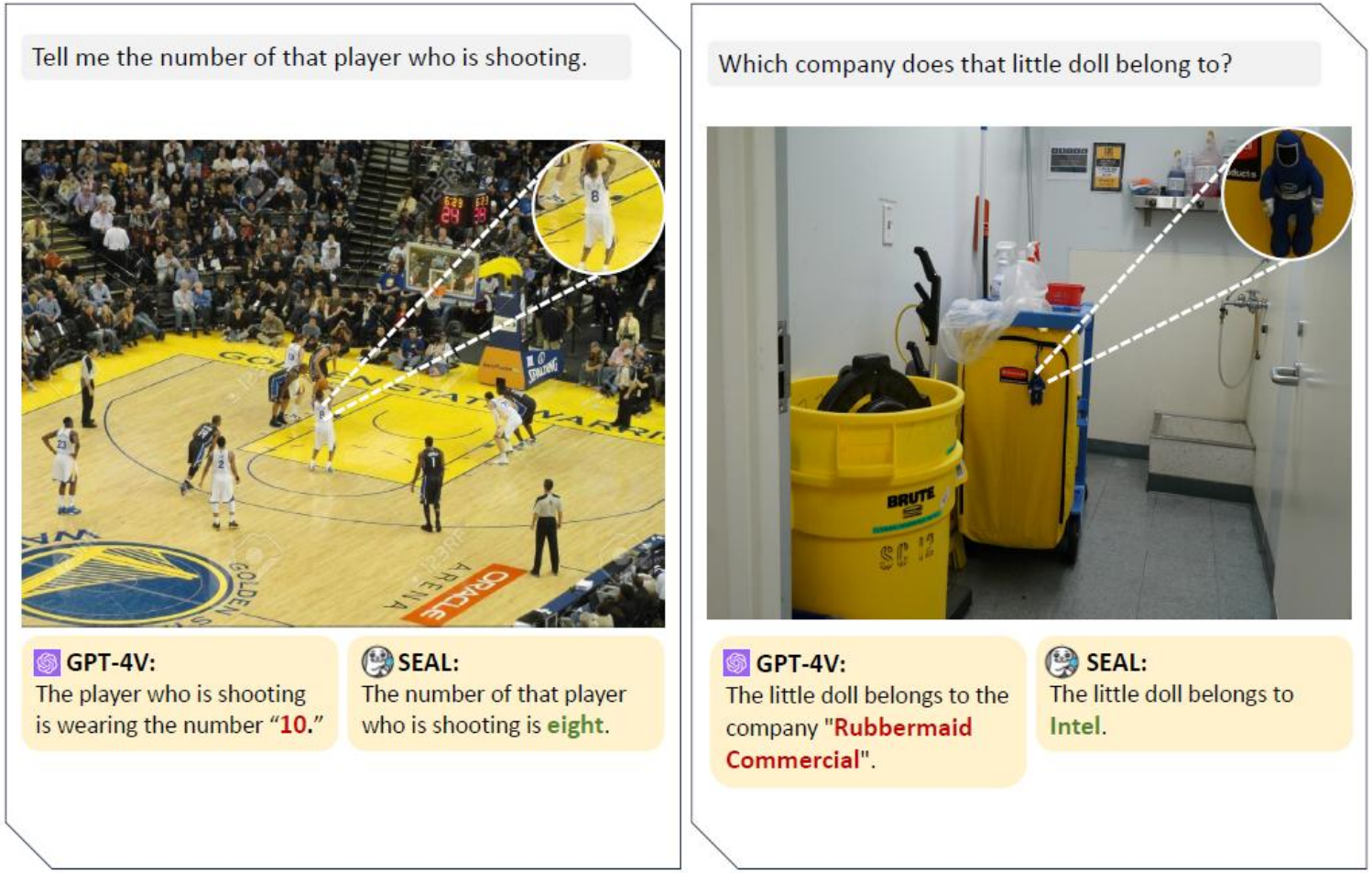
Where to look?

What to look at first?

What to remember?

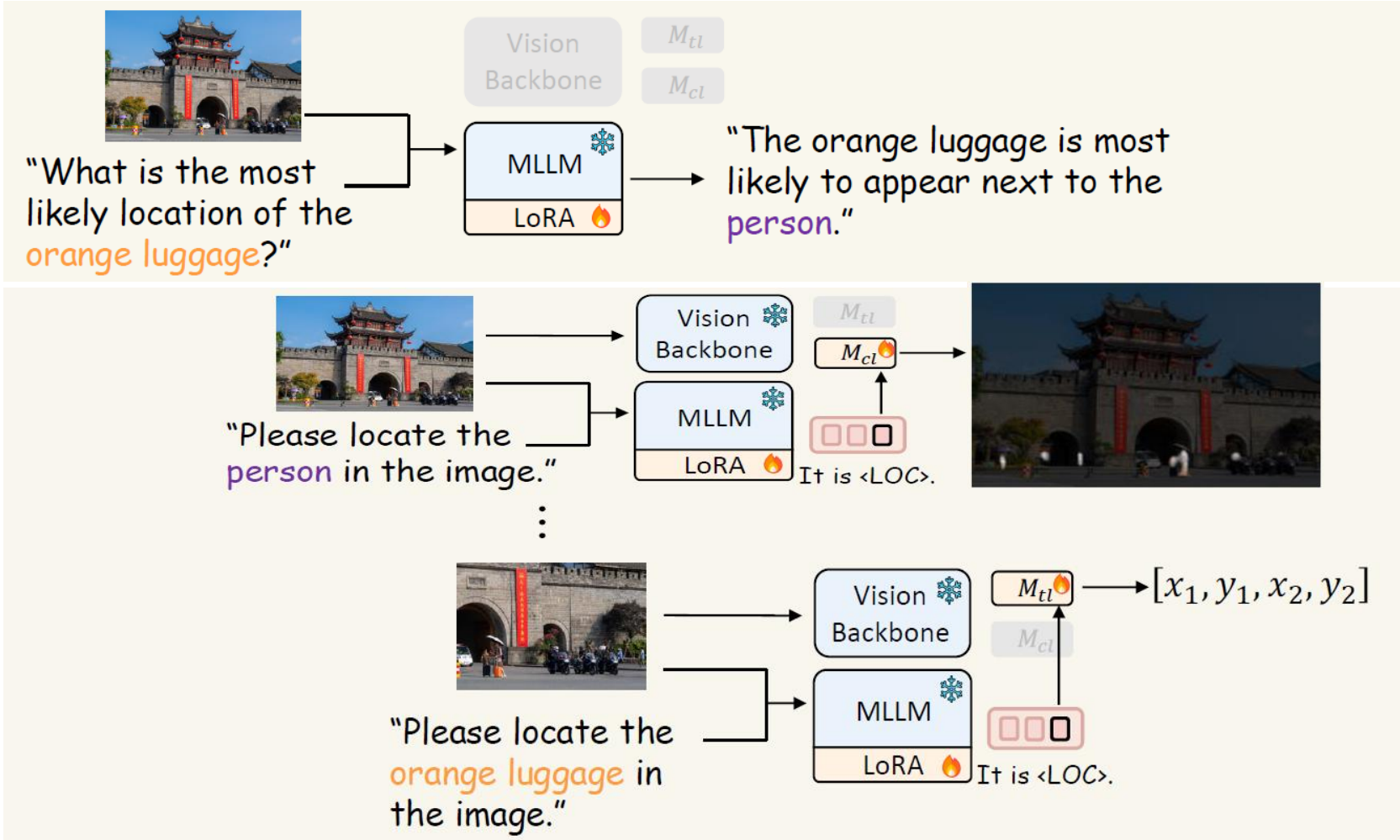
"They Did Not Expect Him", painting by Ilya Repin, 1884

V* Bench dataset



Wu & Xie, V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs, CVPR 2024

V* Visual Search Algorithm

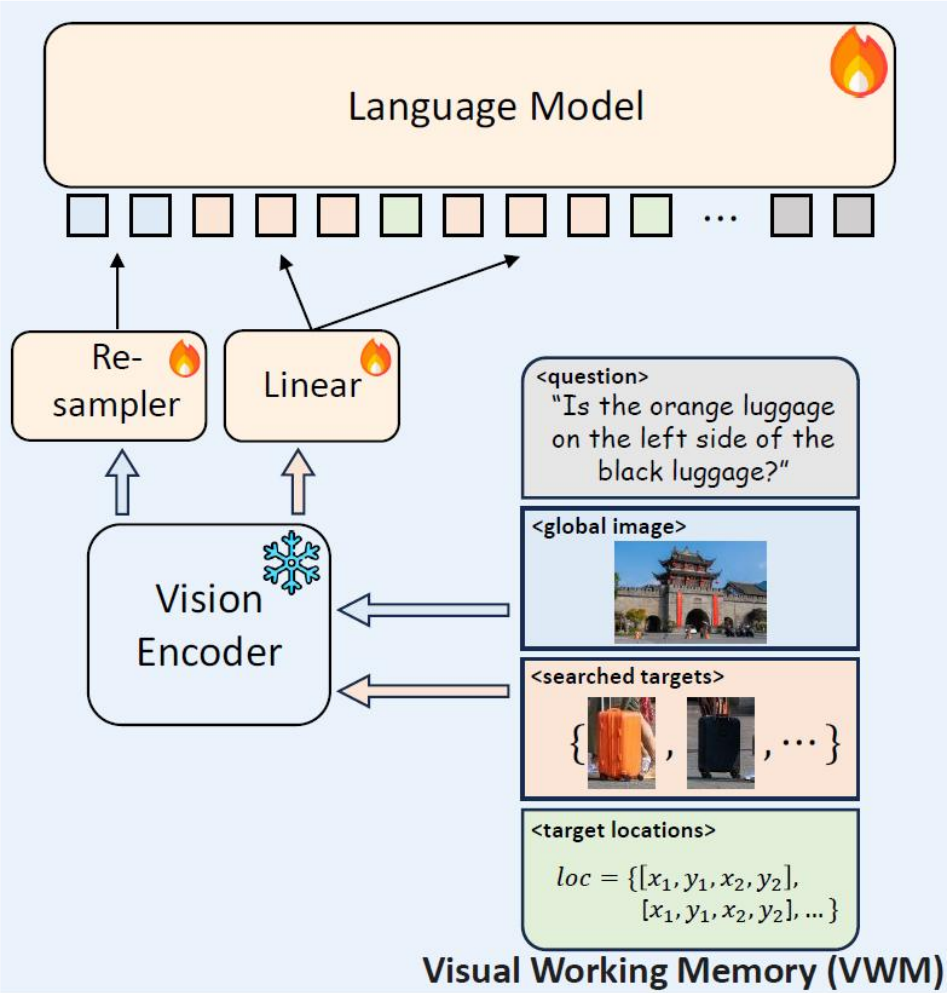


The search prior is semantic, not visual.

Zoomed on the people in the image

Wu & Xie, V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs, CVPR 2024

Results Stored in Visual Working Memory



Where? the spatial domain

Active Perception: the model chooses where in the image to spend its resolution.

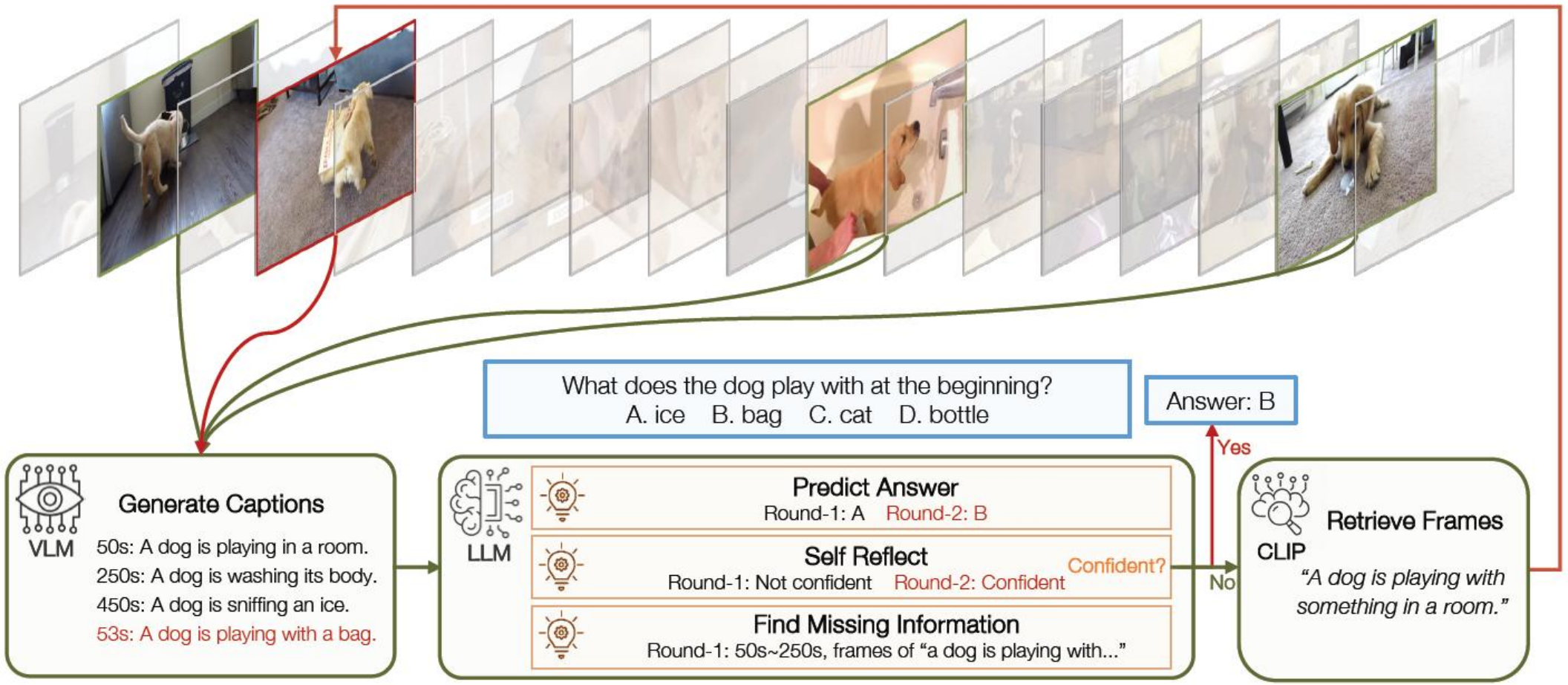
Designed Policy

While the components are learned, the policy is not.

*Researchers designed the rules (e.g., confidence threshold) about **when** to search again.*

Wu & Xie, V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs, CVPR 2024

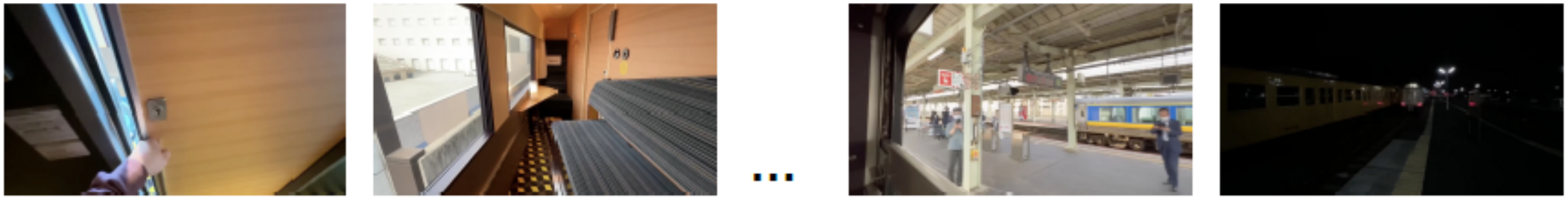
VideoAgent: Frame selection for Video Reasoning



Wang et al., VideoAgent: Long-form Video Understanding with Large Language Model as Agent, 2024

Act2See: Retrieval + Generation

Round 1 Instruction: Provide a step-by-step … (full prompt in Table 1). Question: What is directly to the left of the black phone booth at the station where the train stops in the evening? Choices: 0) …



Gemini 2.5 Pro

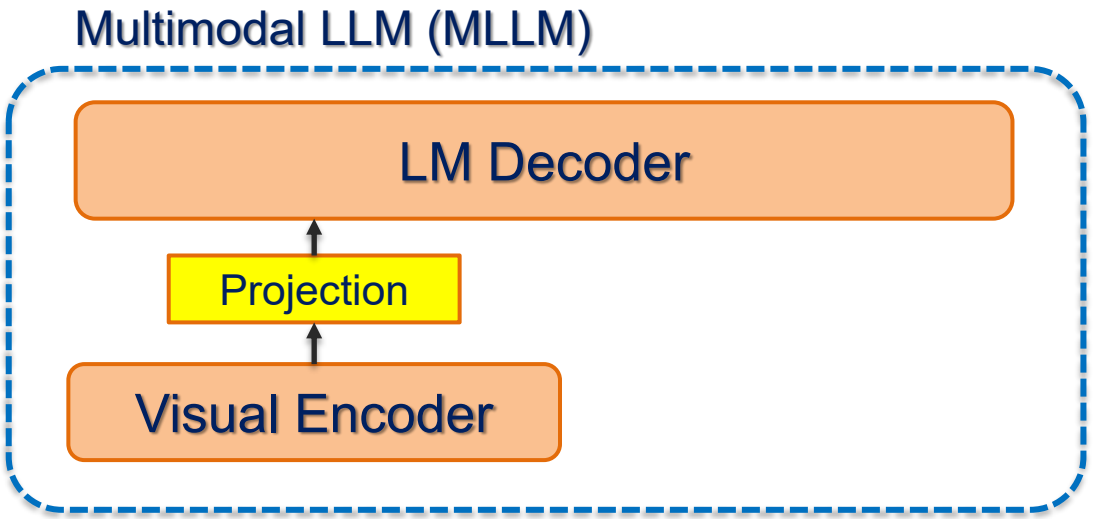
Round 1 Output: ... *<think>...*First, I need to find the scene where the train stops at a station in the evening…. ***<retrieve>*** train stopped at station platform in the evening with black phone booth ***</retrieve>*** *… (the rest after </retrieve> are discarded)*

Ma et al., Act2See: Emergent Active Visual Perception for Video Reasoning, 2025

Act2See: Retrieval + Generation

Ma et al., Act2See: Emergent Active Visual Perception for Video Reasoning, 2025

From Designed to Learned Policy



Same RL, new environment…

Something outside the model executes the action and writes into the context

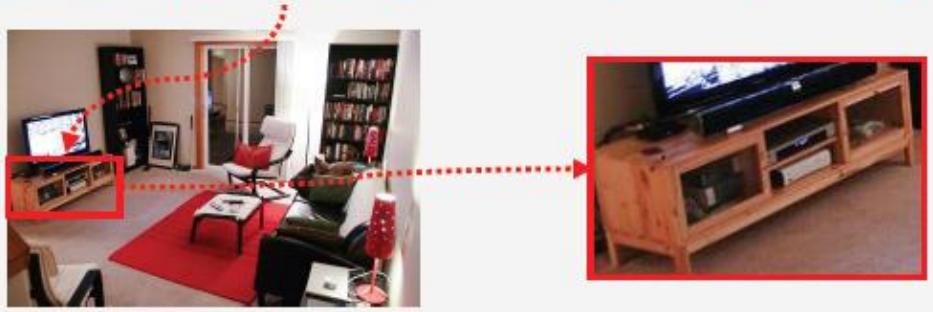
RL term	In an MLLM	with Actions
Policy π_θ	the MLLM itself	*unchanged*
Action a_t	next token	*unchanged* – "crop(…)" is just tokens
State s_t	prompt + image + tokens so far	**+ observations inserted by the environment**
Trajectory	the full response	**multi-turn: generate → observe → generate**
Reward r	score at the end	*unchanged*

Interleaved Multi-modal Chain-of-Thought

[Begin of iMCoT] *The image shows a living room ... To the right side of the image, there is a laptop on a black surface ... There is no visible clock near the laptop or any part of the room that looks like a clock. Therefore, we cannot confirm the presence of a clock to the left of the laptop.*

Since the image alone does not provide sufficient information to answer the question definitively, we need to identify the potential area where the clock might be, if it exists ... I will generate the ROI for this area.

{"bbox_2d": [2, 189, 171, 298], "label": "area potentially containing the clock"}



The zoomed-in image shows part of a TV stand with a CRT television ... there is no visible clock in the provided zoomed-in region, and the user's question still cannot be answered definitively based on this image.

Given this, let's re-evaluate the area to see if we can detect any other potential clock candidates. We may need to adjust or re-examine the area.

{"bbox_2d": [145, 18, 226, 220], "label": " region of the bookshelf "}

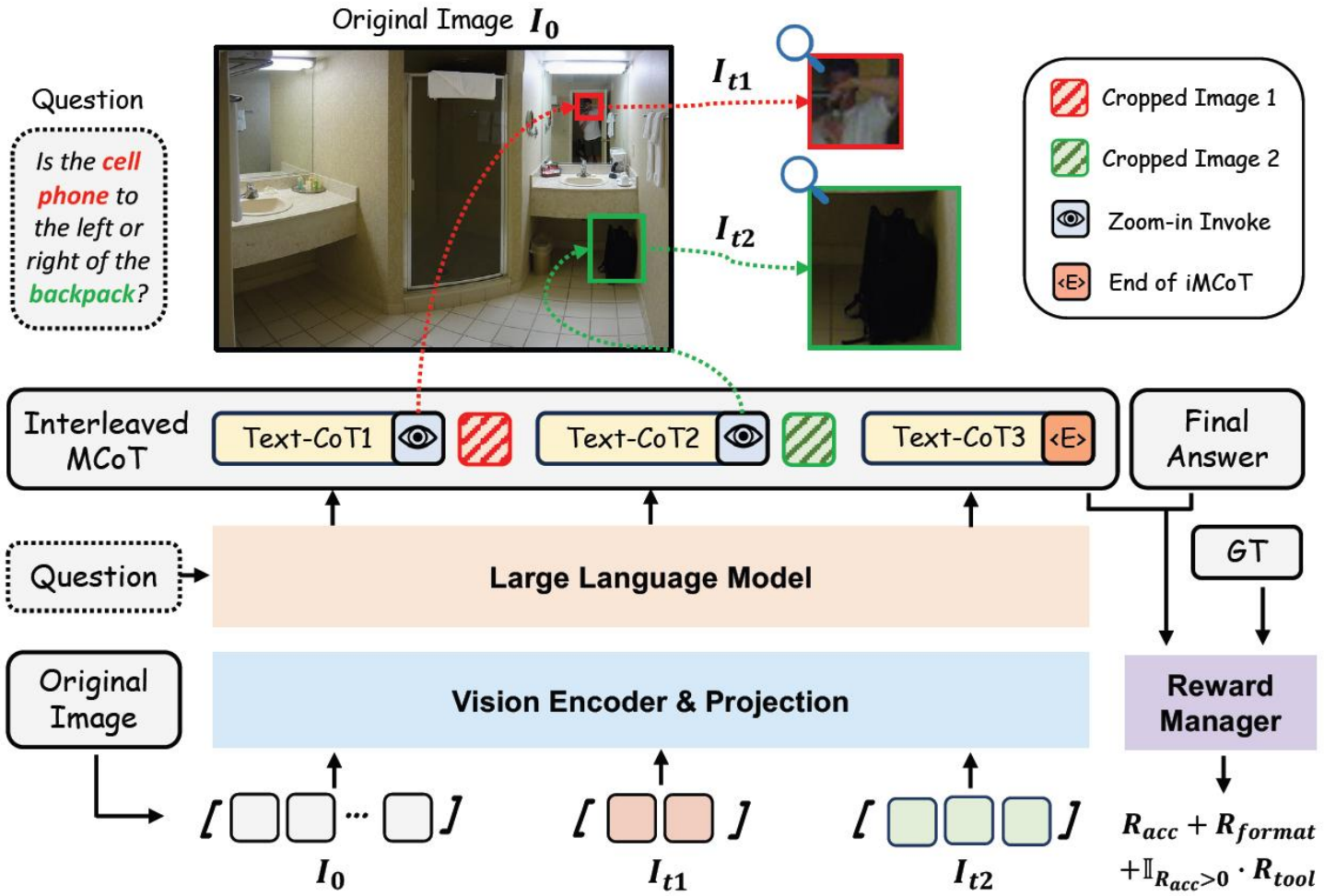
The zoomed-in image now shows a bookshelf next to a TV. On the shelf, there is a round object that appears to have the shape and design of a clock, which makes it a likely candidate being referenced in the original question.

Since we have identified the clock's location in the current zoom-in, this satisfies the condition. ***[End of iMCoT]***

<answer> Yes, the clock is to the left of the laptop. </answer>

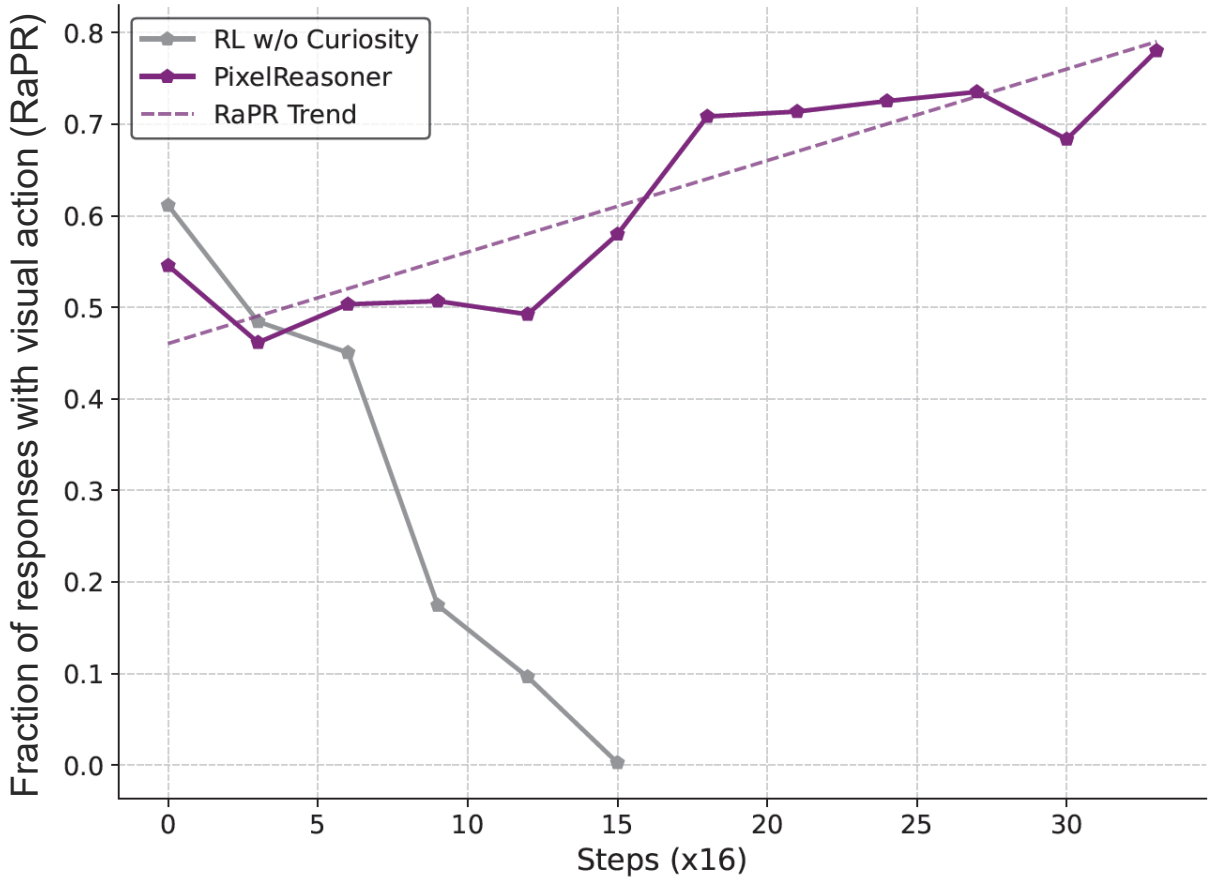
Zheng et al., DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning, 2025

Interleaved Multi-modal Chain-of-Thought



Zheng et al., DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning, 2025

Curiosity-Driven RL



The Learning Trap

Standard RL: the policy learns to **bypass** visual actions in reasoning.

Two Solutions:

Warm-Start Instruction Tuning **lays the foundation** for effectively using visual operations.

Curiosity-Driven RL encourages the **exploration and practice** of visual actions in reasoning

Wang et al., Pixel Reasoner: Incentivizing Pixel-Space Reasoning with Curiosity-Driven Reinforcement Learning, 2025