CSCE 689: Special Topics in Advanced NLP

Lecture 18: Multimodal Reasoning

Kuan-Hao Huang

Fall 2026



Course Project – Midterm Report

- Due: 11/5
- Page limit: 4 pages (excluding reference)
- Format: ACL style
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and expected contributions
  - Evaluation metrics
  - Planned implementation details, including dataset, models, codebases, etc.
  - Current progress (preliminary results, baseline results, etc.)
  - Next steps and expected timeline

Final Presentation

- Each team has 7 minutes for presentation
  - You have to stop once you reach 7 minutes
- The presentation should include
  - The topic you choose
  - Novelty/challenges
  - Your approach
  - Experiments/evaluation
  - Results, findings, insights
- Clarity is important
  - Teach your classmate about your topic
- Time management is also important

Presentation Order

- Final presentation is online

11/30

1. Faaiz Umer , Oseluolemen Inotu
2. Yiming Xiao, Shuning Gu
3. Mingyang Wu, Siyuan Yang
4. Chedi Morchdi
5. Jason Kim, Andrew Vong
6. Akshay Hitesh Shah, Saisugash Senthil
7. Christopher Kim, Yihong Yang
8. Ali Taheri, Damandeep Singh
9. Amir Suhail, Fredy Cortez

12/2

1. Vaishnavi Poti, Ed Stone
2. Taeyoon Kwon, Ming Yan
3. Karan Pramod Sharma, Mrunalini Thamankar
4. Han-Ju Chen
5. Lawrence Wong, Omer Burak Cinar
6. Raphael Zhu, Ruei-Chi Chen
7. Chenghao Qiu
8. Prithvi Patel

Invited Talks

- **Date:** 11/23 online @ Zoom
  - https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1
  - **Talk:** TBD
  - **Speaker:** Da Yin, Founding Member of Technical Staff at NeoCognition

Sparks of Weight-Space Programming:

LLMs can natively read & write neural network weights

READ

Which game was this neural network trained on?

```
W1 row 3 -0.080 0.255 -0.015
W1 row 8 0.127 0.193 0.018 -
W1 row 62 0.531 0.897 0.219 -
W1 row 66 -0.632 0.718 0.238
W1 row 69 -0.199 0.777 -0.881
W1 row 111 -0.608 -0.837 -0.0
W1 row 157 -4.811 0.022 -0.8
```

Astra reads the 17,253 weights, with no tools

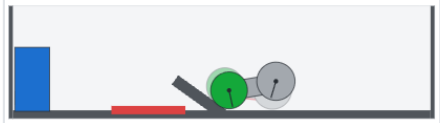


Breakout

99% confident

FIX

A trained policy never gets the car over the strip.

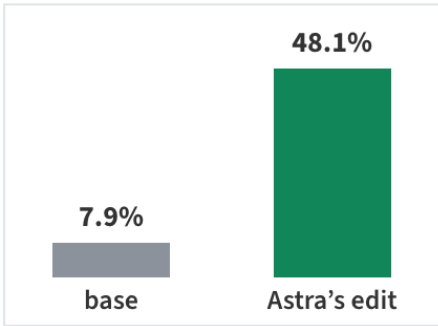


Astra edits 3.5% of the weights, in one shot

100% success, up from 0%

IMPROVE

Make Qwen3.5-0.8B better at math.



Astra edits the weights in a Python kernel, with no training

48.1%

on MATH-500, up from 7.9%

Breakout Pong Pac-Man Freeway

Setup



Gameplay

the policy's 17,253 weights as text

```
W1 row 3  -0.080 0.255 -0.015 -0.141 0.010 -0.170 -0.114 0.160 -0.056 0.088
```

Which game was this neural network trained on?

No tools are allowed.

Show the full prompt ▸

Answer **Game: Breakout** 99% confident · ✓ **Correct** · 846 tokens

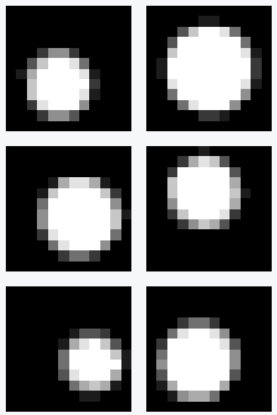
Reasoning

- **Channel 0 — agent:** Strong weights are concentrated on **row 8**, across the playable columns
- At **(8,1)**, weights include **4.066** to hidden unit 15 and **−4.013** to unit 31
- This suggests a **bottom-mounted paddle moving horizontally**
- **Channel 1 — moving objects:** Large, structured weights cover the interior above the paddle
- **Channel 2 — reward targets:** Learned structure is concentrated in **rows 1–3**

Read full reasoning ▸

Circles Squares Triangles

Setup



The model generates

the model's 23,824 weights as text

```
W3 row 29  0.004 0.003 0.003 -0.001 0.007 0.062 0.114 0.082 0.024 0.005 0.00
```

What images does this neural network generate?

No tools or forward passes are allowed.

Show the full prompt ▸

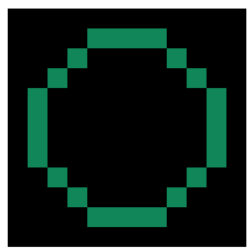
Answer **Images: a single white circle outline on black, with a black interior**

84% confident · ✓ **Correct** · 506 tokens

Reasoning

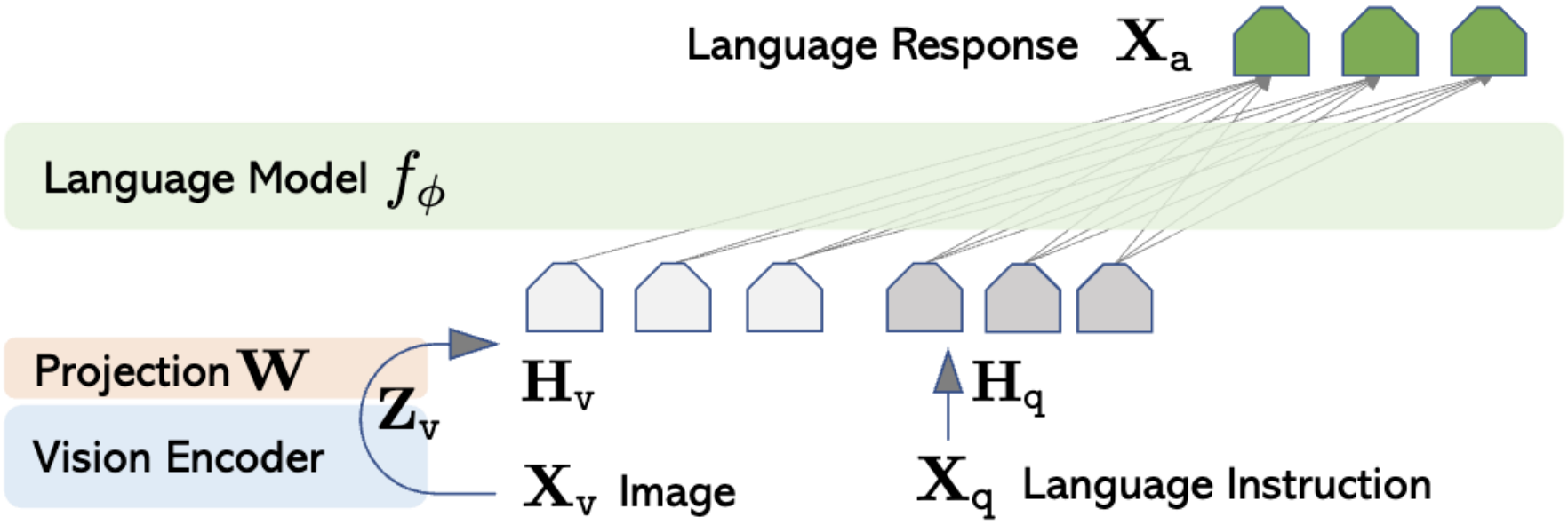
- The output weights contain **curved rim segments**
- This traces an upper-left bend
- Other units describe complementary portions of a rounded boundary
- These are localized bands with weaker or negative weights beside them, rather than uniformly positive filled regions

Read full reasoning ▸



Astra's drawing

Multimodal Large Language Models



Chllanges for Multimodal Reasoning

- Only read image once
  - Query independent
  - No revisit
  - No focus
- Language prior

Mitigating Visual Forgetting via Take-along Visual Conditioning for Multi-modal Long CoT Reasoning

Hai-Long Sun[1,2] *, **Zhun Sun**[4,5] *, **Houwen Peng**[3], **Han-Jia Ye**[1,2] †
[1]School of Artificial Intelligence, Nanjing University
[2]National Key Laboratory for Novel Software Technology, Nanjing University
[3]Tencent [4]Center for Language AI Research, Tohoku University
[5]RIKEN Center for Advanced Intelligence Project

Do Models Look at Images?

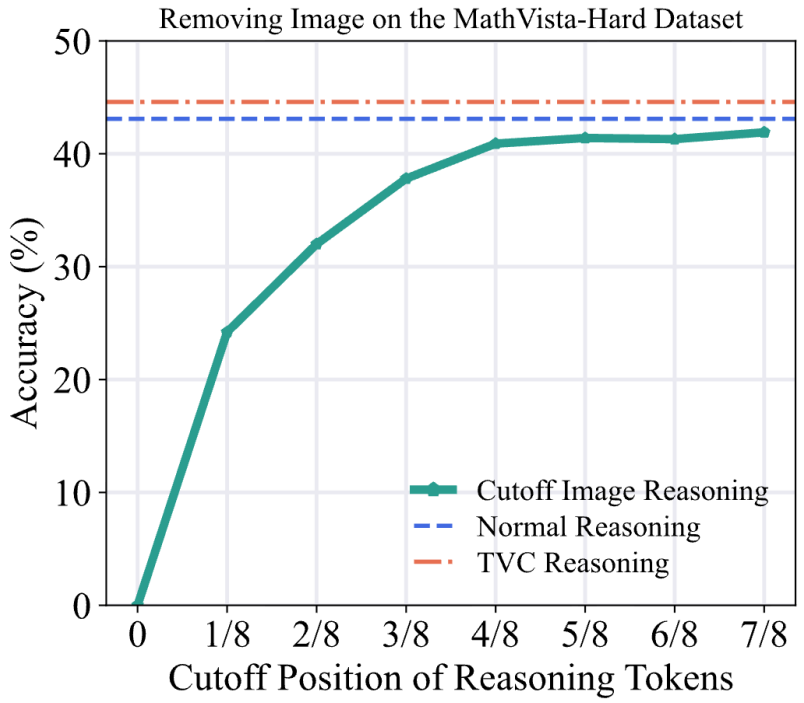
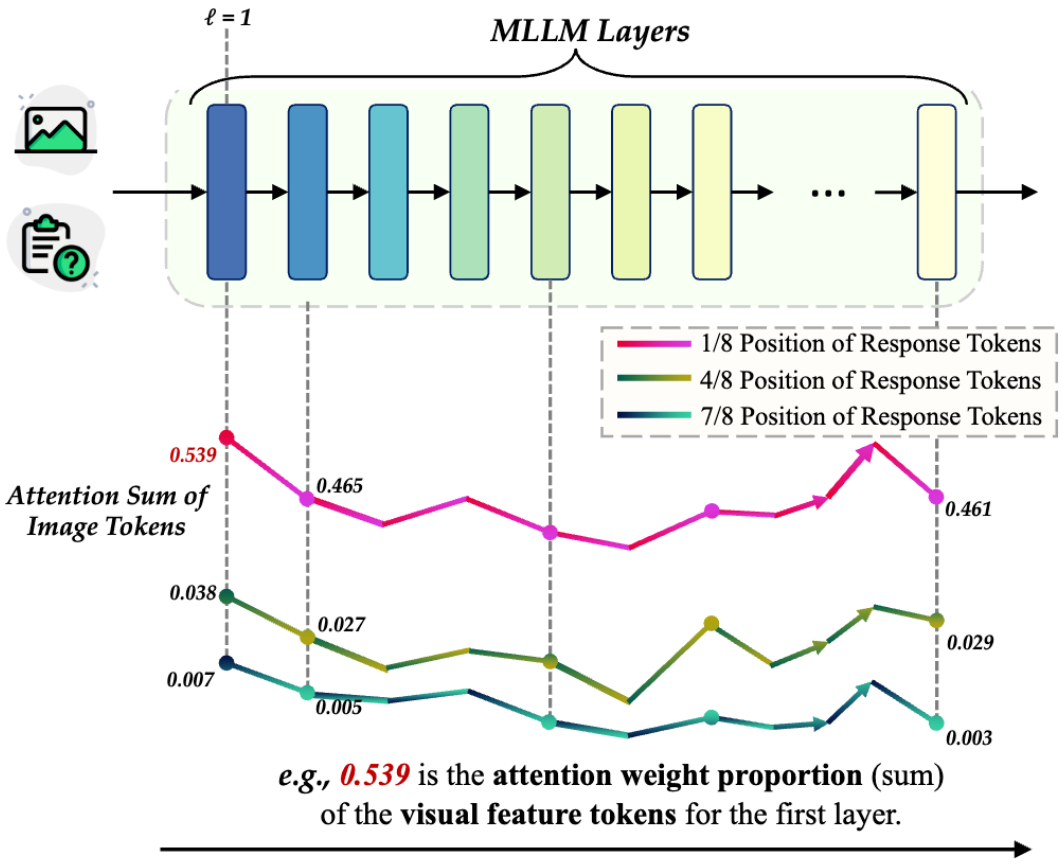
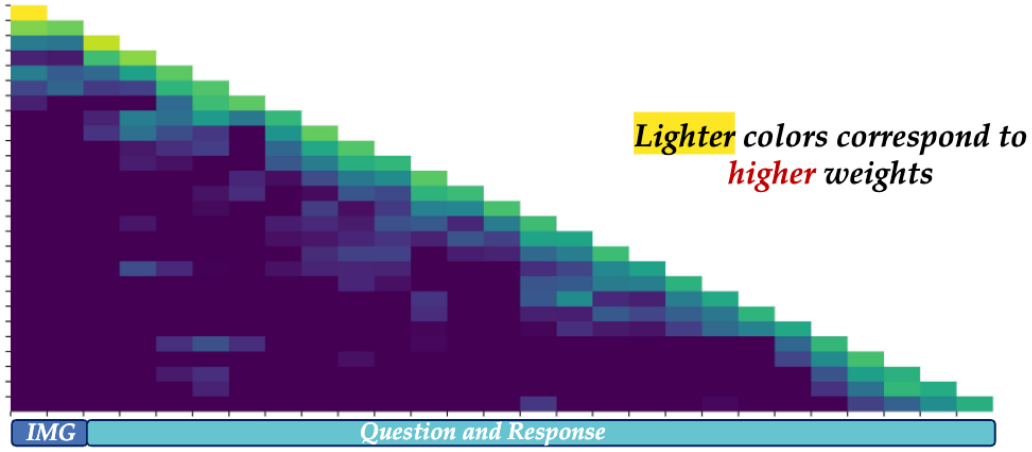


Figure 1: **The visual forgetting phenomenon by removing the image at different reasoning stages.** It shows that by the midpoint of the reasoning process, the model becomes less dependent on the image, causing text-over-relied outputs.

Do Models Look at Images?

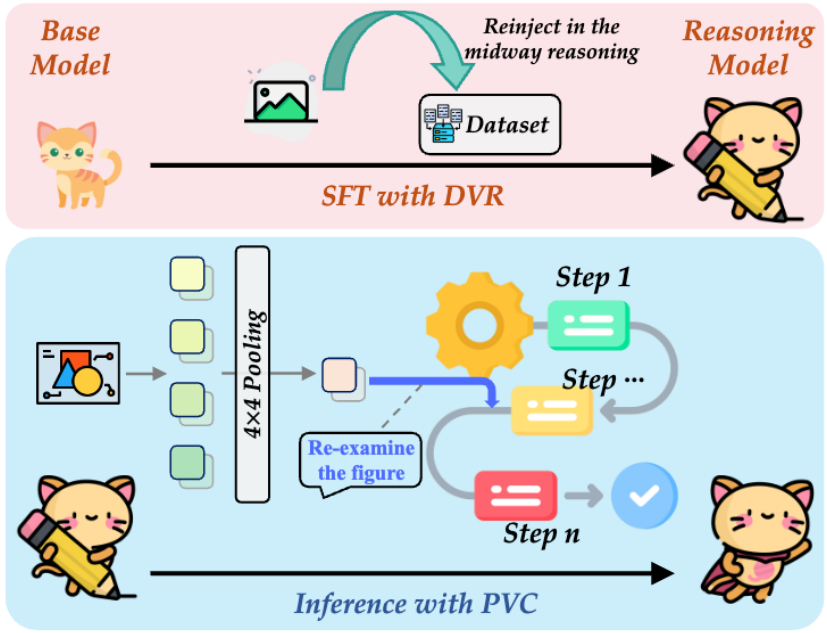


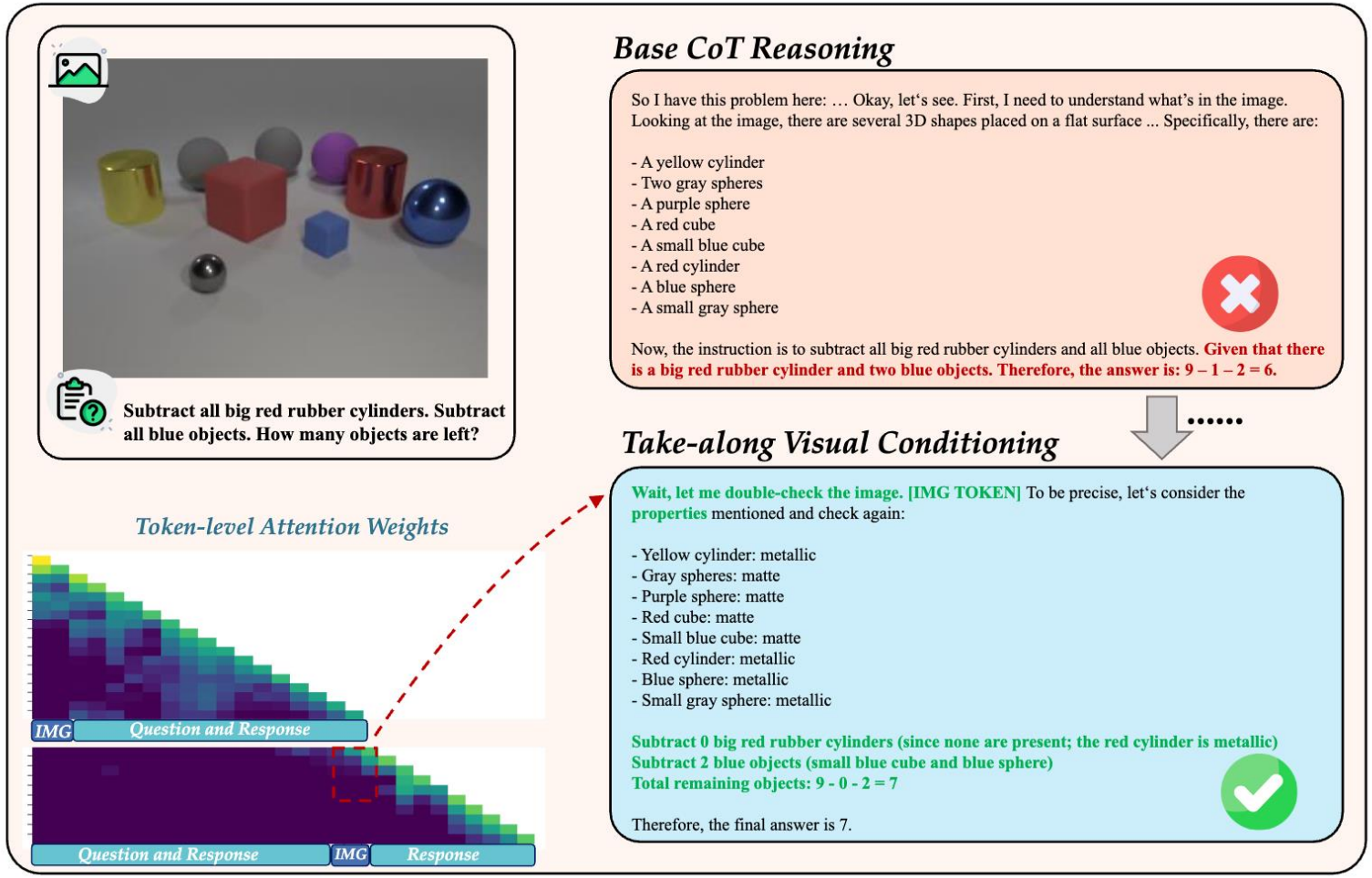
(a) MLLM's Layer-level Attention Weights



(b) MLLM's Token-level Attention Weights

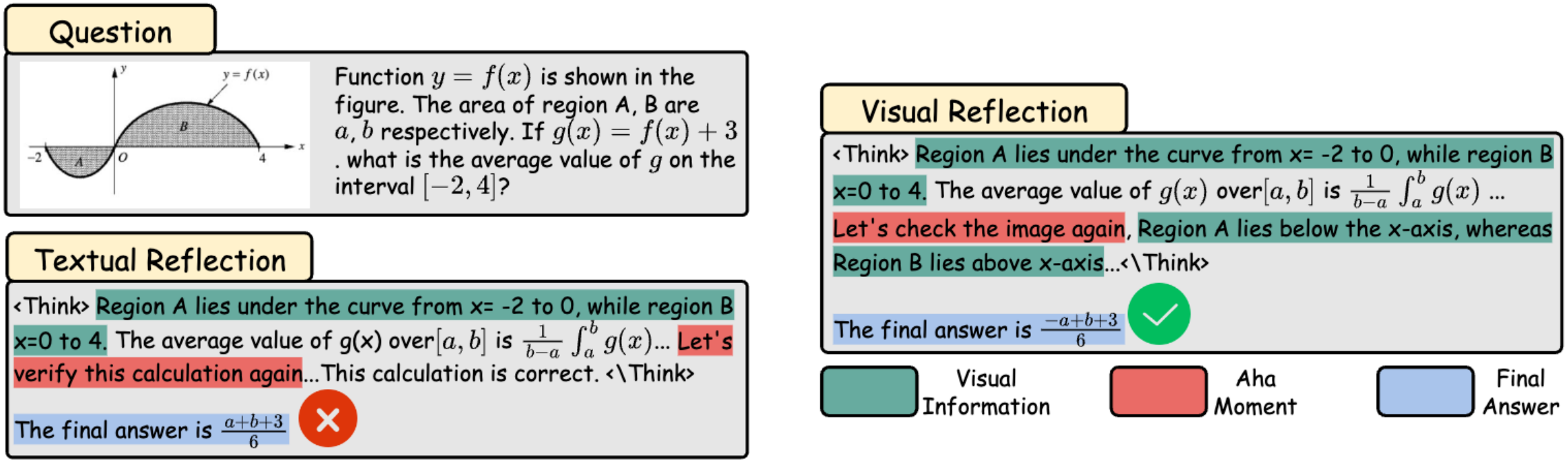
Solution: Re-Examine Images





Look Again, Think Slowly: Enhancing Visual Reflection in Vision-Language Models

Pu Jian[1,2], Junhong Wu[1,2], Wei Sun [1,2], Chen Wang [1,2], Shuo Ren[1], Jiajun Zhang[1,2,3]*
[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Wuhan AI Research
{jianpu2023, wujunhong2021, sunwei2023, chenwang2020, shuo.ren}@ia.ac.cn
jjzhang@nlpr.ia.ac.cn

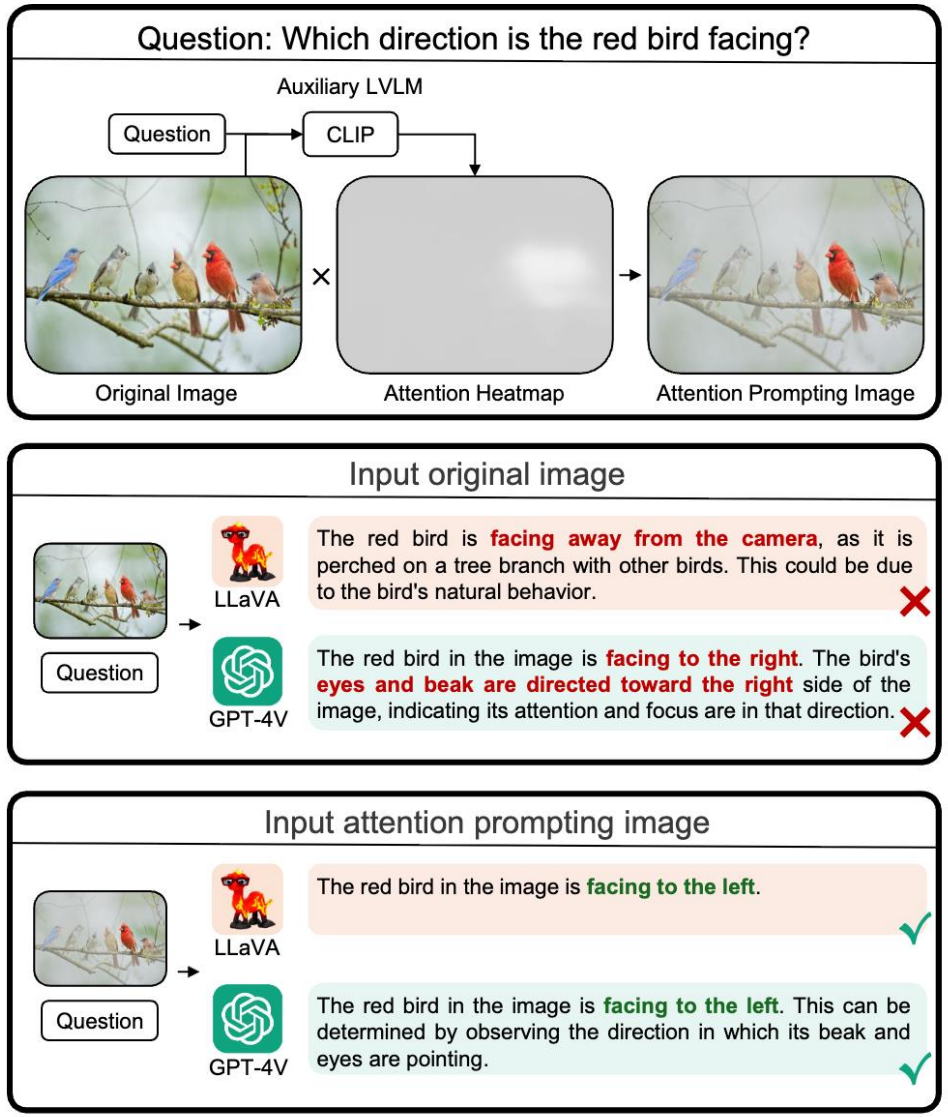


Attention Prompting on Image for Large Vision-Language Models

Runpeng Yu, Weihao Yu†, and Xinchao Wang†

National University of Singapore
{r.yu,weihaoyu}@u.nus.edu, xinchao@nus.edu.sg

Attention Prompting



Divide, Conquer and Combine: A Training-Free Framework for High-Resolution Image Perception in Multimodal Large Language Models

Wenbin Wang[1]*, Liang Ding[2]*, Minyan Zeng[1], Xiabin Zhou[3], Li Shen[4], Yong Luo[1]†, Dacheng Tao[5]

[1]Wuhan University, [2]The University of Sydney
[3]Jiangsu University, [4]Sun Yat-Sen University, [5]Nanyang Technological University
wangwenbin97@whu.edu.cn, liangding.liam@gmail.com, luoyong@whu.edu.cn

Attribute Recognition



What's color of the bike in the image?
A. Yellow ✓ B. Red
C. Black D. Blue



What's the color of the umbrella?
A. Blue ✓ B. Red
C. Black D. Green

OCR



What is the license plate number of the car?
A. RK 879 ✓ B. PK 879
C. RK 789 D. RK 876

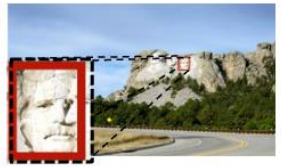


What are the letters written on the green trash bin?
A. VP ✓ B. VD
C. VA D. GP

Visual Prompting

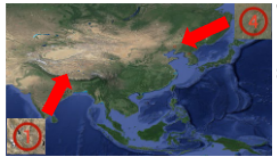


Which president's face is highlighted in red rectangle?
A. George Washington ✓
B. Thomas Jefferson
C. Abraham Lincoln
D. Theodore Roosevelt



Which president's face is highlighted in red rectangle?
A. George Washington
B. Thomas Jefferson
C. Abraham Lincoln
D. Theodore Roosevelt ✓

Map Analysis

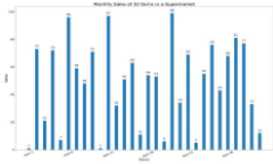


Which two numbered locations belong to the same country?
A. 1 and 4 ✓ B. 1 and 2
C. 2 and 3 D. 3 and 4

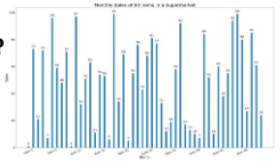


Which two numbered locations belong to the same country?
A. 1 and 4 ✓ B. 1 and 3
C. 1 and 2 D. 2 and 4

Chart Analysis



What is the sum of the sales for Item 1 and Item 6?
A. 97 ✓ B. 94
C. 95 D. 96



What is the sum of the sales for Item 1 and Item 6?
A. 97 ✓ B. 94
C. 95 D. 96

Spatial Relationship



What is the relative position of the red car compared to the black car?
A. To the right of the black car ✓
B. To the left of the black car
C. Behind the black car
D. In front of the black car



What type of structure is visible to the right of the image?
A. An elephant statue ✓
B. A house
C. A bridge
D. A car

HR-Bench

Attribute Recognition



What's color of the bike in the image?

A. Yellow ✓ B. Red

C. Black D. Blue



What's the color of the umbrella?

A. Blue ✓ B. Red

C. Black D. Green

OCR



What is the license plate number of the car?

A. RK 879 ✓ B. PK 879

C. RK 789 D. RK 876



What are the letters written on the green trash bin?

A. VP ✓ B. VD

C. VA D. GP

V^*: Guided Visual Search as a Core Mechanism in Multimodal LLMs

Penghao Wu †
UC San Diego
pew011@ucsd.edu

Saining Xie
New York University
saining.xie@nyu.edu

What is the instrument held by an ape?



GPT-4V:
The ape is holding a **saxophone**.

SEAL:
The instrument held by an ape is a **guitar**.

What is the color of the child's shoes?



GPT-4V:
The child's shoes are **white**.

SEAL:
The child's shoes are **green**.

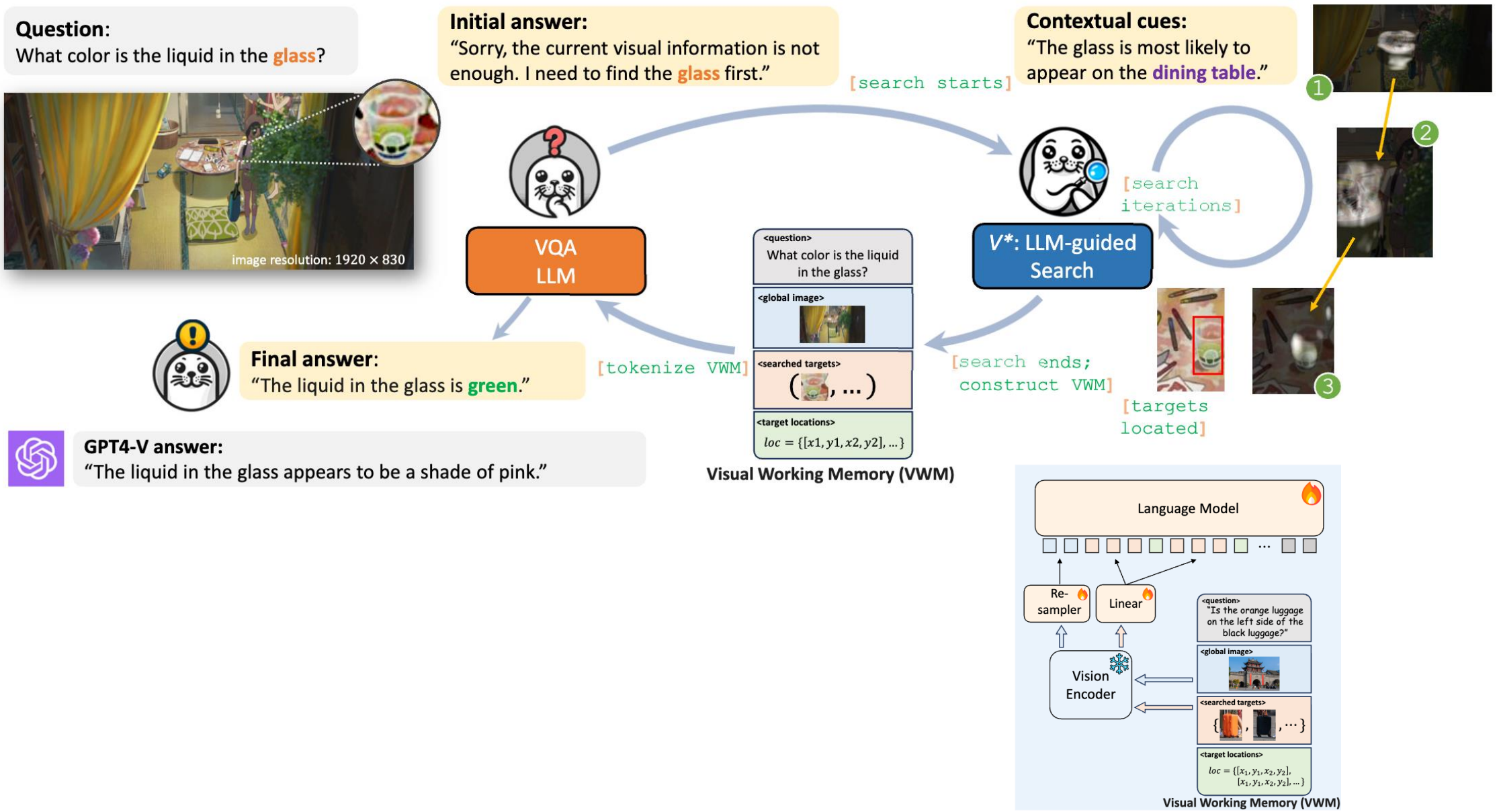
What is the color of the mug?



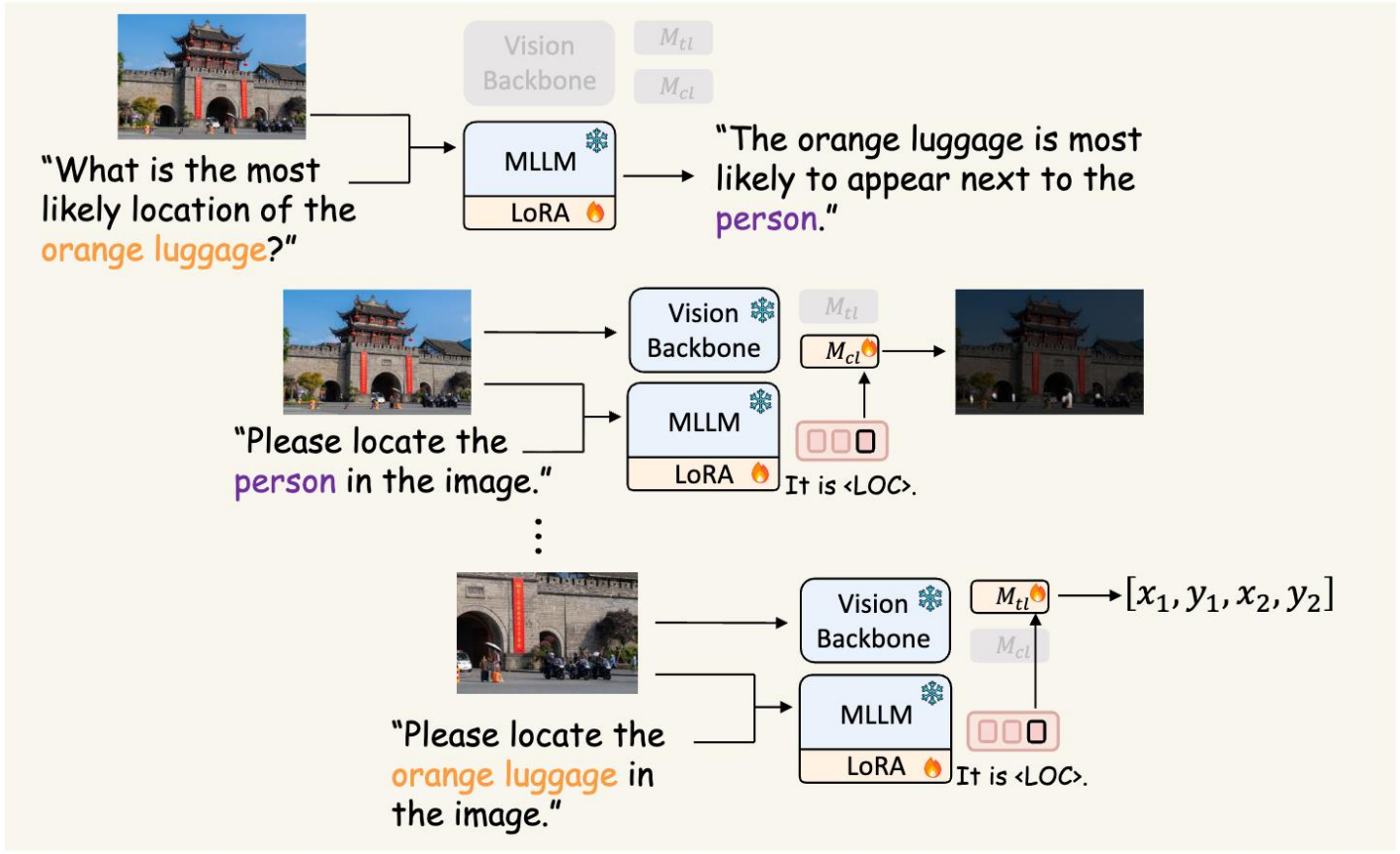
GPT-4V:
The color of the mug in the image is **white**.

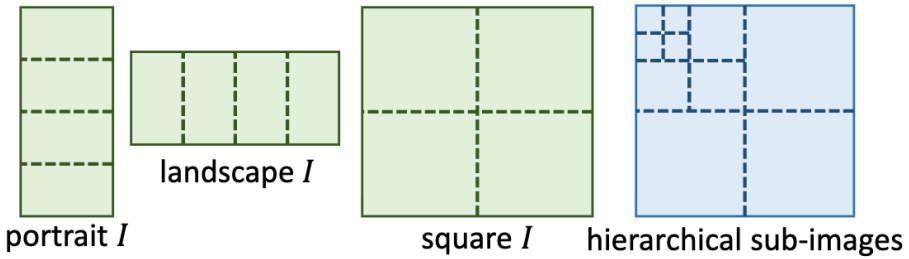
SEAL:
The color of the mug is **blue**.

SEAL Framework and V*



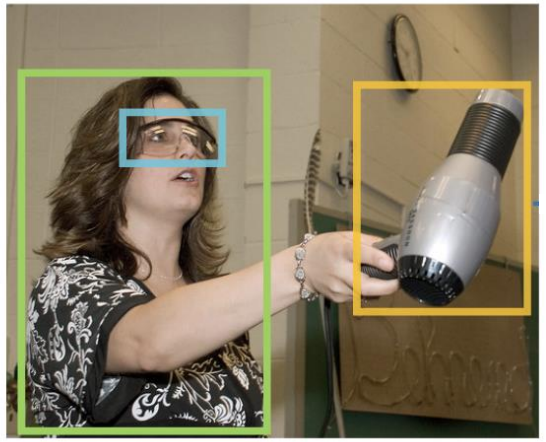
V* Search



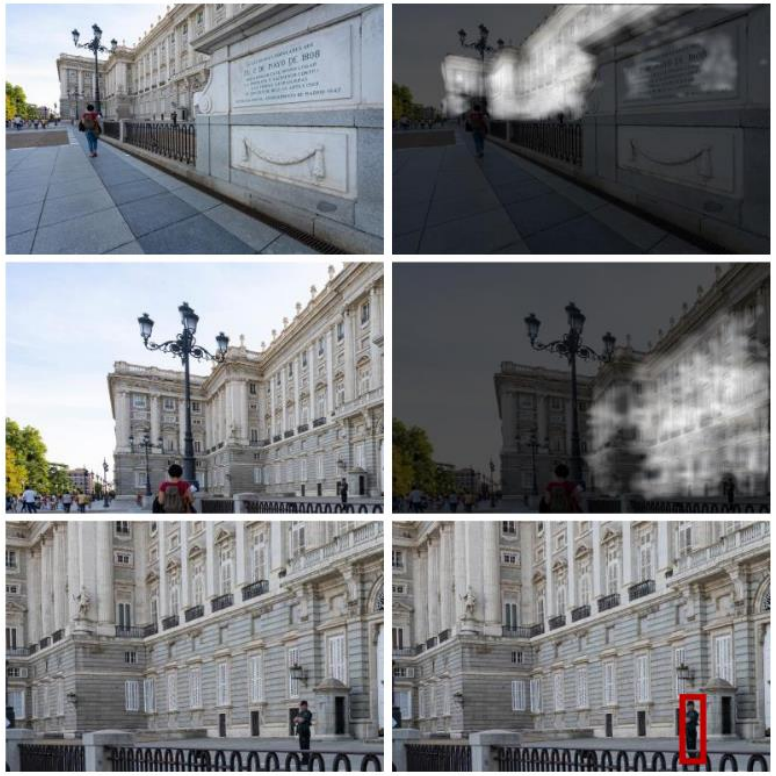


Person. Bicycle ... Hairdryer.

A woman holds a blow dryer, wearing protective goggles



Examples



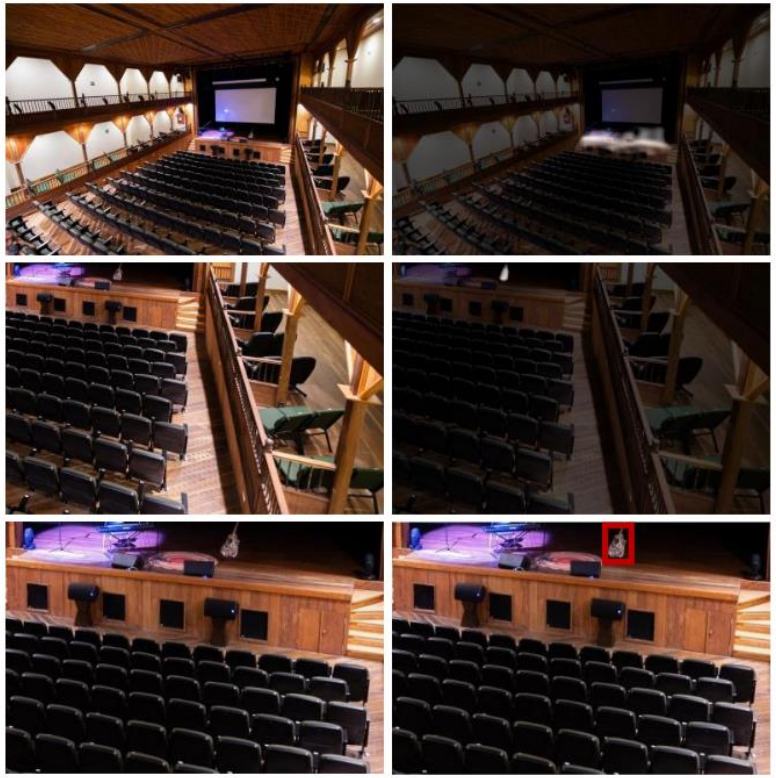
Contextual cue

The guard is most likely to appear next to the building.

Contextual cue

The guard is most likely to appear near the entrance of the building.

Search for the **guard**
(Question: Tell me the guard's detailed posture.)



Contextual cue

The guitar is most likely to appear on the stage.

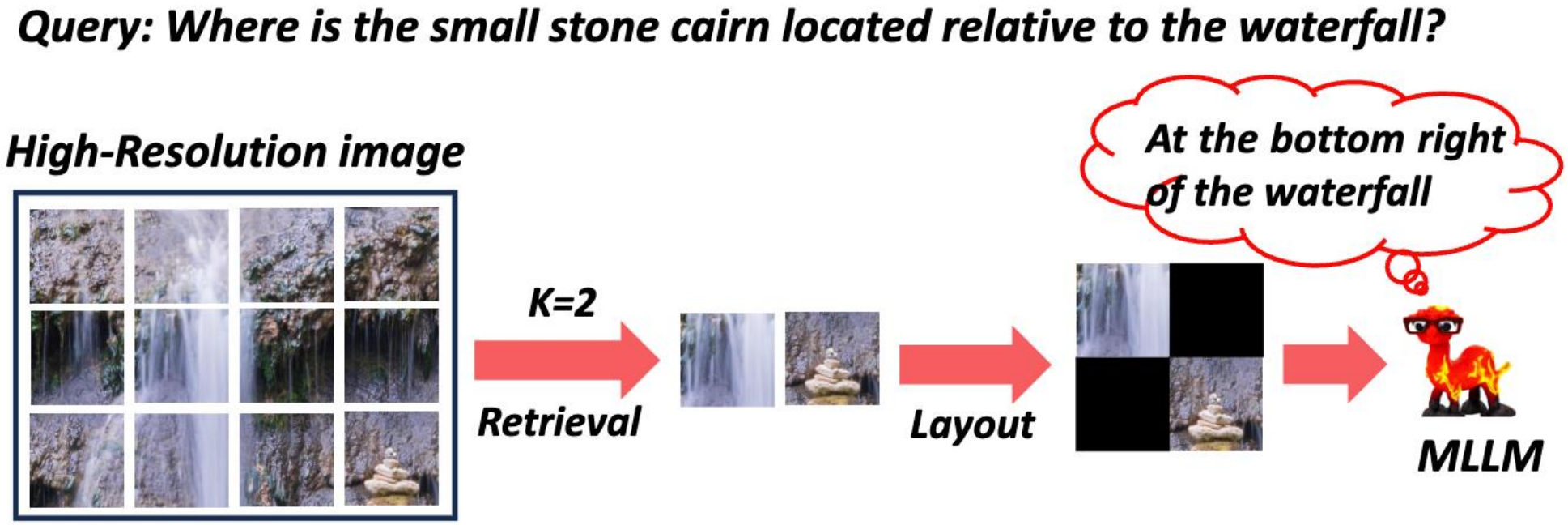
Target-specific cue

Search for the **guitar**
(Question: Is the guitar on the left or right side of the keyboard?)

🔍 Retrieval-Augmented Perception: High-Resolution Image Perception Meets Visual RAG

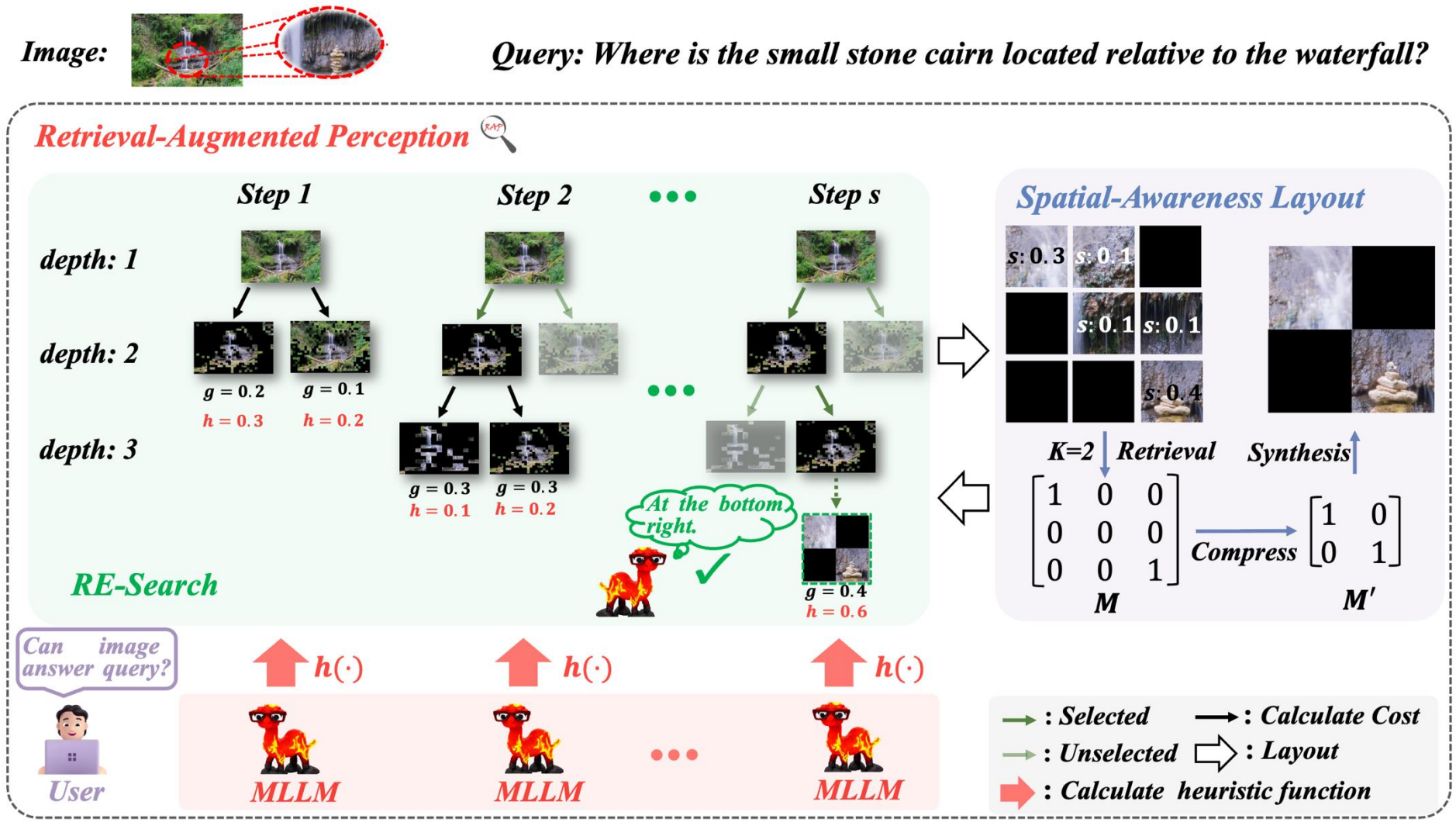
Wenbin Wang [1] **Yongcheng Jing** [2] **Liang Ding** [3] **Yingjie Wang** [2]
Li Shen [4] **Yong Luo** [1] **Bo Du** [1] **Dacheng Tao** [2]

Retrieve Relevant Regions Only



(a) The simple schematic of our *Retrieval-Augmented Perception (RAP)*

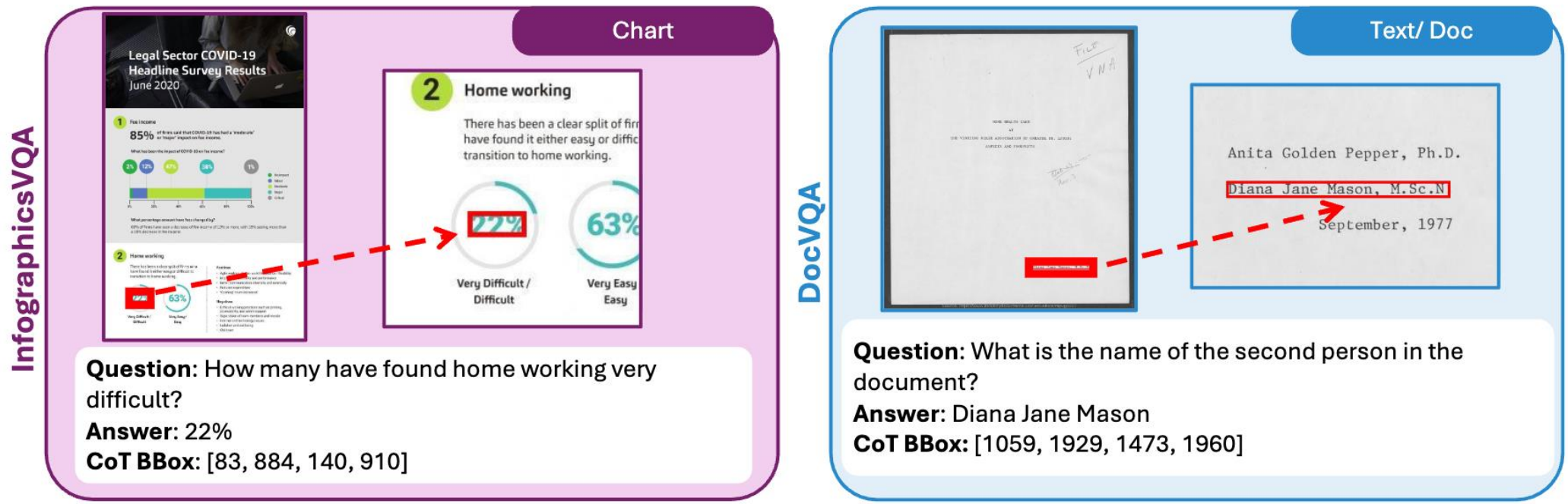
Retrieve Relevant Regions Only



Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning

Hao Shao[1,2] **Shengju Qian**[1] **Han Xiao**[1] **Guanglu Song**[2]
Zhuofan Zong[1] **Letian Wang**[3] **Yu Liu**[2]✉ **Hongsheng Li**[1,4,5]✉

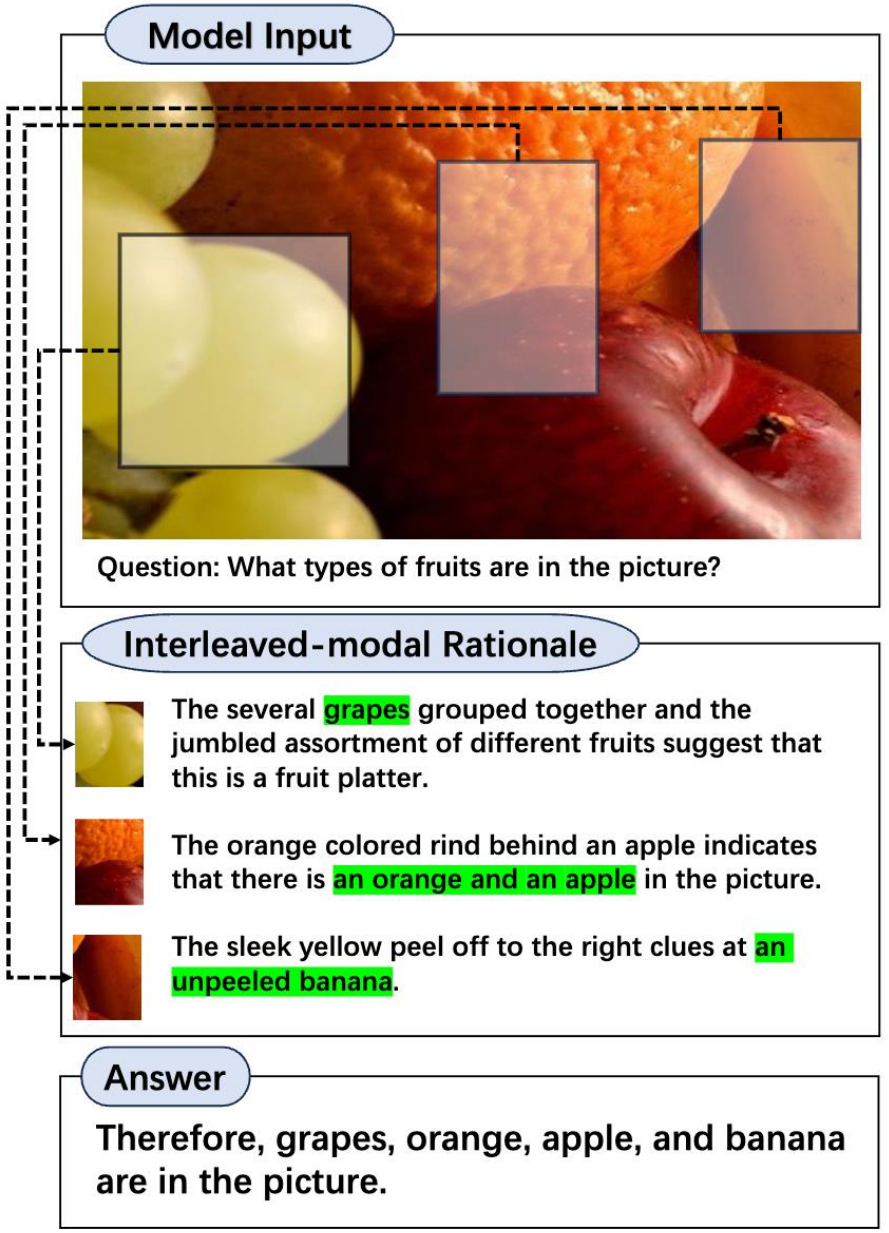
[1]The Chinese University of Hong Kong [2]SenseTime Research
[3]University of Toronto [4]Shanghai Artificial Intelligence Laboratory [5]CPII under InnoHK



Interleaved-Modal Chain-of-Thought

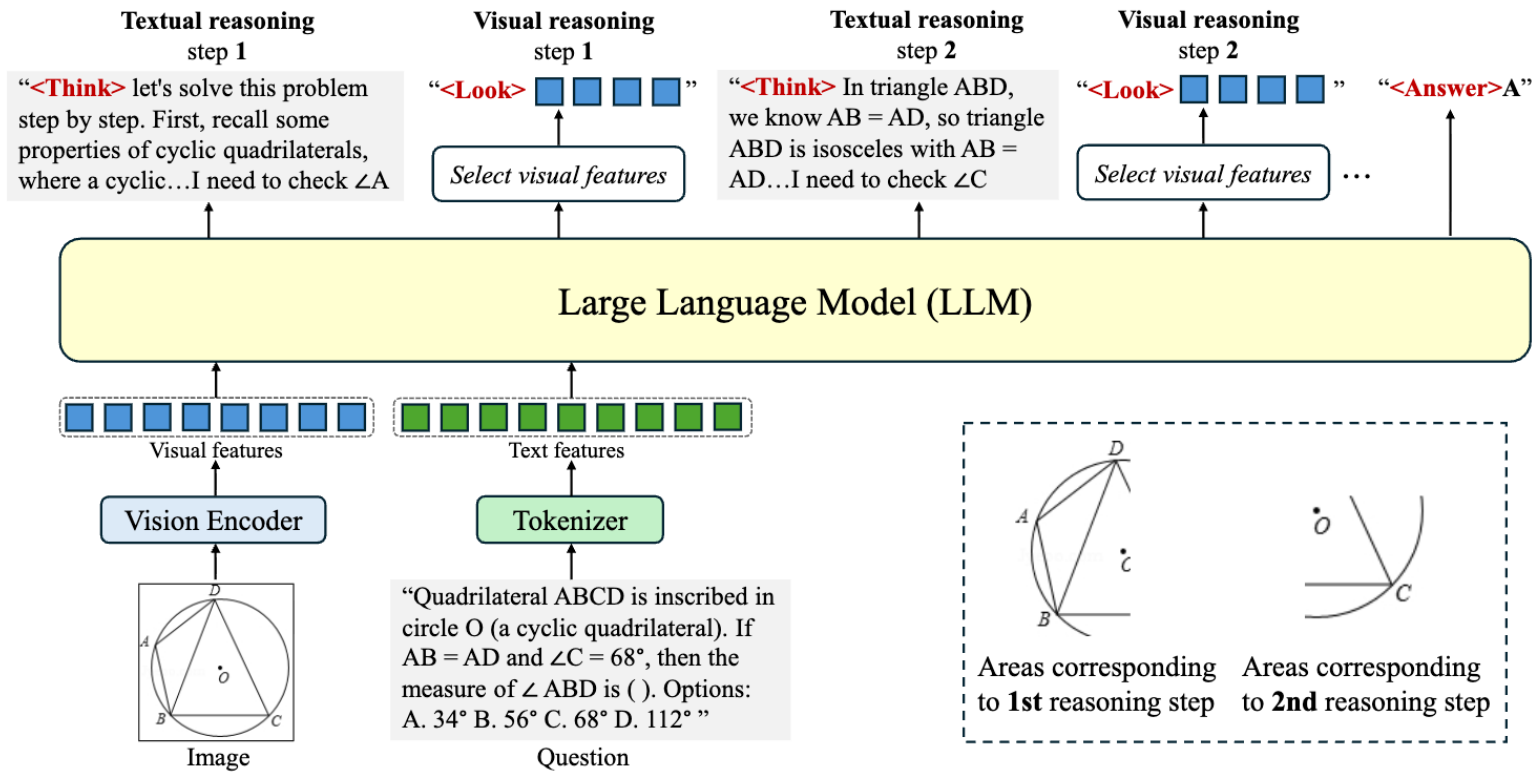
Jun Gao[1], Yongqi Li[2]*, Ziqiang Cao[1]*, Wenjie Li[2]

[1]School of Computer Science and Technology, Soochow University
[2]Department of Computer Science, The Hong Kong Polytechnic University
jgao1106@stu.suda.edu.cn, liyongqi0@gmail.com
zqcao@suda.edu.cn, cswjli@comp.polyu.edu.hk
https://github.com/jungao1106/ICoT



TVI-CoT: Text-Visual Interleaved Chain-of-Thought Reasoning for Multimodal Understanding

Lianyu Hu [1] **Xiaoyu Ma** [1] **Zeqin Liao** [1] ✉ **Yang Liu** [1]



ViperGPT: Visual Inference via Python Execution for Reasoning

Dídac Surís*, Sachit Menon*, Carl Vondrick
Columbia University
viper.cs.columbia.edu

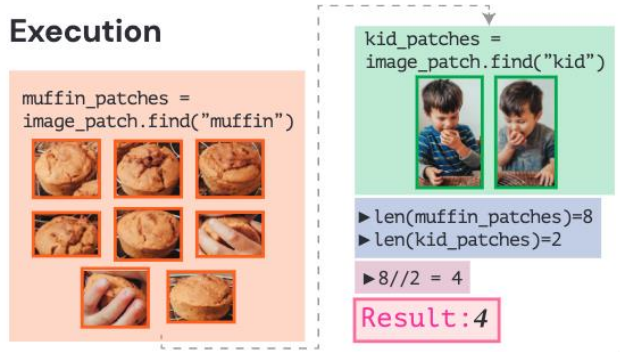
Query: How many muffins can each kid have for it to be fair?



Generated Code

```
def execute_command(image):
    image_patch = ImagePatch(image)
    muffin_patches = image_patch.find("muffin")
    kid_patches = image_patch.find("kid")
    return str(len(muffin_patches) // len(kid_patches))
```

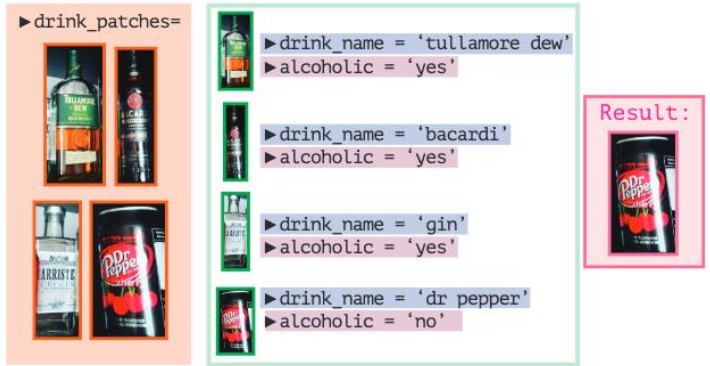
Execution



Query: Drink with zero alcohol



```
def execute_command(image):
    image_patch = ImagePatch(image)
    drink_patches = image_patch.find("drink")
    for drink_patch in drink_patches:
        drink_name = drink_patch.simple_query("What is this?")
        alcoholic = llm_query(f"Does the {drink_name} have alcohol?")
        if alcoholic == "no":
            return drink_patch
    return None
```



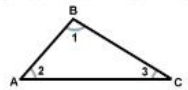
Visual SKETCHPAD: Sketching as a Visual Chain of Thought for Multimodal Language Models

Yushi Hu[*1,2] **Weijia Shi**[*1] **Xingyu Fu**[3] **Dan Roth**[3] **Mari Ostendorf**[1]
Luke Zettlemoyer[1] **Noah A. Smith**[1,2] **Ranjay Krishna**[1,2]

[1]*University of Washington* [2]*Allen Institute for AI* [3] *University of Pennsylvania*

Geometry

Prove the angles of △ABC sum to 180°

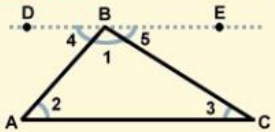


GPT-4o

Start by assuming, for contradiction, that the sum is not 180°. We'll introduce a new angle, ∠4 …

SketchPad + GPT-4o

Draw line DE parallel to AC



∠4 is congruent to ∠2, and ∠5 is congruent to ∠3. ∠1, ∠4, and ∠5 form a straight angle = 180°. So ∠1, ∠2, ∠3 add up to 180°.

(a)

Math Function

Is f(x) an convex function?

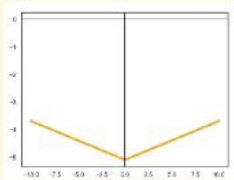
$$f(x) = 0.14\,|x| - 5.09$$

GPT-4o

It is convex for x>0 but concave for x<0

SketchPad + GPT-4o

Plot the function



convex because line between any two points lie above the graph.

(b)

Visual Search

According to the board, what is the type of the shop?



GPT-4o

The sign in the image reads "ceremonial matcha,". it's a cafe.

SketchPad + GPT-4o

Use sliding_window to find the board



I see a board in patch 2 and 3 and will zoom in



The zoomed-in image shows the text "**corepower YOGA**". So it's a yoga studio

(c)

Spatial Reasoning

Are there cookies stacked on other cookies?



GPT-4o

No

SketchPad + GPT-4o

Use segmentation & marking



Cookies 2 and 5 seem to overlap. Need to analyze their depth

Use depth estimation



Warmer colors indicate closer proximity, suggesting they are stacked on cooler-colored ones.

(d)

CogCoM: A Visual Language Model with Chain-of-Manipulations Reasoning

Ji Qi[◇†], Ming Ding[†], Weihan Wang[◇†], Yushi Bai[◇†], Qingsong Lv[†], Wenyi Hong[◇†]
Bin Xu[◇*], Lei Hou[◇], Juanzi Li[◇], Yuxiao Dong[◇], Jie Tang[◇*]
[◇]Tsinghua University [†]Zhipu AI



Q: What is the brand of the plane?

CogCoM: Using GROUNDING(the plane) to find planes, and there are two regions [[255, 217, 312, 245], [193, 113, 481, 208]] that can be found.

With the most notable one, using GROUNDING(the logo in [[193, 113, 481, 208]]) to find the logo, resulting in [[246,142,307,163]].



Using CropZoomIn([[246, 142, 307,163]], 3) to get a new image.

The brand name is AirFrance.

Detailed Recognition



Q: How many wheels of the truck can be seen?

CogCoM: Using GROUNDING(all the individual wheels of the truck) to locate all the wheels in this picture, resulting in the boxes of [[56, 256, 96, 513], [105, 240, 144, 480], …]

Based on the identified boxes of the wheels, the number of visible wheels of the truck is ten.

Counting Objects



Q: What time does this photo show?

CogCoM: Applying GROUNDING(the clock) …, resulting in the region of [[814,531,940,1062]].

Use CropZoomIn([[814,531,940,1062]],2) to crop … and then zoom in it by two times … re-input this new image.

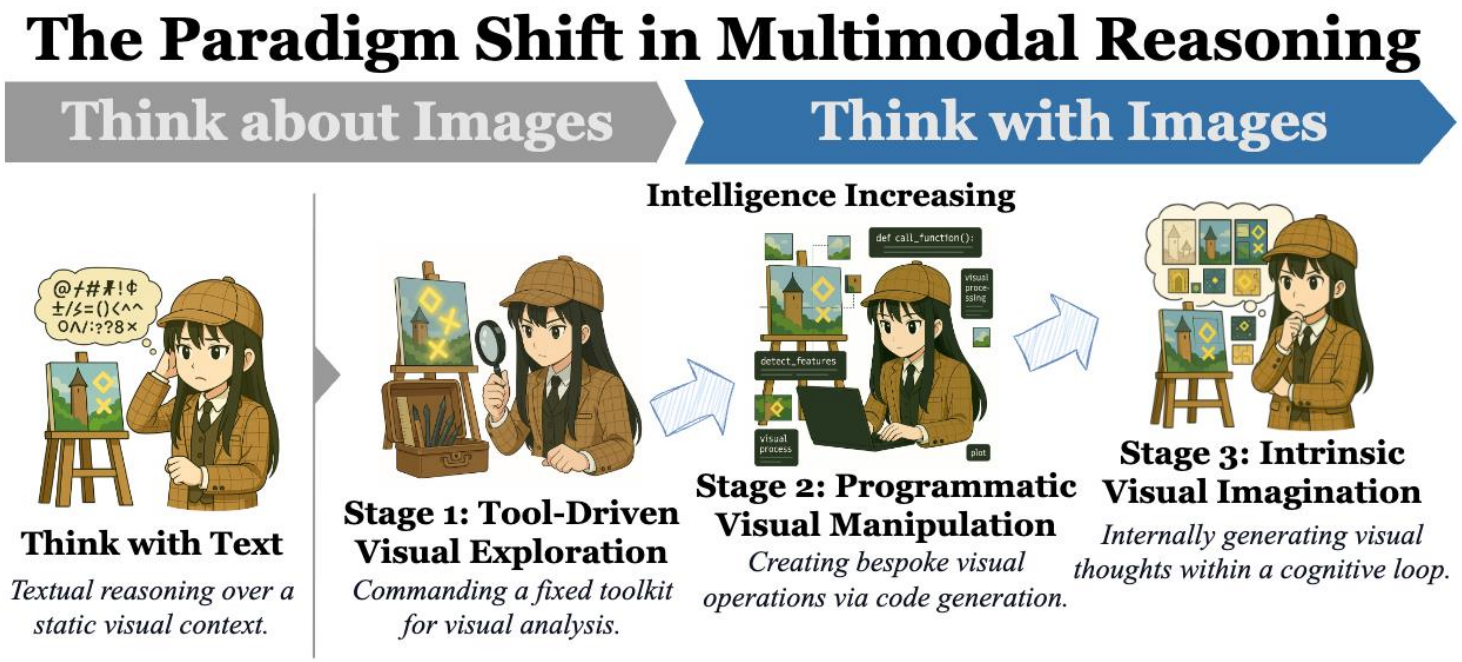


Using ReadTime(the clock in the new image) to re-cognize the time on the clock, and the result is 1:45.

Read Time

Thinking with Images for Multimodal Reasoning: Foundations, Methods, and Future Frontiers

Zhaochen Su[1], **Peng Xia**[2], **Hangyu Guo**[1], **Zhenhua Liu**[1], **Yan Ma**, **Xiaoye Qu**, **Jiaqi Liu**[2], **Yanshu Li**[1], **Kaide Zeng**[2], **Zhengyuan Yang**[3], **Linjie Li**[3], **Yu Cheng**[4], **Heng Ji**[5], **Junxian He**[1], **Yi R. (May) Fung**[1]
[1]The Hong Kong University of Science and Technology
[2]UNC-Chapel Hill, [3]Microsoft, [4]The Chinese University of Hong Kong, [5]UIUC



Questions?