

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

Authors

Penghao Wu

Saining Xie

Affiliations

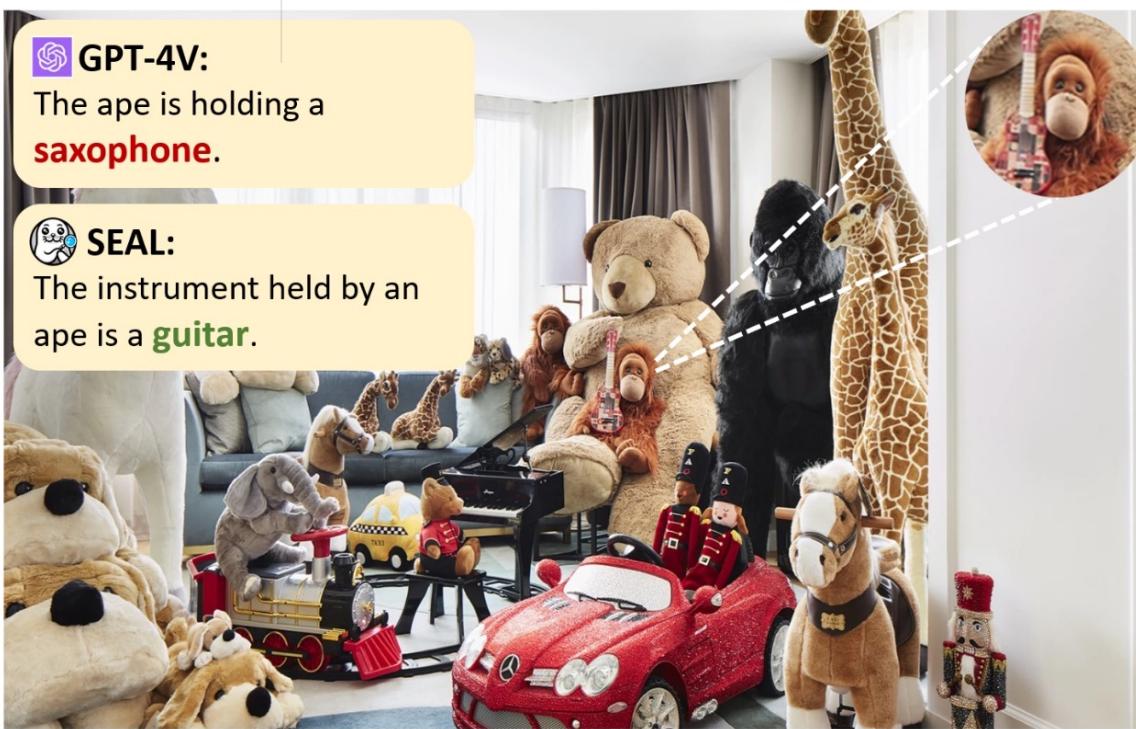
UC San Diego

New York University

Presented by:
John Tan Chong Min

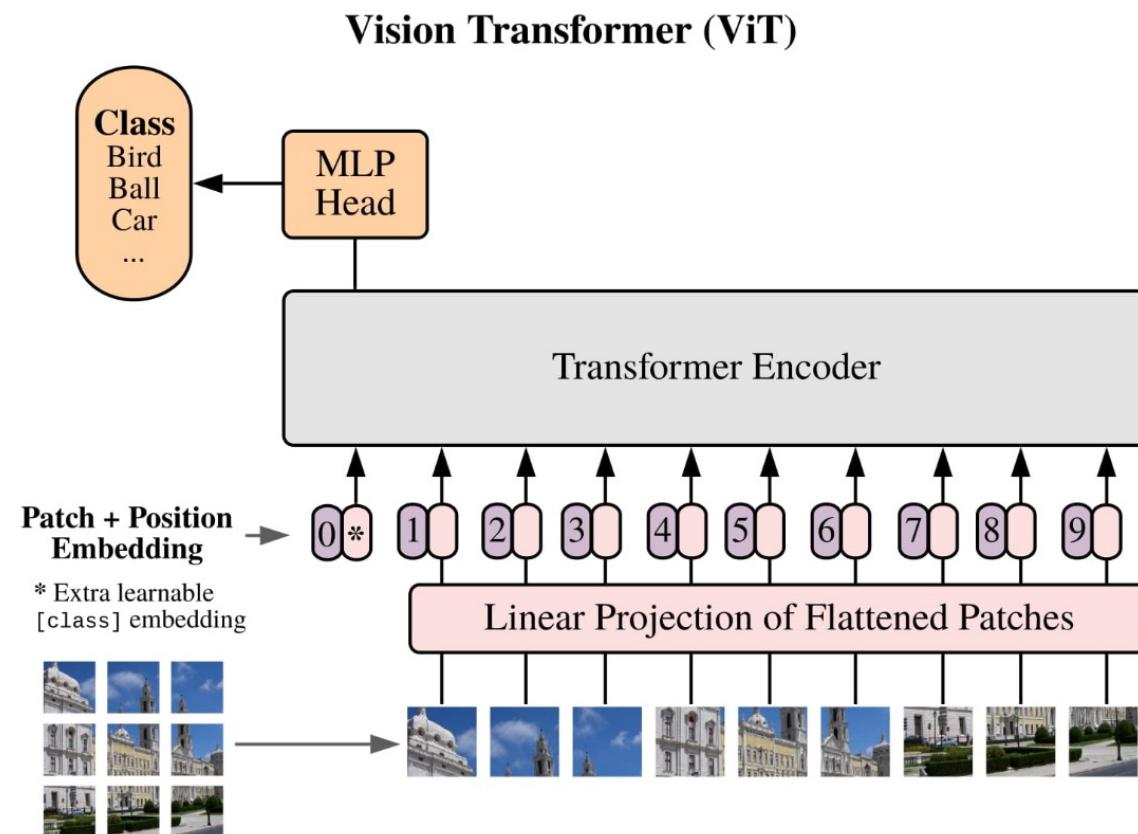
Supposedly better than GPT-4V

Question: What is the instrument held by an ape?



Key issue

- CLIP (Contrastive Language–Image Pre-training) image embeddings based on Vision Transformer (ViT) encodings are lossy!



An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Dosovitskiy et. al. 2021.

Background: Large Language and Vision Assistant (LLaVA)

End-to-end trained large multimodal model that connects a **vision encoder** and an **LLM** for general purpose visual and language understanding

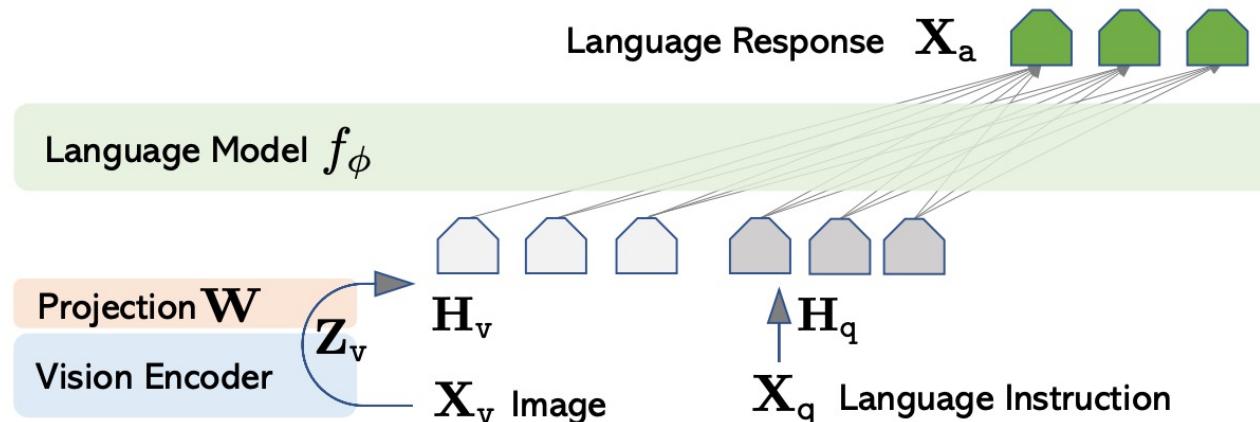
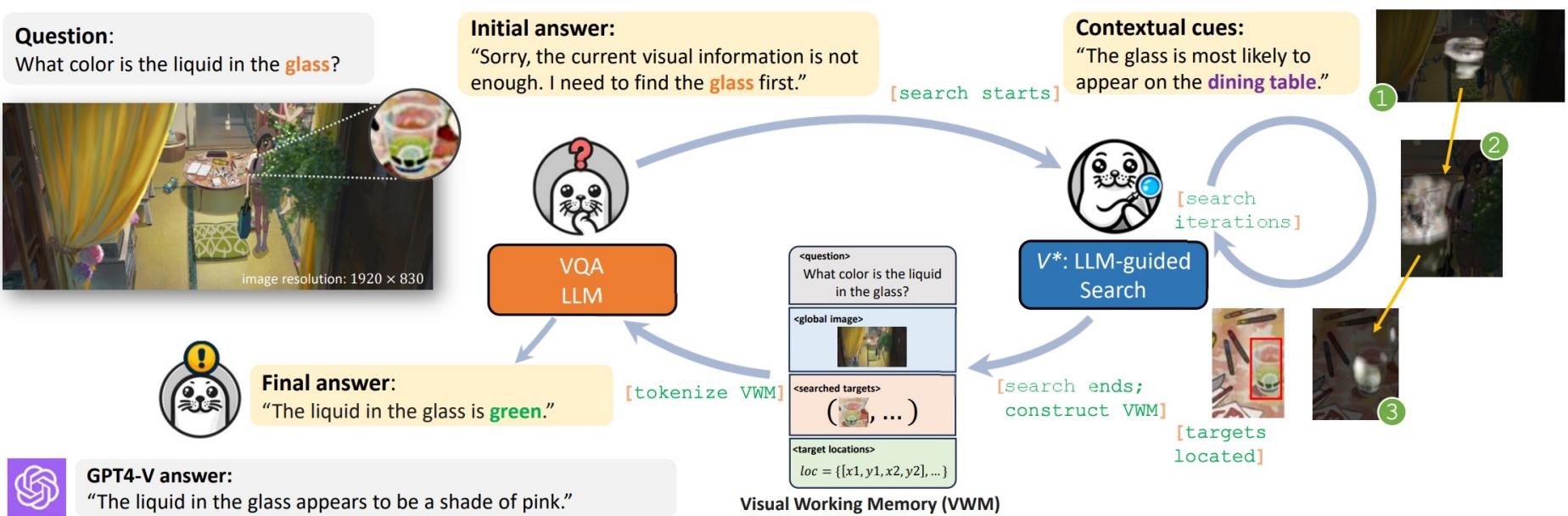


Figure 1: LLaVA network architecture.

Secret Sauce: Multiple Iterations to refine query



Procedure (Check if Visual Search is required)

- The Visual Question Answer (VQA) LLM first evaluates if the encoder's initial (global) visual features suffice for answering the question
- If yes, then answer question (usual Visual QA approach)
- If not, it **explicitly lists all the needed but missing information in the format of a list of target objects.**

Visual Search

- A visual working memory (VWM) is initialised for the **list of target objects**.
- The VWM has four blocks
 - **<question> block:** contains the initial textual question
 - **<global image> block:** contains the initial full image
 - **<searched targets> block:** stores the target object crops after search
 - **<target location> block:** stores the coordinates of the searched targets
- Use a **Visual Search Model** to identify target objects (Any Multimodal LLM (MLLM))
 - Store the cropped target objects and coordinates in VWM



Visual Search Model

- Takes in image tokens, and contextual and location information of query v_{tl}, v_{cl}
- **Target localization decoder:** Outputs a list of coordinates and confidence scores
- **Search cue localization decoder:** Outputs heatmap similar to **Segment Anything** model by META

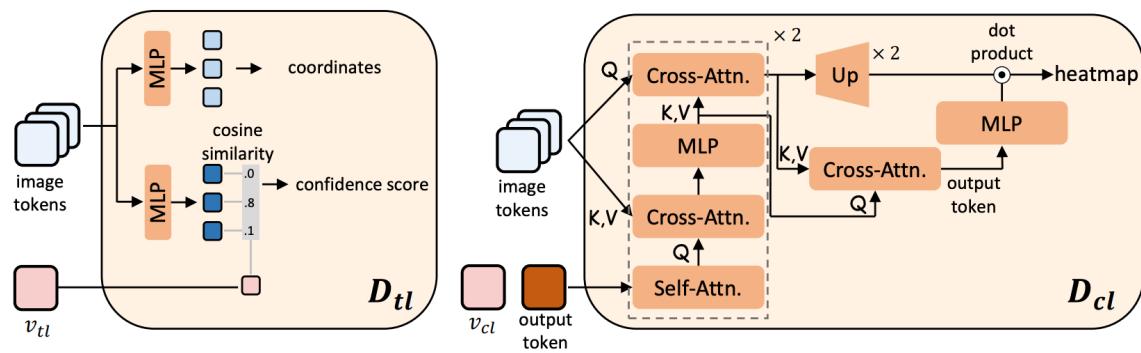
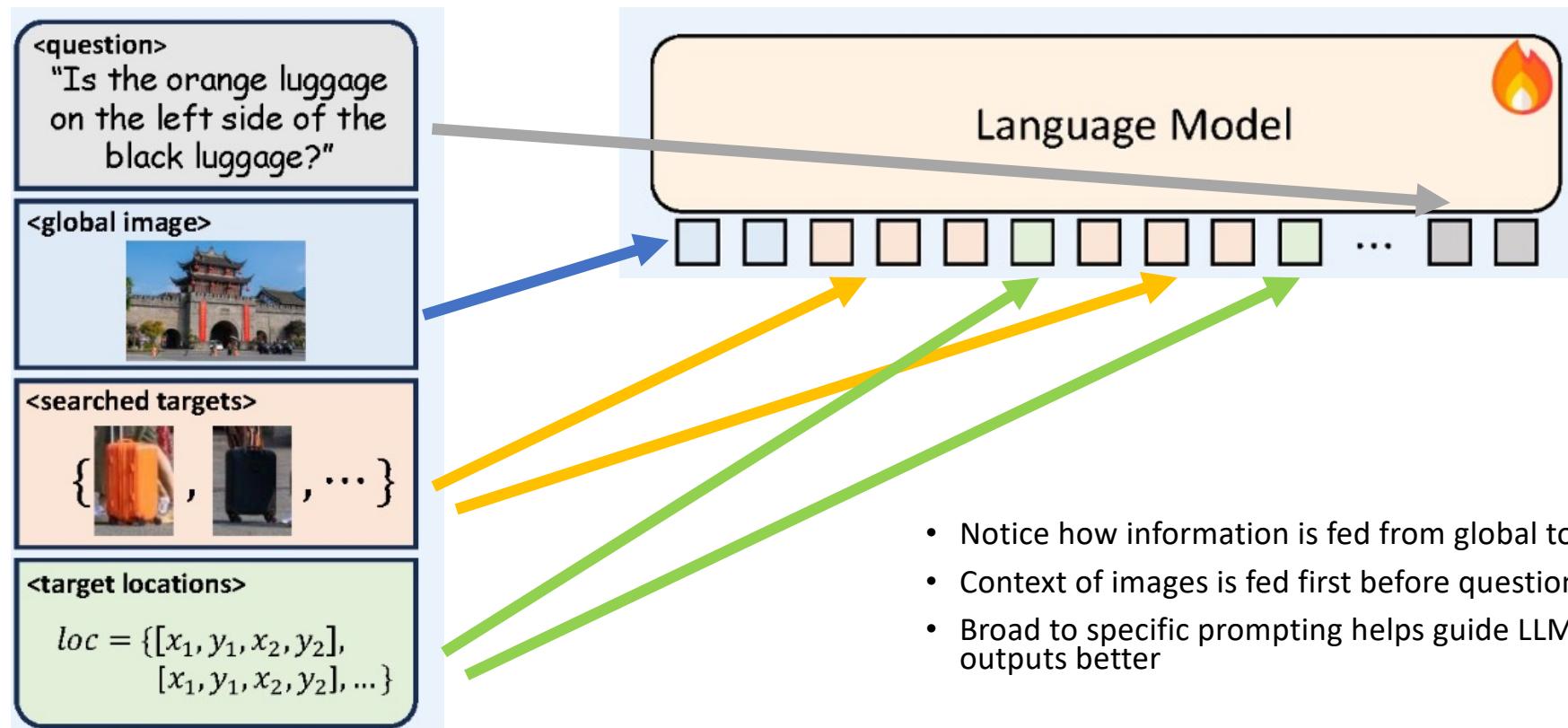


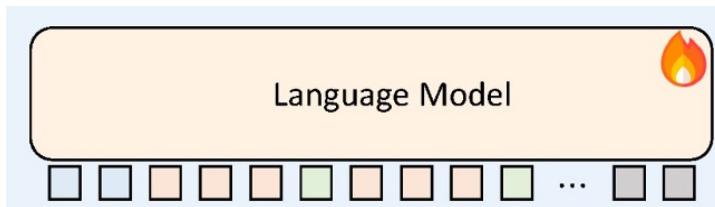
Figure 4. Detailed structure of the *target* localization decoder D_{tl} , and the *search cue* localization decoder D_{cl} .

LlaVA-7B is used as the Visual Search Model

Everything (large and small) is fed into Transformer



What it really means



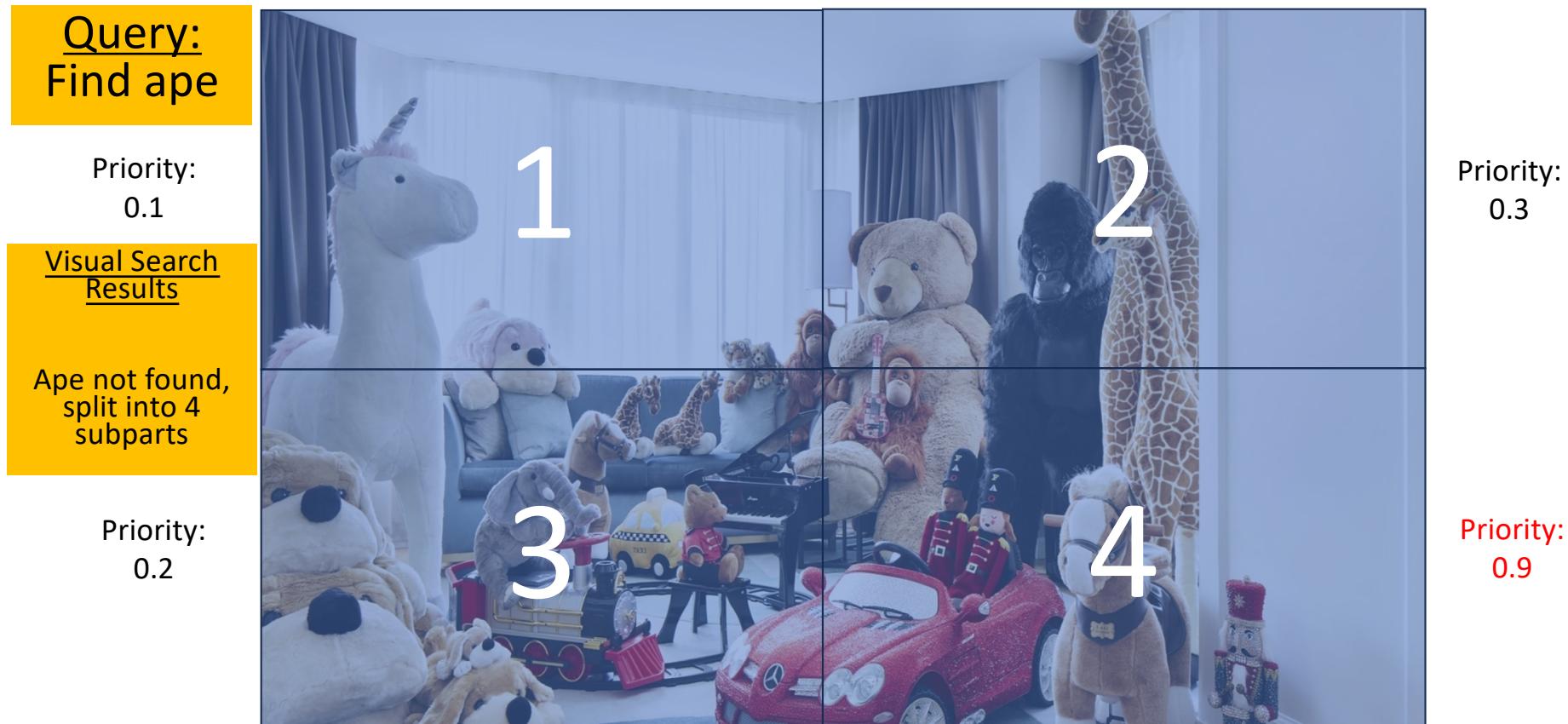
```
<Image>
Additional visual information to focus
on:
{Target Object 1's Name} <Object> at
location [x1, y1, x2, y2];
{Target Object 2's Name} <Object> at
location [x1, y1, x2, y2];
...
Question
```

Here `<Image>` is the feature tokens of the image and `<Object>` is the feature tokens of the target object stored in the VWM.

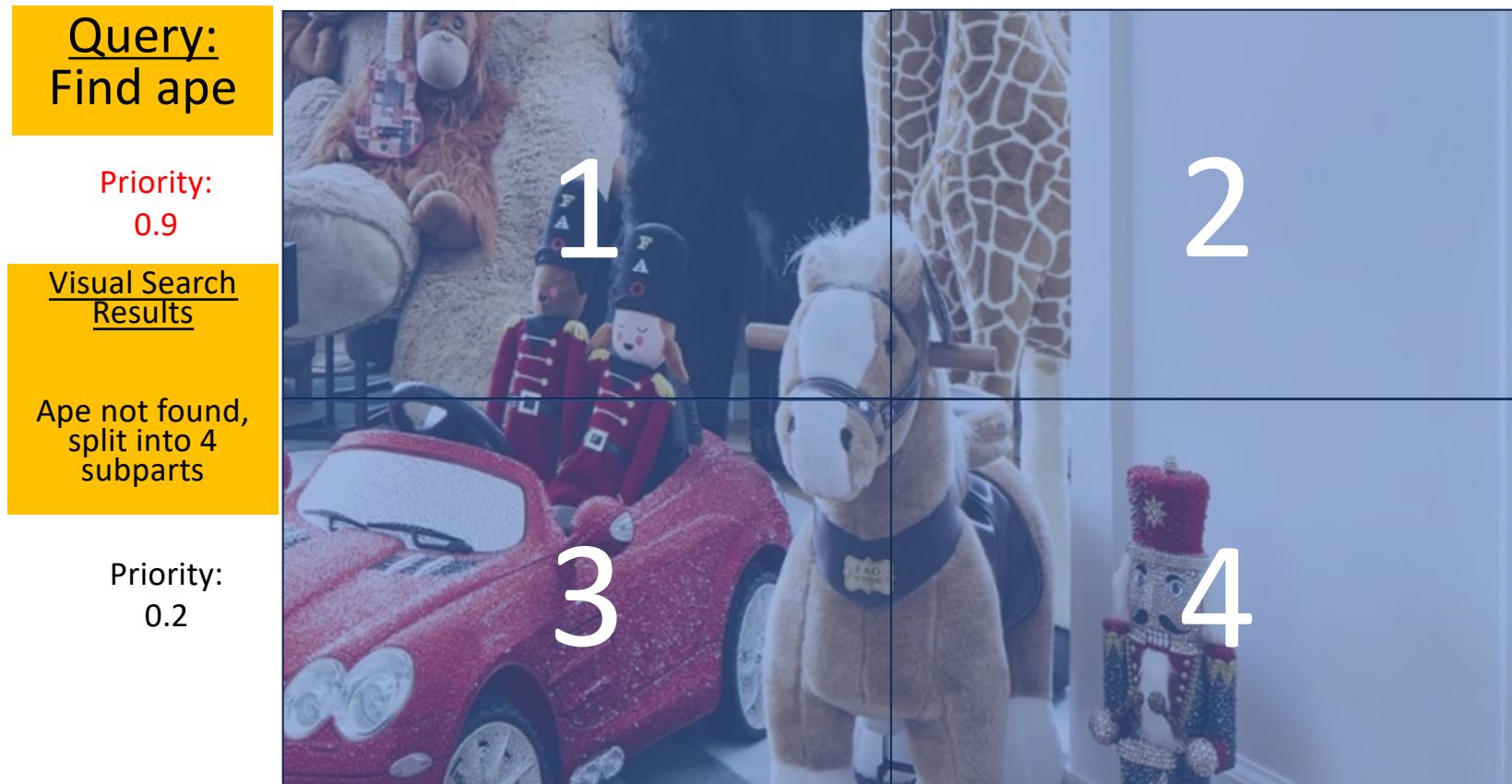
What if I cannot find the target object?

- Ask Visual Search Model guiding questions
 - Ask “What is the most likely location of the <target object> in the image?”
 - Answer of Visual Search Model is textual description <**contextual cue**>
 - Ask “Please locate the <**contextual cue**> in the image”
- Split image into 4 sub-parts
 - Assign priority score to each part based on confidence of the patches from Visual Search Model
 - Add sub-part images to buffer
- Keep searching buffer from highest priority image till target object found or image size too small

Split Image into 4 parts, assign priority



Choose Next Image; Split Image into 4 parts, assign priority



Target Localised! Return cropped image and coordinates

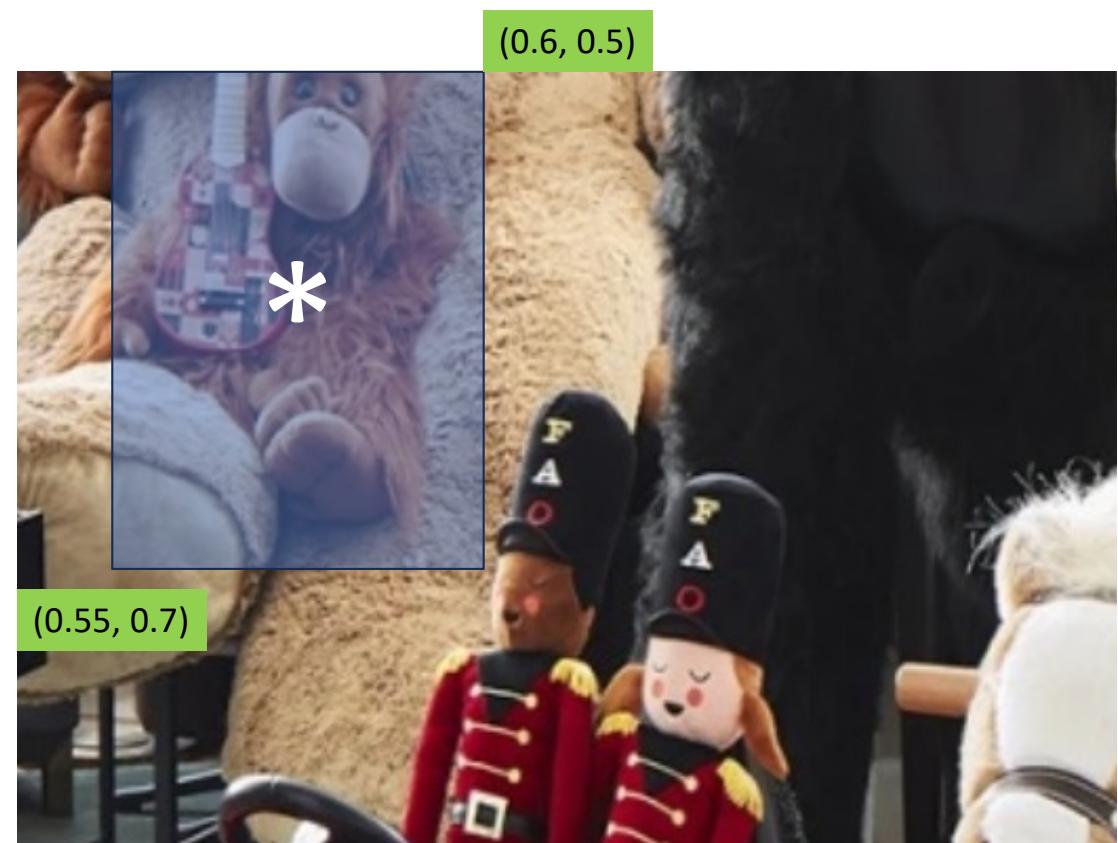
Query:
Find ape

Visual Search Results

Ape localised (Image: *)

Coordinates

- Bottom Left(0.55, 0.7)
- Top Right (0.6, 0.5)



Example showing heatmap

Step 1



Search for the **guard**

Contextual cue

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

Example showing heatmap

Step 2



Search for the **guard**

Contextual cue

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

Example showing heatmap

Step 3

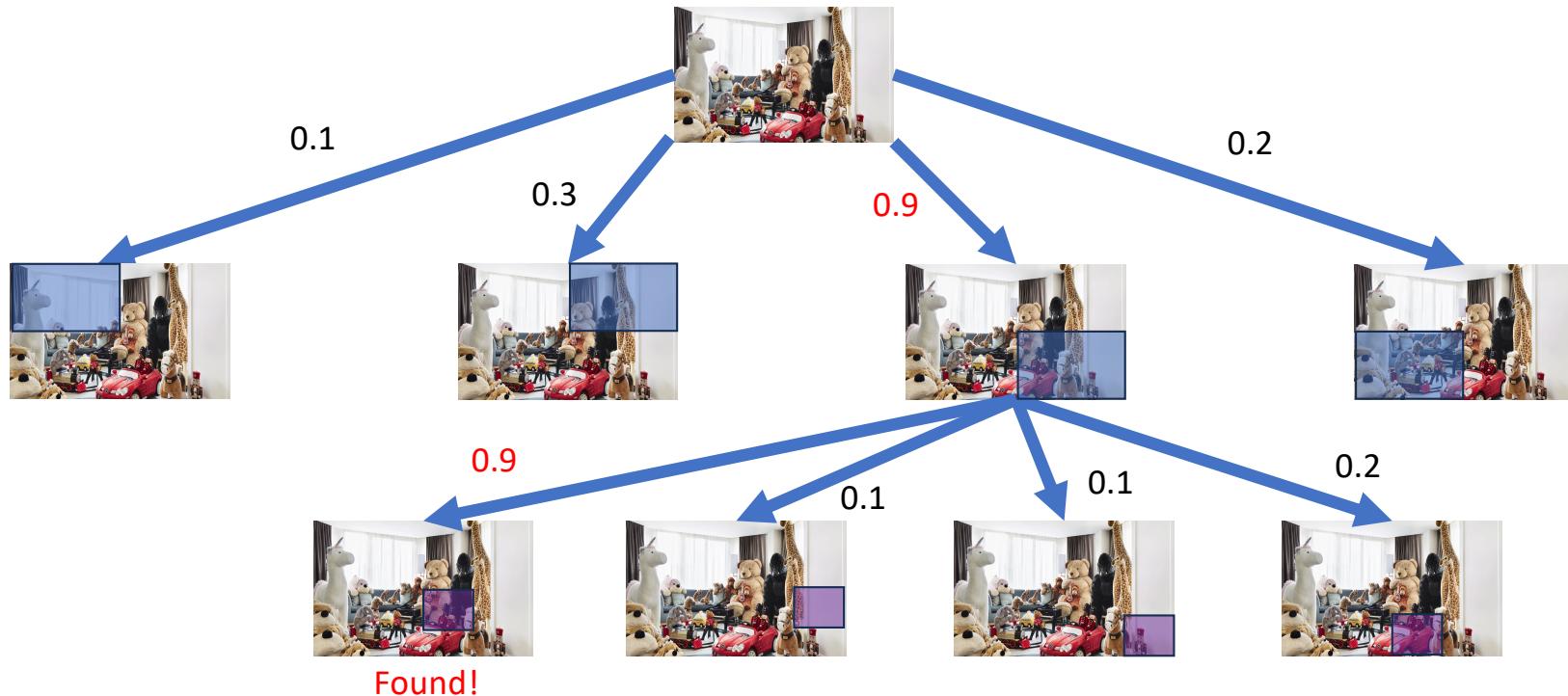


Target Spotted!

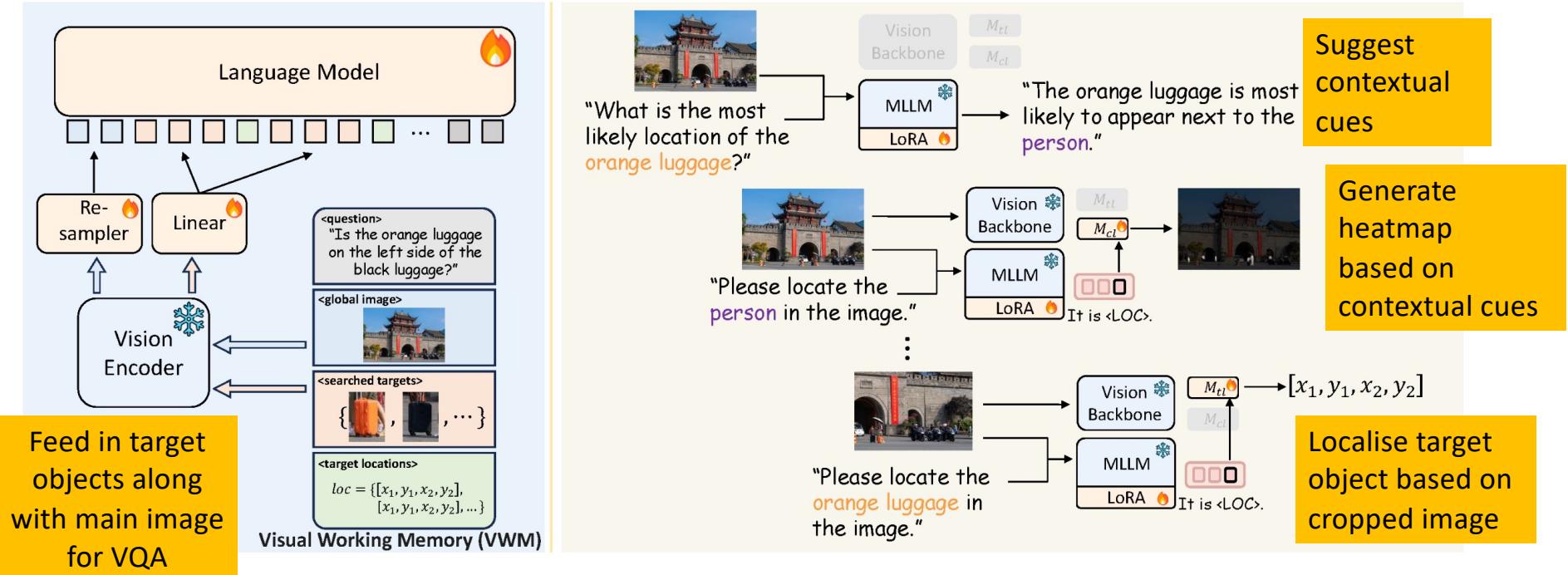
Search for the guard

A form of Best-First Search

- Search Image sub-part with the highest priority first



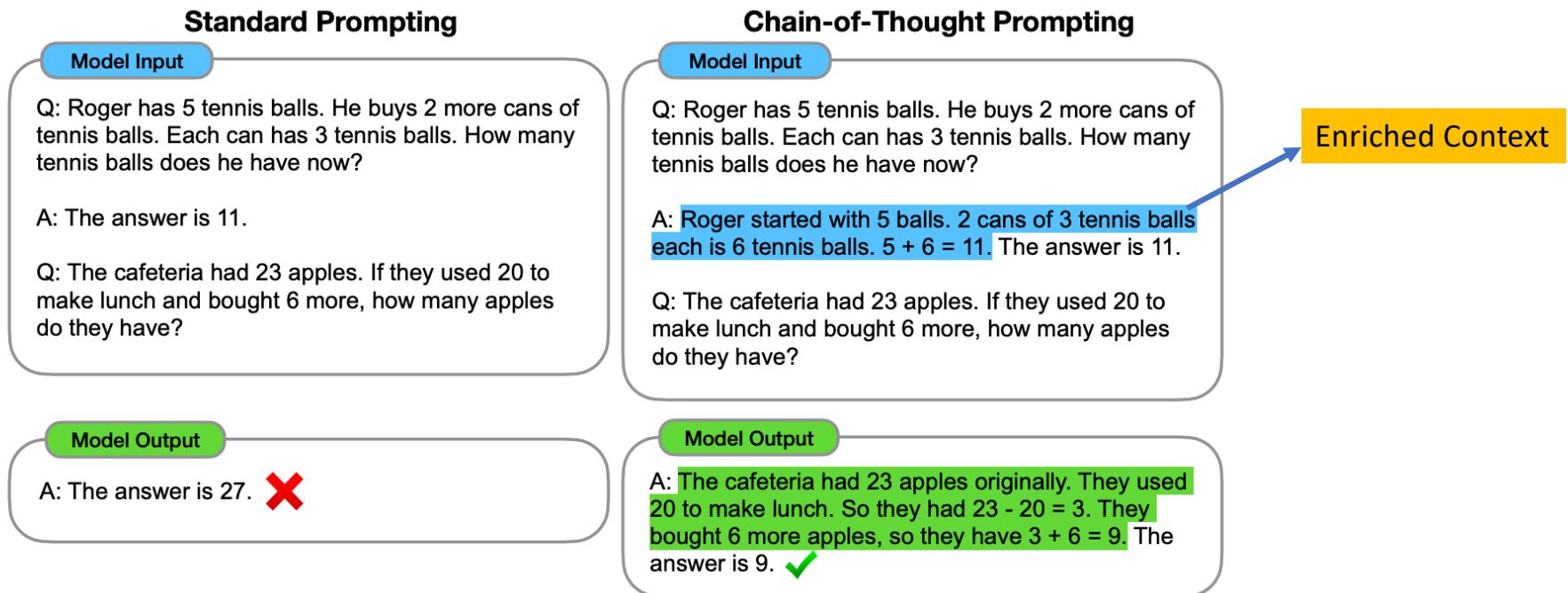
Putting everything together



An instantiation of the proposed SEAL framework. The left section represents the VQA LLM, which utilizes all the data within the Visual Working Memory to respond to questions. On the right, we illustrate the operational pipeline of the V^* visual search algorithm.

Comparison: Chain of Thought

- Chain of Thought is only one pass down the thought chain without multiple cycles of reasoning

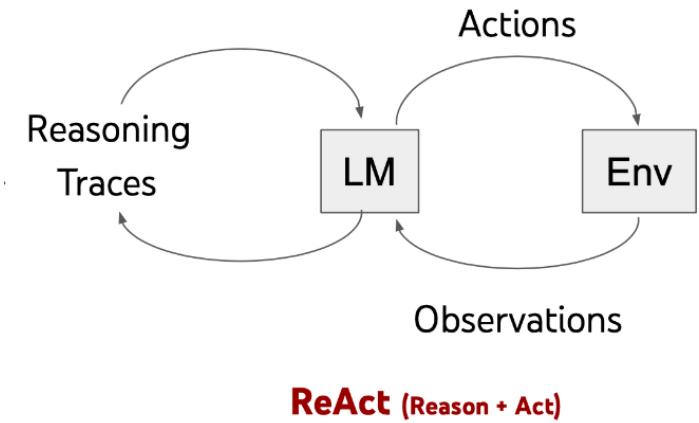
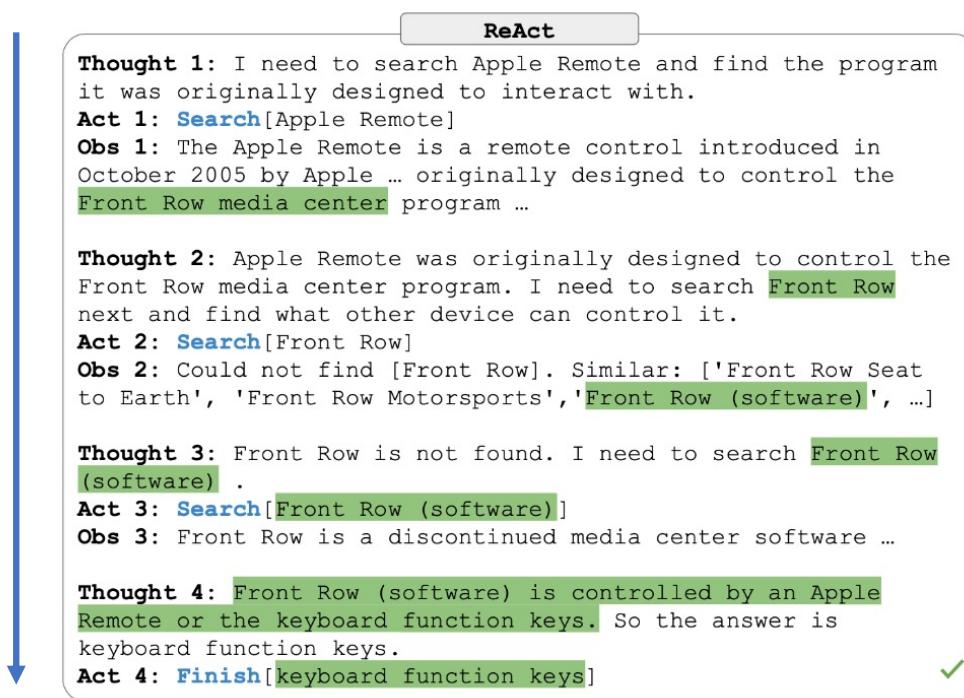


Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Wei et. Al. 2023.

Comparison: ReAct Framework

- ReAct's observational space is not reduced, unlike V*

Multiple Iterations of Reasoning



ReAct: Synergizing Reasoning and Acting in Language Models. Yao et al. 2023.

Results

	Attribute (%)	Spatial (%)	Overall (%)	
Human	98.26	100.00	98.95	
Random Guess	26.73	50.00	35.99	
<i>Open-source end-to-end MLLMs</i>				
BLIP2 [21]	26.95	53.94	37.69	
MiniGPT-4 [58]	30.43	50.00	38.22	
LLaVA [25]	23.47	53.94	35.59	
InstructBLIP [7]	25.21	47.36	34.02	
Otter [20]	26.95	56.57	38.74	
LLaVA-1.5 [24]	43.47	56.57	48.68	
<i>LLM tool-using pipelines</i>				
MM-React [47]	34.78	51.31	41.36	
VisualChatGPT [48]	30.43	48.68	37.69	
Visprog [11]	31.30	56.57	41.36	
<i>Commercial chatbot systems</i>				
Bard [10]	31.30	46.05	37.17	
Gemini Pro [8]	40.86	59.21	48.16	
GPT-4V [31]	51.30	60.52	54.97	
V*	SEAL (Ours)	74.78	76.31	75.39

- Superior performance for a model building on Vicuna / LLaVA
- V* (SEAL) outperforms even GPT-4V
- Showcases that subdividing images and asking relevant contextual questions helps improve accuracy

Could the performance measure be biased to beat GPT-4V?

- Benchmark (V* bench) is self-created based on 191 high-resolution images with an average image resolution of 2246×1582 .
- V* Bench contains two sub-tasks: **attribute recognition** and **spatial relationship reasoning**.
 - The attribute recognition task has 115 samples and requires the model to recognize a **certain type of attribute (eg. color, material)** of an object.
 - The spatial relationship reasoning task has 76 samples and asks the model to determine the **relative spatial relationship between two objects**.
- Most questions on localising small parts of the image AND multiple choice answers favour open-sourced models

GPT-4V works with cropped image!

My hypothesis: GPT-4V can perhaps do even better when given all spatial resolutions of the image

C You



What is the instrument held by an ape

ChatGPT

The ape in the image is holding a brass wind instrument which appears to be a trumpet. The trumpet is a musical instrument commonly used in jazz, classical, and orchestral music, characterized by its bright, penetrating sound.



C You



What is the instrument held by an ape

ChatGPT

The instrument held by the ape in the image appears to be a toy guitar. It's designed to look like an electric guitar, complete with a body, neck, headstock, and what seems to be strings and pickups illustrated on it.



Visual Search may fail XD!

- Prompt made to be less specific

Text Instruction

What is the ape holding?

Input Image



Clear Submit

Direct Answer

Need to conduct visual search to search for: ape.

Visual Search Results

Targets located after search: ape.

Visual Search Results



Final Answer

The ape is holding the teddy bear.

<https://craigwu-vstar.hf.space/>

Cropped Image comparison

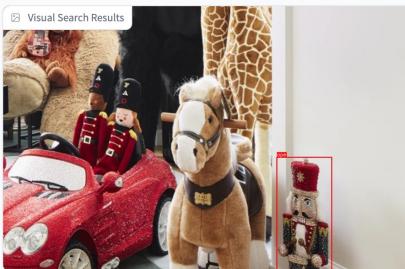
Text Instruction
What is the ape holding?

Input Image 

Clear Submit

Direct Answer
Need to conduct visual search to search for: ape.

Visual Search Results
Targets located after search: ape.

Visual Search Results 

Final Answer
The ape is holding the toy.

c You



What is the ape holding?

ChatGPT

In this image, the ape is holding an electric guitar.



GPT-4V



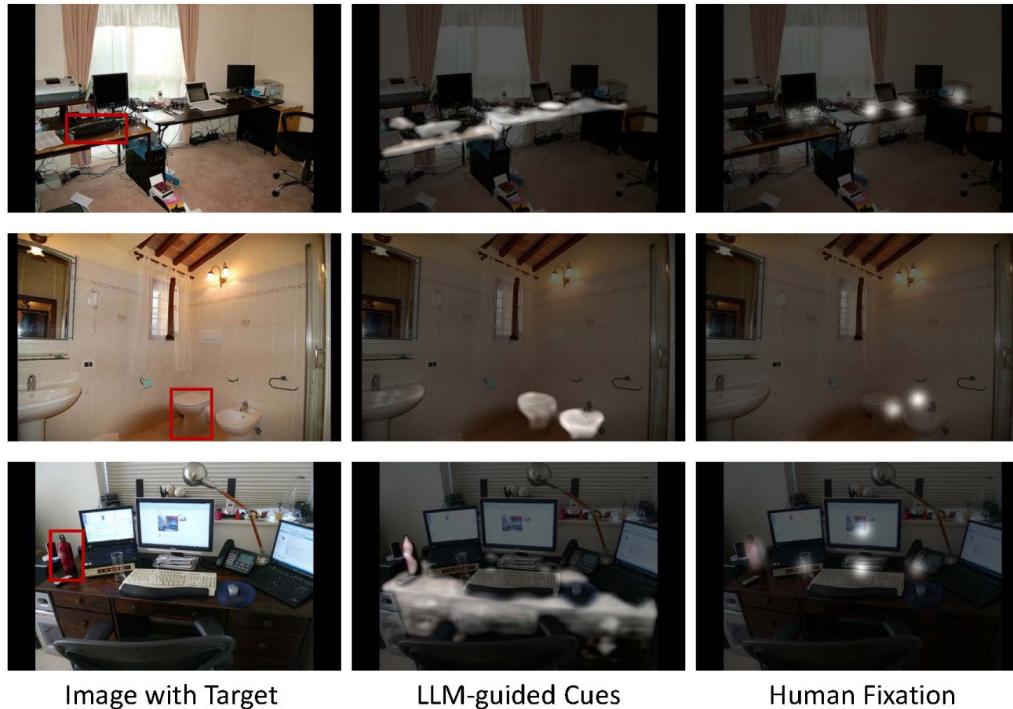
V* with
LlaVa /
Vicuna-7B

Similar to human performance on COCO-18 dataset!

	Search Length ↓
Random-DFS	9.97
Random-BFS	4.90
Sequential-DFS	9.82
Sequential-BFS	4.20
Human Fixation ($\gamma=0.9$)	2.52
Human Fixation ($\gamma=0.8$)	2.70
LLM-guided visual search	2.80

- COCO-Search18 records people's eye fixations when searching for a specific target object in natural scene images
- To evaluate human performance, convert the ground-truth human fixation sequence on each sample to a 2D heatmap and use it as guidance during the search.
- V^* algorithm can achieve similar efficiency to the human fixations

Model attention is similar to human fixation!



Visualization of search cues of our V^* algorithm and human fixation on COCO-Search18. Humans tend to focus on center regions or salient objects while our model focuses on a larger contextual region.

My takeaways

- Iterative focusing of context is a useful way to get better answers
- V* Benchmark is not generic enough for Visual QA, and it is highly likely GPT-4V when using a similar approach of focusing context will do better than Vicuna-7B model used
- CLIP based on Vision Transformers (ViT) is probably the downfall of all image tokenisation-based methods

Questions to Ponder

- V^* does iterative search over one frame of image, how do we do this search over multiple frames, e.g. video
- Can we use the recursive splitting strategy of V^* to enhance ReAct-based models?
- Can the recursive search modify the input modality as well, by moving to a different abstraction space as needed to solve the search problem?
 - Abstraction spaces: e.g. pixel-level, object-level, scene-level
- Why not just start with input images at various scales / abstraction spaces?
 - Can we improve the ViT encoder?