Can Image Enhancement Help the Reasoning for Vision-Language Models?

Zhiyu Xue, Yihao Wu

Background: Vision-Language Models (VLMs)

Closed Source VLMs

GPT-4 [2]

Gemini [3]

Open Source VLMs

LLaVA [4]

Flamingo [5]

MiniGPT4 [6]

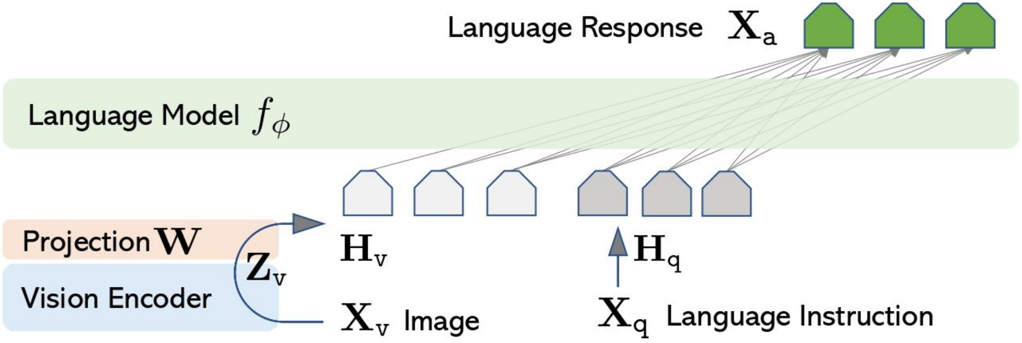
InstructBLIP [7]

MiniCPM–V

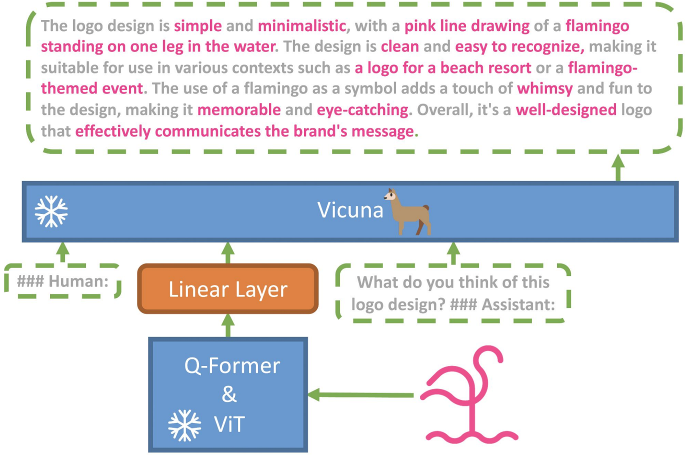
MiniCPM [8]

1. Bordes, Florian, et al. "An Introduction to Vision-Language Modeling." arXiv preprint arXiv:2405.17247 (2024).
2. https://cdn.openai.com/papers/GPTV_System_Card.pdf GPT-4V System Card
3. https://deepmind.google/technologies/gem Gemeni
4. Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2024).
5. Alayrac, Jean-Baptiste, et al. "Flamingo: a visual language model for few-shot learning." Advances in neural information processing systems (2022)
6. Zhu, Deyao, et al. "Minigpt-4: Enhancing vision-language understanding with advanced large language models." arXiv preprint arXiv:2304.10592 (2023).
7. Dai, Wenliang, et al. "Instructblip: Towards general-purpose vision-language models with instruction tuning." Advances in Neural Information Processing Systems 36 (2024).
8. https://github.com/OpenBMB/MiniCPM-V, MiniCPM

Background: Open Source VLMs



LLaVA

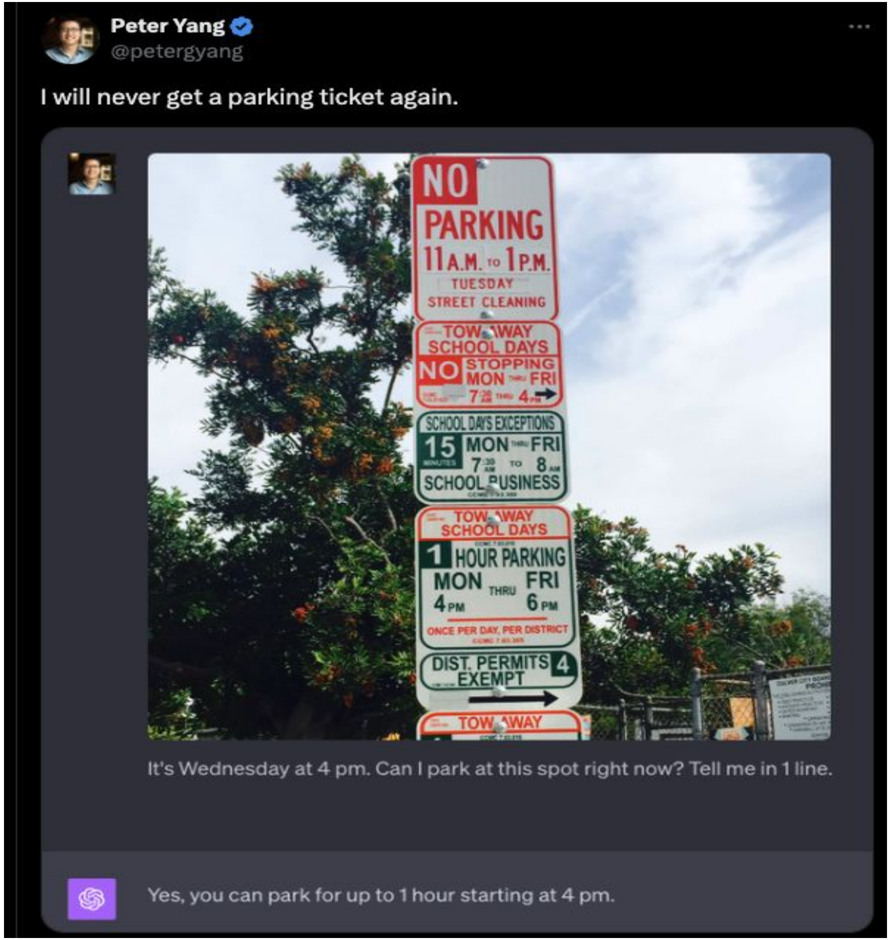


The logo design is simple and minimalistic, with a pink line drawing of a flamingo standing on one leg in the water. The design is clean and easy to recognize, making it suitable for use in various contexts such as a logo for a beach resort or a flamingo-themed event. The use of a flamingo as a symbol adds a touch of whimsy and fun to the design, making it memorable and eye-catching. Overall, it's a well-designed logo that effectively communicates the brand's message.

What do you think of this logo design? ### Assistant:

Mini-GPT4

Background: Visual Question Answering (VQA)



Peter Yang
@petergyang

I will never get a parking ticket again.

It's Wednesday at 4 pm. Can I park at this spot right now? Tell me in 1 line.

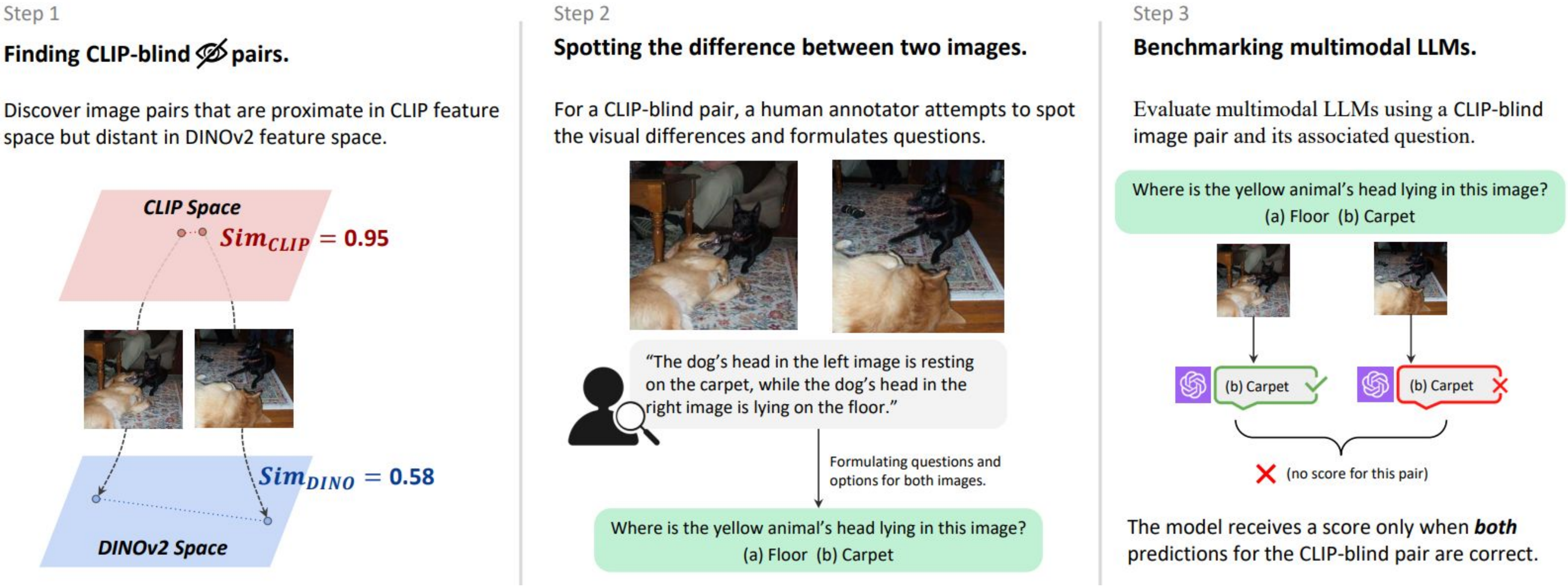
Yes, you can park for up to 1 hour starting at 4 pm.

https://x.com/petergyang/status/1707169696049668472

It seems that GPT4 is super excellent at VQA, even for some complicated tasks.

However, there are still many failure cases that are easy to human.

Background: MMVP Baseline



Step 1

Finding CLIP-blind pairs.

Discover image pairs that are proximate in CLIP feature space but distant in DINOv2 feature space.

Step 2

Spotting the difference between two images.

For a CLIP-blind pair, a human annotator attempts to spot the visual differences and formulates questions.

"The dog's head in the left image is resting on the carpet, while the dog's head in the right image is lying on the floor."

Formulating questions and options for both images.

Where is the yellow animal's head lying in this image?
(a) Floor (b) Carpet

Step 3

Benchmarking multimodal LLMs.

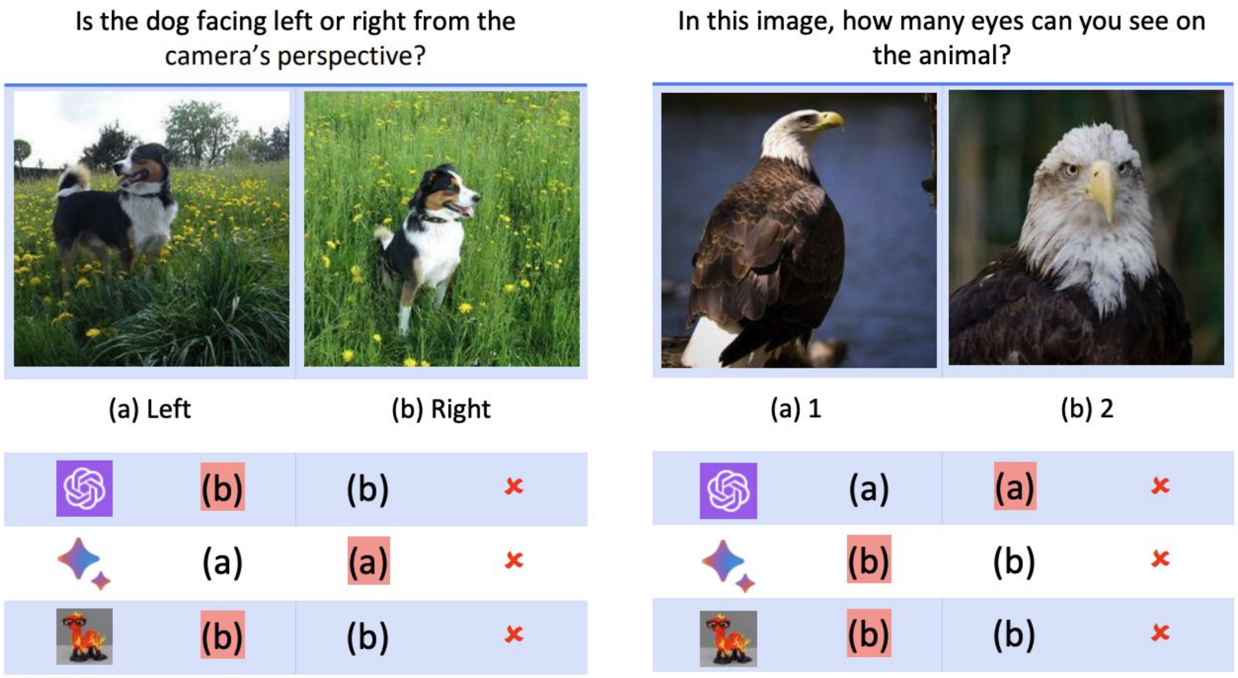
Evaluate multimodal LLMs using a CLIP-blind image pair and its associated question.

Where is the yellow animal's head lying in this image?
(a) Floor (b) Carpet

The model receives a score only when ***both*** predictions for the CLIP-blind pair are correct.

[1] Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." arXiv preprint arXiv:2401.06209 (2024).

Background: MMVP Baseline



Is the dog facing left or right from the camera's perspective?

(a) Left (b) Right

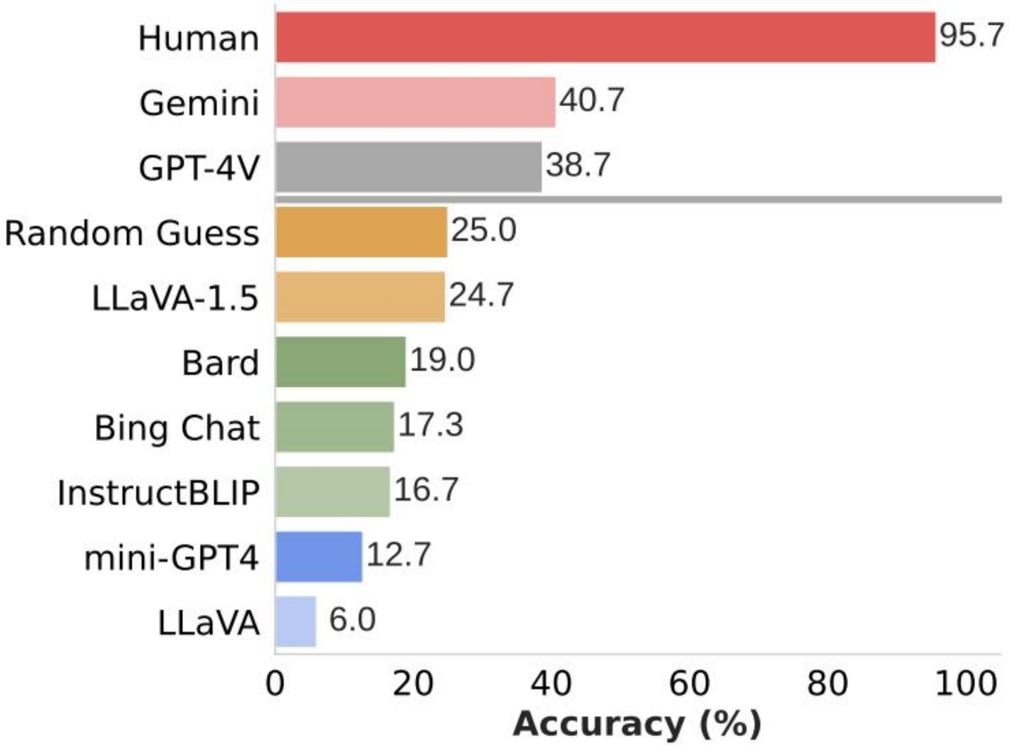
In this image, how many eyes can you see on the animal?

(a) 1 (b) 2

MMVP Baseline[1]

[1] Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." arXiv preprint arXiv:2401.06209 (2024).

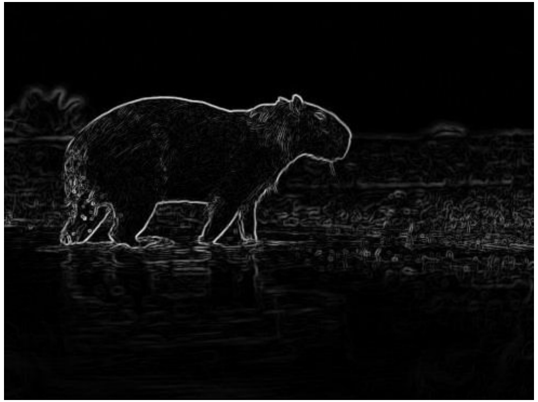
Background: MMVP Baseline



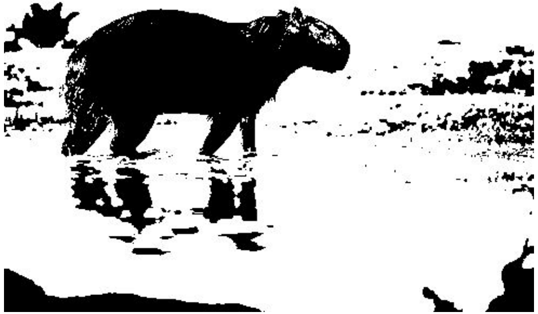
[1] Tong, Shengbang, et al. "Eyes wide shut? exploring the visual shortcomings of multimodal llms." arXiv preprint arXiv:2401.06209 (2024).

Background: Image Enhancement by Traditional Digital Image Processing Methods

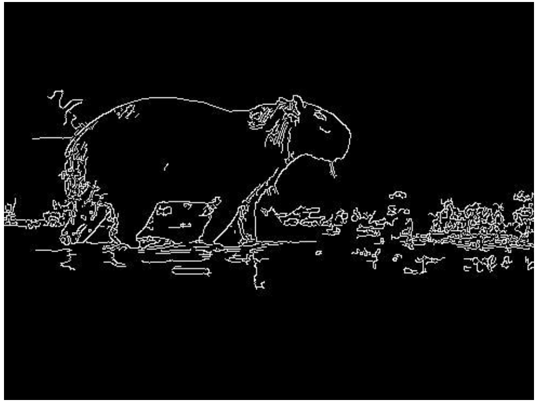
Original Capybara



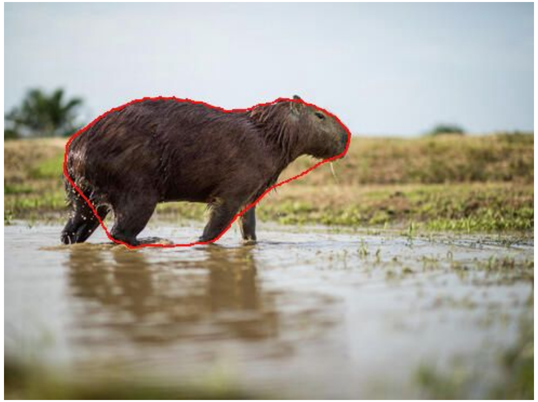
Sobel Detector



Thresholding



Canny Detector



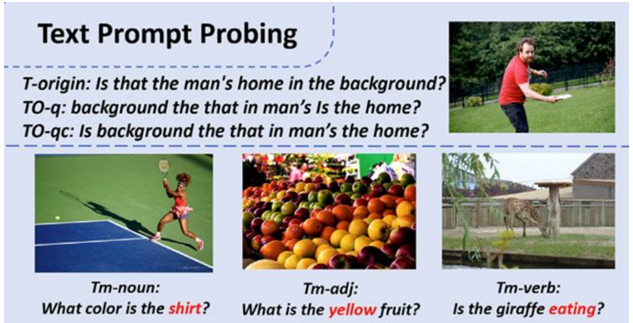
Active Contour Tracing

Potential Benefits of Using Traditional Image

- **Enhancing Image Contrast and Details:** Highlighting key areas of the image, easier for the model to detect and recognize

E-origin: [Empty] Question: What color is the boat?

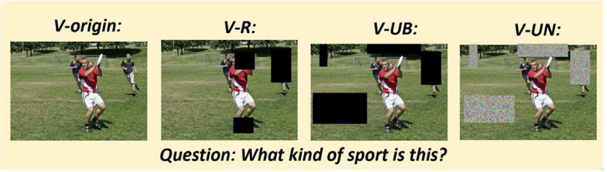
- **Multi-Scale Analysis:** Processing of images at different scales, extracting multi-level information, highly beneficial for understanding complex scenes



T-origin: Is that the man's home in the background?
TO-q: background the that in man's Is the home?
TO-qc: Is background the that in man's the home?

Tm-noun: What color is the shirt?
Tm-adj: What is the yellow fruit?
Tm-verb: Is the giraffe eating?

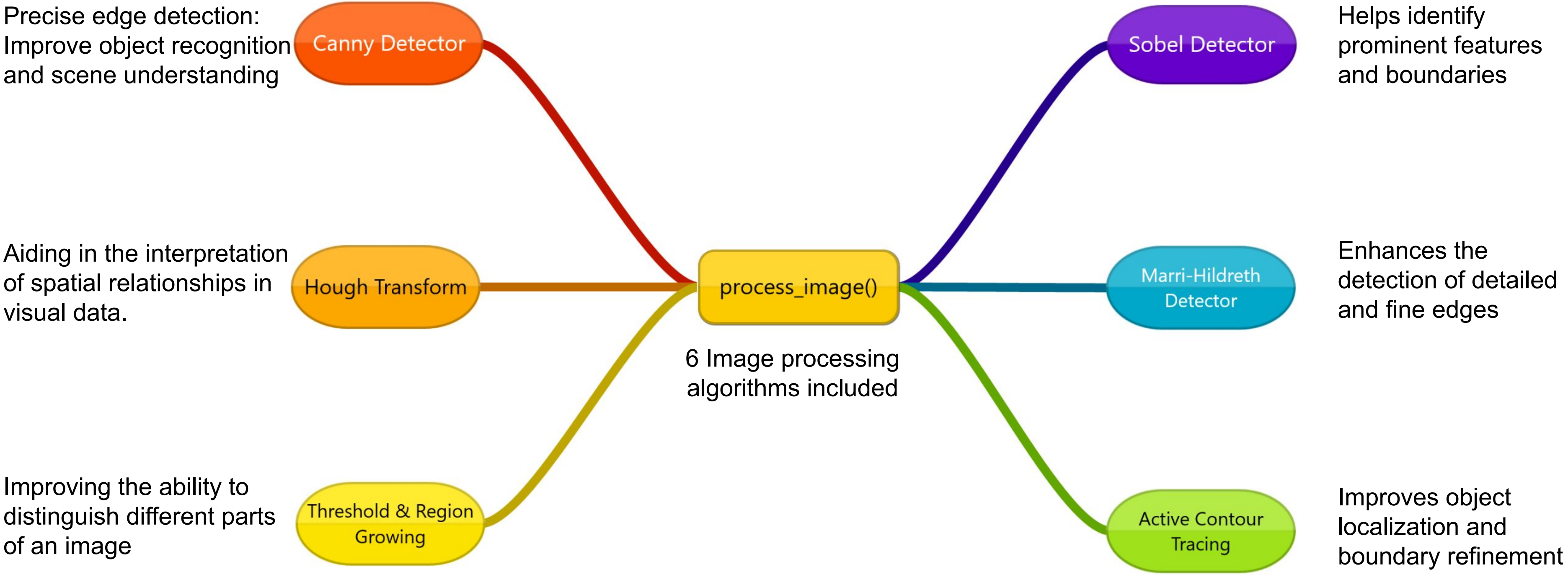
- **Improving Image Segmentation:** Enabling the model to more precisely analyze and process the image content, thereby improving the performance



Question: What kind of sport is this?

Sun, R., Lei, T., Chen, Q., Wang, Z., Du, X., Zhao, W., & Nandi, A. K. (2022). Survey of Image Edge Detection. *Frontiers in Signal Processing*, *2*. https://doi.org/10.3389/frsip.2022.826967

process_image(input_image, method)



process_image()

6 Image processing algorithms included

Canny Detector: Precise edge detection: Improve object recognition and scene understanding

Hough Transform: Aiding in the interpretation of spatial relationships in visual data.

Threshold & Region Growing: Improving the ability to distinguish different parts of an image

Sobel Detector: Helps identify prominent features and boundaries

Marri-Hildreth Detector: Enhances the detection of detailed and fine edges

Active Contour Tracing: Improves object localization and boundary refinement

This function provides a series of image processing algorithms.
Calling different algorithms as needed and pass the corresponding parameters when calling.

Algorithm Evaluation

- Dataset: **BSDS300 (Berkeley Segmentation Dataset and Benchmark)** Widely used for evaluating image processing algorithms, especially for edge detection and segmentation.

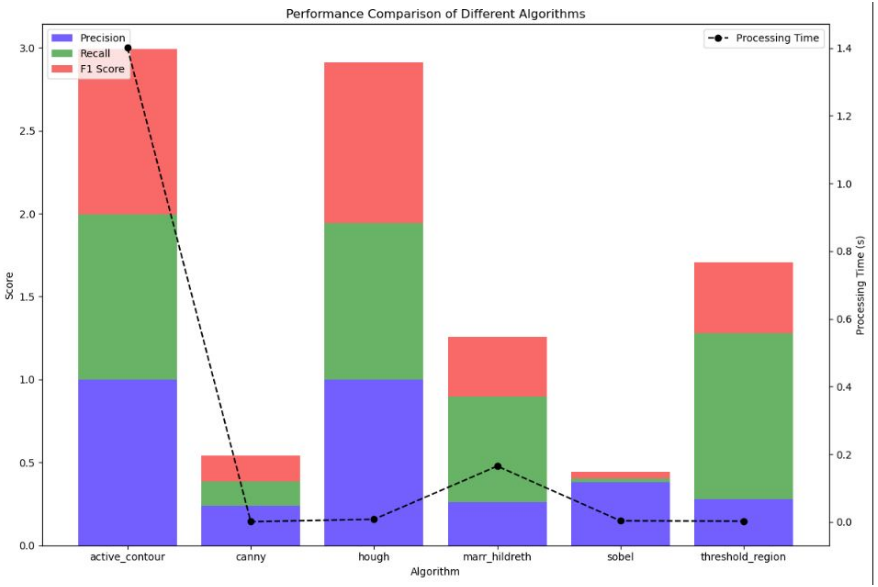
- Ideal for evaluating image processing algorithms due to its versatility, high quality annotations, standard evaluation benchmarks, and widespread use
- 100 images in test set are used

Test Results

Hough transform
Performs excellently across all performance metrics
Highly suitable for scenarios requiring precise edge detection.

Active Contour
Best performance, achieving top results in almost all metrics
longer processing time

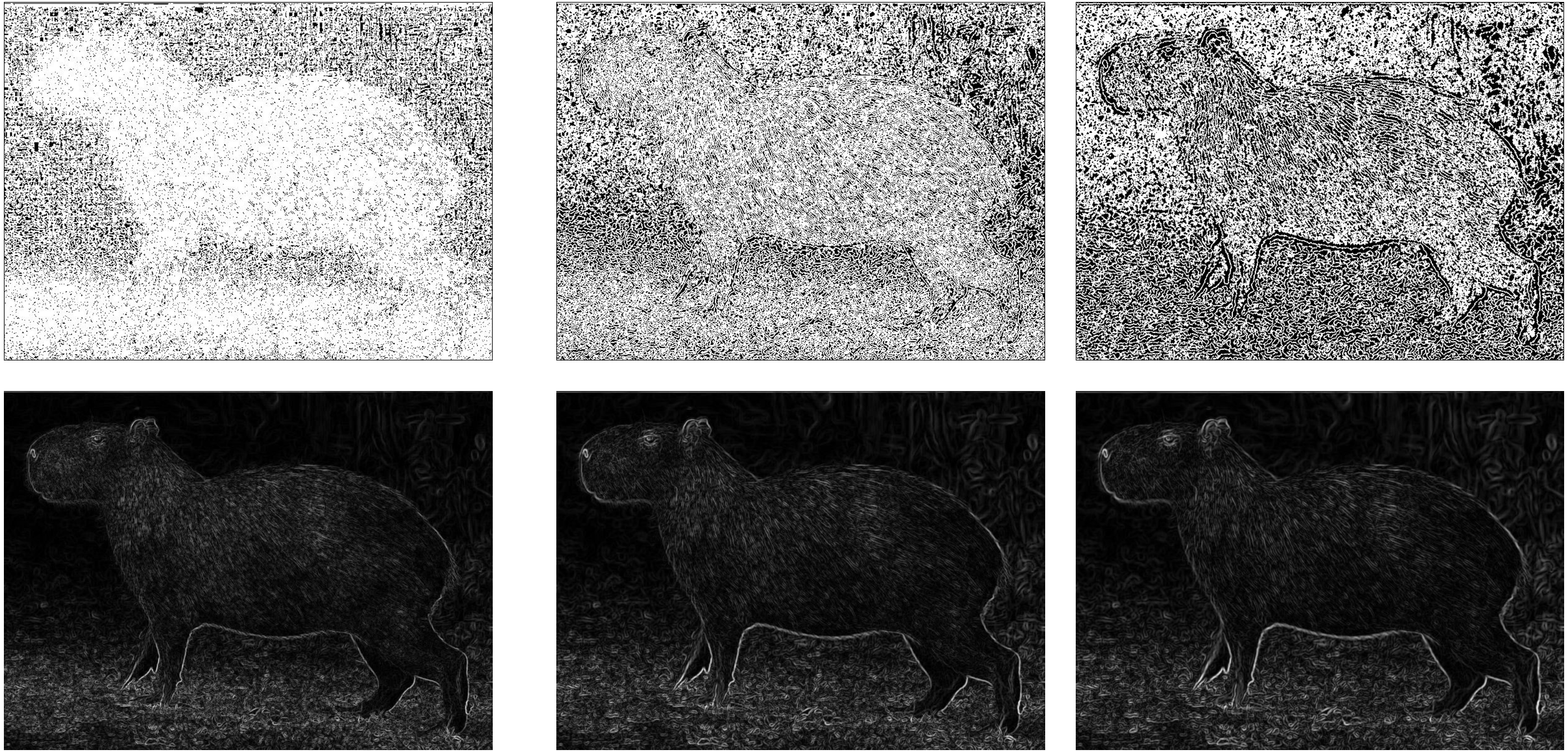
Algorithm	Precision	Recall	F1 Score	Processing Time (s)
Active_contour	0.995647	0.994959	0.99747	1.400346
Canny	0.237822	0.149056	0.155135	0.000999
Hough	0.984124	0.942315	0.969973	0.007831
Marr_hildreth	0.261623	0.636124	0.360671	0.164979
Sobel	0.379839	0.023172	0.042099	0.003503
Threshold_region	0.280154	0.998114	0.428306	0.002500



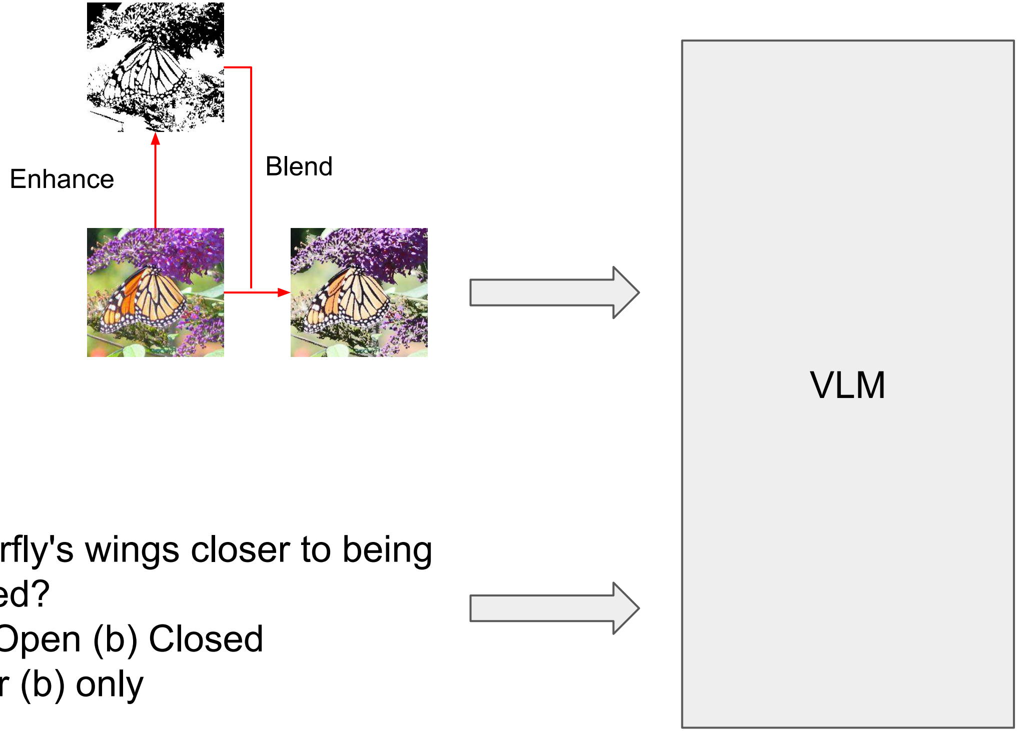
Canny algorithm
It has a significant advantage in speed, although its precision and F1 score are not as high as Hough and Active Contour. It may be the better choice for real-time processing applications.

Sobel and Threshold
Average performance in precision and F1 score but have faster processing speeds, making them suitable for quick preliminary analysis.

Test Results



Our Idea



Are the butterfly's wings closer to being open or closed?
Options: (a) Open (b) Closed
Answer (a) or (b) only

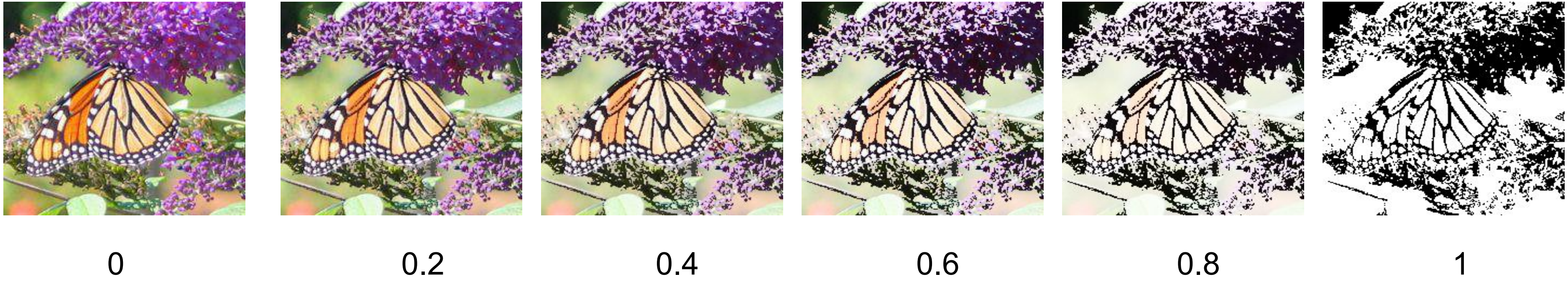
Motivation: Enhance the visual features of the input image may boost the reasoning ability of VLMs

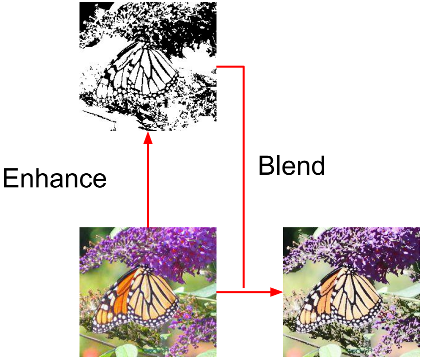
Results

	original image	sobel	canny	marr_hildreth	threshold_region	active_contour
LLaVANeX-Mixtral-7B	**38.00%**	33.33%	33.33%	22.00%	35.33%	37.33%
LLaVA-LLaMA2-7B	**14.66%**	13.33%	13.33%	12.00%	13.33%	12.66%
LLaVA-LLaMA2-13B	20.66%	21.33%	**22.66%**	18.66%	19.33%	20.66%
MiniCPM-LLaMA3-7B	**34.66%**	26.00%	26.66%	12.00%	24.66%	28.66%
GPT4	**42.67%**	X	X	X	40.67%	38.00%

Random Guess: 25%

Blend Factor





Results for Different Blend Factors

LLaVANeX-Mixtral-7B	0 (original image)	0.2	0.4	0.6	0.8	1
canny	**38.00%**	31.33%	30.66%	29.33%	23.33%	16.00%
sobel	**38.00%**	38.00%	36.00%	30.00%	23.33%	15.33%
threshold_region	38.00%	38.00%	**40.00%**	34.66%	28.66%	15.33%
active_contour	**38.00%**	35.33%	36.66%	36.66%	38.00%	38.00%

Credits

Yihao Wu:

- Image processing algorithm design, coding
- Testing, benchmarking
- Test results analyzation

Zhiyu Xue:

- Implement VLM models
- Evaluation pipeline on MMVP with image enhancement

Thank you for Listening!