Visual Haystacks: A Vision-Centric Needle-In-A-Haystack Benchmark

Tsung-Han (Patrick) Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta,
Joseph E. Gonzalez, Trevor Darrell, David M. Chan

UC Berkeley

ICLR 2025

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

ICLR
International Conference On Learning Representations

Many visual problems requires long-context reasoning

Applications: Large-Scale Visual Question Answering

Many problems require answering questions about large **sets of images.**

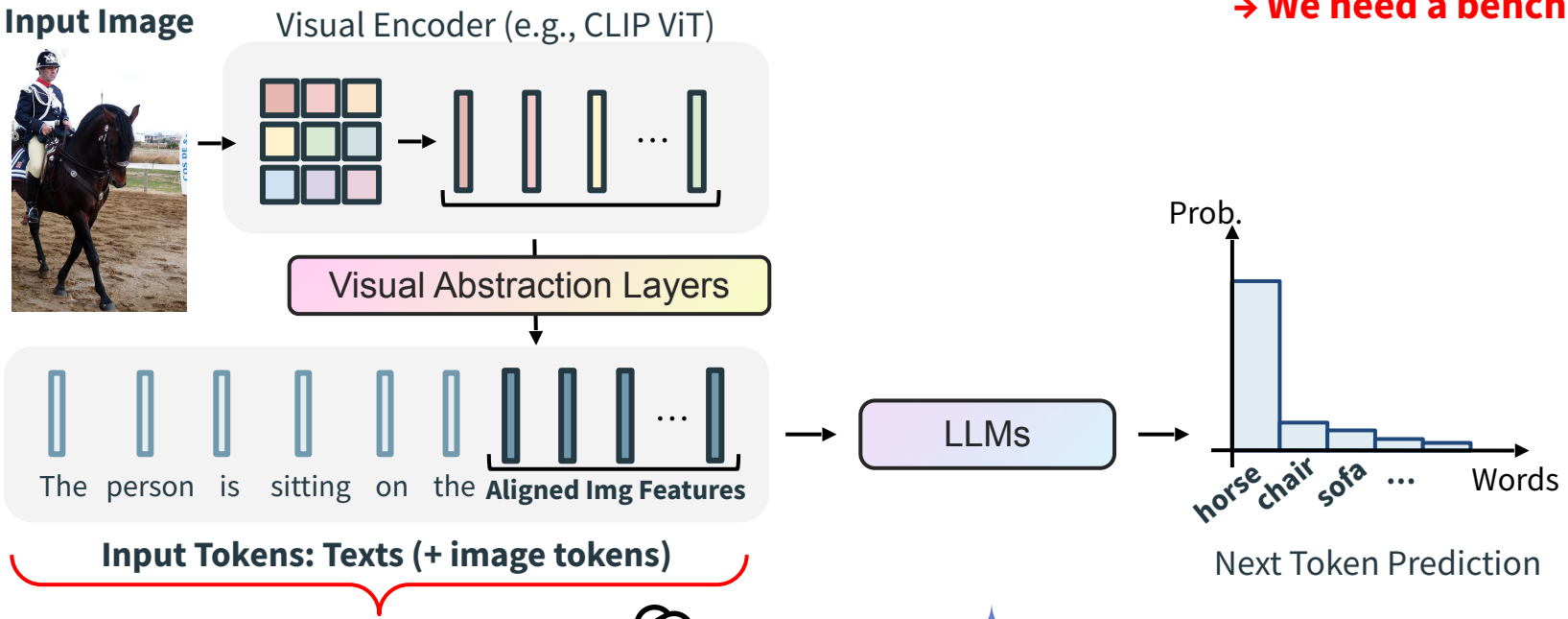
- Answering questions about photo albums and collections.
- Analyzing patterns in medical imagery.
- Monitoring climate change and deforestation from satellite images.
- Understanding consumer behavior from retail surveillance footage.
- etc.

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Background: Large Multimodal Models (LMMs)

Does taking more visual tokens lead to improved results in large-scale multi-image QA?

→ We need a benchmark!

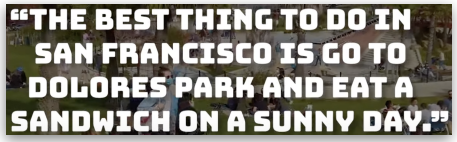


Large Context Window GPT-4o: 128K Gemini 1.5: 2M

Needle-In-A-Haystack (NIAH) Benchmark

NIAH is a common benchmark to evaluate long-context models in retrieval and reasoning.

Random information



Paul Graham Essays

Q: What is the best thing to do in SF?

Information to retrieve (**Needle**)

Collection of data (**Haystack**)

Test if model can correctly retrieve the information and answer the question.

https://www.youtube.com/watch?v=KwRRuiCCdmc

Prior Visual NIAH

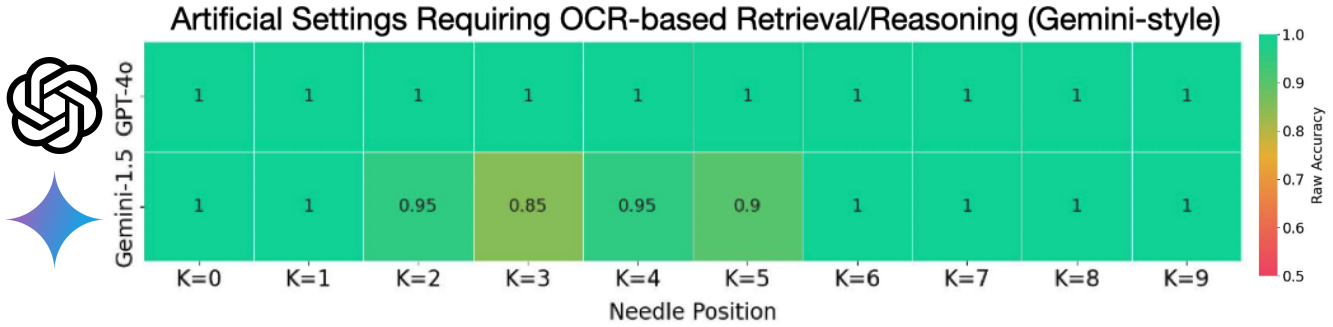
Gemini-style Challenge

Insert the image to the haystack

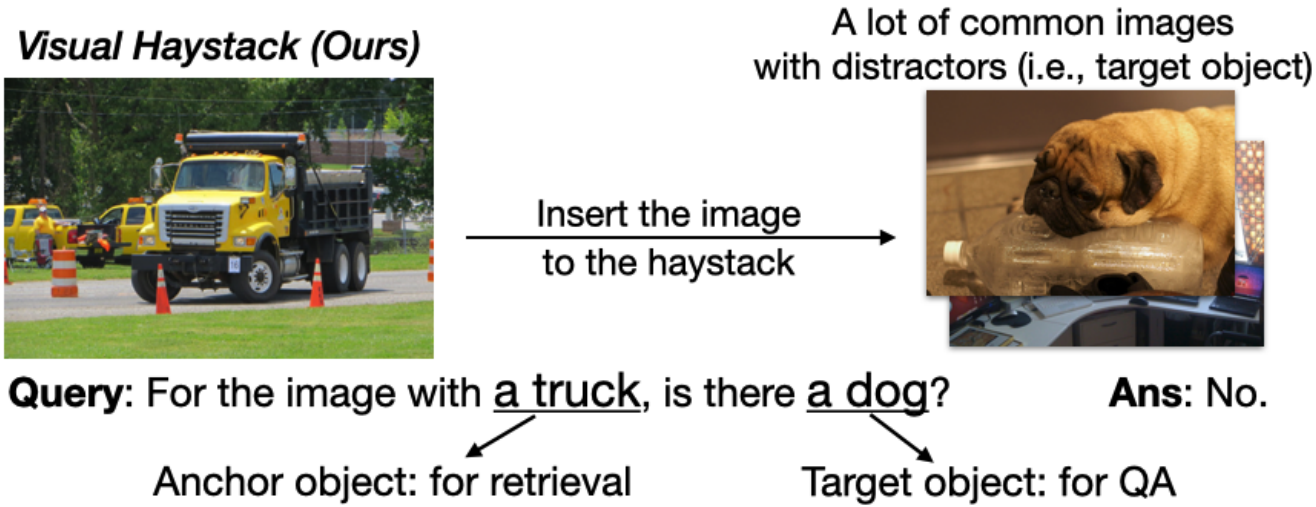
A lot of common images

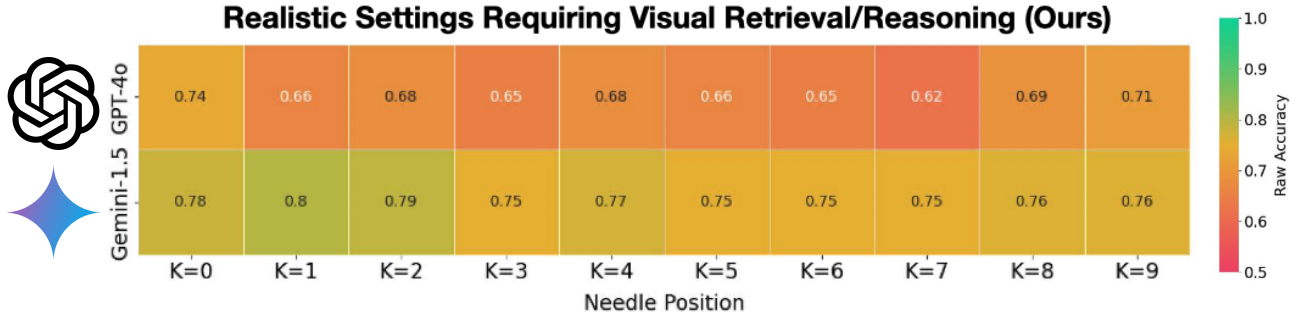
Query: What is the secret word?

Ans: {needle word}.

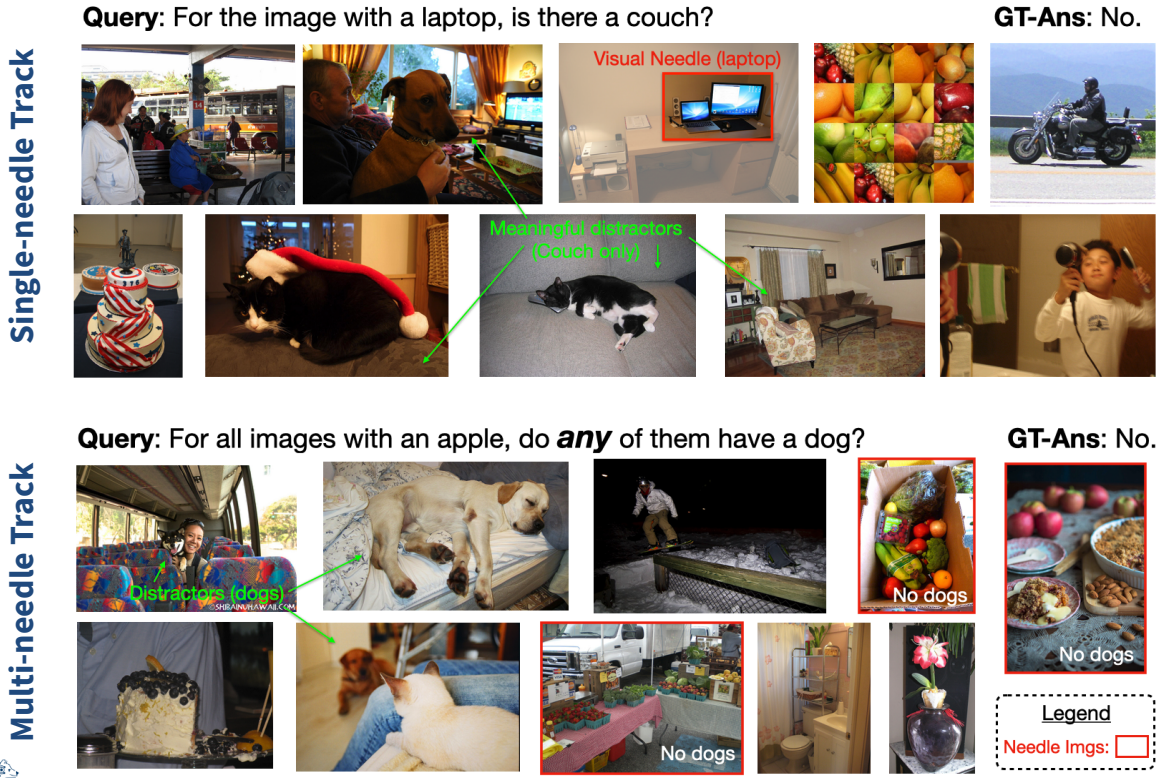


Visual Haystacks (VHs)





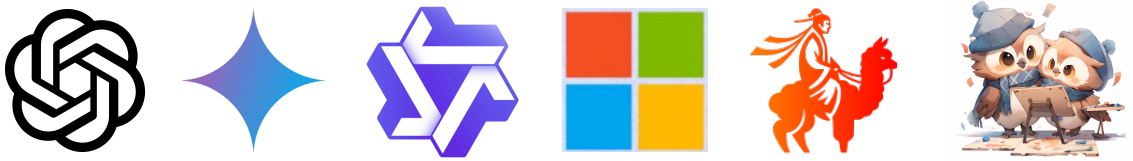
Simple LMM unit-test with in-domain images/texts



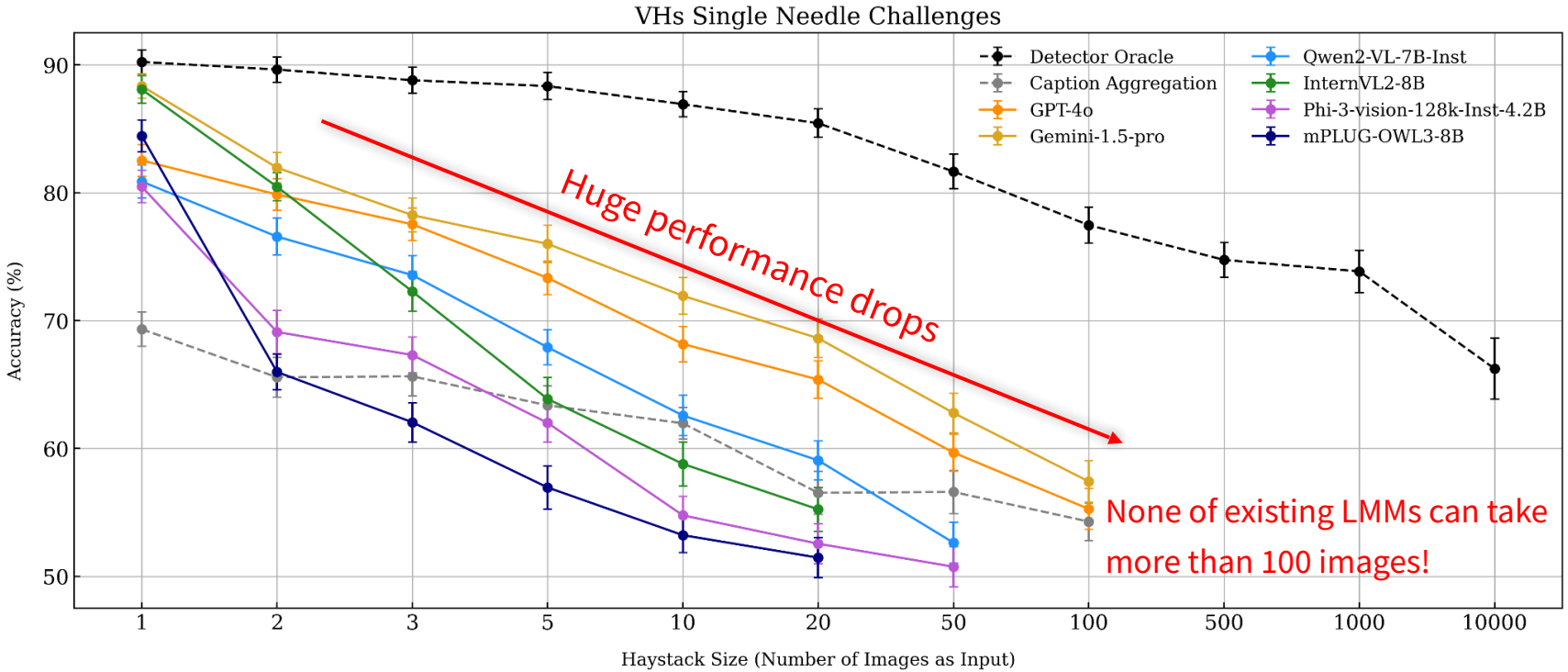
- COCO images
- 1,000 diverse test cases
- Binary and easy question
- 50% Acc if randomly guessing

VHs should be easy, huh?

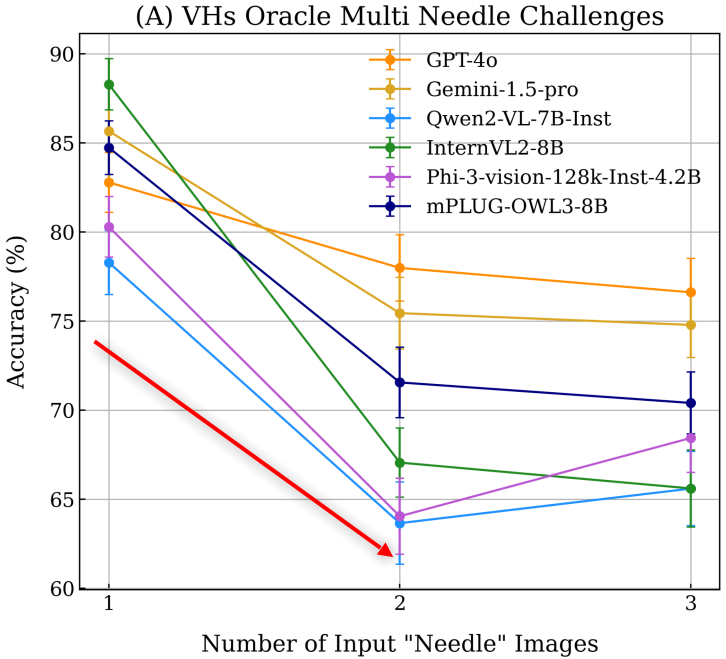
Can LMMs handle long-context “visual” inputs?

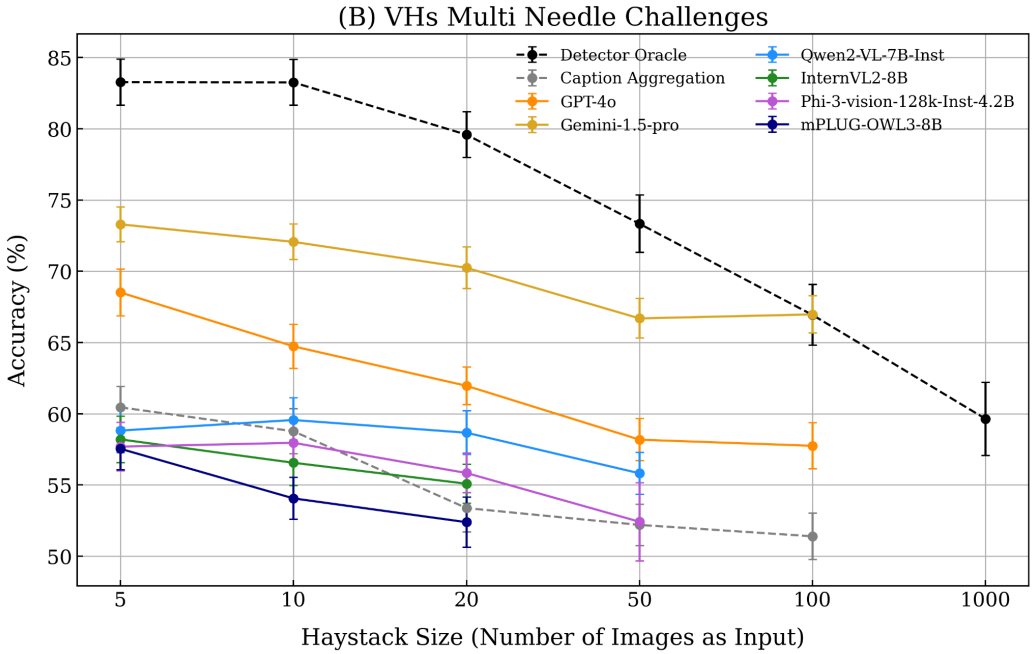


Finding #1: Susceptible to visual distractors



Finding #2: Difficulty in cross-image reasoning

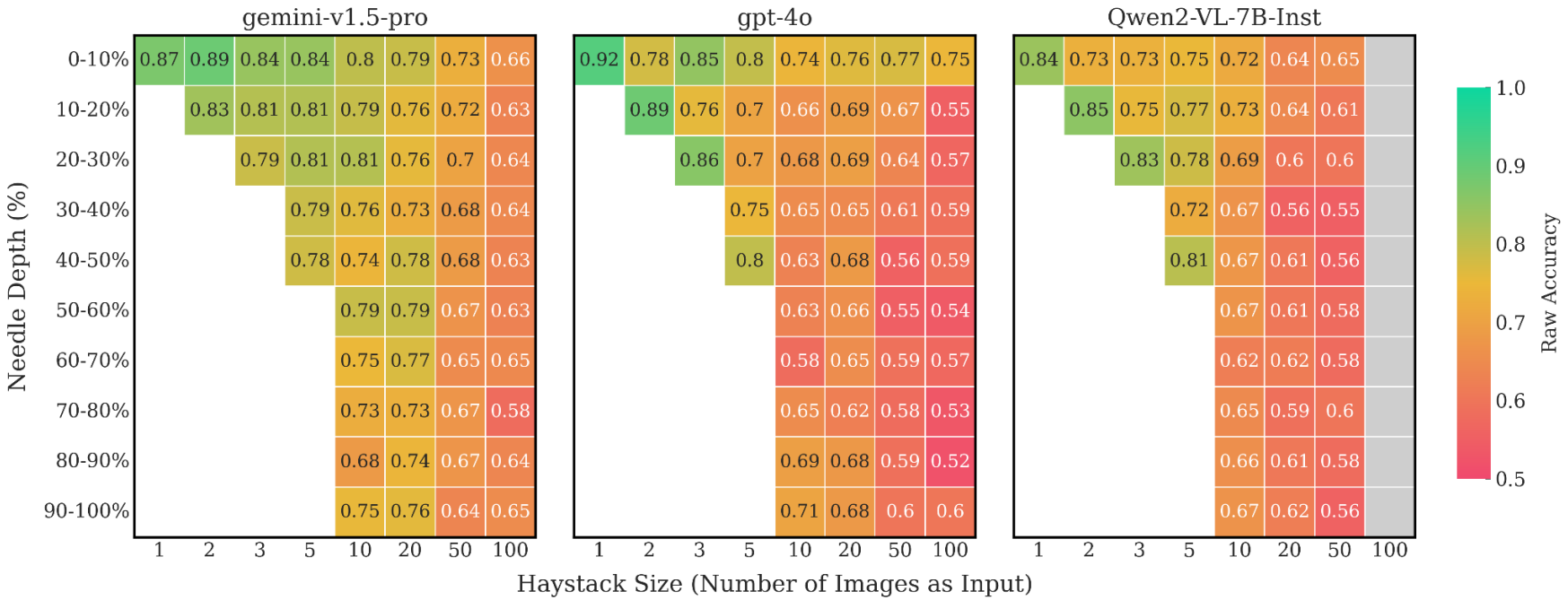




Huge performance drops when doing cross-image reasoning

There's room for improvement in how LMMs do long-context "retrieval" and "reasoning"

Finding #3: Positional Biases

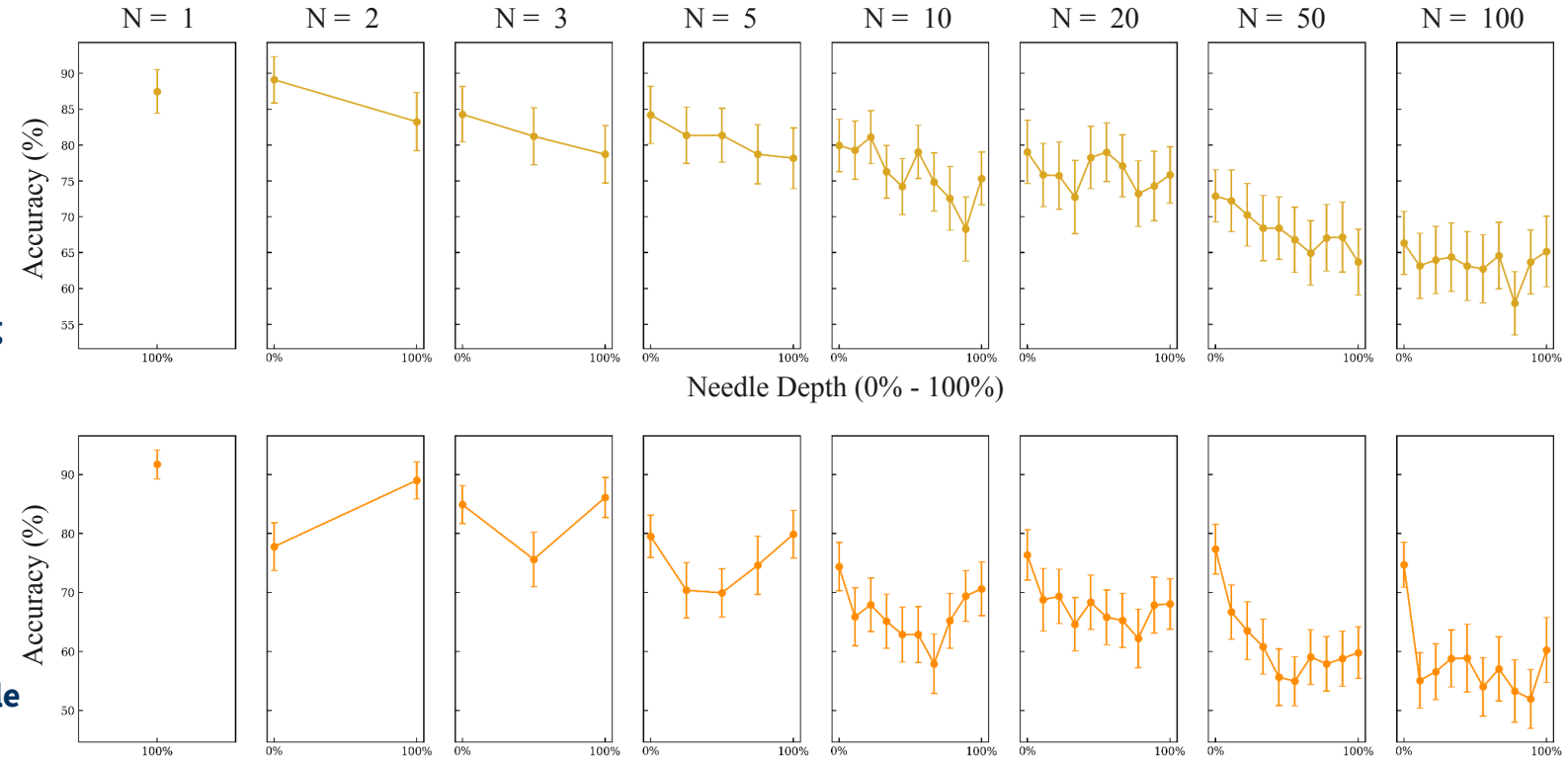


Huge performance drop if the key information is not placed at the optimal position

Take a closer look...

Prefers images at the beginning

Lost-in-the-middle



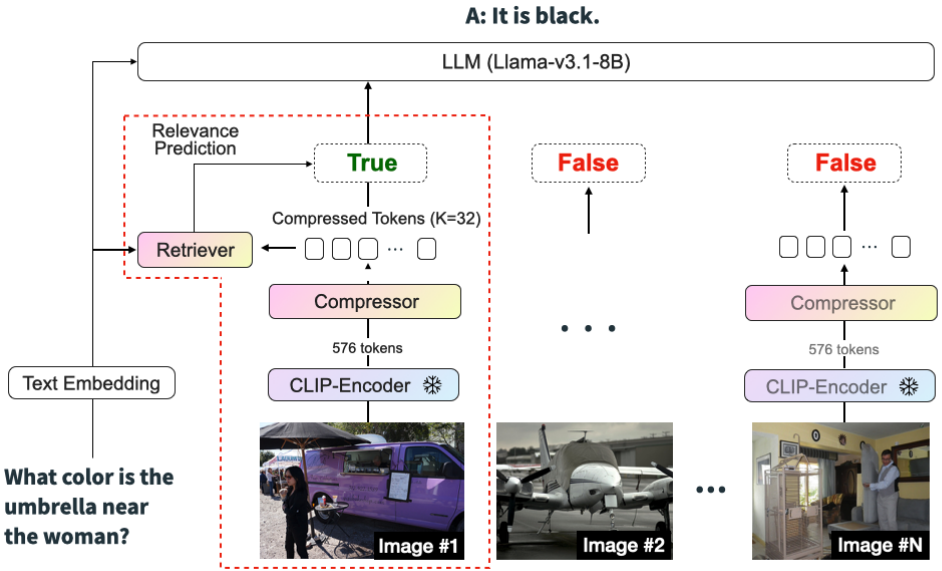
BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Can we do better at large-scale multi-image QA?

MIRAGE: Multi-Image Retrieval Augmented GEneration

We develop the first **visual-RAG** framework that can…

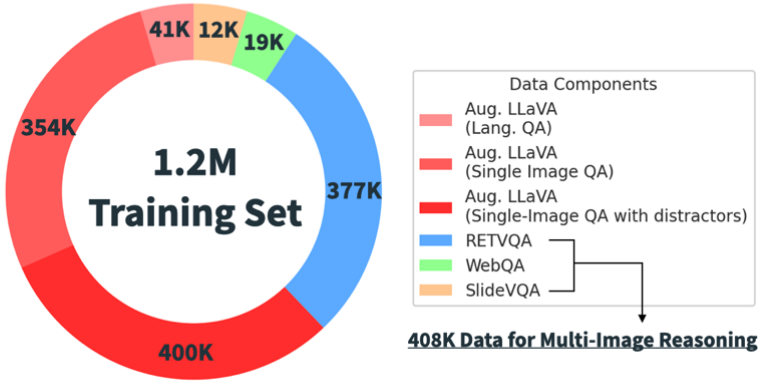
(1) retrieve and reason across multiple images (2) scale to **tens of thousands of images**



(A) MIRAGE: A simple LMM framework with a compressor and a retriever for MIQA tasks

Number of Images/Question: 0~30

Number of Relevant Images/Question: 1~3

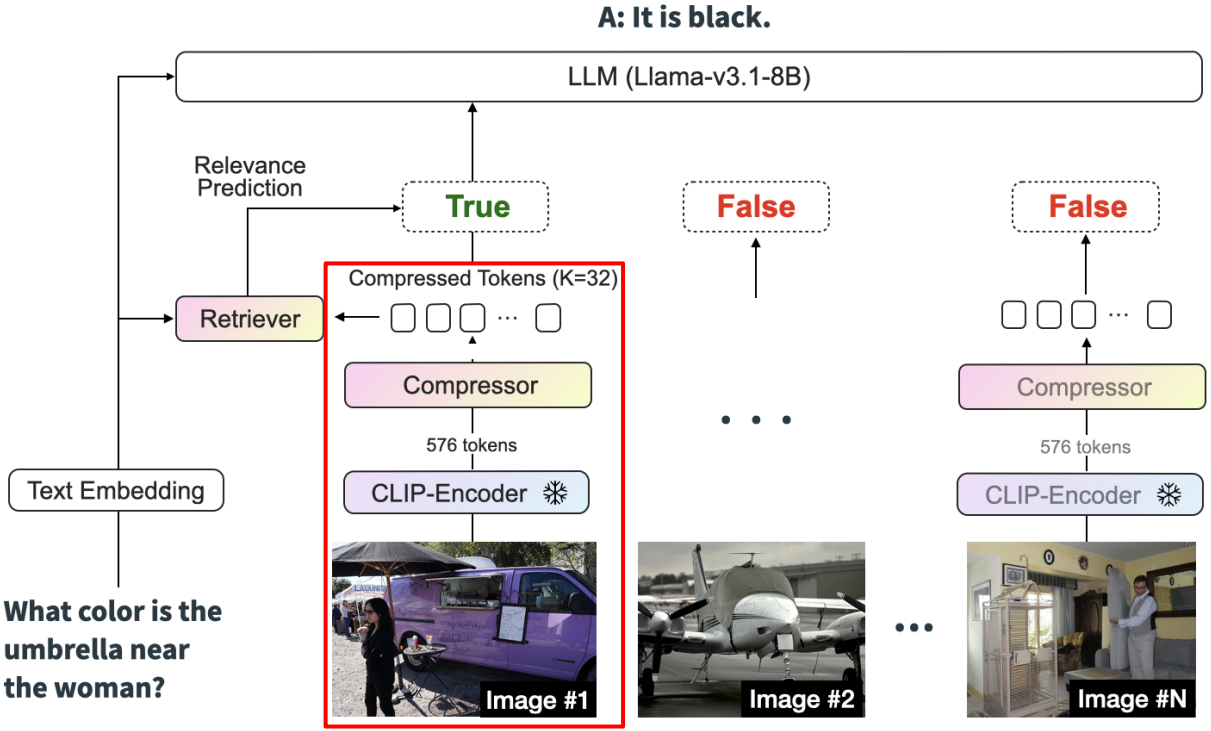


(B) Visualization of our augmented instruction tuning data

MIRAGE: Multi-Image Retrieval Augmented GEneration

Stage 1: Compress visual tokens

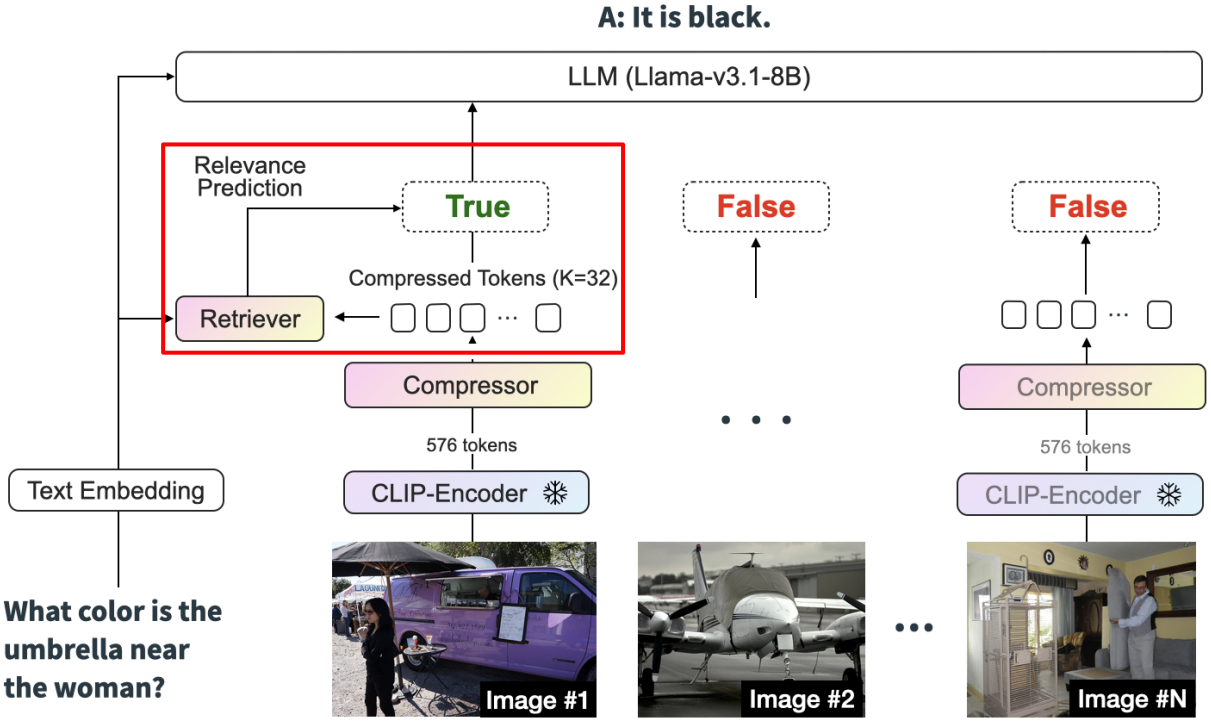
MIRAGE reduces the number of tokens from 576 to 32 (18x reduction, meaning that **we can take 18x more images!**)



MIRAGE: Multi-Image Retrieval Augmented GEneration

Stage 2: Drop irrelevant images

MIRAGE uses a retriever trained in-line with the LLM fine-tuning, to predict if an image will be relevant, and **dynamically drop** irrelevant images.



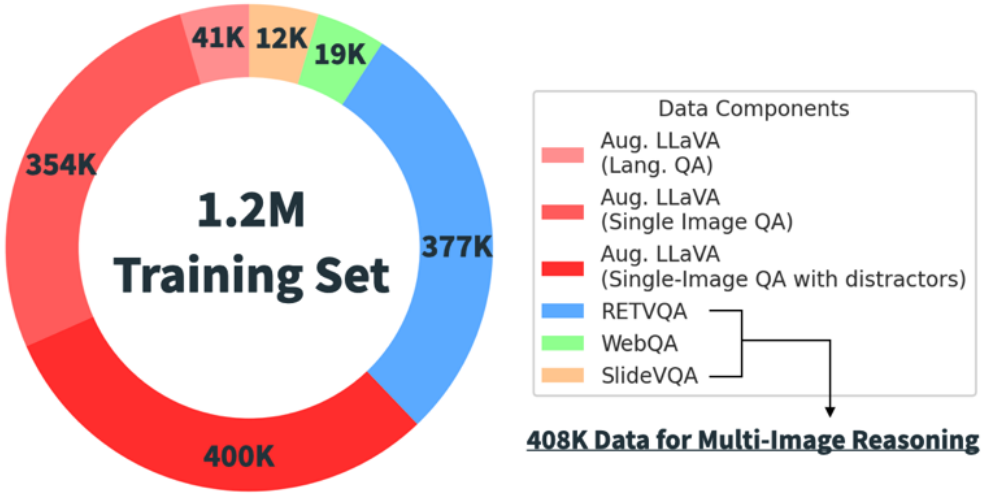
MIRAGE: Multi-Image Retrieval Augmented GEneration

Stage 3: Instruction Finetuning

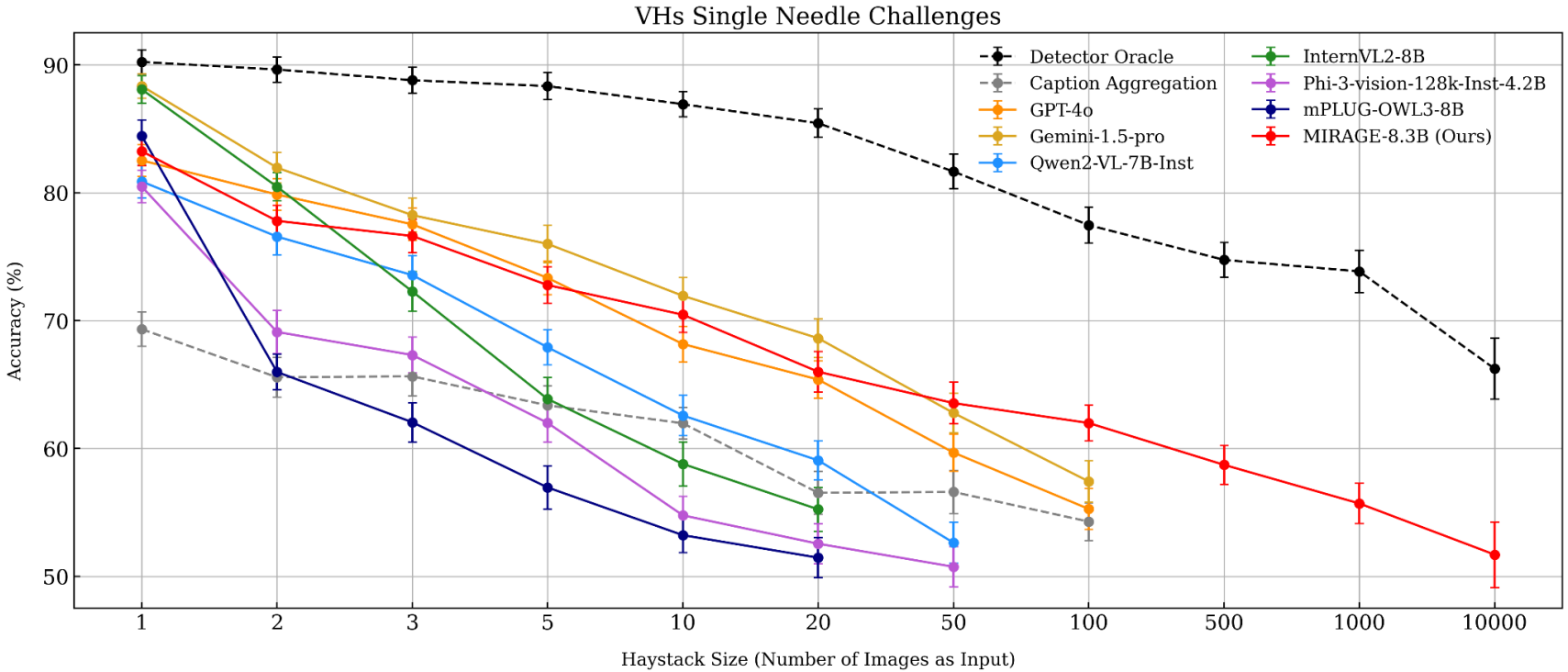
MIRAGE augments existing single-image training data with multi-image reasoning data, and synthetic multi-image reasoning data.

Number of Images/Question: 0~30

Number of Relevant Images/Question: 1~3

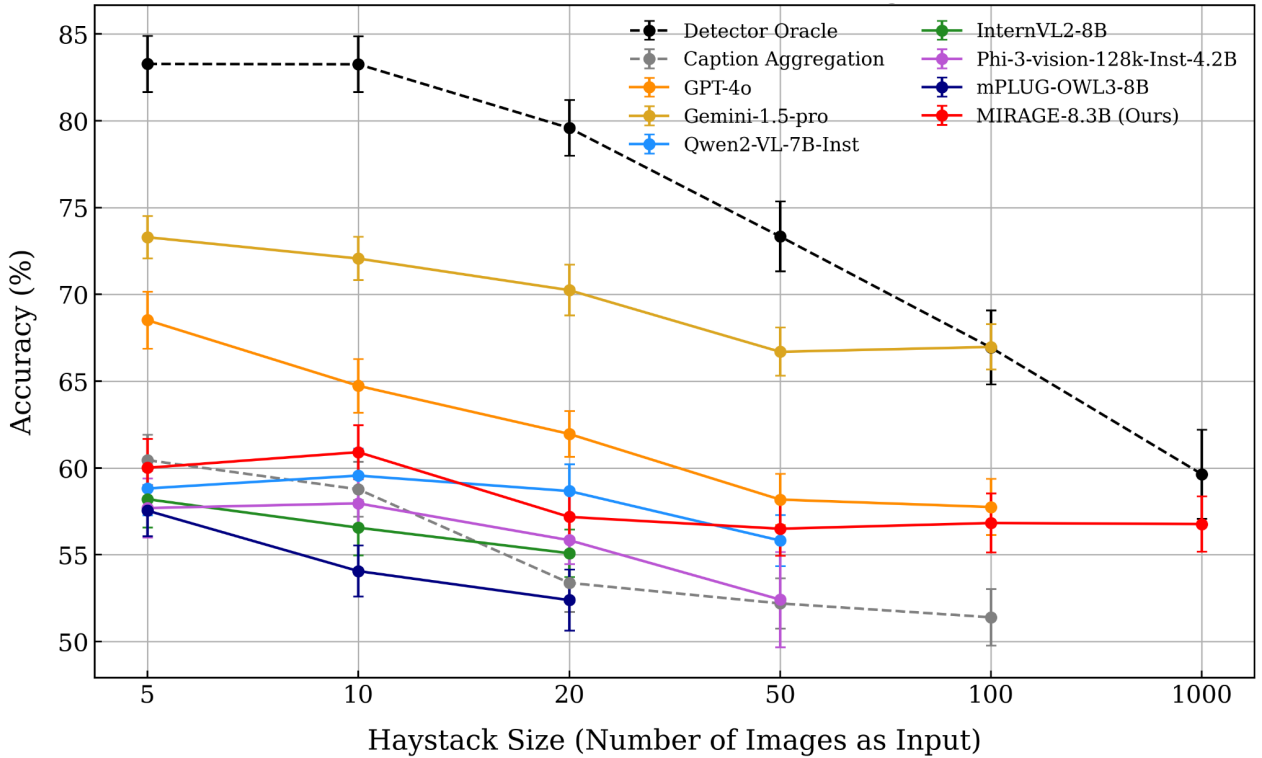


Visual Haystacks – Single-needle Track



Scale to 10K images and surpass GPT-4o / Gemini 1.5 Pro when taking 50+ images

Visual Haystacks – Multi-needle Track



There's still room for improvement in both "data" and "model design"

Takeaways

➔ We introduce a vision-centric, realistic long-context benchmark for LMMs, called Visual Haystacks (VHs), much better than existing ones.

Takeaways

- ➔ We introduce a vision-centric, realistic long-context benchmark for LMMs, called Visual Haystacks (VHs), much better than existing ones.
- ➔ Current LMMs are prone to visual distractions, struggle with cross-image reasoning, and exhibit positional biases.

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Takeaways

- ➔ We introduce a vision-centric, realistic long-context benchmark for LMMs, called Visual Haystacks (VHs), much better than existing ones.
- ➔ Current LMMs are prone to visual distractions, struggle with cross-image reasoning, and exhibit positional biases.
- ➔ We present MIRAGE, the pioneering visual-RAG framework capable of scaling up to 10K images and mitigate some of existing challenges.

Takeaways

- ➔ We introduce a vision-centric, realistic long-context benchmark for LMMs, called Visual Haystacks (VHs), much better than existing ones.
- ➔ Current LMMs are prone to visual distractions, struggle with cross-image reasoning, and exhibit positional biases.
- ➔ We present MIRAGE, the pioneering visual-RAG framework capable of scaling up to 10K images and mitigate some of existing challenges.
- ➔ VHs and MIRAGE are not endpoints but starting points that invite the community to deepen and expand the exploration of long-context LMMs.

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Project Page

Takeaways

- ➔ We introduce a vision-centric, realistic long-context benchmark for LMMs, called Visual Haystacks (VHs), much better than existing ones.
- ➔ Current LMMs are prone to visual distractions, struggle with cross-image reasoning, and exhibit positional biases.
- ➔ We present MIRAGE, the pioneering visual-RAG framework capable of scaling up to 10K images and mitigate some of existing challenges.
- ➔ VHs and MIRAGE are not endpoints but starting points that invite the community to deepen and expand the exploration of long-context LMMs.

Thanks for listening!

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH