“Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Sanjeev Raja

September 23, 2024

Berkeley
UNIVERSITY OF CALIFORNIA

Motivation

VQAv2:
General Visual Reasoning

Q: What is the mustache made of?
A: Bananas

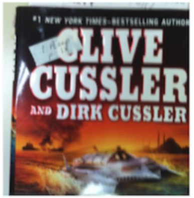
TextVQA:
Text-Based Reasoning

Q: What is the price of the bananas per kg? A: $11.98

GQA:
Spatial Reasoning

Q: What is the animal sitting on the sidewalk? A: Bear

VizWiz:
Unanswerable Questions



Q: What is the name of this book? A: Unanswerable

Considerable progress in **single-image VQA** with VLMs

“Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models”

Motivation

But SOTA models still struggle on Multi-Image VQA (MIQA)

Video (4000 frames): Valley-Instruct-73k + Video-Instruct 100k

t = 0s　　　　　t = 10min

User: Could you provide a brief summary of the employee's actions? **Assistant**: In the video, an employee prepares a sub. After assembling the bread, ham, pepperoni, salami, and cheese, he toasts the sub in the oven...

Context: 1M
Tokens: 200M
Examples: 173K

“WORLD MODEL ON MILLION-LENGTH VIDEO AND LANGUAGE WITH BLOCKWISE RING ATTENTION”

Needle in a Haystack (NIAH) Benchmarks

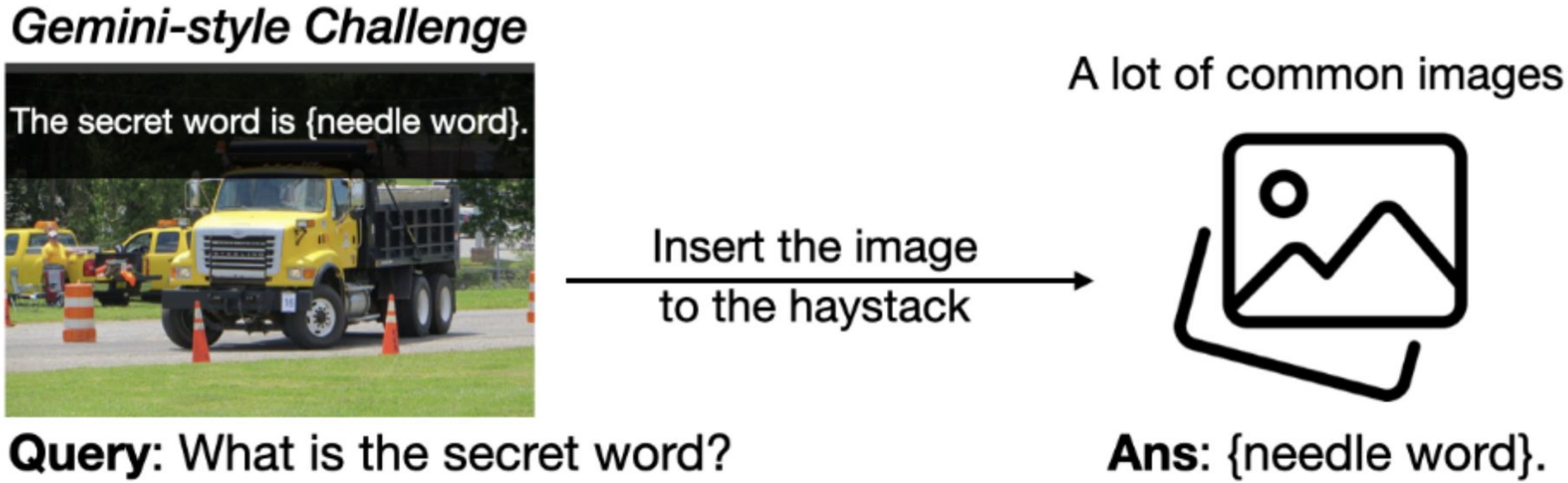
Information to Retrieve (Needle)

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

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Motivation

Existing benchmarks focus excessively on **textual reasoning and OCR capabilities**



Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Improving Existing Benchmarks

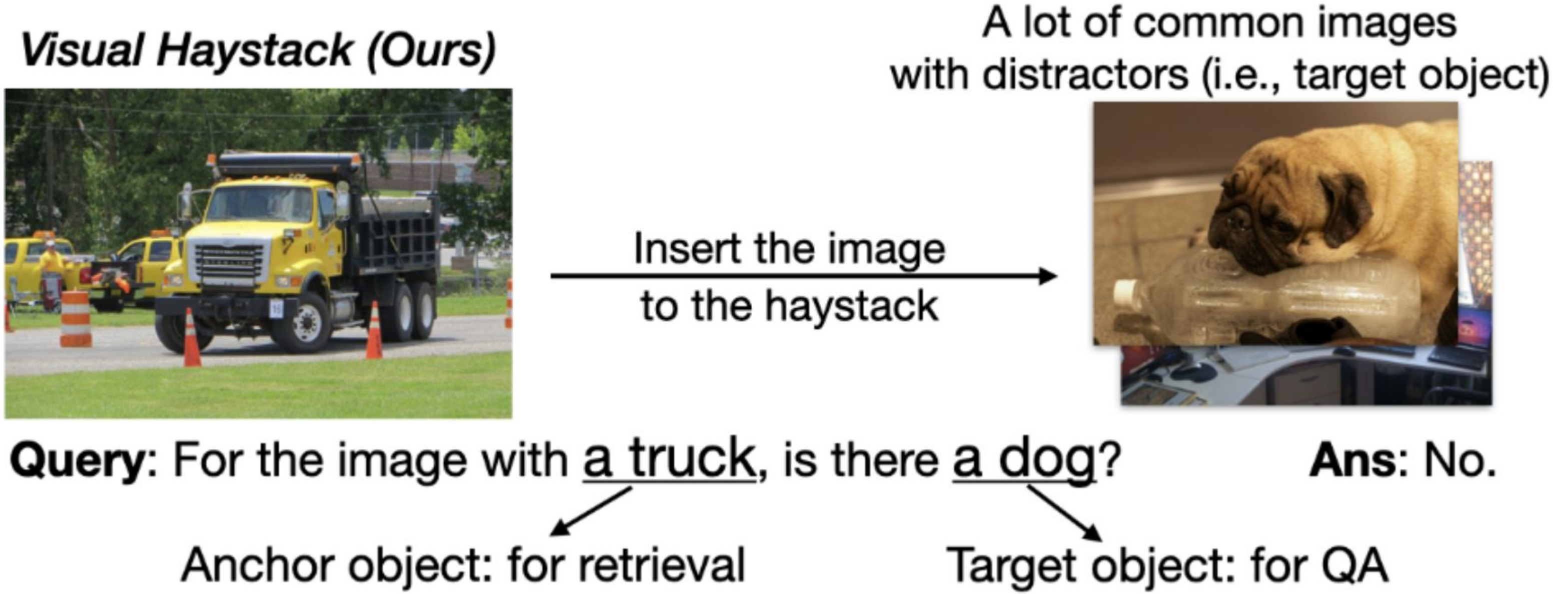
A good benchmark for visual long context learning should require a model to

1) **Retrieve** relevant image(s) from a vast collection
2) **Reason visually** over the retrieved image(s)

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

Visual Haystack Benchmark



Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

Visual Haystack Benchmark

Two variants

1) **Single-Needle:** “For the image with <anchor> object, is there <target> object?”
2) **Multi-Needle:** ““For all images with <anchor> object, do all/any of them contain <target> object?”

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Dataset Construction

Images sourced from COCO dataset

Each haystack contains

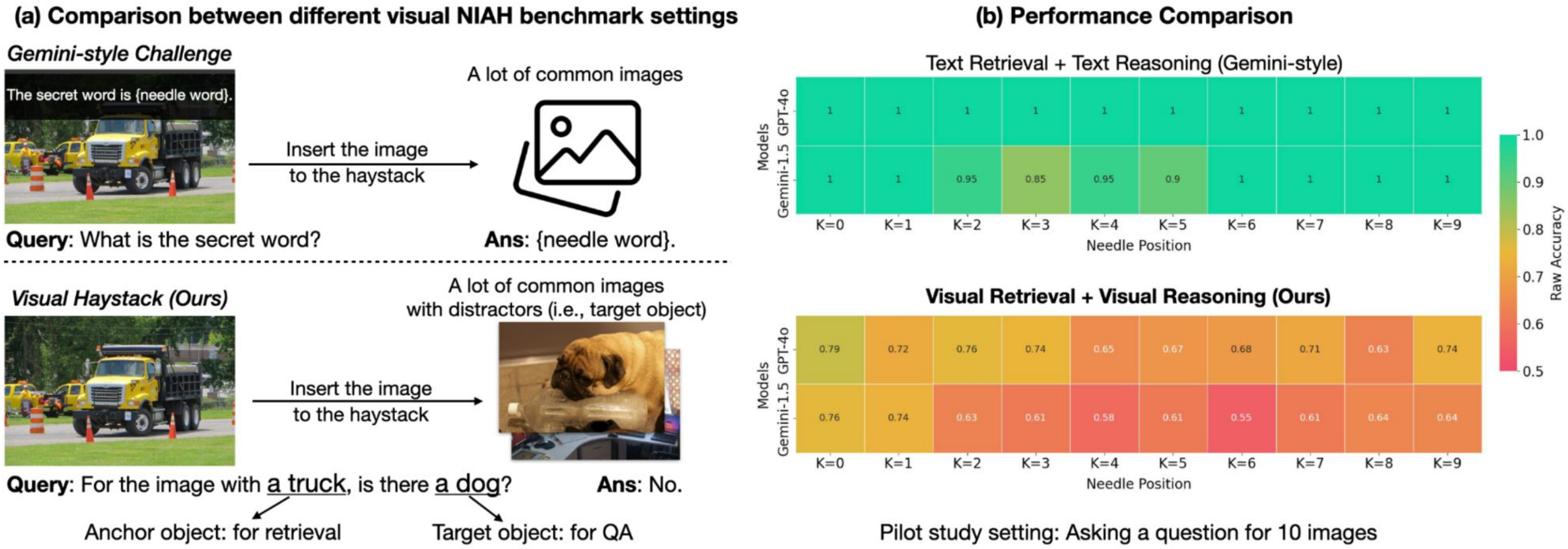
1) 1-5 needle images (with anchor)
2) 1-10k negative images (no anchor, some with target)

Total size: 880 single-needle, 1k multi-needle questions

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

Task is Challenging for SOTA Models



Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

Performance as a Function of Haystack Size

Method	Tokens/Img	N=1 (Oracle)	N=3	N=5	N=10	N=50	N=100	N=1K	N=10K
Naive									
Question Only (LLama3)	-	0.52	-	-	-	-	-	-	-
Caption-Based (LLaVA + LLama3)	576	0.79	0.67	0.69	0.68	0.59	E	E	E
LMM									
LLaVA-v1.5-7B	576	**0.87**	0.70	E	E	E	E	E	E
Claude-3 Opus	≈64	0.67	0.54	0.51	0.47	E	E	E	E
Gemini-1.5	≈258	**0.87**	0.73	0.68	0.64	0.58	**0.59**	E	E
GPT-4o (low-res)	≈85	0.82	0.68	0.68	0.64	0.57	0.53	E	E

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

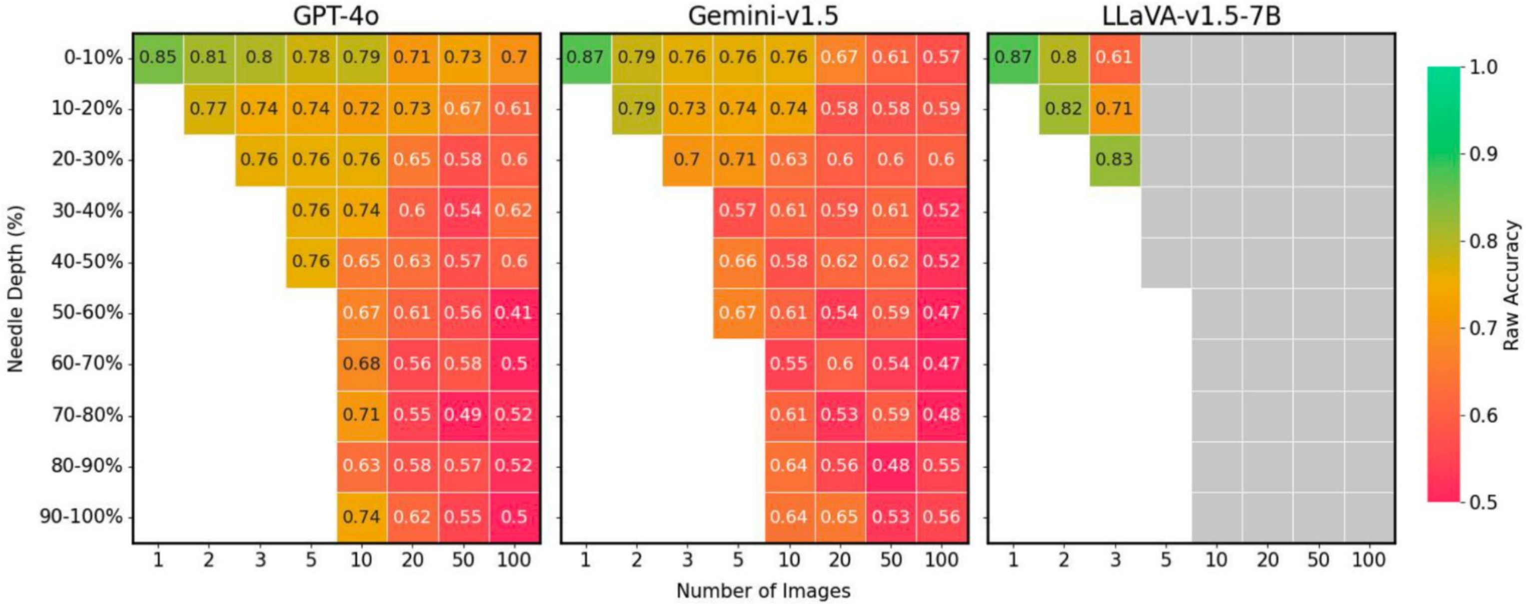
Performance as a Function of Haystack Size

Takeaways

1) **Captioning** with LlaVA and then using LLM to answer the questions improves performance but is slow.
2) **Context-length** issues, particularly with open-source models

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

Positional Bias in Models



Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

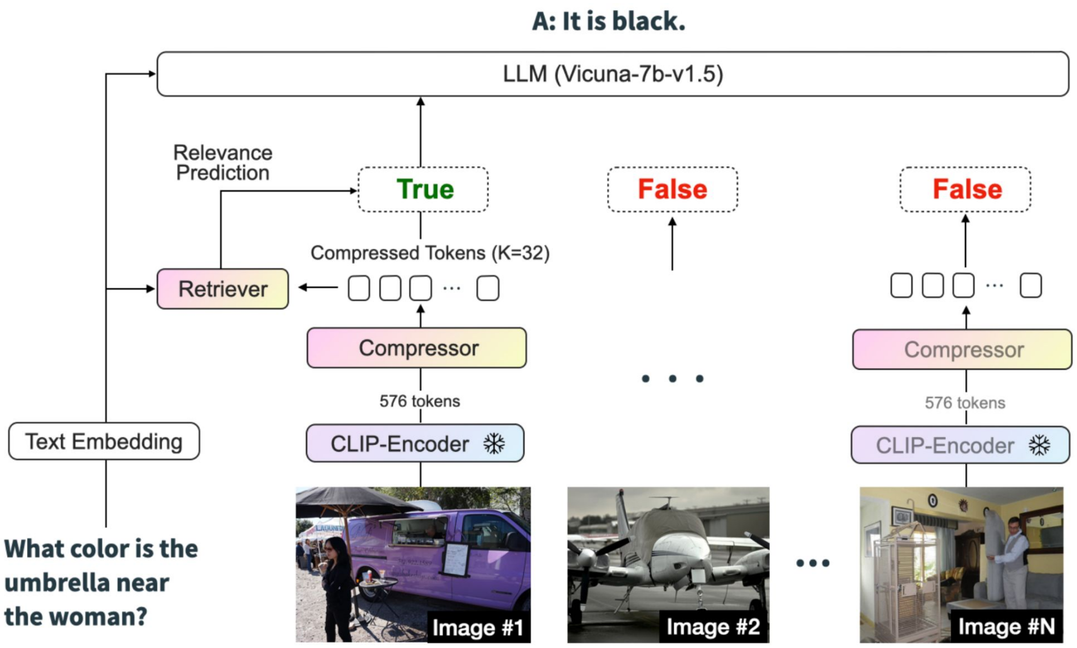
Berkeley
UNIVERSITY OF CALIFORNIA

Takeaways

1) **SOTA** models struggle with visual reasoning over many images
2) **Captioning** with LlaVA and then using LLM to answer the questions improves performance but is slow.
3) **Context-length** issues, particularly with open-source models

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

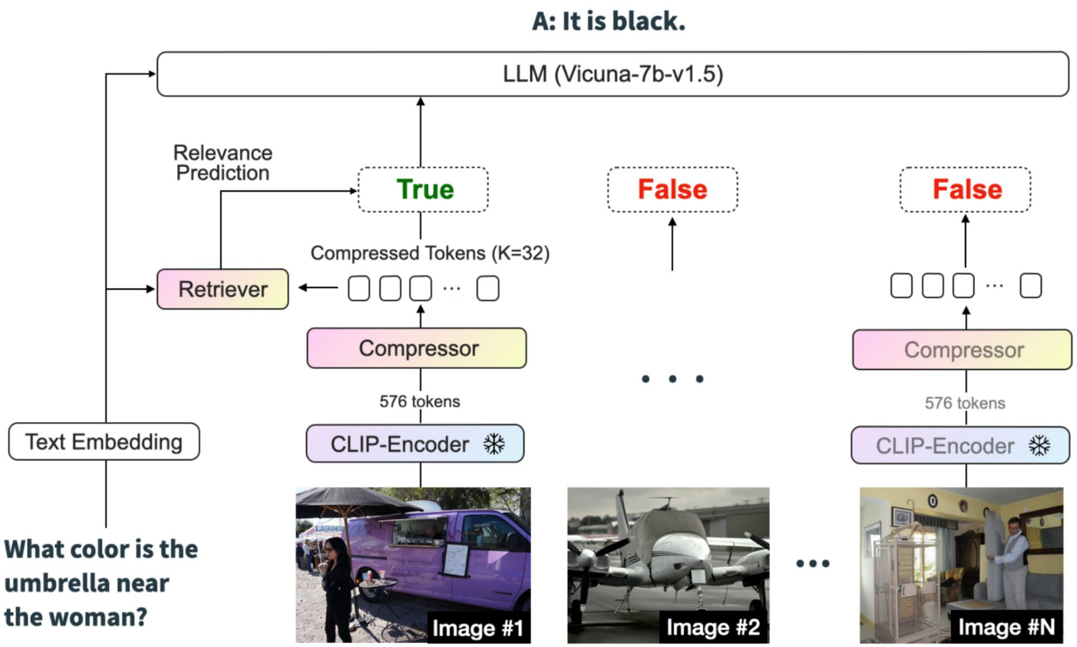
MIRAGE: Multi-Image Retrieval Augmented Generation



Authors proposed solution to improve performance on Visual Haystacks

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

MIRAGE: Multi-Image Retrieval Augmented Generation



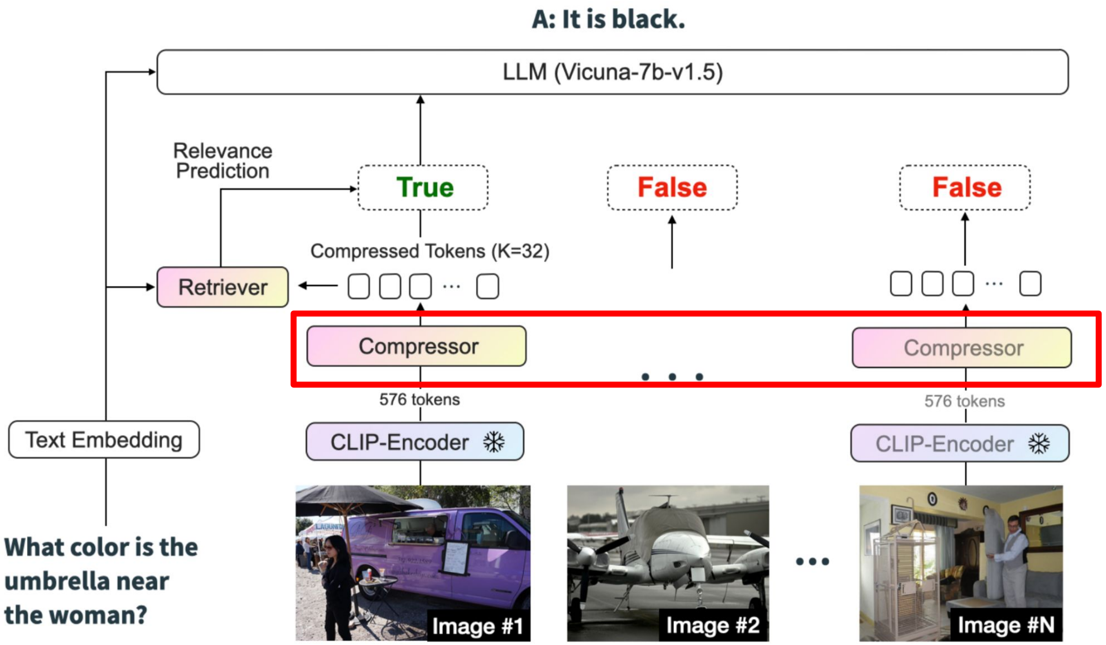
Authors proposed solution to improve performance on Visual Haystacks

1. Image Compressor
2. Retriever (relevance predictor)
3. LLM

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

MIRAGE: Multi-Image Retrieval Augmented Generation



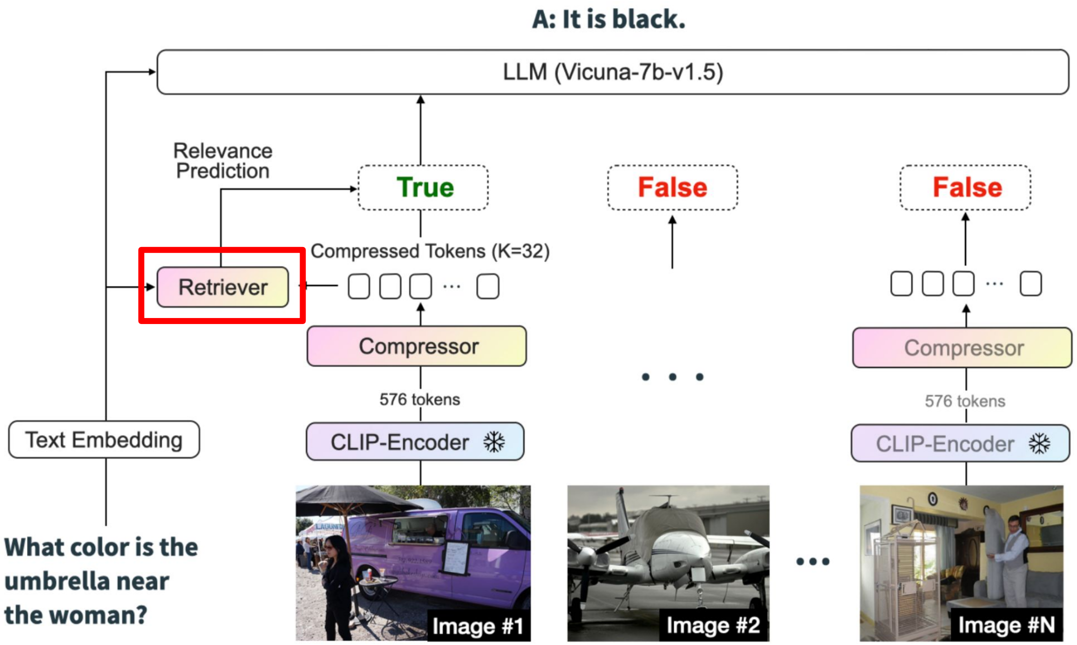
1. Image Compressor

Q-former: compress from 576 to 32 tokens/image using 32 learned query vectors

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

Berkeley UNIVERSITY OF CALIFORNIA

MIRAGE: Multi-Image Retrieval Augmented Generation



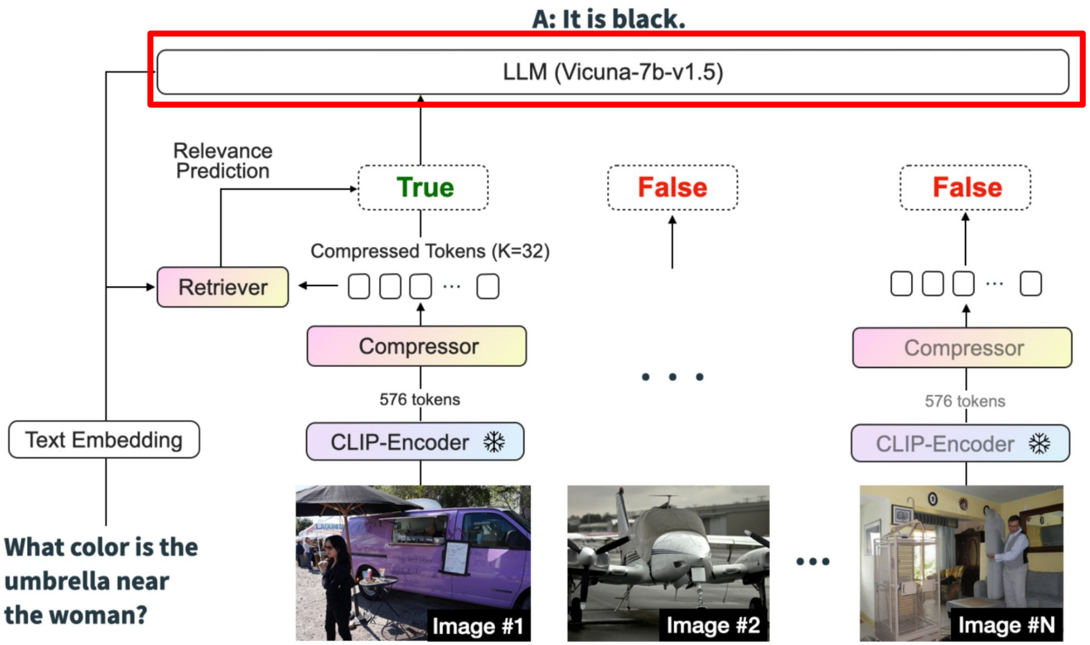
2. Retriever

MLP which takes in **image tokens and query** and predicts a relevance score from 0 to 1 - at inference only relevant (>0.5) images are forwarded to the LLM

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

MIRAGE: Multi-Image Retrieval Augmented Generation



3. LLM (finetuned from **LLaVA-v1.5-7B)**

LLM answers the question using the text query and only the relevant image tokens.

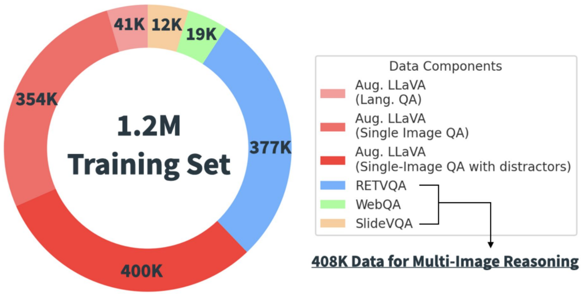
Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

MIRAGE: Multi-Image Retrieval Augmented Generation

Training procedure

1. **Pretraining:** train the compressor alone on ShareGPT data

2. **Finetuning:** train retriever+LLM on a dataset containing single and multi-question VQA
    a. Adapt LlaVA single-image training data



Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

MIRAGE: Results

Method	MIQA (RetVQA)			Single-Image QA							
	Recall	Precision	VQA Acc.	VQAv2	GQA	Vizwiz	TextVQA	POPE	MMB	MMB-CN	MM-Vet
Qwen-VL-Chat [2]	-	-	0.0*	78.2	57.5	38.9	61.5	-	60.6	56.7	-
LLaVA-v1.5-7B [26]	-	-	30.6	78.5	62.0	50.0	58.2	85.9	64.3	58.3	31.1
GPT-4o [32]	-	-	34.6	-	-	-	-	-	-	-	-
GPT-4 [33]	-	-	-	77.2	-	-	78.0	-	-	-	-
Gemini-v1.5 [42]	-	-	32.2	73.2	-	-	73.5	-	-	-	-
LWM [28]	-	-	-	55.8	44.8	11.6	18.8	75.2	-	-	9.6
MI-BART [34]	-	-	76.5	-	-	-	-	-	-	-	-
MIRAGE (Ours)	80.5	49.9	70.8	70.0	55.2	40.1	46.3	83.4	57.6	48.8	25.8

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA

MIRAGE: Results

Q-Former approach is effective at compressing tokens while retaining performance

Method	Tokens/Img	VQAv2	GQA	Vizwiz	TextVQA	POPE	MMB	MMB-CN	MM-Vet
Original LLaVA	576	78.5	62.0	50.0	58.2	85.9	64.3	58.3	31.1
3x3 Max-Pooling	64	68.7	56.2	41.3	48.5	83.0	59.2	49.3	24.3
Global Avg. Pooling	1	62.5	51.3	37.7	45.5	79.6	55.0	45.5	18.9
MIRAGE/Q-Former (Ours)	32	72.8	56.6	48.0	47.1	83.9	61.5	55.0	27.3

Table 4: Exploration of various token reduction methods. We can see that the Q-former is most efficient at reducing the number of tokens while retaining most of the general QA performance.

Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

MIRAGE: Takeaways

1. Outperforms GPT-4o across all settings
2. Outperforms Gemini in the multi-needle setting
3. Underperforms in oracle (N=1) performance due to token compression
4. Shines at retrieval - enables much longer context than is possible with GPT-4o or Gemini alone.

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Conclusions/Takeaways

This paper introduces **Visual Haystacks,** a new benchmark for MIQA that requires retrieval and reasoning over large numbers of images

Many **SOTA models struggle** considerably with this task

A **RAG-based technique** is introduced to improve performance via token compression and image relevance classification

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Discussion Questions

1. How much of MIRAGE's success is due to having MIQA training data rather than the proposed architecture? It would be interesting to see how this architecture performs without being fine-tuned on the MIQA datasets, which could be a fairer comparison.

2. After retrieval in what order are the images provided to the LMM? Will sorting the images based on the relevance scores (in ascending order due to LLaVA-1.5-7B positional bias) improve the results?

3. Visual haystack is still a synthetic benchmark. Are there other ways to obtain non-synthetic long-context data with questions?

4. Is it harsh to say that MIRAGE doesn't really solve the long-context reasoning problem, but essentially converts it to a short-context problem via a preprocessing step? What if many/most images are relevant to the query? In that case, how can we more fundamentally address the limitations of VLMs in learning over long contexts?

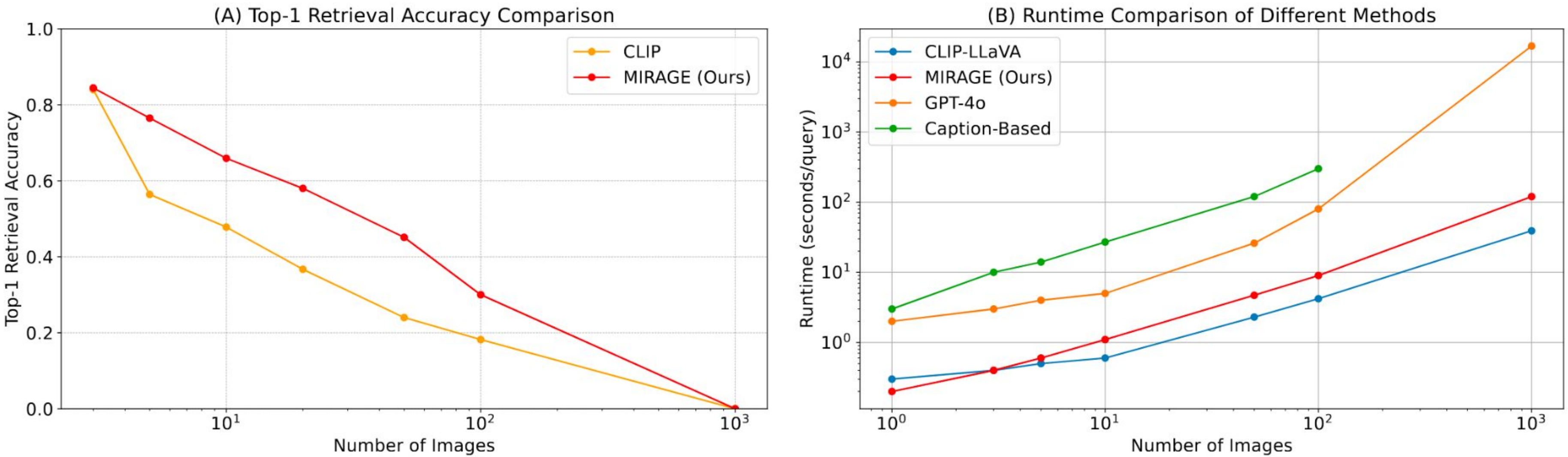
Thanks!

Multi-Needle Performance

Method		Oracle	N=5	N=10	N=50	N=100	N=1K	N=10K
Naive	Question Only (LLama3)	0.48	-	-	-	-	-	-
	Caption-Based (LLaVA + LLama3)	0.70	**0.70**	**0.66**	**0.56**	E	E	E
LMM	Claude-3 Opus	0.55	0.49	0.48	E	E	E	E
	Gemini-1.5	0.56	0.51	0.54	0.50	**0.52**	E	E
	GPT-4o (low-res)	**0.71**	0.65	0.63	0.49	**0.52**	E	E
RAG-based	MIRAGE (Ours)	0.57	0.56	0.54	0.51	0.50	**0.48**	**0.49**

Visual Haystacks: Answering Harder Questions About Sets of Images” (Wu, et. al)

Retriever Ablations



Visual Haystacks: Answering Harder Questions About Sets of Images" (Wu, et. al)

Berkeley
UNIVERSITY OF CALIFORNIA