Fitting Cognify to VLM applications

Benjamin Xia, Qi Qi, Nitya Agarwal, Jan Szczekulski

CSE 291 | March 2025

Content

1 What is the problem, and why is it important?

2 New Workflows

3 New Cogs

4 New Search Option

5 Evaluation

6 Issues?

7 Q&A

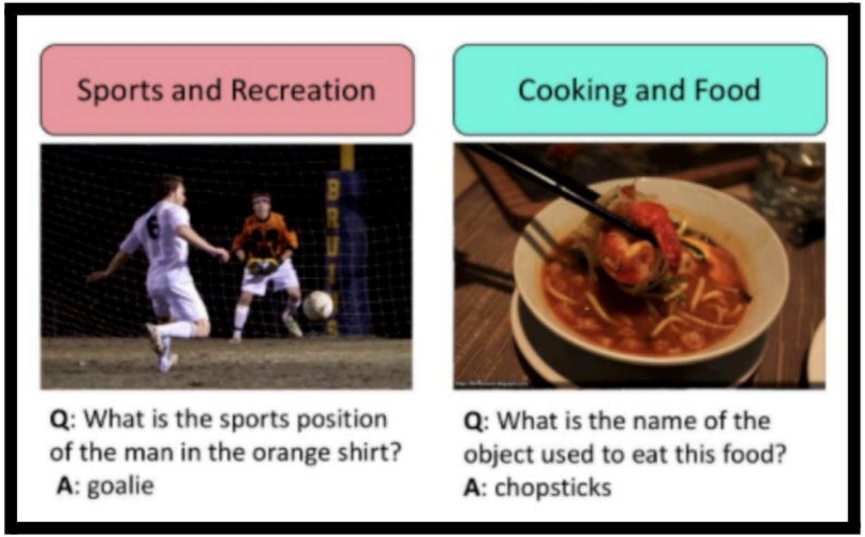
Cognify

Cognify = GenAI Workflow Optimization

https://github.com/GenseeAI/cognify

GenseeAI

Current Cognify Performance on Vision Tasks



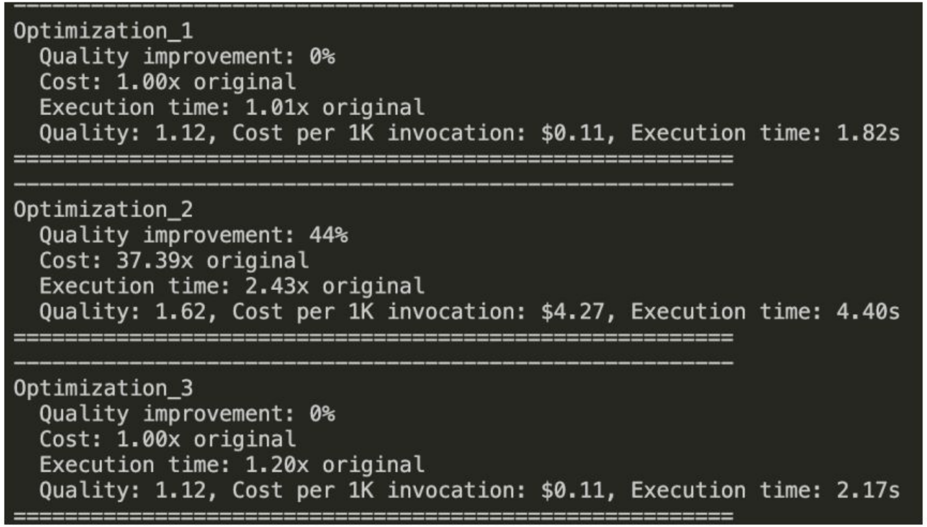
Sports and Recreation

Q: What is the sports position of the man in the orange shirt?
A: goalie

Cooking and Food

Q: What is the name of the object used to eat this food?
A: chopsticks

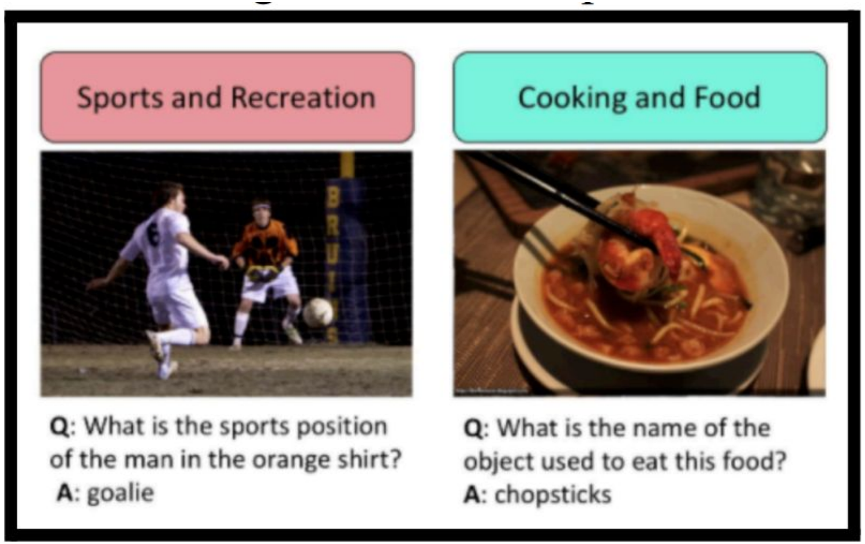
Figure 1 - VQA Dataset Example



```
-----------------------------------------------------------------
Optimization_1
  Quality improvement: 0%
  Cost: 1.00x original
  Execution time: 1.01x original
  Quality: 1.12, Cost per 1K invocation: $0.11, Execution time: 1.82s
=================================================================
-----------------------------------------------------------------
Optimization_2
  Quality improvement: 44%
  Cost: 37.39x original
  Execution time: 2.43x original
  Quality: 1.62, Cost per 1K invocation: $4.27, Execution time: 4.40s
=================================================================
-----------------------------------------------------------------
Optimization_3
  Quality improvement: 0%
  Cost: 1.00x original
  Execution time: 1.20x original
  Quality: 1.12, Cost per 1K invocation: $0.11, Execution time: 2.17s
=================================================================
```

Figure 2 - Optimization Results

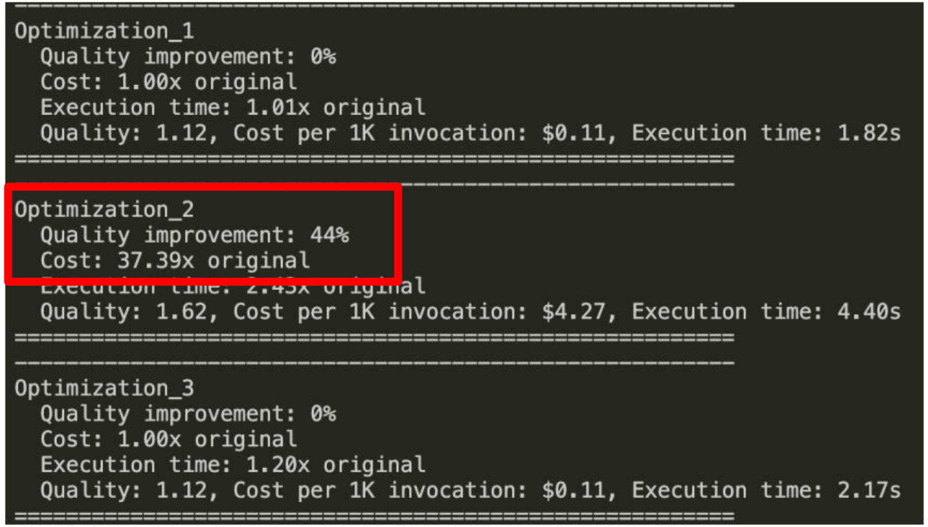
Current Cognify Performance on Vision Tasks



Q: What is the sports position of the man in the orange shirt?
A: goalie

Q: What is the name of the object used to eat this food?
A: chopsticks

Figure 1 - VQA Dataset Example



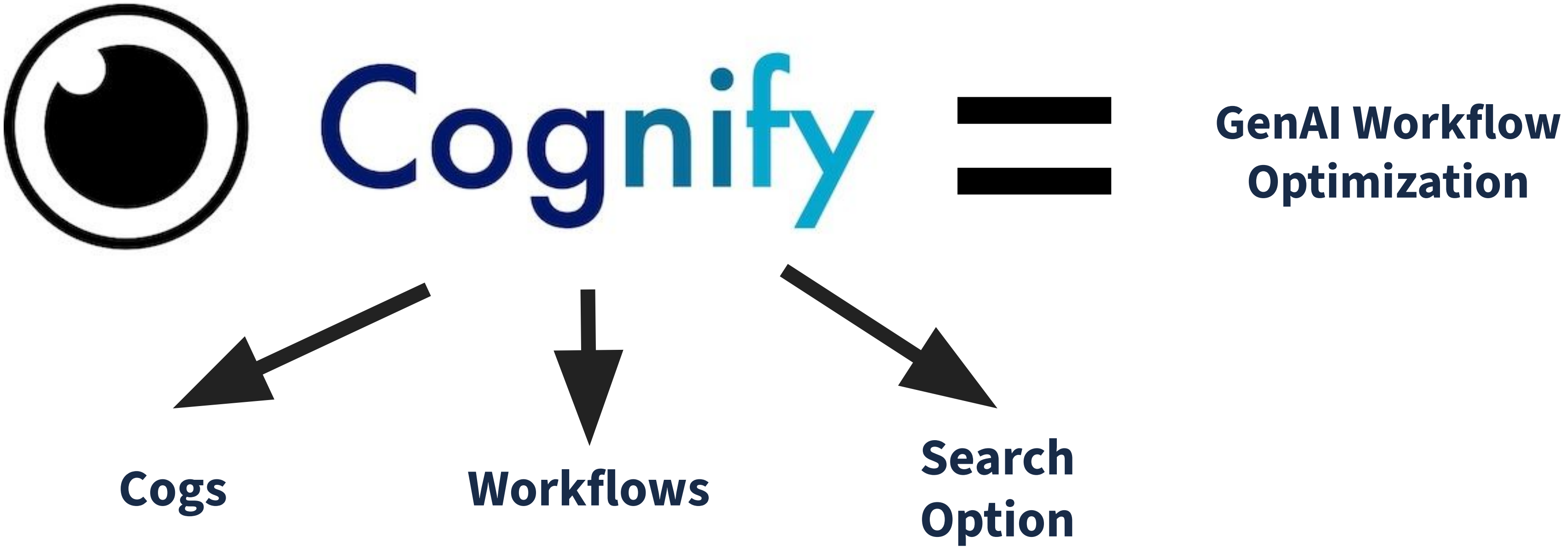
```
--------------------------------------------------------
Optimization_1
  Quality improvement: 0%
  Cost: 1.00x original
  Execution time: 1.01x original
  Quality: 1.12, Cost per 1K invocation: $0.11, Execution time: 1.82s
========================================================
--------------------------------------------------------
Optimization_2
  Quality improvement: 44%
  Cost: 37.39x original
  Execution time: 2.43x original
  Quality: 1.62, Cost per 1K invocation: $4.27, Execution time: 4.40s
========================================================
--------------------------------------------------------
Optimization_3
  Quality improvement: 0%
  Cost: 1.00x original
  Execution time: 1.20x original
  Quality: 1.12, Cost per 1K invocation: $0.11, Execution time: 2.17s
========================================================
```

Figure 2 - Optimization Results

Can we do better?

Cognify = GenAI Workflow Optimization

Can we do better?



Cognify = GenAI Workflow Optimization

Cogs

Workflows

Search Option

New VLM Workflows! (to evaluate on)

Image Classification

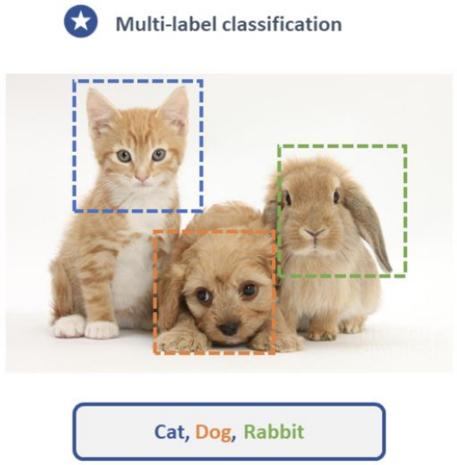
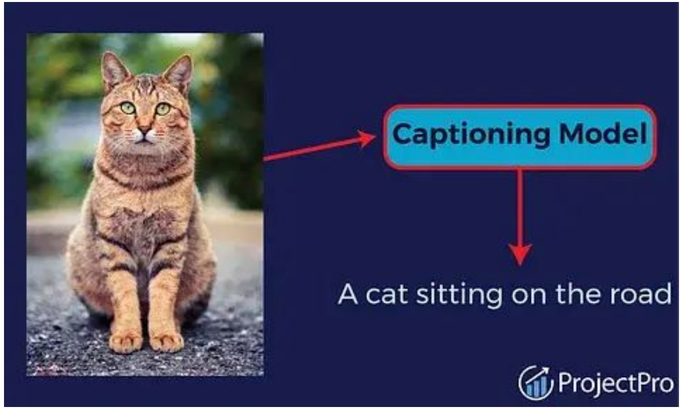


Image Captioning (Description)

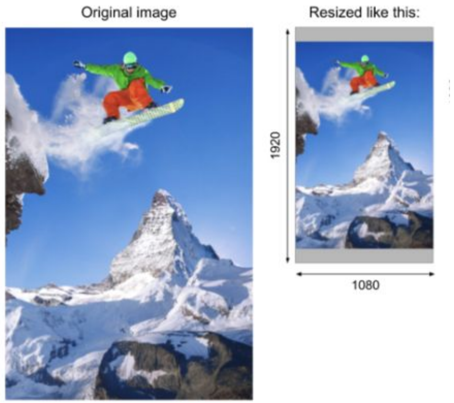


Visual Question Answering

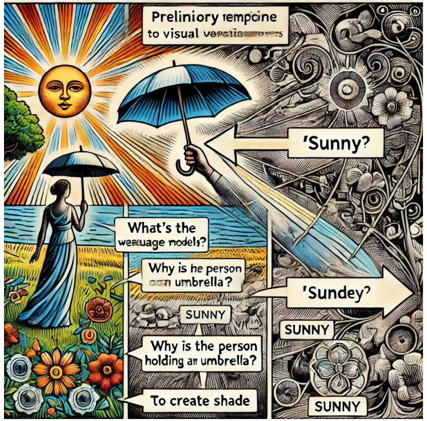
How many slices of pizza are there?
Is this a vegetarian pizza?

New VLM-Specific Cogs!

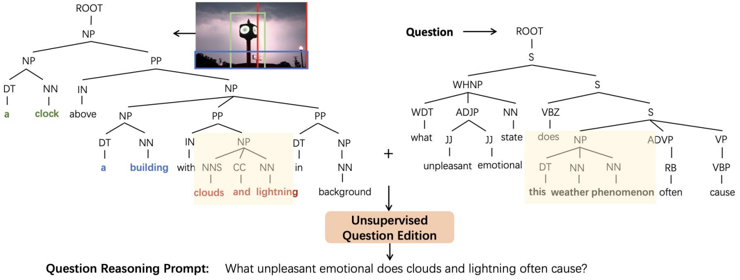
Image Compression



Vision Planning

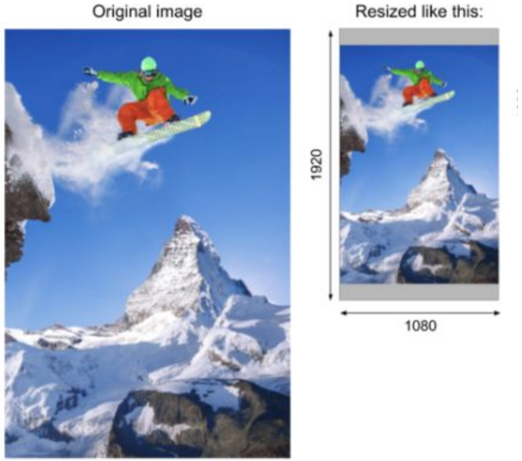

Preliniory rempcine to visual veretionwers
'Sunny?
What's the wenuage nodels?
Why is he person umbrella?
'Sundey?
SUNNY
SUNNY
Why is the person holding an umbrella?
To create shade
SUNNY

Query Disambiguation



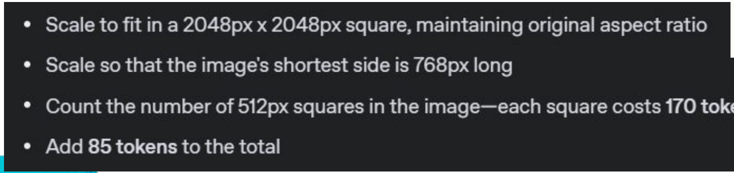
Question Reasoning Prompt: What unpleasant emotional does clouds and lightning often cause?

New VLM Cog - Image Compression

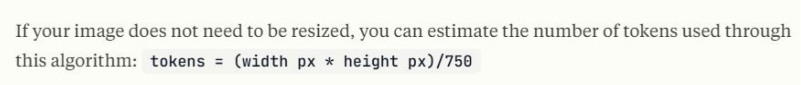


Issue:
Images get tokenized when input into VLM. Tokens cost money.

Solution:
Test few different image sizes to find a relatively cheap one which will not reduce the quality of workflow output.



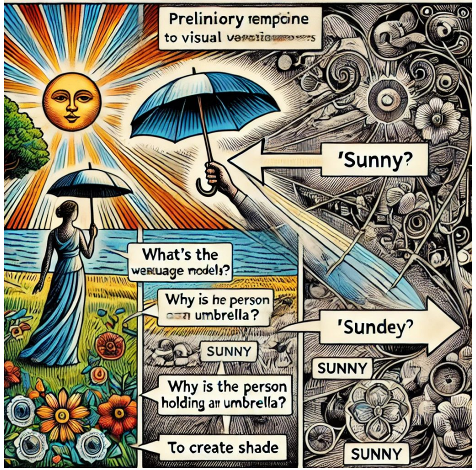
- Scale to fit in a 2048px x 2048px square, maintaining original aspect ratio
- Scale so that the image's shortest side is 768px long
- Count the number of 512px squares in the image—each square costs **170 tokens**
- Add **85 tokens** to the total



If your image does not need to be resized, you can estimate the number of tokens used through this algorithm: `tokens = (width px * height px)/750`

Bigger images = Bigger cost

New VLM Cog - Vision Planning


Preliniory rempoine to visual ve
'Sunny?
What's the weruage rodels?
Why is he person umbrella?
'Sundey?
SUNNY
SUNNY
Why is the person holding an umbrella?
To create shade
SUNNY

gpt made this, please don't read the text

Issue:

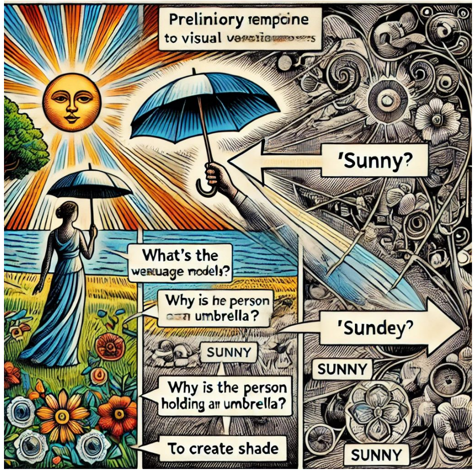
VLM responses will sometimes be more influenced by the query, ignoring the content of the image[1]

Solution:

Inject more information about the image as extra context.

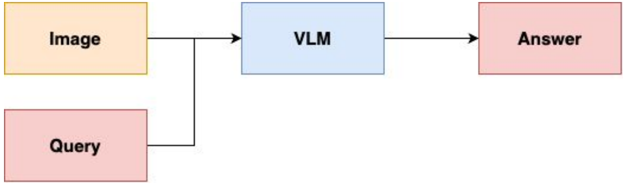
[1] Wang, Z., Chen, C., Li, P., & Liu, Y. (2023). Filling the image information gap for vqa: Prompting large language models to proactively ask questions. arXiv preprint arXiv:2311.11598.

New VLM Cog - Vision Planning



gpt made this, please don't read the text

Vanilla VLM Workflow:



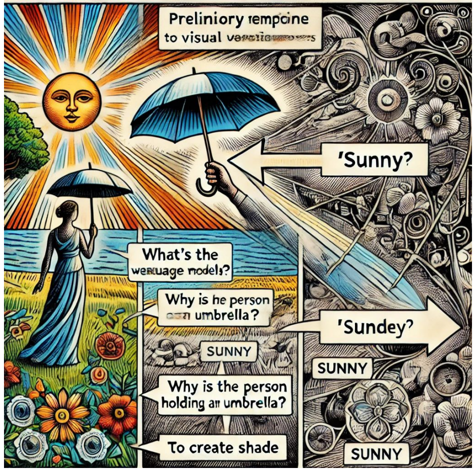
Image

Query

VLM

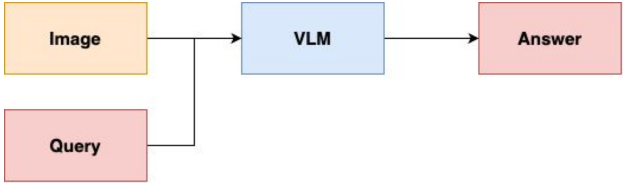
Answer

New VLM Cog - Vision Planning

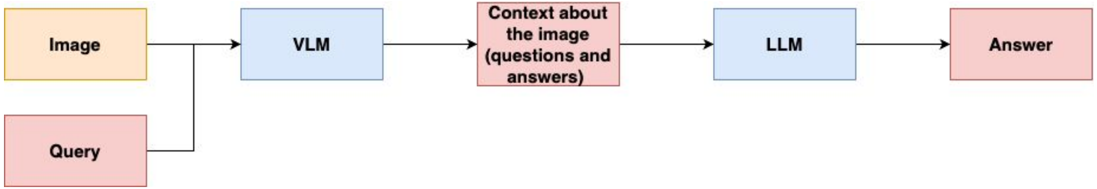

Preliniory tempoine to visual ve
'Sunny?
What's the weεuage models?
Why is he person umbrella?
SUNNY
'Sundey?
SUNNY
Why is the person holding an umbrella?
To create shade
SUNNY

gpt made this, please don't read the text

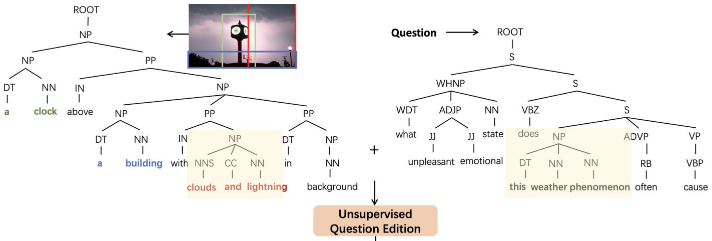
Vanilla VLM Workflow:



VLM Workflow with Vision Planning Cog:



New VLM Cog - Query Disambiguation



Question Reasoning Prompt: What unpleasant emotional does clouds and lightning often cause?

Issue:

The queries in VLM often reference objects in the image, which makes answering them harder

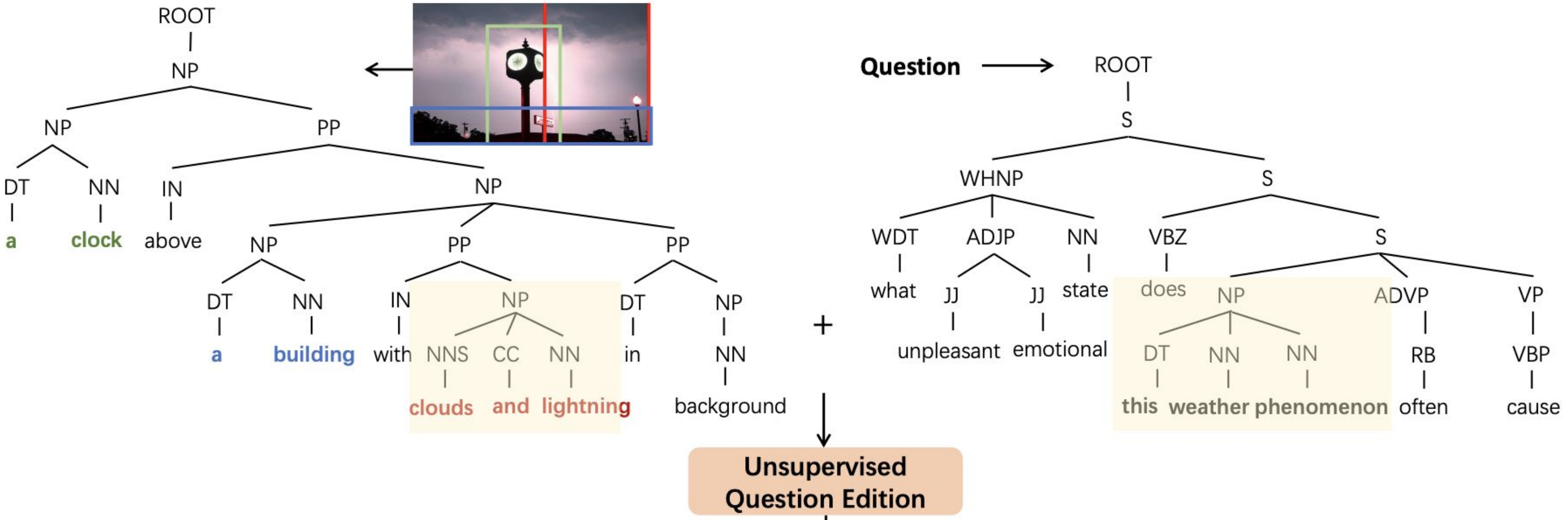
Solution:

Disambiguate query.

Example:

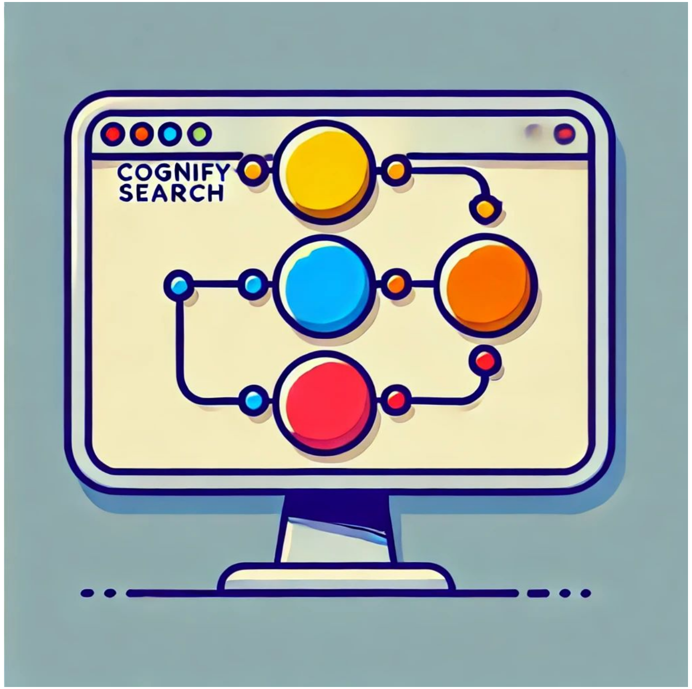
Query: “What did person in the image invent?”
Image Caption: “Einstein standing on a stand”
Disambiguated query: “What did Einstein invent?”

New VLM Cog – Query Disambiguation



Question Reasoning Prompt: What unpleasant emotional does clouds and lightning often cause?

New Search Option!



Issue:
All of our Cogs are applicable to only VLM-related task

Solution:
Create a new separate search **option** that can be applied to VLM tasks

Details:
One Layer Search.
Search created based on our (somewhat short) experience.
Search combines our newly created Cogs, together with Few Shot learning and Model Selection Cogs

Evaluation per Cog - Summary

Applications

Cogs

	Image Classification	Image Captioning (Description)	Visual Question Answering
Image Compression	Decreases cost	Decreases cost	Decreases cost
Vision Planning	?	?	?
Query Disambiguation	?	?	?

Evaluation

- ● Vanilla Cognify
- ● FewShot + ModelSelect + Image Comp.+ Vision Plan
- ● Image Comp. + Vision Plan
- ● Image Comp. + FewShot + ModelSelect
- ● Image Compression Only

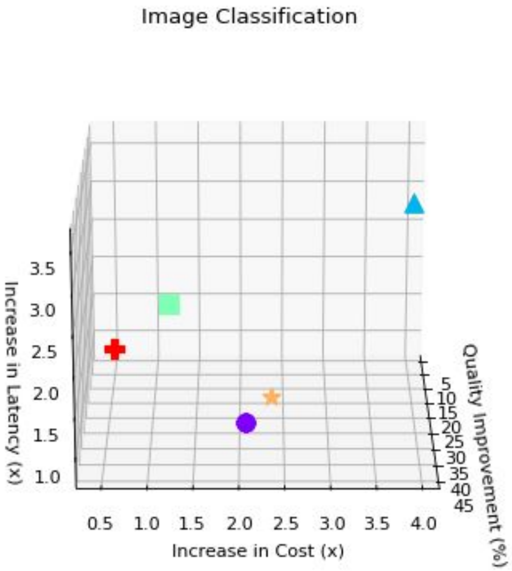


Image Classification

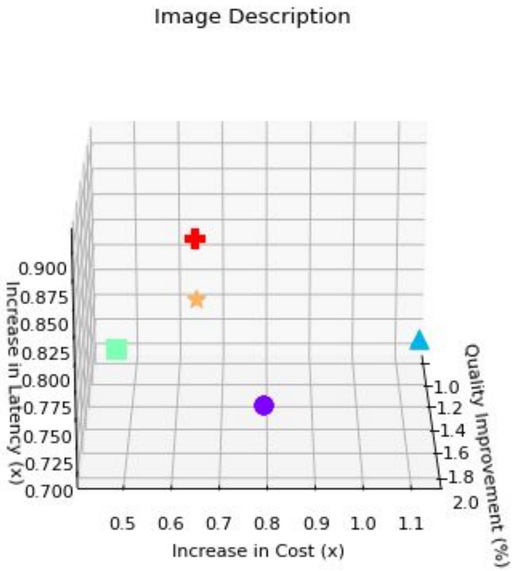
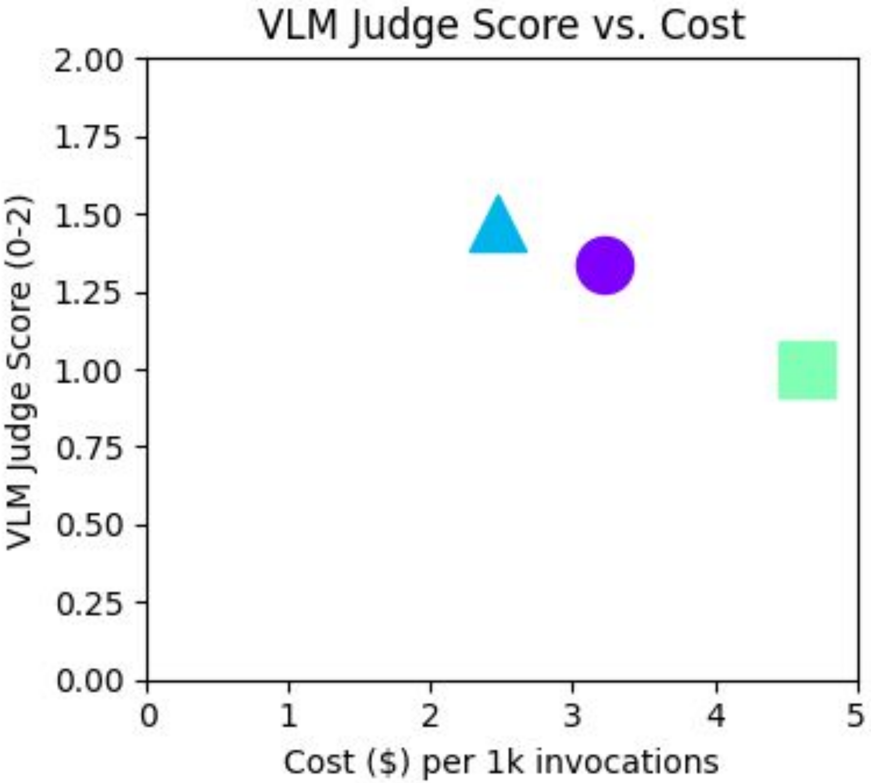


Image Description

Evaluated on Google Open Images Dataset V6

Evaluation - VQA

- ● Vanilla Cognify
- ● Image Compression + Vision Planning+ Vanilla Cognify
- ● Image Compression + Vision Planning Only

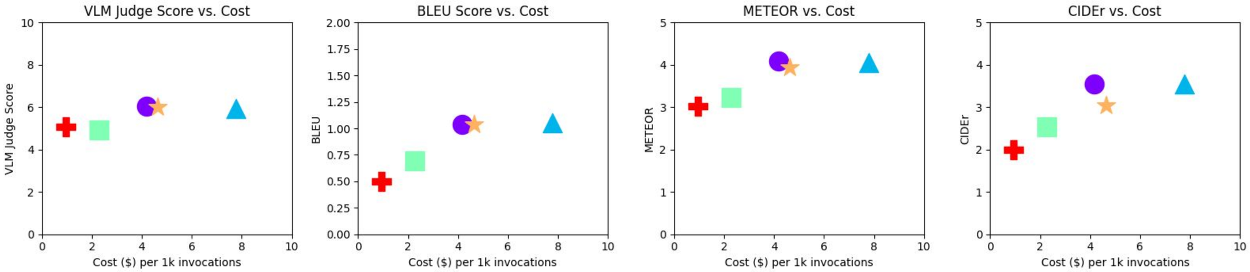


Evaluated on OK-VQA

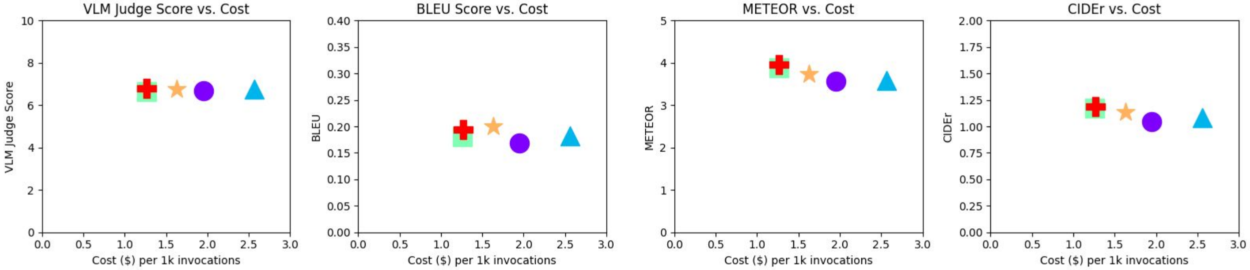
Evaluation - Other applications

- 🟣 Vanilla Cognify
- 🔵 Image Comp.+ Vision Plan + Vanilla Cognify
- 🟢 Image Comp. + Vision Plan Only
- 🟠 Image Comp. + Vanilla Cognify
- 🔴 Image Compression Only

Classification



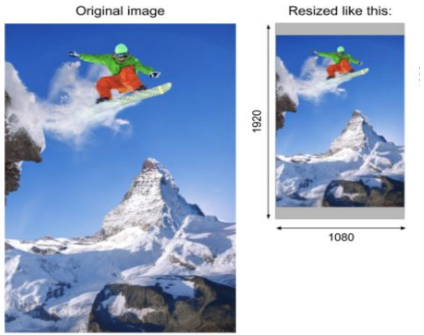
Description



Evaluated on Google Open Images Dataset V6

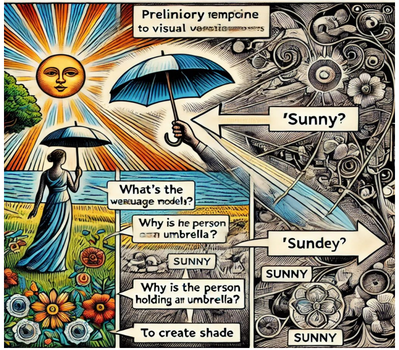
Evaluation per Cog - Summary

Image Compression



Cost & Latency

Vision Planning

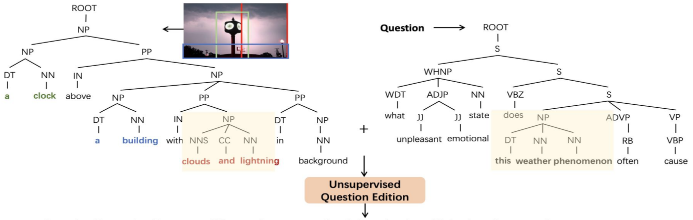


Preliniory rempoine to visual ve[illegible]
'Sunny?
What's the weruage nodels?
Why is he person [illegible] umbrella?
'Sundey?
SUNNY
SUNNY
Why is the person holding an umbrella?
To create shade
SUNNY

Mixed results for Quality

(slightly positive)

Query Disambiguation



Question Reasoning Prompt: What unpleasant emotional does clouds and lightning often cause?

Quality

(Still experimenting)

Challenges

- There was a **sharp learning curve** for the code-base since the documentation is user-based and not contribution focused

- **GPT API calls are costly**, and have a token rate limit per minute - intervening with our evaluation efforts. **Using our own OpenAI API credits limited depth of testing.**

- **Coming up with Cog ideas isn't straightforward**. We've read through 10s of VLM-related research papers. All of our Cogs are grounded in singular or multiple research papers - cited here: Image Compression Cog, Vision Planning Cog, Query Disambiguation Cog.

- **Not every idea works**, we have to implement and more importantly experiment, especially do ablation studies.

Next Steps

- Finish cog integration with previous cogs
  - Some cogs may be closely integrated with each other
  - Verify with other existing example workflows

- Open a PR to contribute to open source project

Q&A