



DFG



small data
MORE FROM LESS



Open-ended VQA benchmarking of Vision-Language models by exploiting Classification datasets and their semantic hierarchy



Simon Ging *



María A. Bravo *



Thomas Brox

University of Freiburg



Poster session

Halle B, Thu 9 May, 4:30 p.m.

iclr.cc/virtual/2024/poster/19102

[CV] Vision

COMPUTER VISION University of Freiburg

* Equal Contribution



ICLR
International Conference On
Learning Representations

Motivation

Vision-Language Models

Vision-language research requires understanding of **vision** and **language**



Motivation

Vision-Language Models

Vision-language research requires understanding of **vision** and **language**



Captioning



The image shows a person sitting on a sandy beach, with three large dogs. The person is looking towards the sea.

Motivation

Vision-Language Models

Vision-language research requires understanding of **vision** and **language**



Visual Question Answering

How many dogs are in the image?



There are three dogs.

Captioning



The image shows a person sitting on a sandy beach, with three large dogs. The person is looking towards the sea.

Motivation

Vision-Language Models

Vision-language research requires understanding of **vision** and **language**



Captioning



The image shows a person sitting on a sandy beach, with three large dogs. The person is looking towards the sea.

Visual Question Answering

How many dogs are in the image?



There are three dogs.

Visual Dialog

Where is the scene taking place?



The scene is taking place on a sandy beach with the **ocean** in the background.

Does the **ocean** have strong waves?



The waves in the ocean appear moderate, not particularly strong.

Motivation

Vision-Language Models

Vision-language research requires understanding of **vision** and **language**



Captioning

Make a short description of the image.



The image shows a person sitting on a sandy beach, with three large dogs. The person is looking towards the sea.

Visual Question Answering

How many dogs are in the image?



There are three dogs.

Visual Dialog

Where is the scene taking place?



The scene is taking place on a sandy beach with the **ocean** in the background.

Does the **ocean** have strong waves?



The waves in the ocean appear moderate, not particularly strong.

Motivation

Vision-Language Models

Vision-language research requires understanding of **vision** and **language**



Captioning

Visual Question Answering

Visual Dialog

Our goal:

Assess **VLMs'** performance
in an **open-ended VQA** setup

Make a short description of the image.



The image shows a person sitting on a sandy beach, with three large dogs. The person is looking towards the sea.



The waves in the ocean appear moderate, not particularly strong.

ng



aking place on a
with the **ocean** in
d.

rong



Motivation

Visual Question Answering

Challenges in evaluation of Open-ended VQA

Ambiguous object



What's this? (*Label: Porcupine*)

Model output: A tree with no leaves

Motivation

Visual Question Answering

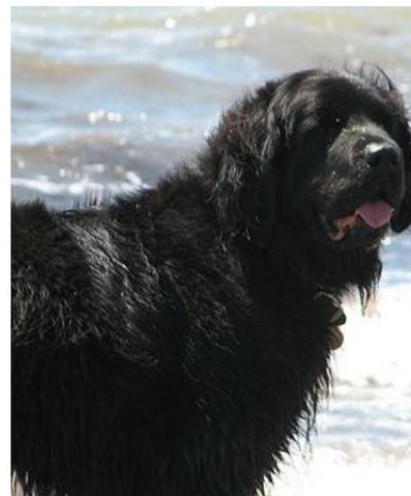
Challenges in evaluation of Open-ended VQA

Ambiguous object



What's this? (*Label*: Porcupine)
Model output: A tree with no leaves

Unknown label granularity



What's this? (*Label*: Newfoundland dog)
Model output: A black dog standing in the water

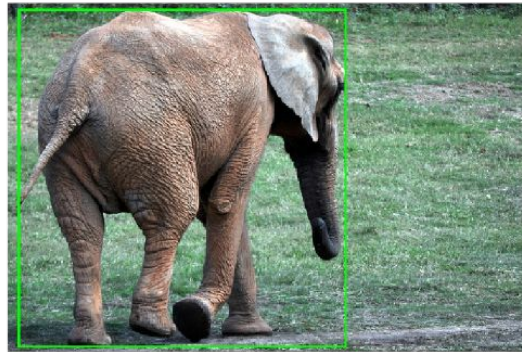
Open-ended Visual Question Answering

oVQA benchmark

Objects



Dataset: ImageNet
Question: What's this?
Label: cougar



Dataset: COCO
Question: What's this?
Label: elephant

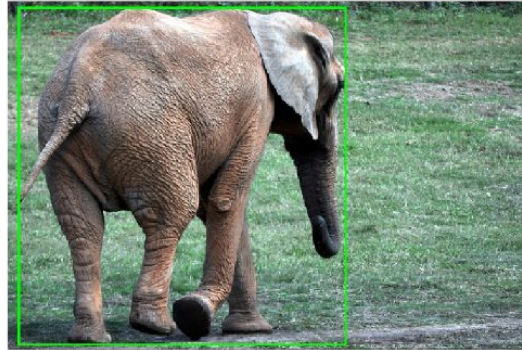
Open-ended Visual Question Answering

oVQA benchmark

Objects



Dataset: ImageNet
Question: What's this?
Label: cougar



Dataset: COCO
Question: What's this?
Label: elephant

Actions



Dataset: ActivityNet
Question: What activity is this?
Label: playing drums

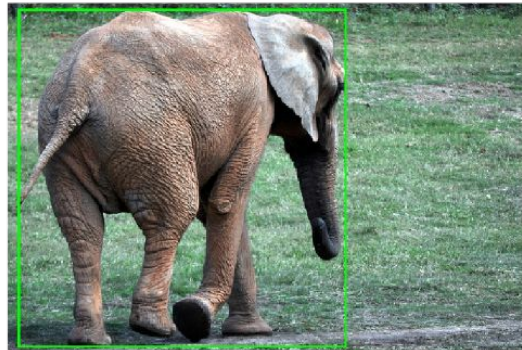
Open-ended Visual Question Answering

oVQA benchmark

Objects



Dataset: ImageNet
Question: What's this?
Label: cougar



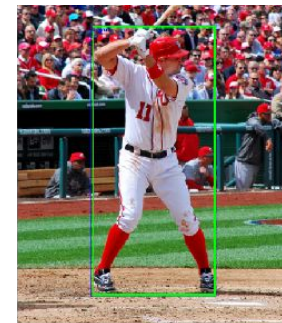
Dataset: COCO
Question: What's this?
Label: elephant

Actions



Dataset: ActivityNet
Question: What activity is this?
Label: playing drums

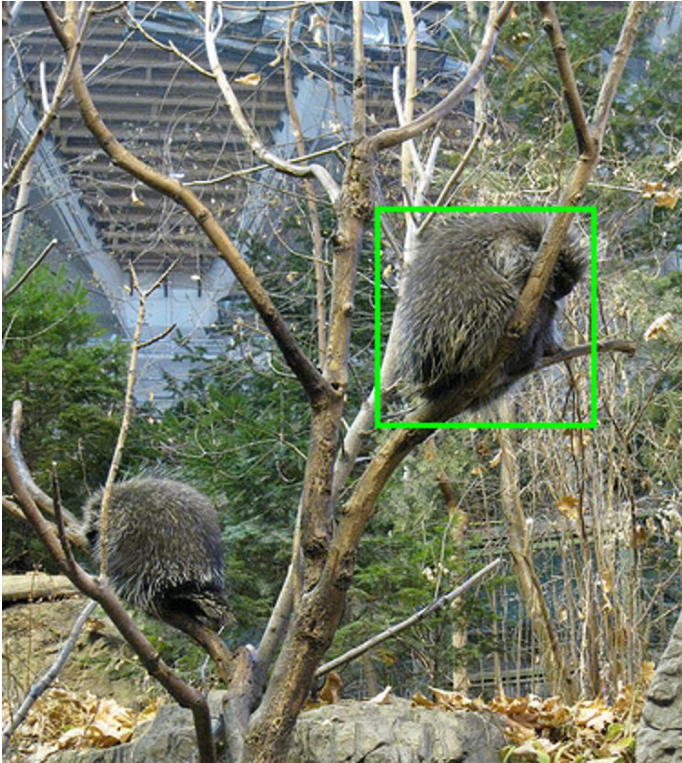
Attributes



Dataset: OVAD
Question: What is the position of the person?
Label: standing / upright / vertical

oVQA Benchmark

Visual guidance

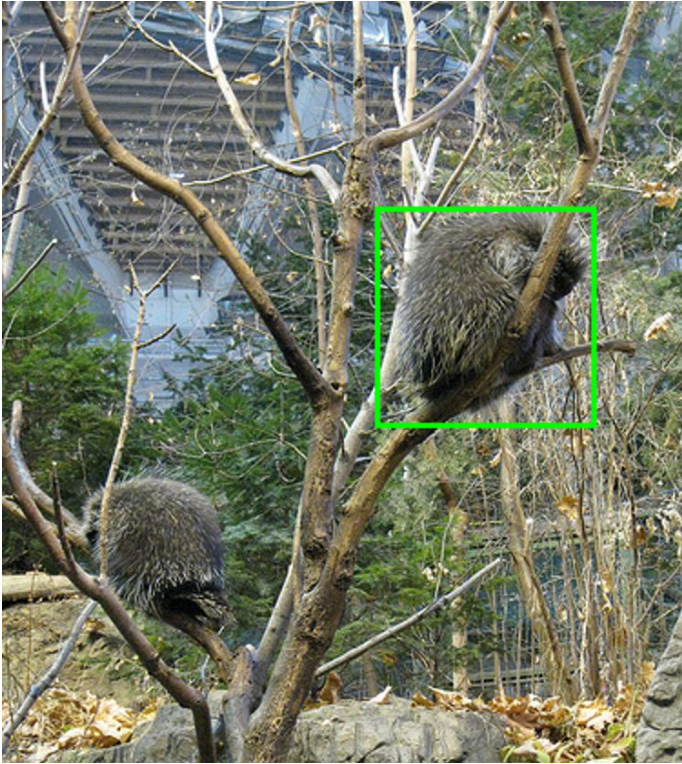


What's this?
Label: Porcupine

Model output: A tree with no leaves

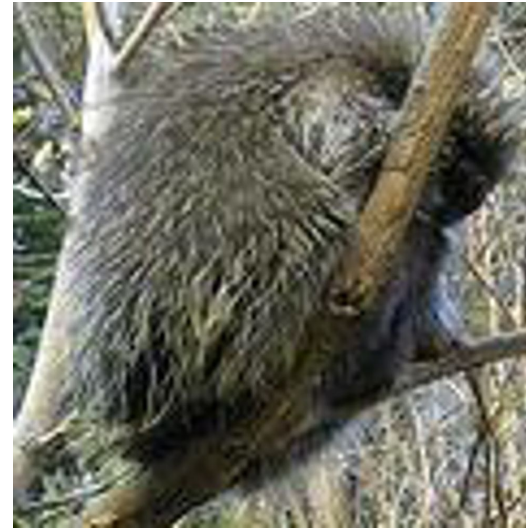
oVQA Benchmark

Visual guidance



Model output: A tree with no leaves

What's this?
Label: Porcupine



Model output: A porcupine

oVQA Benchmark

Follow-up question



Label: Newfoundland dog

What's this?



**A black dog standing
in the water**



oVQA Benchmark

Follow-up question



Label: Newfoundland dog

What's this?



**A black dog standing
in the water**



Parent Hierarchy

dog / domestic dog

canine / canid

carnivore

placental

mammal / mammalian

...

entity

oVQA Benchmark

Follow-up question



Label: Newfoundland dog

What's this?



**A black dog standing
in the water**



Parent Hierarchy

dog / domestic dog
canine / canid
carnivore
placental
mammal / mammalian
...
entity



What type of **dog** is this ?



Newfoundland dog



oVQA Benchmark

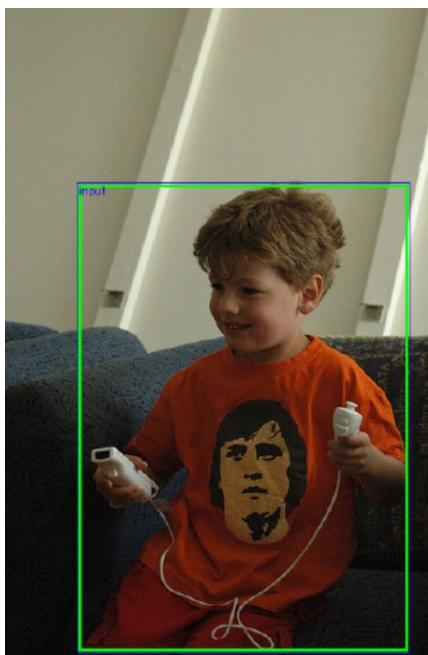
Choosing the correct metric for binary classification

Which is the maturity of the person?



Label:
young

Synonyms:
{ young, baby, child, kid }



“Exact match”:
Answer matches the
label?



child



He's a kid.



oVQA Benchmark

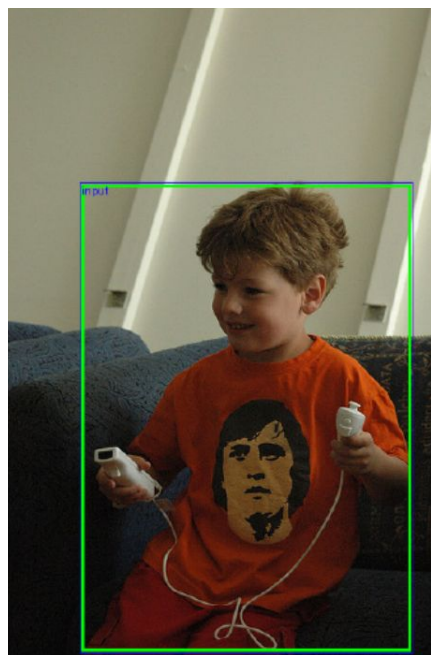
Choosing the correct metric for binary classification

Which is the maturity of the person?



Label:
young

Synonyms:
{ young, baby, child, kid }



child



He's a kid.



“Exact match”:
Answer matches the
label?

“Contains” metric:
Any synonym is contained
in the answer?

oVQA Benchmark

Choosing the correct metric for multi-class classification

What's this?

Model output: a mountain lion



Label:
cougar



ExactMatch:
answer matches label
exactly



Contains:
label is contained in
answer



ClipMatch:
most similar label in Clip
text space



ClipM Top-5 similarities:

cougar (0.792)

lion (0.682)



snow leopard (0.636)

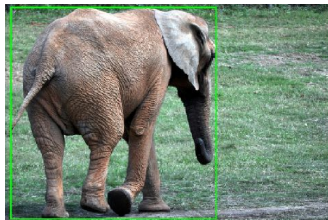
lynx (0.527)

leopard (0.511)

oVQA Benchmark

Sub-benchmarks

	Object-oVQA (COCO)	Object-oVQA (ImageNet)
# Classes	80 objects	1000 objects
Follow-up		
Size	36,800 crops	50,000 images
Question ex.	What is in the image?	What is in the image?



oVQA Benchmark





Sub-benchmarks

	Object-oVQA (COCO)	Object-oVQA (ImageNet)	Activity-oVQA (ActivityNet)
# Classes	80 objects	1000 objects	200 activities
Follow-up	✘	✔	✔
Size	36,800 crops	50,000 images	7,700 frames
Question ex.	What is in the image?	What is in the image?	What is happening in the image?



oVQA Benchmark

Sub-benchmarks

	Object-oVQA (COCO)	Object-oVQA (ImageNet)	Activity-oVQA (ActivityNet)	Attribute-oVQA (OVAD)
# Classes	80 objects	1000 objects	200 activities	117 attributes
Follow-up	✗	✓	✓	✗
Size	36,800 crops	50,000 images	7,700 frames	14,300 crops
Question ex.	What is in the image?	What is in the image?	What is happening in the image?	What is the position of the person?
				

Motivation

Vision-Language Models



Multi-purpose VLM

Make a short description of the image.



The image shows a person sitting on a sandy beach, with three large dogs.

- **BLIP-2 FlanT5 XL**
- **BLIP-2 OPT**

Finetuned Visual Question Answering models

How many dogs are in the image?



There are three dogs.

- **BLIP**
- **X2-VLM**

Dialog and instruction models

Where is the scene taking place?



The scene is taking place on a sandy beach with the **ocean** in the background.

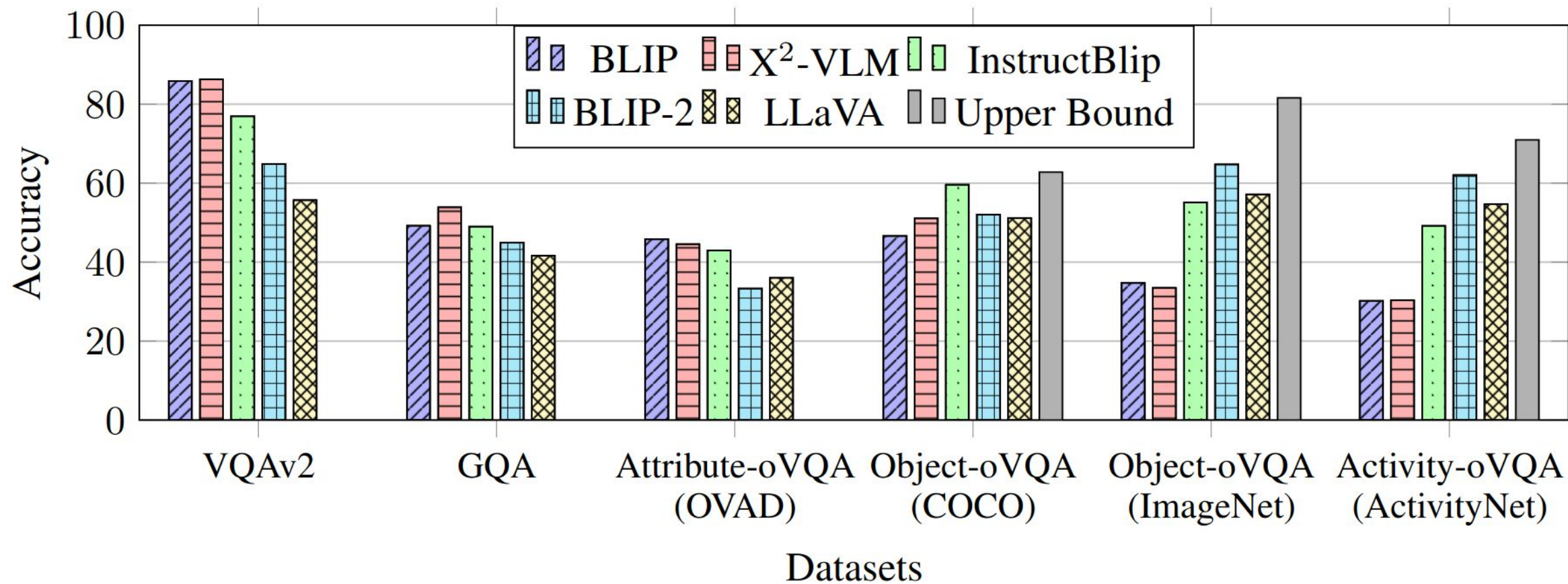
Does the **ocean** have strong waves?



The waves in the ocean appear moderate, not particularly strong.

- **LLaVA**
- **InstructBlip**

Model results



Qualitative Examples

Objects (ImageNet)



correct answer
wrong answer

Question: What's this?

Label: Dalmatian

BLIP-2 OPT output: it's a **dalmatian**

LLaVA output: The image features a large black and white dog laying down on the floor, possibly on a carpet.

Follow-up Question: What type of dog is this?

LLaVA output: The dog in the image is a **Dalmatian**.

Qualitative Examples

Objects (ImageNet)



Question: What's this?

Label: Dalmatian

BLIP-2 OPT output: it's a **dalmatian**

LLaVA output: The image features a large black and white dog laying down on the floor, possibly on a carpet.

Follow-up Question: What type of dog is this?

LLaVA output: The dog in the image is a **Dalmatian**.

Attributes (OVAD)



correct answer
wrong answer

Question: How many people are present in the image?

Label: individual / one / single / 1 / sole / alone

BLIP_{vqa} output: **one**

BLIP-2 OPT output: **None**.

InstructBLIP T5 output: **2**

LLaVA output: **There are two people present in the image.**

X²-VLM_{vqa} L output: **one**

Metrics

User study

What type of donut is on the top right?

Label: chocolate iced glazed



chocolate glazed donut

I rate **5/5**.



2000 model predictions evaluated.

Metrics

User study

What type of donut is on the top right?

Label: chocolate iced glazed



chocolate glazed donut

I rate **5/5**.



Metric *	Pearson Corr
GPT-4 _{10-shot}	0.972
Llama2 _{5-shot}	0.919
Cont	0.906
EM	0.525
LERC	0.827
ROUGE	0.717

* More metrics in the paper

2000 model predictions evaluated.

Metrics

User study

- **LLMs** outperform classical metrics
- **Contains metric** improves over learned metrics and translation metrics



What are the vegetables to the left of the bowl that is to the left of the cookies?

Label: **carrots**

	Output	Label	EM	Cont	LLaMA-2	GPT-4
	carrots	carrots	1.00	1.00	1.00	1.00
The vegetables to the left of the bowl are carrots and green beans.		carrots	0.00	1.00	1.00	0.25

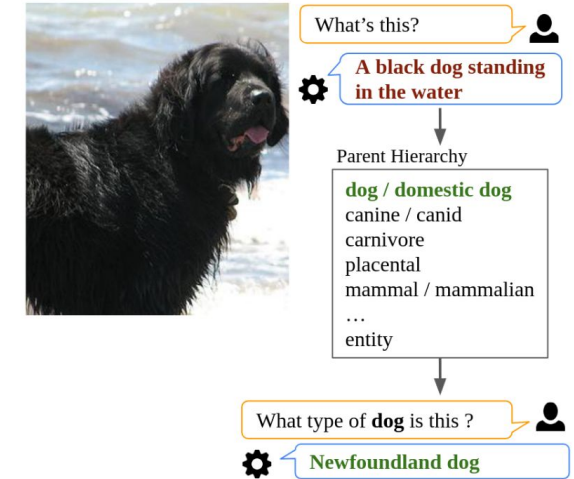
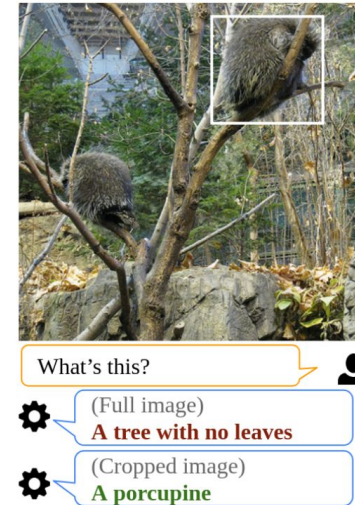
Metric *	Pearson Corr
GPT-4 _{10-shot}	0.972
Llama2 _{5-shot}	0.919
Cont	0.906
EM	0.525
LERC	0.827
ROUGE	0.717

* More metrics in the paper

Contributions

oVQA: A new benchmark for diagnosing Text-VLM performance in an open-ended VQA setup


- Remove ambiguities
- Ask follow-up questions



Contributions

oVQA: A new benchmark for diagnosing Text-VLM performance in an open-ended VQA setup

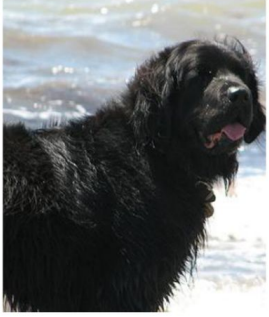
- Remove ambiguities
- Ask follow-up questions
- Use provably strong metrics



What's this?

(Full image)
A tree with no leaves

(Cropped image)
A porcupine



What's this?

A black dog standing in the water

Parent Hierarchy

- dog / domestic dog
- canine / canid
- carnivore
- placental
- mammal / mammalian
- ...
- entity

What type of dog is this ?

Newfoundland dog

oVQA benchmark



Dataset: VQAv2
Question: Where is the cat?
Label: on desk (x4), desk (x3), center of picture, at home, on table

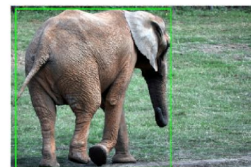


Dataset: GQA
Question: What is the spoon made of?
Label: metal

Objects



Dataset: ImageNet
Question: What's this?
Label: cougar



Dataset: COCO
Question: What's this?
Label: elephant

Actions



Dataset: ActivityNet
Question: What activity is this?
Label: playing drums

Attributes

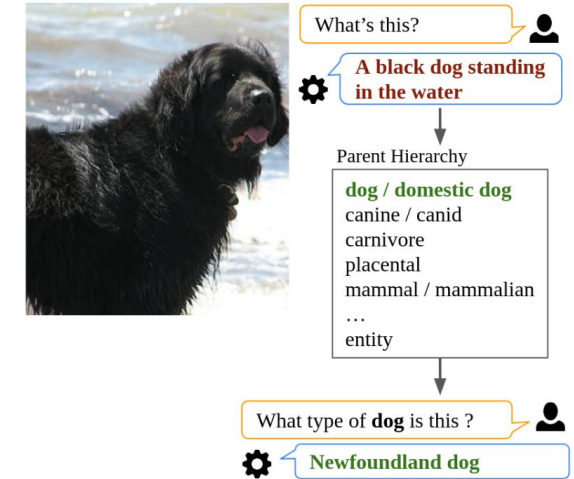
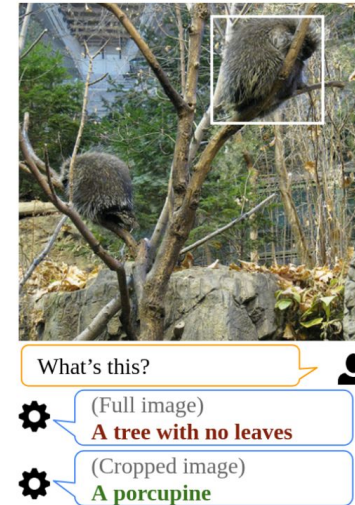


Dataset: OVAD
Question: What is the position of the person?
Label: standing / upright / vertical

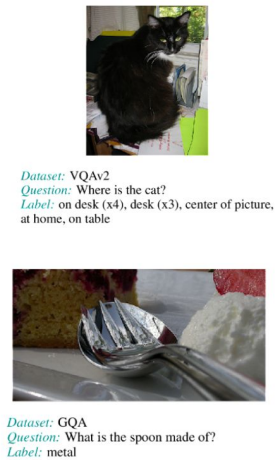
Contributions

oVQA: A new benchmark for diagnosing Text-VLM performance in an open-ended VQA setup

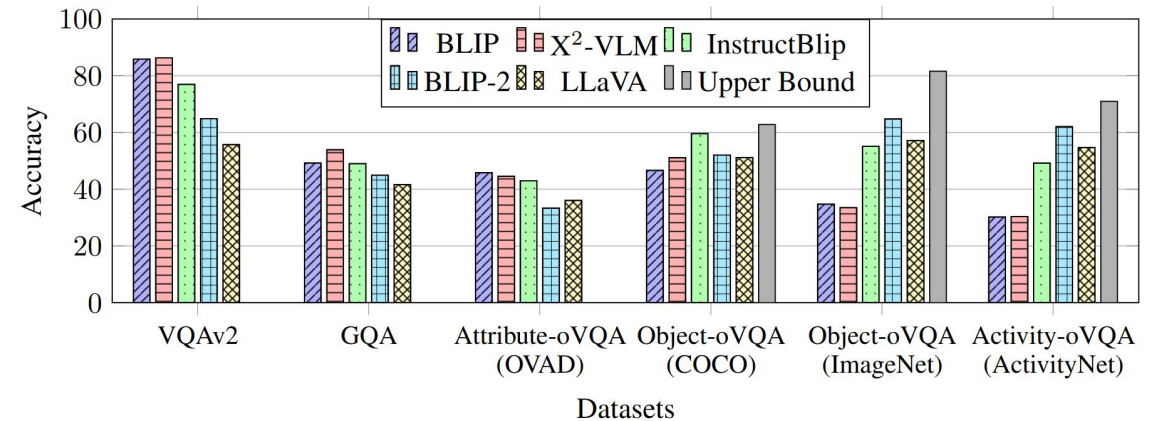
- Remove ambiguities
- Ask follow-up questions
- Use provably strong metrics



oVQA benchmark



Objects





DFG



small data
MORE FROM LESS



Open-ended VQA benchmarking of Vision-Language models by exploiting Classification datasets and their semantic hierarchy



Simon Ging *



María A. Bravo *



Thomas Brox

University of Freiburg



Poster session

Halle B, Thu 9 May, 4:30 p.m.

iclr.cc/virtual/2024/poster/19102

[CV] Vision

COMPUTER VISION University of Freiburg

* Equal Contribution



ICLR
International Conference On
Learning Representations