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

COMPUTER VISION University of Freiburg

Simon Ging* María A. Bravo* Thomas Brox {gings, bravoma, brox}@cs.uni-freiburg.de *Equal contribution

ICLR 2024

Motivation

Visual Dialog

Where is the scene taking place?

It 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.

Visual Question Answering

How many dogs are in the image?

There are three dogs.

How many dogs are in the image?

3 dogs.

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.

Text-generative Vision-Language Models show impressive capabilities, but their evaluation is challenging.

Evaluating **open-ended Visual Question Answering (oVQA)** requires flexible metrics that capture the richness of natural language. It requires non-ambiguous questions which point to semantically concrete answers and a well structured benchmark to uncover the strengths and weaknesses of the models.

Our **oVQA Benchmark:**

- **Fairly** compares Text-VLMs with discriminative VLMs.
- Identifies and resolves **question ambiguities**.
- Probes for **detailed strengths and weaknesses**.
- Uses a **proper evaluation** for free-form answers.

Challenges in oVQA → oVQA Benchmark Solution

1. Missing **fine-grained evaluation** of model strengths → Stratified benchmark for **objects**, **actions**, and **attributes**

2. Ambiguity of refered object → **Visual Guidance**

What's this?

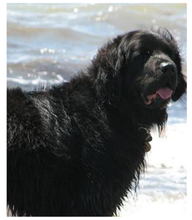
Label: **Porcupine**

crop

A porcupine

A tree with no leaves

3. Unknown answer granularity → **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 ?

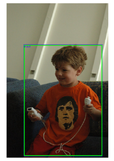
Newfoundland dog

4. Weaknesses of existing **metrics** → Apply metrics that are **better suited for open-ended answer generation**.

In **binary classification** we evaluate whether the model output matches the label. We use the **Contains** metric with synonyms to judge open-ended answers.

In **multi-class classification** we can additionally use the label space. With the **ClipMatch** metric, we select and judge the nearest answer in the label space.

Which is the maturity of the person?



Label: **young**

Synonyms: **{young, baby, child, kid}**

***"Exact match"**: Answer matches the label?*

***"Contains" metric**: Any synonym is contained in the answer?*

child ✗ ✓

He's a kid. ✗ ✓

What's this?

a mountain lion

Label: **cougar**

***ExactMatch**: answer matches label exactly* ✗

***Contains**: label is contained in answer* ✗

***ClipMatch:** most similar label in Clip text space* ✓

ClipMatch Top-5 similarities:

0.79 cougar
0.68 lion
0.64 snow leopard
0.53 lynx
0.51 leopard

Datasets and Experimental Setup

We transform **classification datasets** into VQA benchmarks and compare them to established VQA benchmarks.

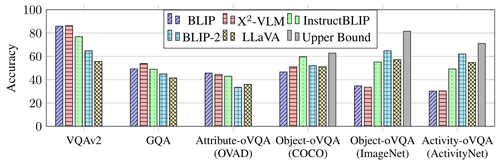
oVQA classification benchmarks

	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?

Classical VQA benchmarks

	VQAv2	GQA
# Images	41k	0.4k
# Questions	214k	13k
# Answers	2.1M	13k
Question ex.	Where is the cat?	What is the spoon made of?

Results



Pretrained generic models are stronger in **classifying fine-grained objects and activities** (last two bar plots).

Models **trained for VQA** are stronger in **classifying coarse concepts and answering common VQA questions** (first three bar plots).

Model	INet-Crop	COCO obj	ActivityNet
CLIP ViT-L-14	77.90	55.20	68.76
CLIP EVA 01-g-14	**81.60**	62.81	**71.14**
BLIP-2	69.20	57.16	61.03
BLIP-2 (COCO ft)	65.21	**62.98**	59.59

Upper Bound. Zero-shot Retrieval using discriminative VLMs.

	Method	ClipM	Cont	EM
1st Question	$BLIP_{vqa}$	24.51	10.22	9.24
	X^2-VLM_{vqa} L	23.67	9.42	8.84
	BLIP-2 OPT	**57.10**	**35.49**	0.87
	BLIP-2 T5	54.50	30.05	1.77
	InstructBLIP V	39.69	20.03	**19.30**
	LLaVA	45.07	23.18	0.00
Follow-up Q.	$BLIP_{vqa}$	34.68	15.00	13.85
	X^2-VLM_{vqa} L	33.50	13.69	12.92
	BLIP-2 OPT	**67.22**	**40.31**	2.54
	BLIP-2 T5	64.71	34.43	4.81
	InstructBLIP V	55.10	27.38	**26.40**
	LLaVA	57.12	27.59	0.00

Object-oVQA ImageNet.

	Method	ClipM	Cont	EM
1st Question	$BLIP_{vqa}$	23.73	7.30	6.70
	X^2-VLM_{vqa} L	24.21	6.92	6.34
	BLIP-2 OPT	**55.08**	14.66	7.10
	BLIP-2 T5	53.39	**15.70**	7.07
	InstructBLIP V	40.64	13.61	**12.51**
	LLaVA	46.75	8.76	0.00
Follow-up Q.	$BLIP_{vqa}$	30.16	8.65	8.05
	X^2-VLM_{vqa} L	30.35	8.24	7.65
	BLIP-2 OPT	60.96	16.14	7.68
	BLIP-2 T5	**62.02**	**18.09**	8.84
	InstructBLIP V	49.18	15.09	**13.99**
	LLaVA	54.67	10.68	0.00

Activity-oVQA ActivityNet.

Method	ClipM	Cont	EM
$BLIP_{vqa}$	46.58	26.04	21.74
X^2-VLM_{vqa} L	51.09	25.00	22.50
BLIP-2 OPT	47.62	33.59	0.46
BLIP-2 T5	51.97	32.50	0.10
InstructBLIP V	**59.58**	27.52	**26.50**
LLaVA	51.13	**45.08**	0.00

Object-oVQA COCO
(Common Objects in Context)

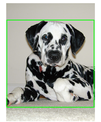
Method	Cont Syn	EM Syn
$BLIP_{vqa}$	**45.70**	36.99
X^2-VLM_{vqa} L	44.56	39.13
BLIP-2 OPT	21.89	5.23
BLIP-2 T5	33.45	15.73
InstructBLIP V	42.92	**40.42**
LLaVA	36.02	0.00

Attribute-oVQA OVAD
(Open-Vocabulary Attribute Detection)

Model	VQAv2 test-dev EM	VQAv2 val EM	VQAv2 val Cont	GQA testdev balanced EM	GQA testdev balanced Cont
BLIP-2 OPT	51.32	51.51	52.66	32.21	32.93
BLIP-2 T5	**62.80**	**63.13**	**64.80**	**44.03**	**44.88**
LLaVA	0.00	0.00	55.66	0.00	41.58
$BLIP_{vqa}$	77.50	*83.69*	*85.81*	47.04	49.20
X^2-VLM_{vqa} B	79.64	*83.05*	*84.00*	52.54	53.62
X^2-VLM_{vqa} L	**81.34**	*85.33*	*86.29*	**52.80**	**53.90**
InstructBLIP T5	*74.75*	*73.22*	*74.29*	47.61	48.46
InstructBLIP V	*77.54*	*76.12*	*76.92*	48.34	48.97

Classical VQA Benchmarks.
Top: zero-shot, bottom: finetuned.
Italic: Evaluated data is part of training.

Qualitative Results



ImageNet oVQA
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.

COCO oVQA
Question: What's this? *Label:* surfboard
$BLIP_{vqa}$ *output*: dog
BLIP-2 T5 *output*: a photo of a man holding a surfboard
InstructBLIP T5 *output*: surfboard
LLaVA *output*: The image is a blurry photo of a person holding a surfboard (...)
X^2-VLM_{vqa} L *output*: man

OVAD oVQA
Question: Which texture does the dining table have? *Label:* sleek / smooth
$BLIP_{vqa}$ *output*: smooth
BLIP-2 OPT *output*: it feels like wood, but it's not wood, it's laminate
InstructBLIP T5 *output*: wood
LLaVA *output*: The dining table's surface is made of wood (...)
X^2-VLM_{vqa} L *output*: smooth

ActivityNet oVQA
Question: What activity is this?
Label: beach volleyball
BLIP-2 OPT *output*: volleyball
Hierarchy: playing volleyball
Follow-up Question: What type of *playing volleyball* is this?
BLIP-2 OPT *follow-up output*: beach volleyball

(Legend: correct answer, wrong answer)

Metrics User Study

Pearson correlation between humans and metrics shows:

- **LLMs** outperform all other metrics.
- **Contains** metric outperforms learned metrics and translation metrics.

Example **metric output**:

What are the vegetables to the left of the bowl that is to the left of the cookies?
Label: **carrots**

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

Metric	Pearson corr.
$GPT\text{-}4_{10\text{-shot}}$	**0.972**
$GPT\text{-}4_{0\text{-shot}}$	0.947
$Llama2_{5\text{-shot}}$	0.919
$Llama2_{0\text{-shot}}$	0.798
Contains	0.906
ExactMatch	0.525
BertScore	0.223
BLEURT-20	0.554
BEM	0.627
LERC	0.827
BLEU-1	0.637
BLEU-4	0.023
CIDEr	0.597
METEOR	0.689
ROUGE	0.717

Human evaluation pipeline. We evaluate 2000 metric predictions each with 3 human workers.

What type of donut is on the top right?

Label: **chocolate iced glazed**

chocolate glazed donut

I rate **5/5**.

Conclusions / Contributions

- Our work establishes a framework for fair evaluation of Text-VLMs by removing ambiguities and asking follow-up questions.
- It allows fair comparison of classical Text-VLMs, Image-dialog models, and discriminative VLMs.
- Our oVQA benchmark enables diagnosing Text-VLM performance on object, action, and attribute classification.
- Our User study shows *Contains* metric and LLM metrics to be strong compared to existing NLP metrics.
- Our benchmark reveals an imbalance in model strengths, with pretrained generic models excelling in fine-grained classification, and VQA-trained models performing better in answering common VQA questions.

Our poster on iclr.cc:
iclr.cc/virtual/2024/poster/19102

Code for the benchmark and metrics:
github.com/lmb-freiburg/ovqa

Acknowledgements. This work was funded by the German Research Foundation (DFG) - 401269959, 417962828 and 499552394 (SFB 1597 - Small Data). This work was supported by the German Academic Exchange Service (DAAD) - 57440921 Research Grants - Doctoral Programmes in Germany, 2019/20. The authors acknowledge support from the state of Baden-Württemberg through bwHPC.

universität freiburg

DAAD small data MORE FROM LESS DFG