Advancing Vision-Language Models for Open-vocabulary Recognition and Generative Evaluation

María Alejandra Bravo Sarmiento

May 30, 2025

PhD Defense. Albert-Ludwigs-Universität Freiburg
Erstgutachter und Betreuer: Prof. Dr. Thomas Brox
Zweitgutachter: Prof. Dr. Phillip Isola
Beisitzer: Prof. Dr. Abhinav Valada
Vorsitzer: Prof. Dr. Matthias Teschner

Motivation


Caroline Jayne Church
I Am a Big Sister!
¡Soy una hermana mayor!
A bilingual book
Un libro bilingüe
SCHOLASTIC


Here's a clean one found by me!
¡Aquí hay uno limpio que encontré!

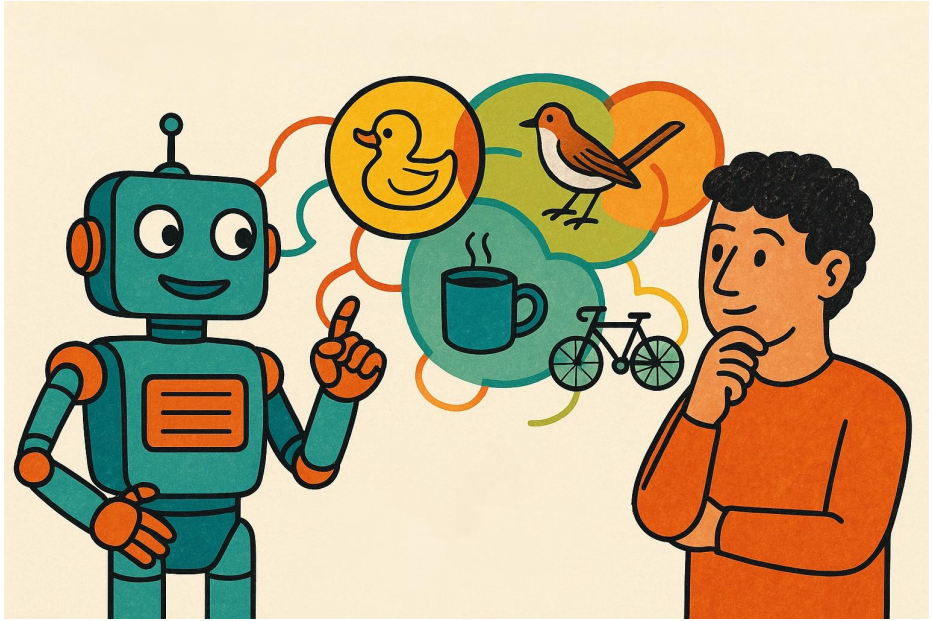
Motivation


Caroline Jayne Church
I Am a Big Sister!
¡Soy una hermana mayor!
A bilingual book
Un libro bilingüe
SCHOLASTIC


Here's a clean one found by me!
¡Aquí hay uno limpio que encontré!

Goals of Visual Language Models

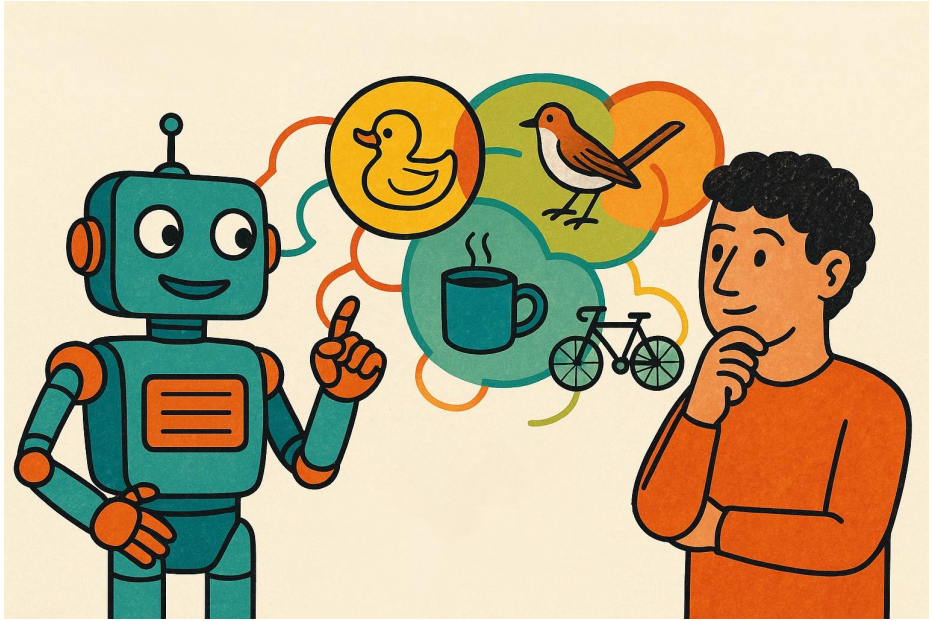
- **Understand more like humans**
  - novel objects
  - fine-grained attributes
  - relationships



DALL·E 3, ChatGPT OpenAI. (May 2025).

Goals of Visual Language Models

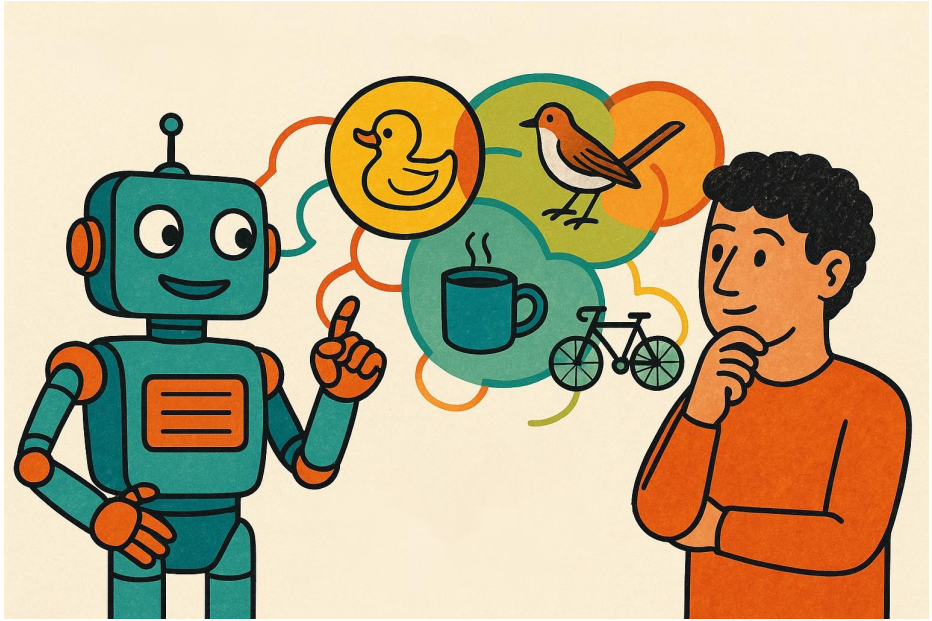
- **Understand more like humans**
  - novel objects
  - fine-grained attributes
  - relationships
- **Communicate more like humans**
  - open-ended
  - precise



DALL·E 3, ChatGPT OpenAI. (May 2025).

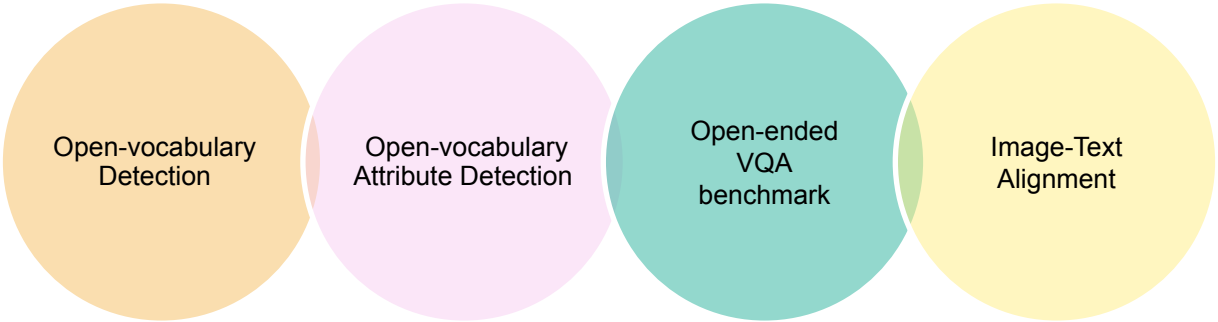
Goals of Visual Language Models

- **Understand more like humans**
  - novel objects
  - fine-grained attributes
  - relationships
- **Communicate more like humans**
  - open-ended
  - precise
- **Align better with humans**
  - match human preferences
  - align human values

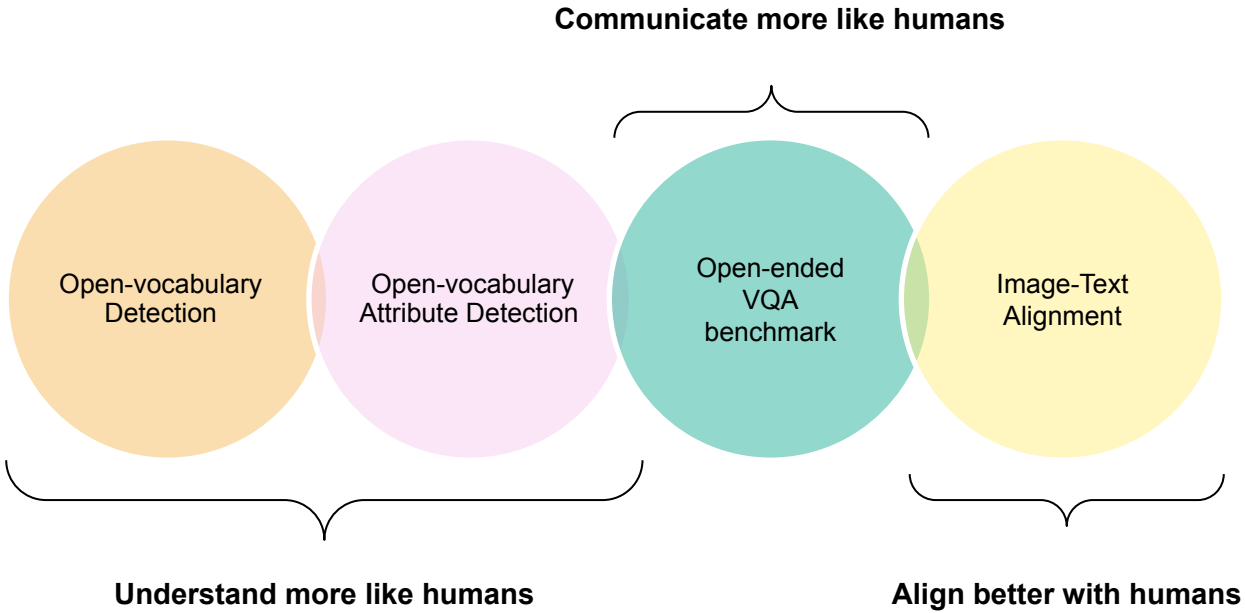


DALL·E 3, ChatGPT OpenAI. (May 2025).

Overview


Open-vocabulary Detection
Open-vocabulary Attribute Detection
Open-ended VQA benchmark
Image-Text Alignment

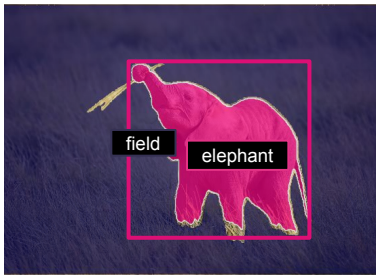
Overview


Communicate more like humans
Open-vocabulary Detection
Open-vocabulary Attribute Detection
Open-ended VQA benchmark
Image-Text Alignment
Understand more like humans
Align better with humans

Vision Language Models for Visual Recognition

A gray baby elephant in a field

VLM



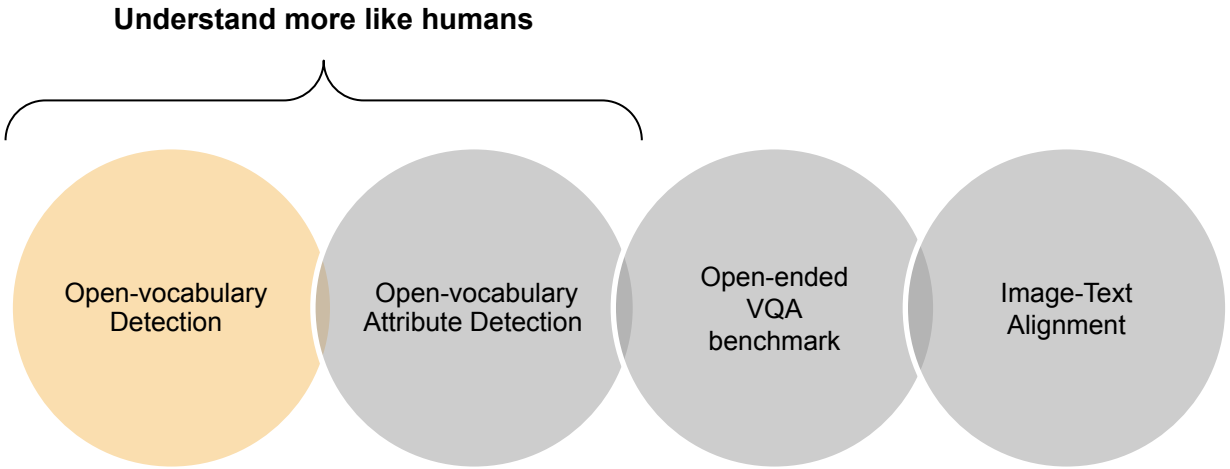
Classes: elephant, field

Beyond a closed-set of categories

Use Image-Text pairs as supervision

Use Natural Language to query classes

Overview

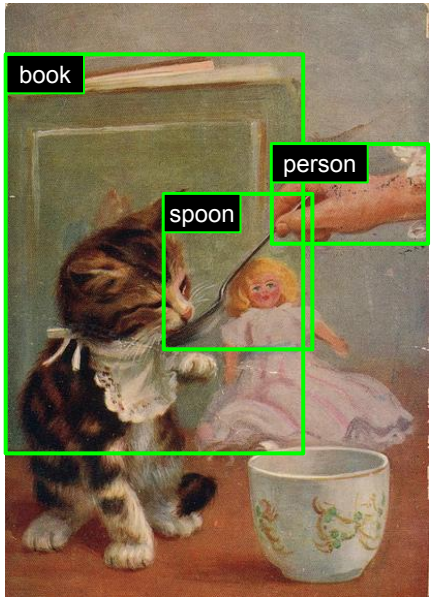


Localized Vision-Language Matching for Open-vocabulary Object Detection

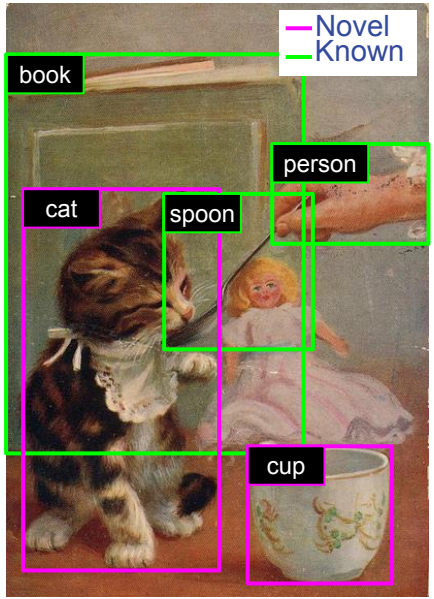
Maria A. Bravo, Sudhanshu Mittal, Thomas Brox

GCPR 2022

Open-vocabulary Object Detection (OVD)



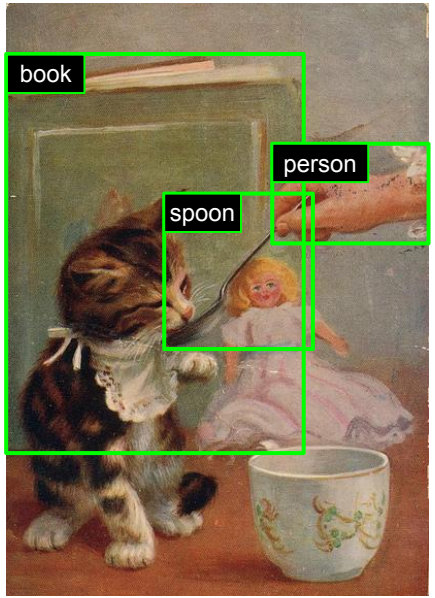
Classical Object Detection



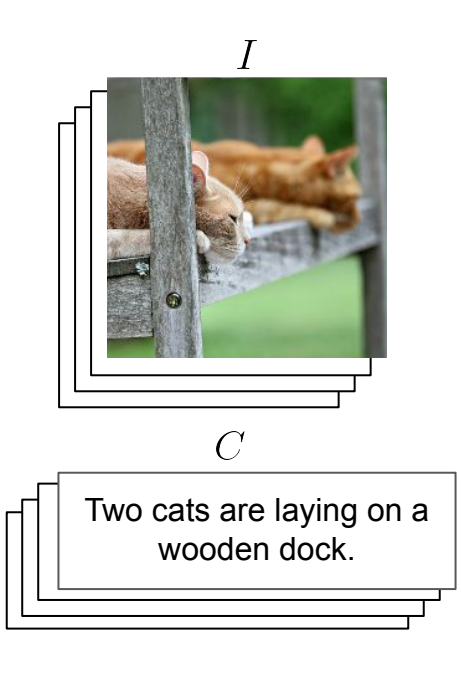
Open-vocabulary Object Detection

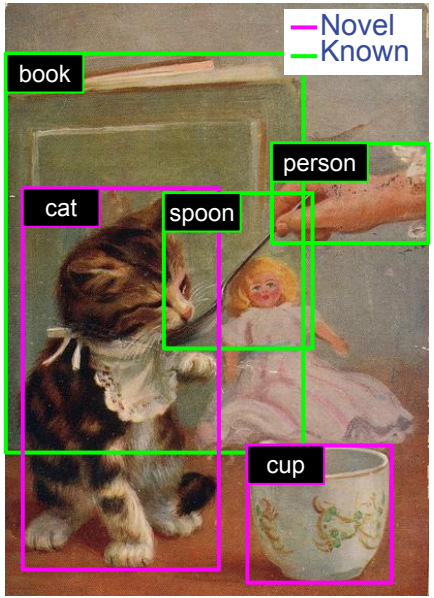
OVR. Zaraian et al. 2021

Open-vocabulary Object Detection (OVD)



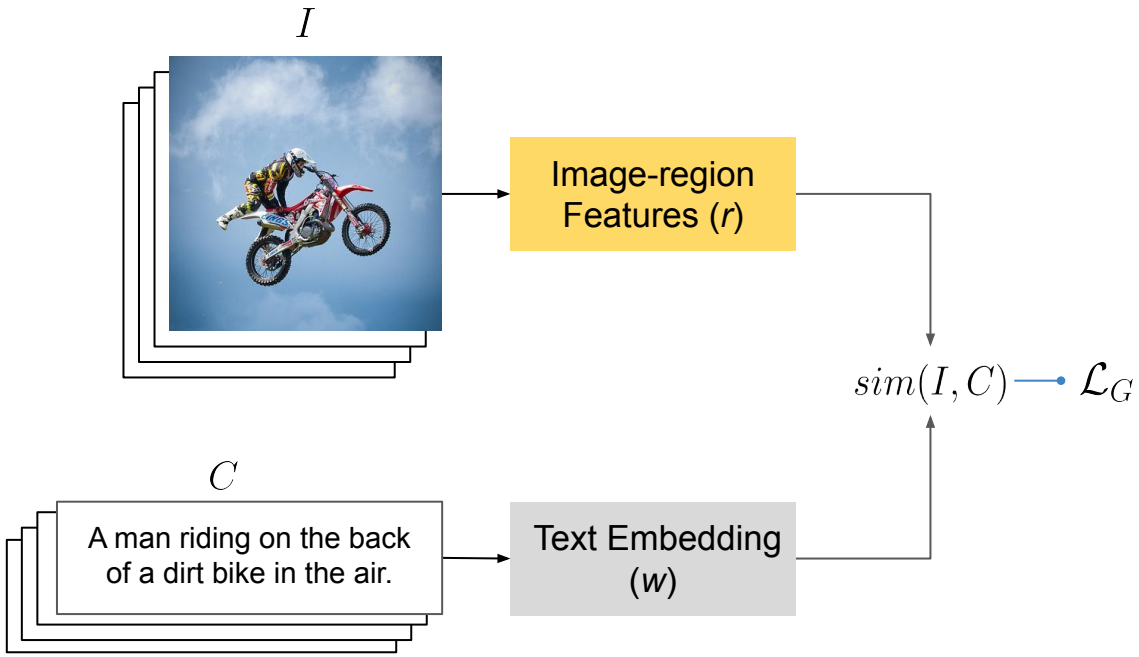
Classical Object Detection



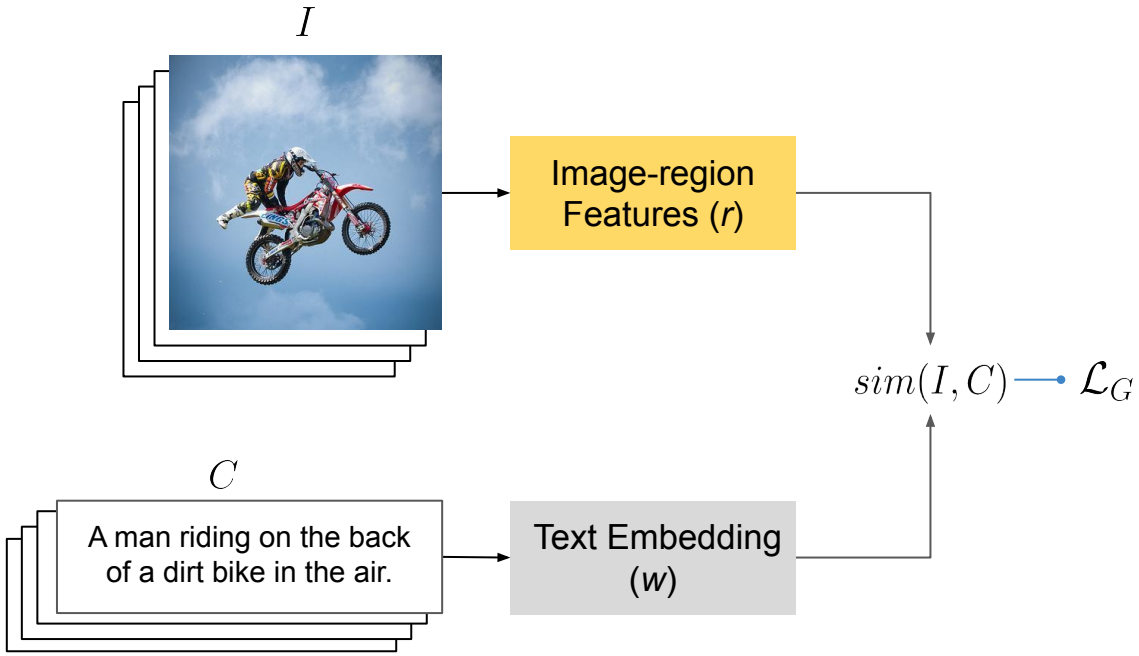


Open-vocabulary Object Detection

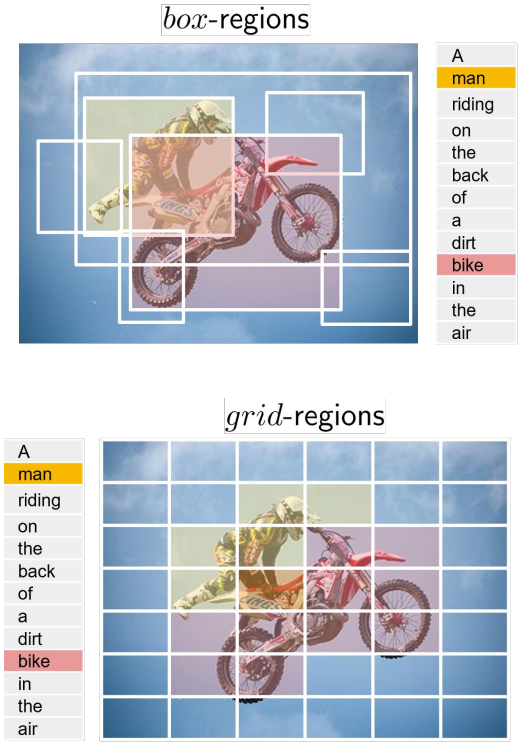
Localized Vision-Language Matching for OVD



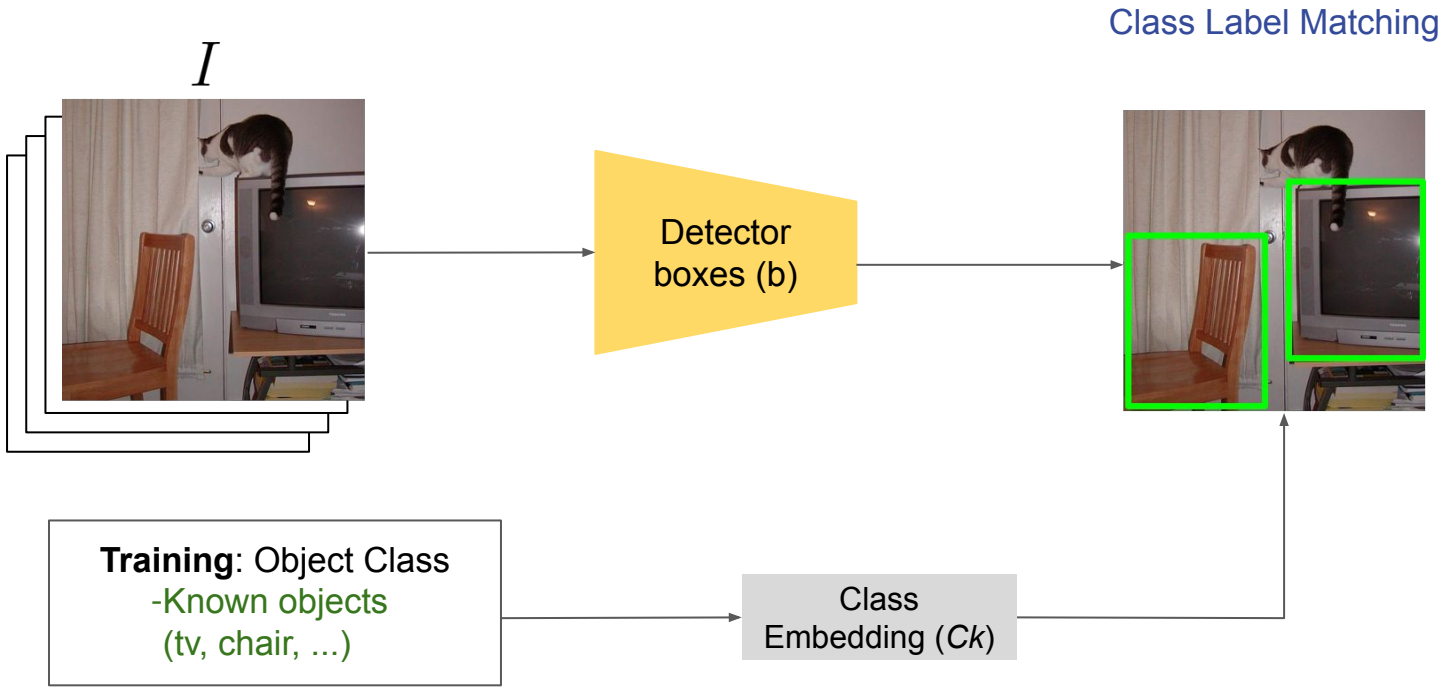
Localized Vision-Language Matching for OVD



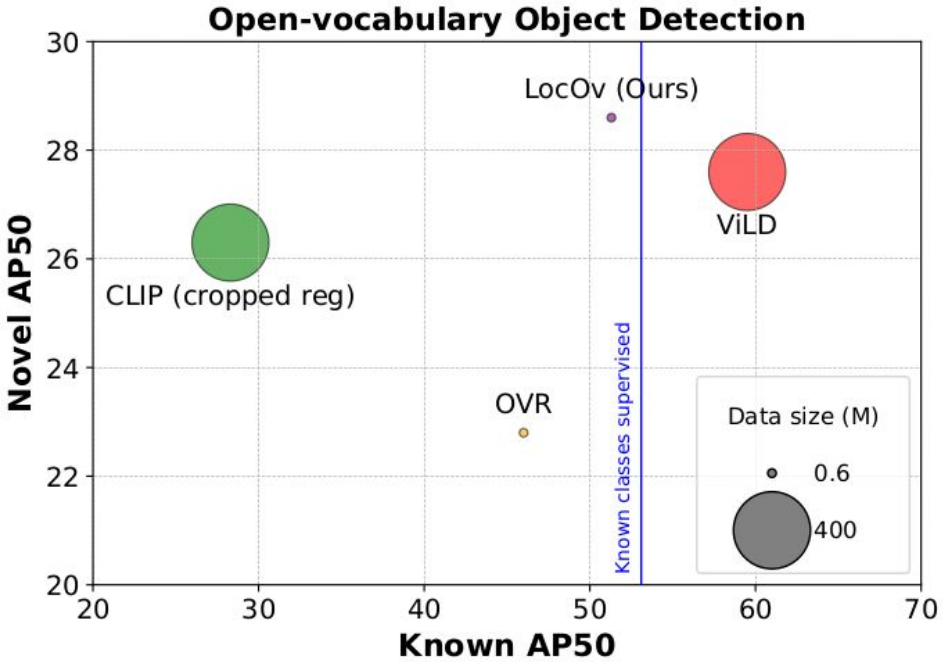
Contrastive Grounding loss



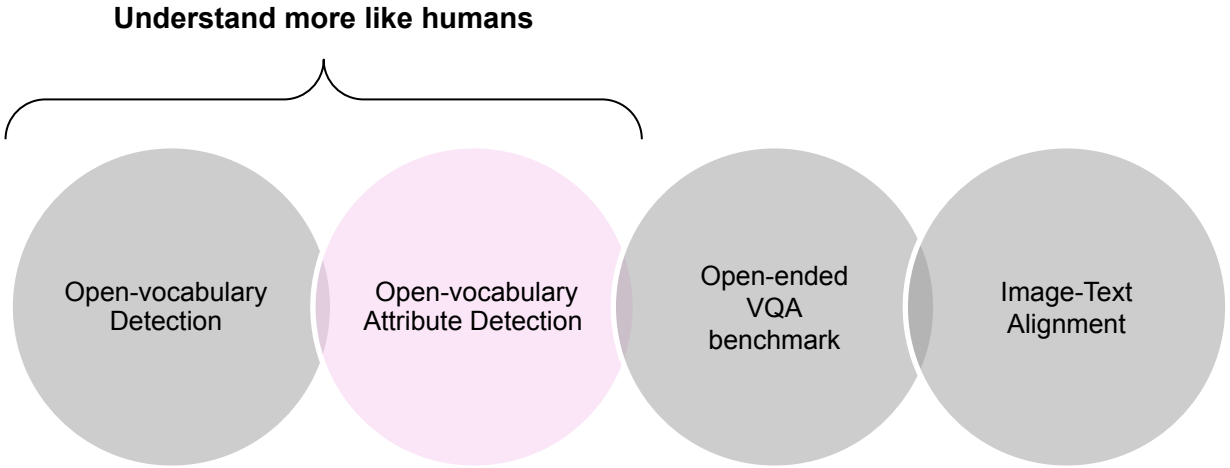
Specialized Task Tuning for OVD



Results: Open-vocabulary Detection



Overview

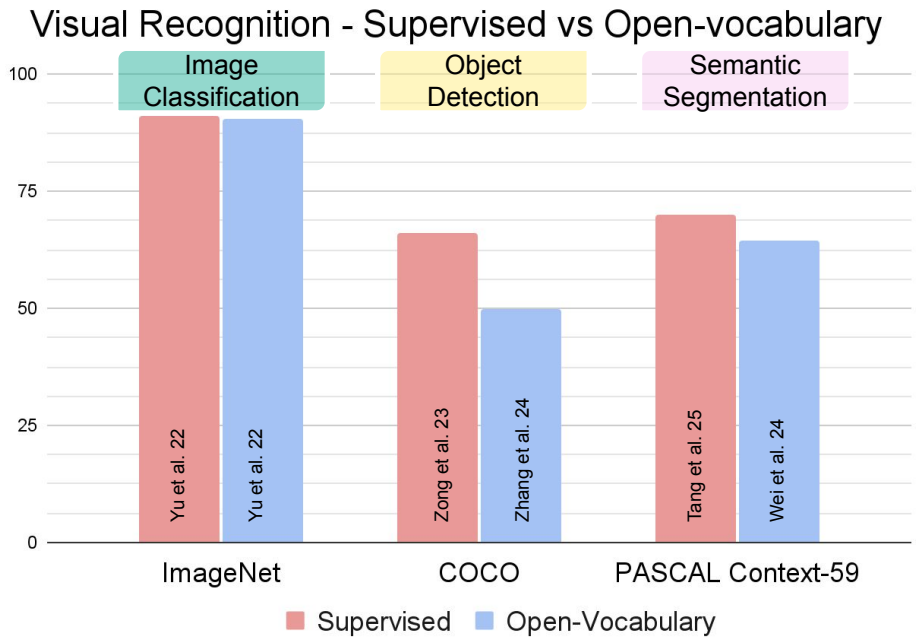


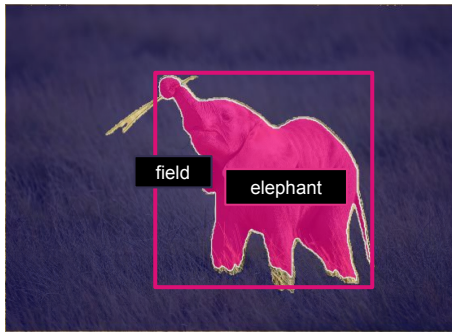
Open-vocabulary Attribute Detection
https://ovad-benchmark.github.io

Maria A. Bravo, Sudhanshu Mittal, Simon Ging, Thomas Brox
CVPR 2023

Vision Language Models for Open Vocabulary Recognition

Noun concepts

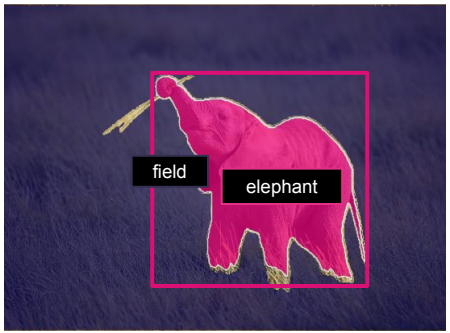




Classes: elephant, field

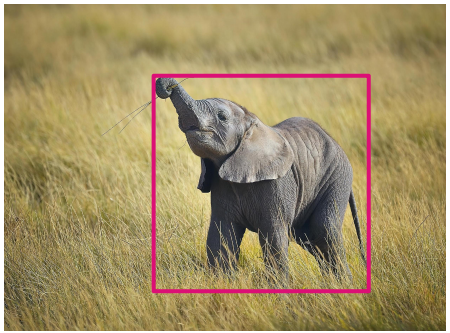
Vision Language Models for Open Vocabulary Recognition

Noun concepts



Classes: elephant, field

Attribute concepts



A gray baby elephant
in a field

Object class: elephant

Attributes
color quantity: one
color: gray
group: single
maturity: baby
pattern: plain
position: vertical
size: small
state: dry
texture: rough
tone: light

Significance of attributes in an object's identity

A **white** and a **spotted** horse on a field of grass.

Red traffic signal in the middle of a **wide** street.

Images from COCO Captions. Chen et al. 2015

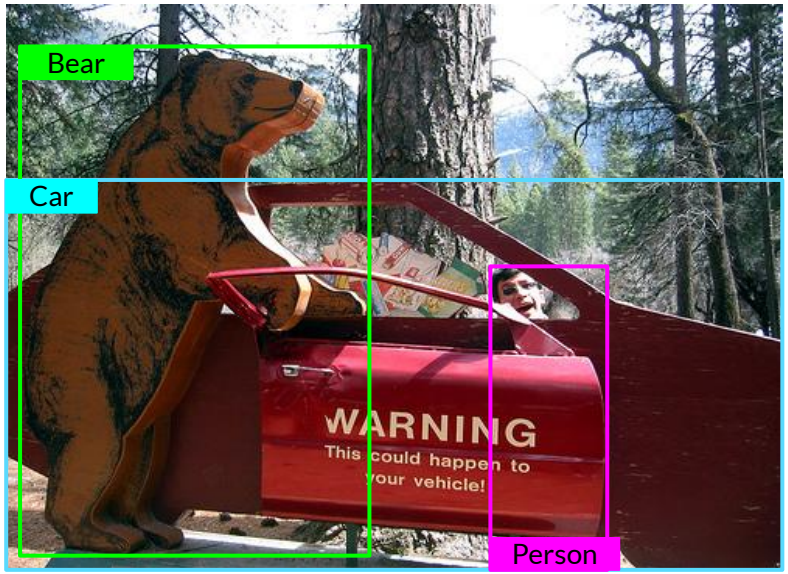
OVAD task: Open-vocabulary Attribute Detection

Object class: bear

Attributes
color quantity: two
color: black and brown
group: single
material: wood
maturity: adult
position: upright
size: big
state: dry
texture: smooth
tone: dark

Solving the **OVAD** task:

(1) Detect all object classes.
(2) Detect their attributes.



Object class: car

Attributes
color quantity: one
color: red
group: single
material: wood
optical prop: opaque
patterns: lettered
state: piece / cut
texture: smooth
tone: dark

Object class: person

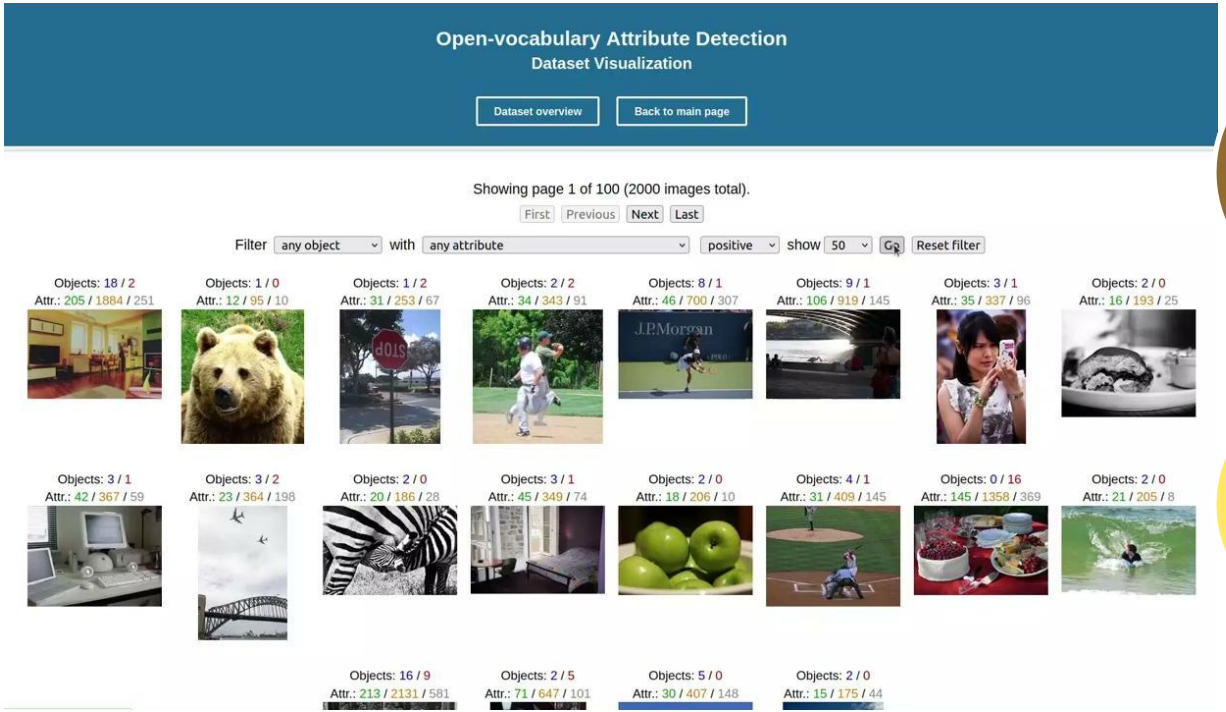
Attributes
face exp: surprise
group: single
hair color: black
hair length: short
hair tone: dark
hair type: straight
maturity: adult
position: upright
clothes color: white

OVAD Benchmark

Human annotated test dataset

Based on MSCOCO Images

96.8 attributes / instance



2,000 Test Images

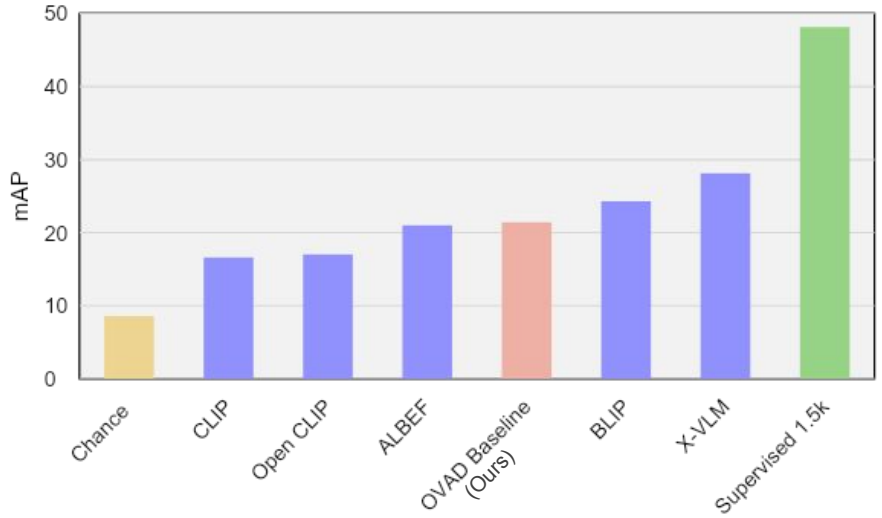
117 Attribute Categories

80 Object Categories

14.3k Object Instances

1.4M Attribute Annotations

VLM Evaluation on Attributes (OVAD)

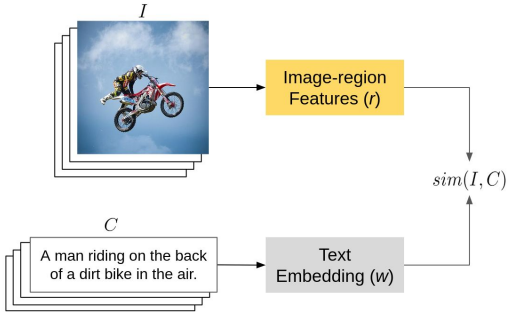


- Large pre-trained VLMs fail to capture fine-grained information.
- X-VLM model performs best due to localized alignment during training.

Contributions

LocOV:

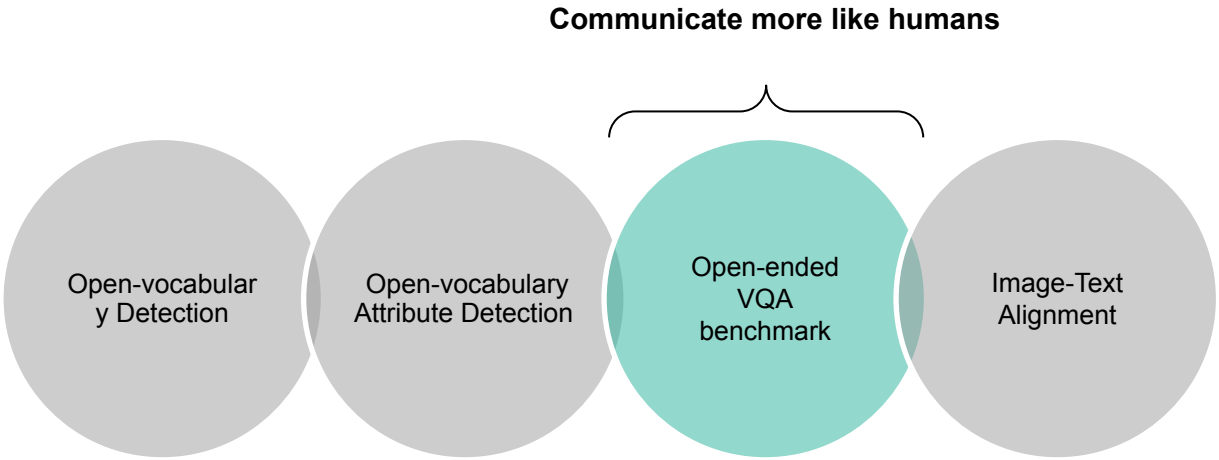
- Proposed a new method that exploits part-wise alignment of visual and text features



OVAD:

- Proposed a new OVAD task
- Introduced a new OVAD benchmark
- Evaluated SOTA methods and highlighted their weaknesses in attribute recognition

Overview



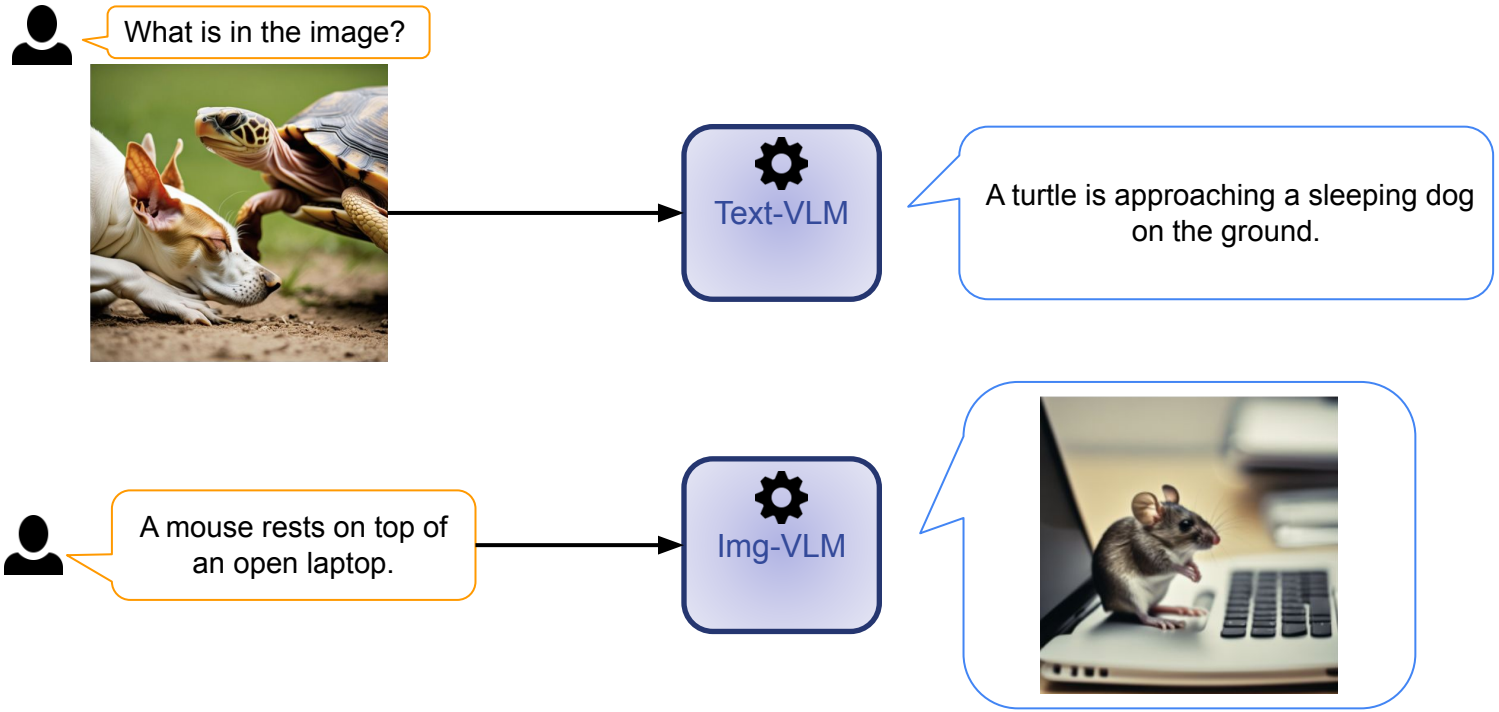
Open-ended VQA benchmarking of Vision-Language models by exploiting Classification datasets and their semantic hierarchy

Simon Ging*, Maria A. Bravo*, Thomas Brox

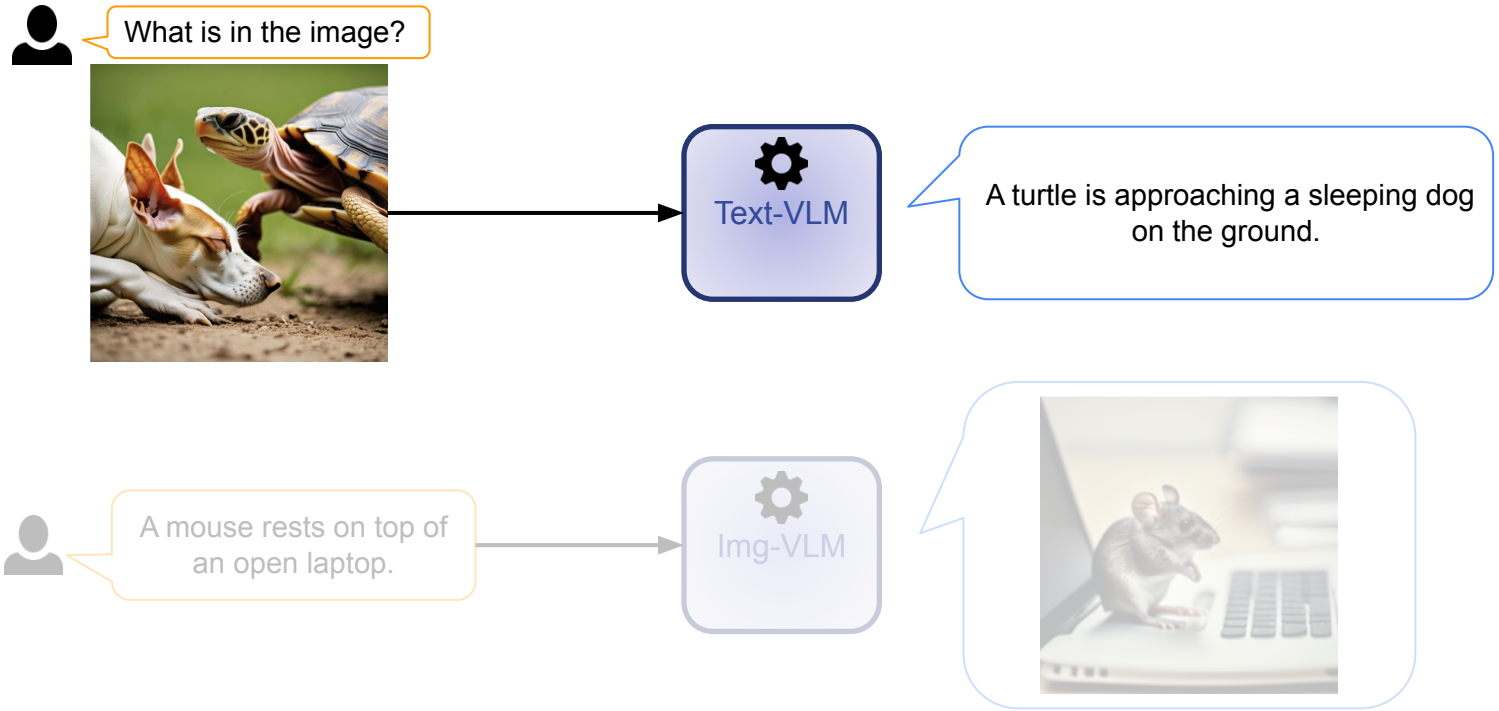
Spotlight ICLR 2024

* Equal Contribution

Vision Language Models for Generation



Vision Language Models for Generation



VLMs for Text Generation

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 in the image.

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.

Who is in the image?

The image includes a person and three dogs on a beach.

VLMs for Text Generation: Unified Evaluation

Visual Question Answering

How many dogs are in the image?

There are three dogs in the image.

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.

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.

Who is in the image?

The image includes a person and three dogs on a beach.

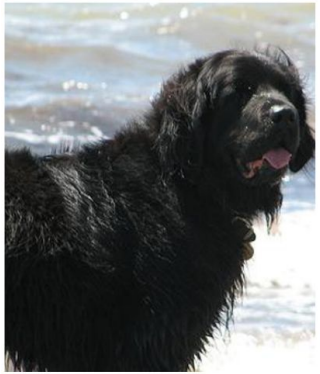
Visual Question Answering (VQA)

Challenges in evaluation of Open-ended VQA

Ambiguous object

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

Unknown label granularity

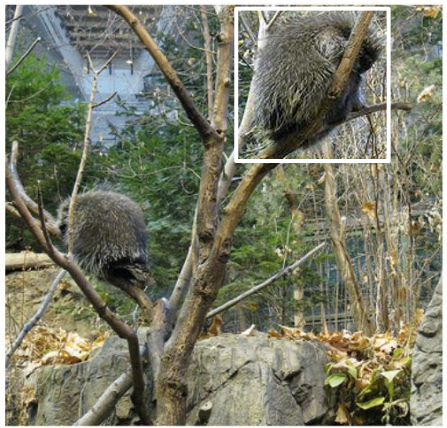


What's this? (*Label:* Newfoundland dog)
Model output: A black dog standing in the water

oVQA Benchmark

Visual Guidance

Ambiguous object



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

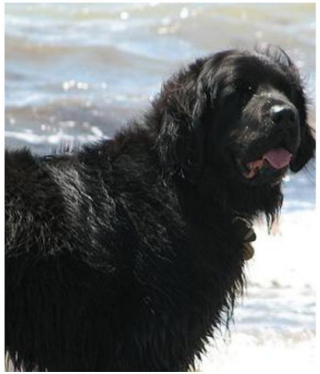
crop

What's this? (*Label:* Porcupine)
Model output: A porcupine

oVQA Benchmark

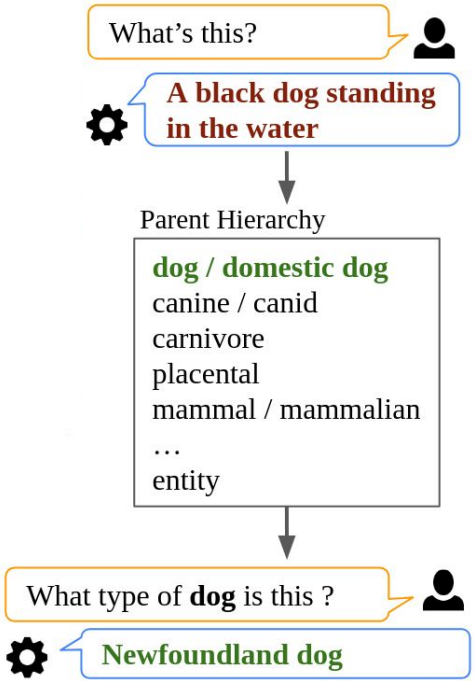
Follow-up Question

Unknown label granularity



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

What type of **dog** is this?
Model output: Newfoundland dog



Open-ended Visual Question Answering Benchmark

Objects

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

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

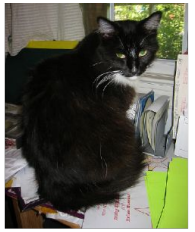
Actions

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

Classical VQA



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

Model results

VQA fine-tuned Models

BLIP X^2-VLM

Instruction-tuned Models

InstructBlip

Generic image-text Models

BLIP-2 LLaVA

Model results

VQA fine-tuned Models

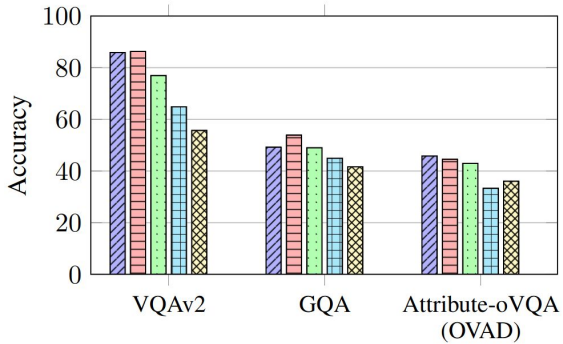
BLIP X^2-VLM

Instruction-tuned Models

InstructBlip

Generic image-text Models

BLIP-2 LLaVA



For VQA and attribute datasets:

- VQA fine-tuned Models > Generic image-text Models

Model results

VQA fine-tuned Models

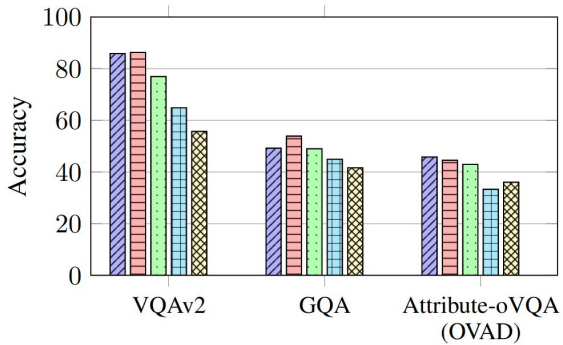
BLIP X^2-VLM

Instruction-tuned Models

InstructBlip

Generic image-text Models

BLIP-2 LLaVA



Attributes (OVAD)

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.

correct answer
wrong answer

For VQA and attribute datasets:

- VQA fine-tuned Models > Generic image-text Models

Model results

VQA fine-tuned Models

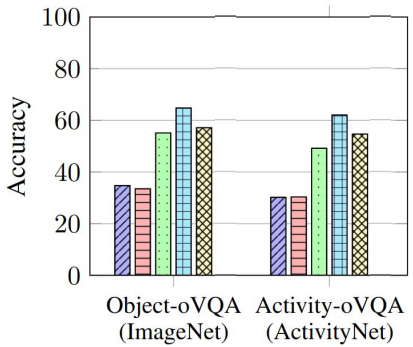
BLIP X^2-VLM

Instruction-tuned Models

InstructBlip

Generic image-text Models

BLIP-2 LLaVA



For fine-grained Objects and Activities

- VQA fine-tuned Models < Generic image-text Models

Model results

VQA fine-tuned Models

BLIP X^2-VLM

Instruction-tuned Models

InstructBlip

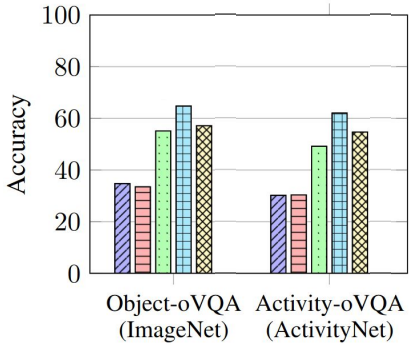
Generic image-text Models

BLIP-2 LLaVA

Activities (ActivityNet)

Question: What is this?
Label: playing blackjack
BLIP-2 OPT *output*: it's a blackjack
X^2-VLM$_{vqa}$ L *output*: table

correct answer
wrong answer



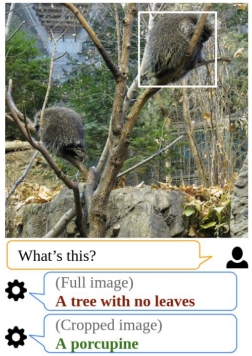
For fine-grained Objects and Activities

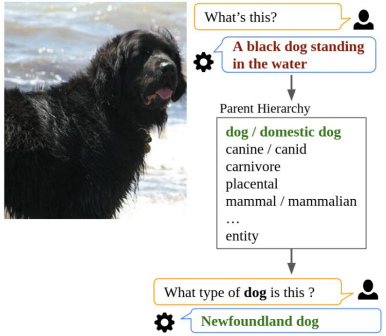
- VQA fine-tuned Models < Generic image-text Models

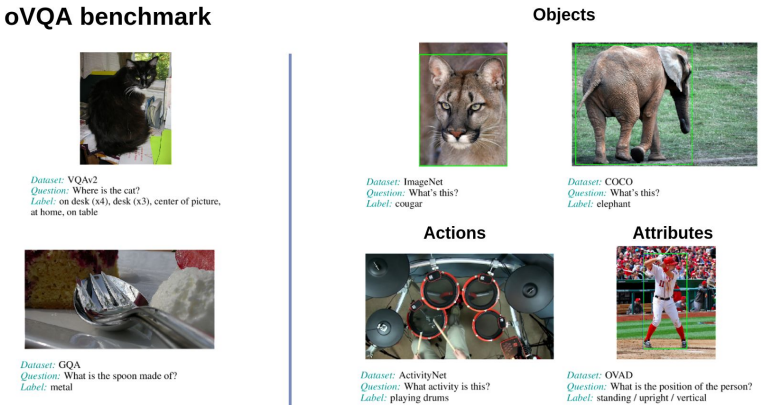
Contributions

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

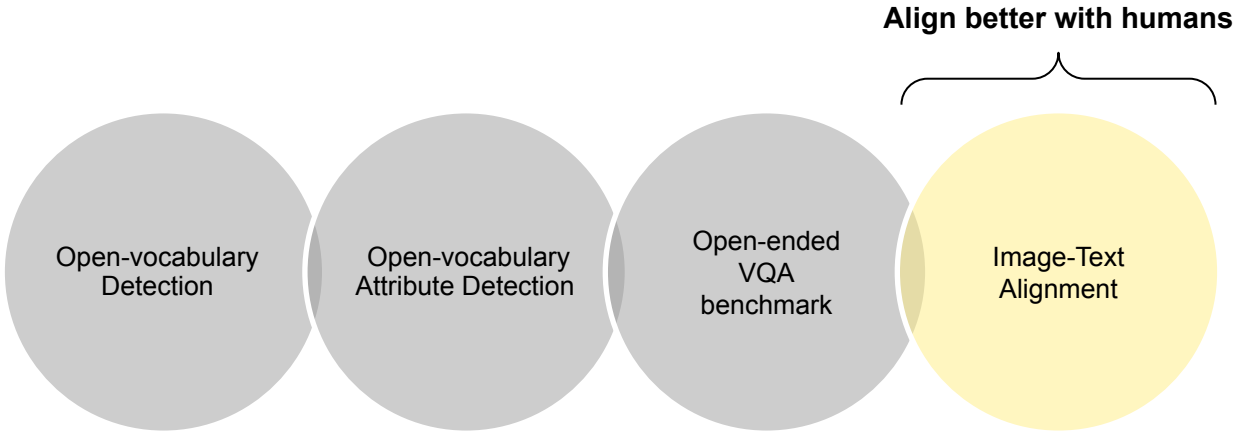
- Remove ambiguities with visual guidance
- Ask follow-up questions
- Use verified strong metrics
- Unified evaluation of Text-VLMs







Overview

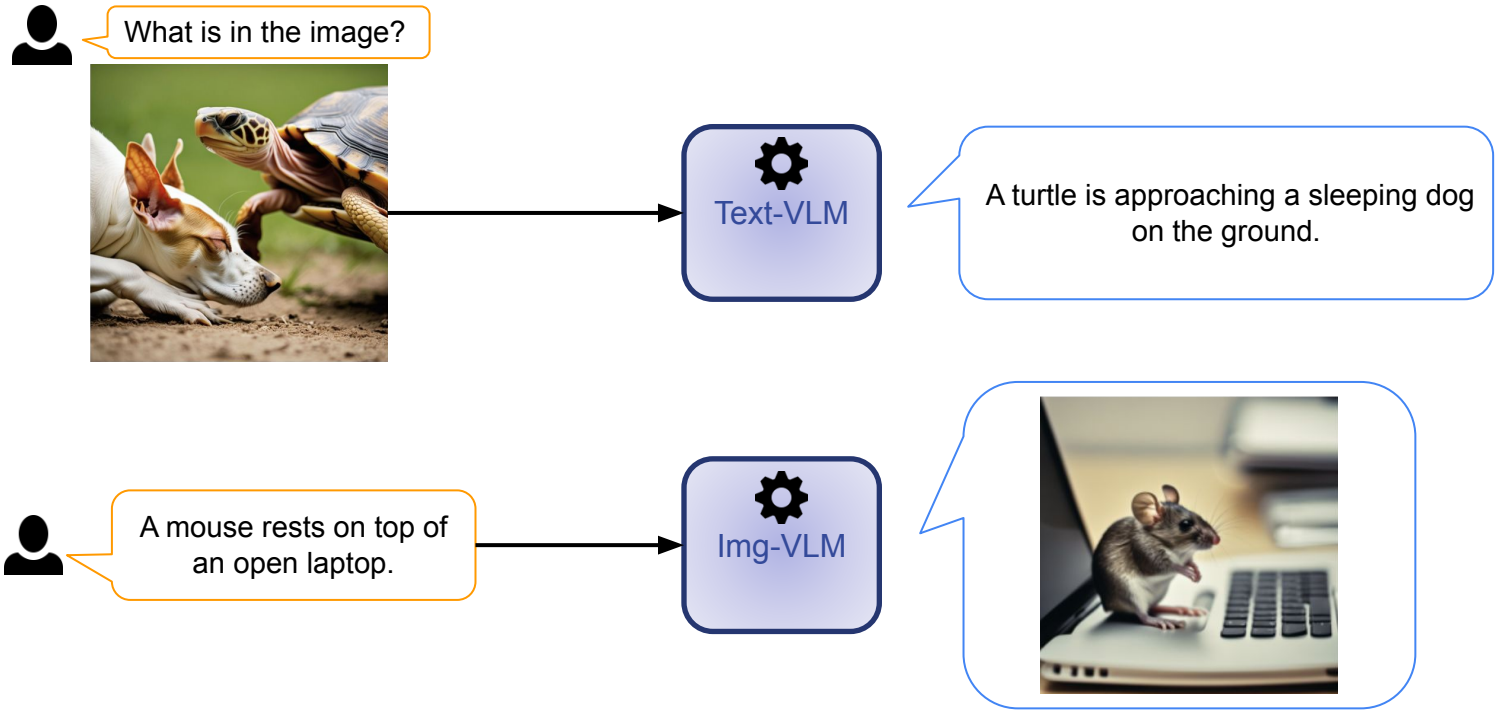


Text-Image Concept Human Alignment Dataset and Metric

Maria A. Bravo

Ack. Betty Mohler, Ali Jahanian, Phillip Isola
NeurIPS 2025 (Workshop)

Vision Language Models for Generation



Text-Image Alignment

Which image is **best described** by the reference text?

A pig is above a table.

Text-Image Alignment

Human guidance

Measuring vision-language alignment requires compositional reasoning skills, and a holistic analysis of both parts.

Which image is **best described** by the reference text?

A pig is above a table.

Text-Image Alignment

Measuring vision-language alignment requires compositional reasoning skills, and a holistic analysis of both parts.

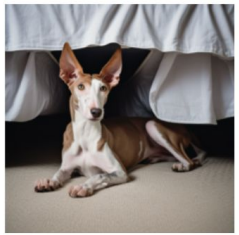
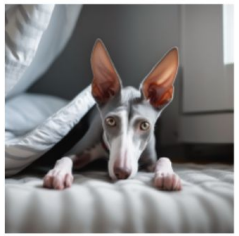
Human guidance

A pig is above a table.

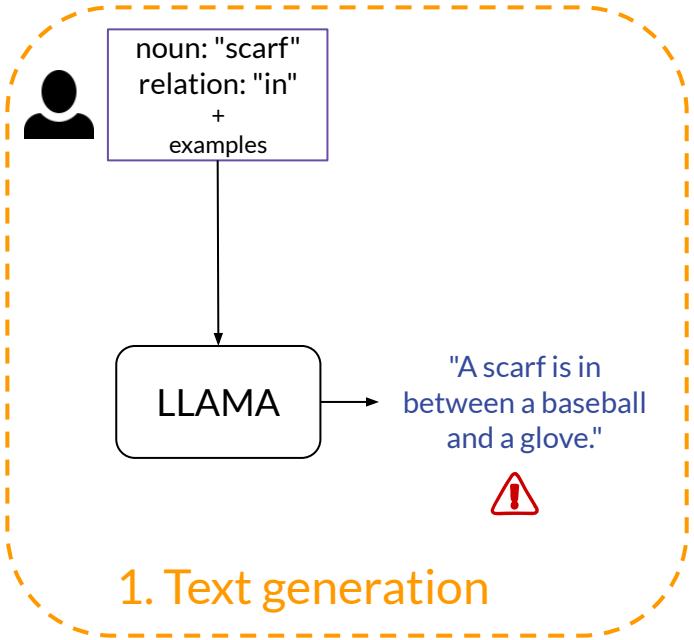
Beautiful iris flowers blooming, yellow and white.

This flower next to a white daffodil looks very pretty.

TIAlign Dataset

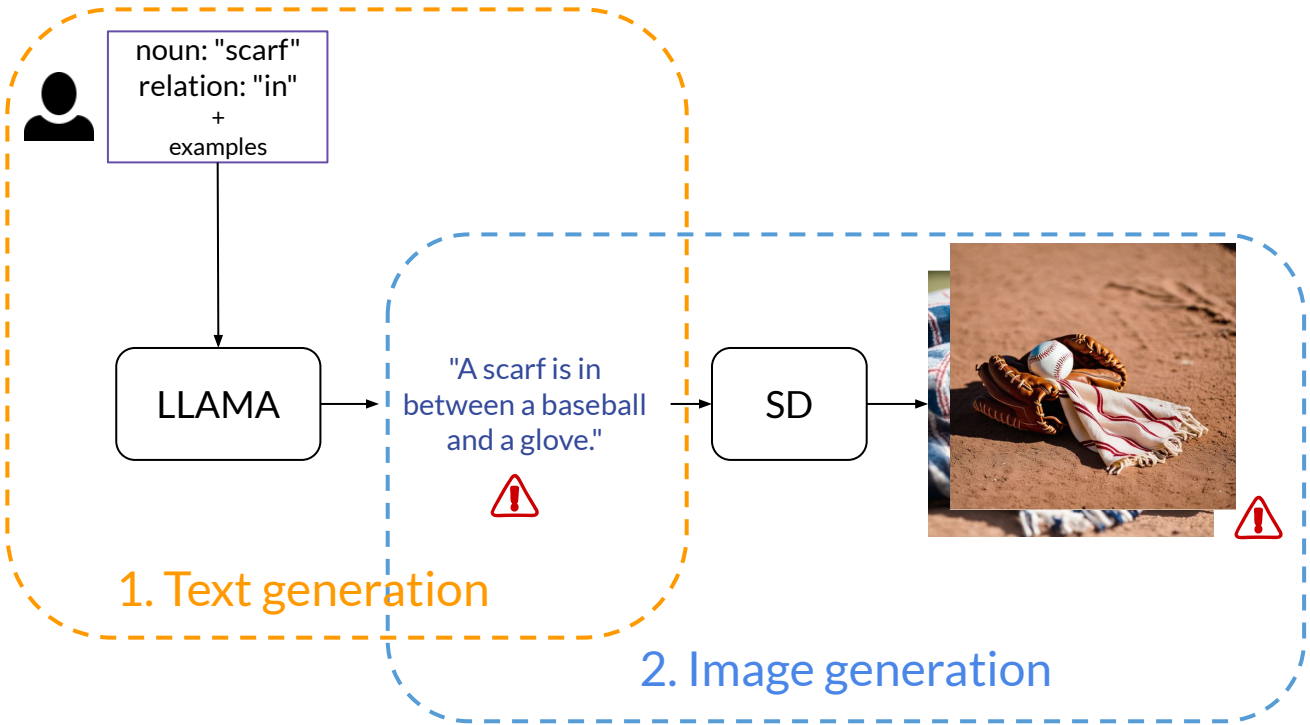
Attribute	Activity	Relation	Definition	Scene
	The picture shows a big orange crab with sharp claws floating on some water with little bubbles around it.		A radiator is a device that transfers heat from a hot fluid to air.	
A slice of garlic bread with purple garlic butter is served on a wooden cutting board.		A gray Ibizan Hound is lying under a bed.		An elderly woman stands on a heliport overlooking a vast city.
	An organism is swimming in the ocean.		Device transfers heat from hot fluid to air.	

Dataset Generation



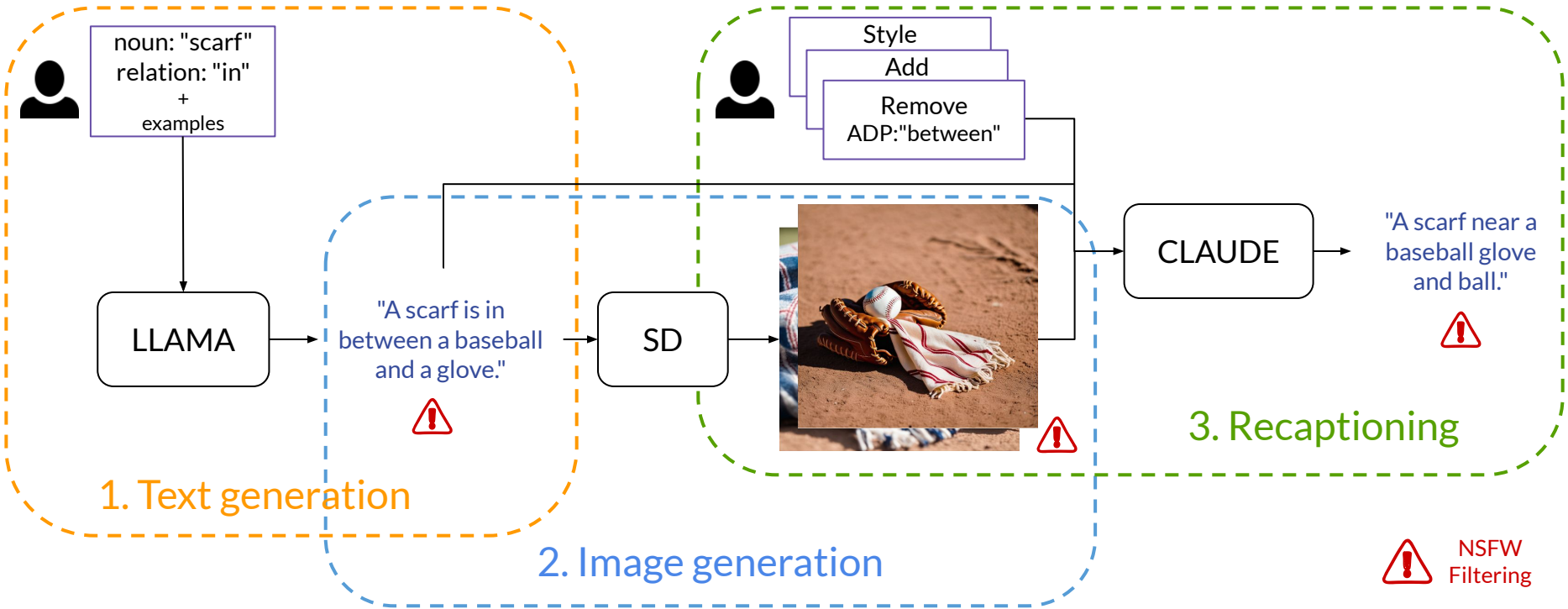
NSFW Filtering

Dataset Generation



NSFW Filtering

Dataset Generation



Data Annotation

Two-alternative forced choice test

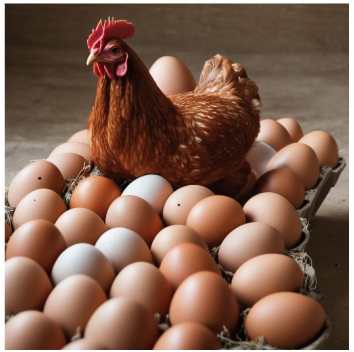
Which image is **best described** by the reference text?

A

Text

A brick hen sits on some eggs.

B



Human

Data Annotation

Two-alternative forced choice test

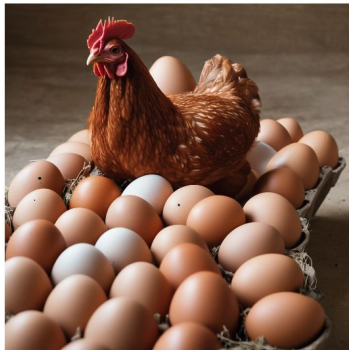
Which image is **best described** by the reference text?

A

Text

A brick hen sits on some eggs.

B



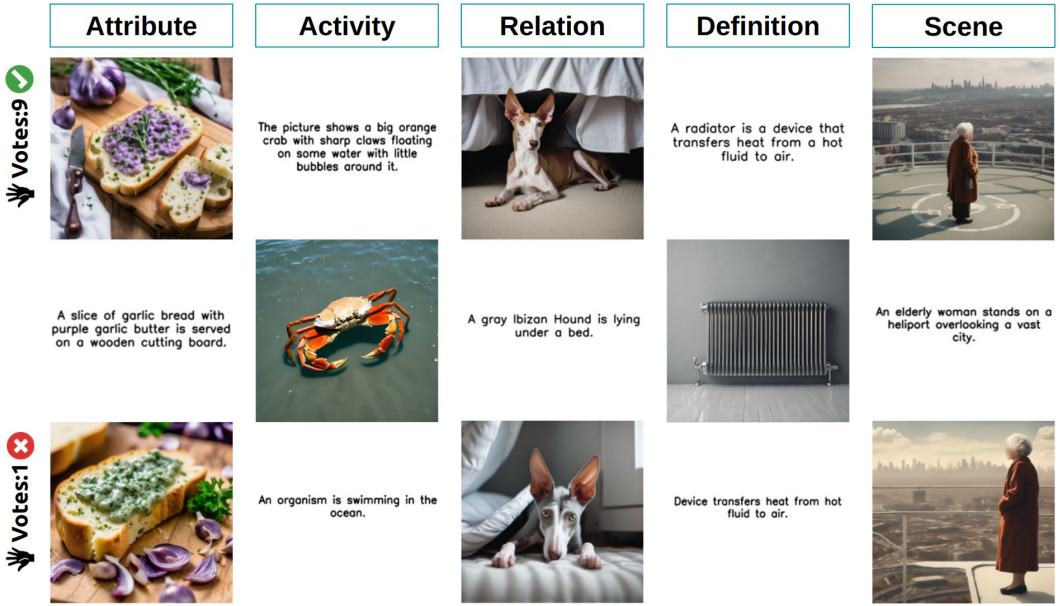
Human

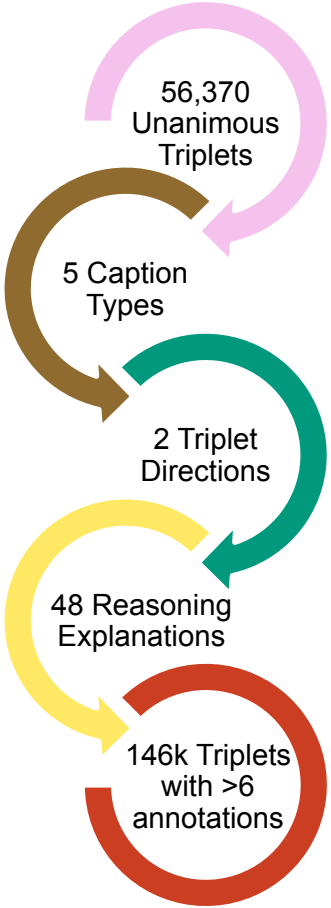
Reasons

- [ ] Object(s) Better
- [ ] Part(s) of Object Better
- [ ] Object: Whole preferred to a part
- [ ] Less Extra Objects
- [ ] Color Attribute Better
- [ ] State Attribute Better
- [ ] Scale Attribute Better
- [ ] Shape Attribute Better
- [ ] Surface Properties Attribute Better
- [x] Action(s) Better
- [ ] Spatial Relation Better
- [ ] Discrete Counting Better

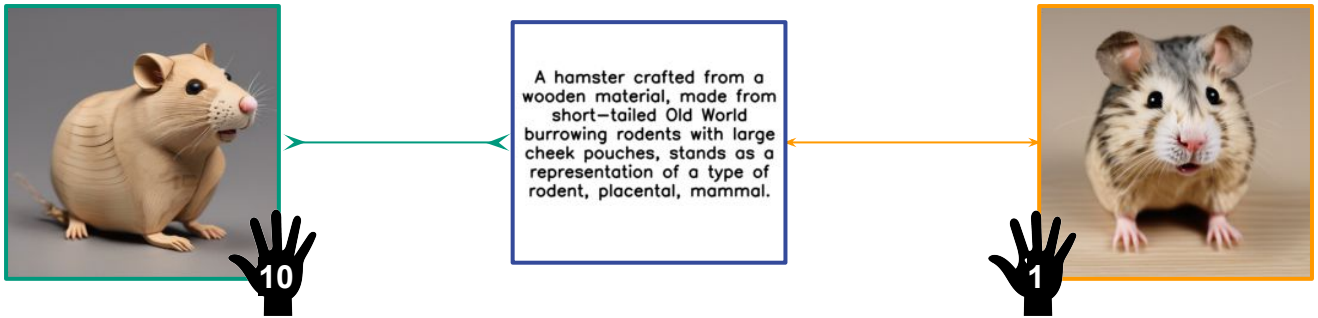
...

TIAlign Dataset

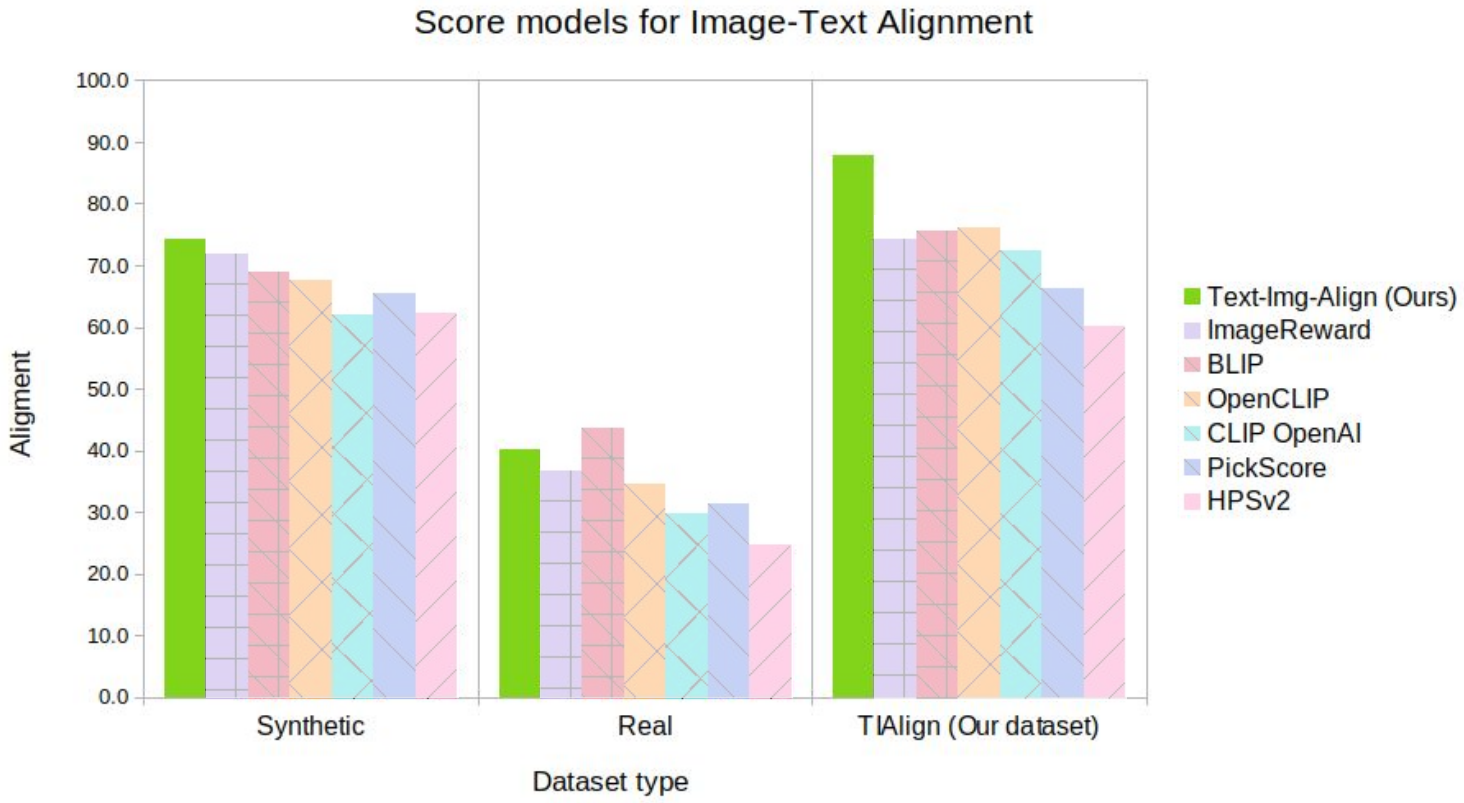




Alignment Score Model



Results for text-image Alignment



Contributions

TIAlign:

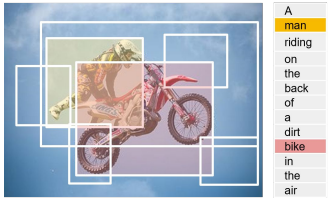
- Diverse large dataset with preference scores
- High human agreement
- Reasoning annotations

Text-Img-Align Score Model:

- Trained on synthetic data
- Outperforms other methods on synthetic data
- Reduces the gap with models trained on real data

Summary

Open-vocabulary detection
→ part-wise alignment



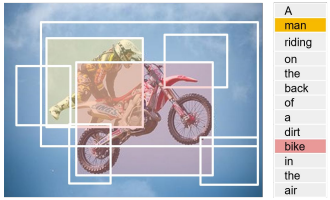
Open-vocabulary Detection

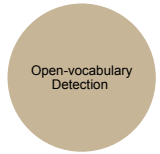
GCPR 2022

Summary

Open-vocabulary detection
→ part-wise alignment

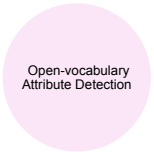
GCPR 2022

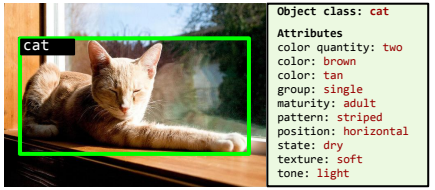




Attribute recognition
→benchmarking and finding the limitations

CVPR 2023

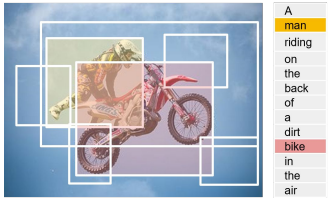




Summary

Open-vocabulary detection
→ part-wise alignment

GCPR 2022

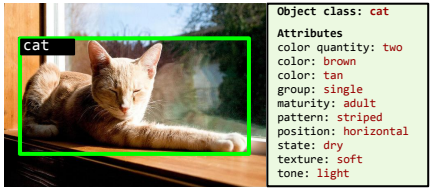


Open-vocabulary Detection

Attribute recognition
→benchmarking and finding the limitations

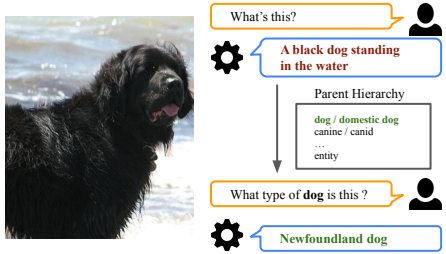
CVPR 2023

Open-vocabulary Attribute Detection



Text Generative evaluation
→strategies for evaluation

ICLR 2024



Open-ended VQA benchmark

Summary

Open-vocabulary detection
→ part-wise alignment

GCPR 2022

Attribute recognition
→benchmarking and finding the limitations

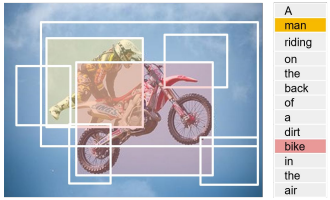
CVPR 2023

Text Generative evaluation
→strategies for evaluation

ICLR 2024

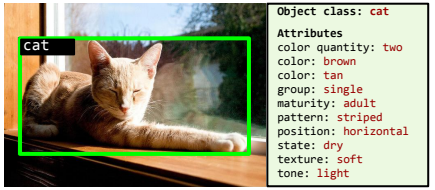
Alignment learning
→learning from human feedback

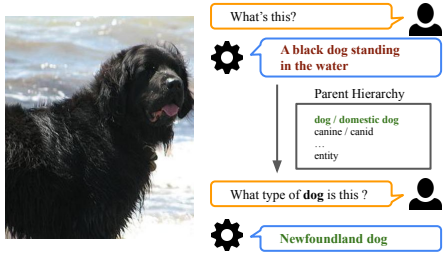
Neurips 2024 (Workshop)



Open-vocabulary Detection

Open-vocabulary Attribute Detection





Open-ended VQA benchmark

Image-Text Alignment

Thank you!


Thomas Brox


Sudhanshu Mittal


Simon Ging


Silvio Galesso

[o] Vision

COMPUTER VISION University of Freiburg

universität freiburg

Questions?

Open-vocabulary detection

perceiving what hasn't been explicitly labeled before

→ part-wise alignment

GCPR 2022

Attribute recognition

understanding subtle visual semantics

→benchmarking and finding the limitations

CVPR 2023

Text Generative evaluation

measuring open-ended outputs in human-relevant terms

→strategies for evaluation

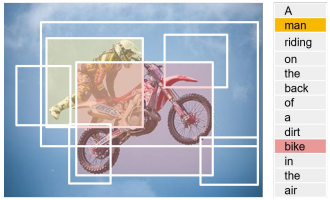
ICLR 2024

Alignment learning

anticipating what humans value or prefer

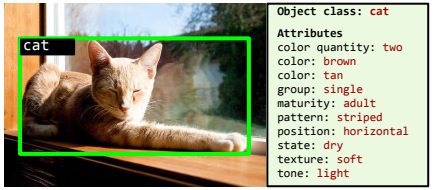
→learning from human feedback

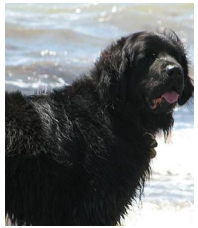
Neurips 2024 (Workshop)

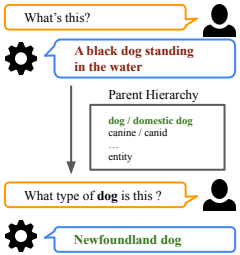


Open-vocabulary Detection

Open-vocabulary Attribute Detection







Open-ended VQA benchmark

Image-Text Alignment