VLM Benchmarks

&

Evaluation Metrics

PRESENTED BY
BHOMIK SHARMA

Objective

- Enlists Benchmarks (legacy → recent) across VLM sub-tasks
  - VQA
  - OCR/Doc
  - Math
  - Charts
  - Grounding
  - Retrieval
  - Video
  - Hallucination).

A) VQA Benchmarks

1. VQAv2:(2017, Virginia Tech, Army Research, Georgia Institute)
    - 443K train, 214K val and 453K test (question, image) pairs.
    - 10 ground truth answers per question
    - Automatic evaluation metric

$$\text{Acc}(ans) = \min\left\{\frac{\#\text{humans that said } ans}{3}, 1\right\}$$

Is the TV on?

yes

no

How many pets are present?

2

1

What time of day is it?

night

noon

Does the man have a foot in the air?

yes

no

What sign is this?

handicap

one way

Is the computer a laptop or a desktop?

desktop

laptop

Are any benches occupied?

no

yes

What color are the wall tiles?

blue

brown

What is the dog wearing?

life jacket

collar

How many skiers are there?

2

1

How many doughnuts have sprinkles?

3

2

What task is the man performing?

talking on phone

eating

What number is on the train?

7907

8551

What is sitting in the window?

bird

clock

What is this device?

train

airplane

What is the girl reaching into?

bucket

apples

What room is photographed?

kitchen

bathroom

What is the weather like here?

cloudy

sunny

What is she holding?

tennis racket

pot

How many surfers are there?

3

1

2. GQA(Graph Question Answering)(2017, Stanford)

- Why it's widely used:
  - GQA became the successor to VQAv2 for testing compositional reasoning, grounding, and consistency. Many VLMs (like LLaVA, BLIP-2, Flamingo) benchmark on it.
- Task:
  - Models answer compositional, graph-structured questions over real images (from Visual Genome).
- Evaluation:
  - Accuracy of answers.
  - Additional diagnostics: consistency, plausibility, grounding (does the model stay logically coherent across related questions).
- Strengths:
  - Focuses on multi-step reasoning (e.g., "What color is the book on the table next to the chair?").
  - Has detailed scene graph annotations.
- Limitations:
  - Still tied to synthetic question templates → less natural than human-written benchmarks.

GQA

1. What is the **woman** to the right of the **boat** holding? umbrella
2. Are there **men** to the left of the **person** that is holding the **umbrella**? no
3. What color is the **umbrella** the **woman** is holding? purple

VQA

1. Why is the person using an umbrella?
2. Is the picture edited?
3. What's the color of the umbrella?

GQA

1. Is that a **giraffe** or an **elephant**? giraffe
2. Who is feeding the **giraffe** behind the **man**? lady
3. Is there any **fence** near the **animal** behind the **man**? yes
4. On which side of the image is the **man**? right
5. Is the **giraffe** is behind the **man**? yes

VQA

1. What animal is the lady feeding?
2. Is it raining?
3. Is the man wearing sunglasses?

GQA

1. Is the **person**'s **hair** brown and long? yes
2. What **appliance** is to the left of the **man**? refrigerator
3. Is the **man** to the left or to the right of a **refrigerator**? right
4. Who is in front of the **appliance** on the left? man
5. Is there a **necktie** in the picture that is not red? yes
6. What is the **person** in front of the **refrigerator** wearing? suit
7. What is hanging on the **wall**? picture
8. Does the **vest** have different color than the **tie**? no
9. What is the color of the **shirt**? white
10. Is the color of the **vest** different than the **shirt**? yes

VQA

1. Does this man need a haircut?
2. What color is the guys tie?
3. What is different about the man's suit that shows this is for a special occasion?

3. MMBench (2024, Shanghai, NTU, Chinese University, NUS)

- Why it’s widely used:
  - MMBench is one of the newer, comprehensive, leaderboard-driven benchmarks. Hugely popular in the post-ChatGPT multimodal era (2023–2025). Nearly all recent VLM papers report MMBench scores.
- Task:
  - Multi-choice questions covering diverse skills: perception, reasoning, math, text recognition, commonsense, charts, etc.
- Evaluation:
  - Multiple-choice accuracy, often with LLM-as-judge settings for ambiguous outputs.
- Strengths:
  - Curated for broad multimodal coverage (image/text tasks beyond classical VQA).
  - Actively maintained with leaderboards.
  - Strong adoption across Chinese & global AI research labs.
- Limitations:
  - Some dependence on multiple-choice format may not reflect real open-ended usage.
  - LLM-as-judge evaluation introduces variability.

Attribute Recognition

Q: What is the shape of this object?
A. Circle
B. Triangle
C. Square
D. Rectangle
GT: A

Q: what is the color of this object?
A. Purple
B. Pink
C. Gray
D. Orange
GT: D

Celebrity Recognition

Q: Who is this person
A. David Beckham
B. Prince Harry
C. Daniel Craig
D. Tom Hardy
GT: B

Q: Who is this person
A. Benedict Cumberbatch
B. Idris Elba
C. Ed Sheeran
D. Harry Styles
GT: A

Object Localization

Q: How many apples are there in the image? And how many bananas are there?
A. 4 apples and 2 bananas
B. 3 apples and 3 banana
C. 2 apples and 4 bananas
D. 4 apples and 1 bananas
GT: A

Q: Which corner is the juice?
A. Up
B. Down
C. Left
D. Right
GT: D

OCR

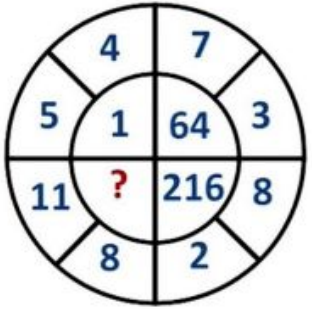
Q: What does this outdoor billboard mean?
A. Smoking is prohibited here.
B. Something is on sale.
C. No photography allowed
D. Take care of your speed.
GT: B

Q: What does this picture want to express?
A. We are expected to care for green plants.
B. We are expected to care for the earth.
C. We are expected to stay positive.
D. We are expected to work hard.
GT: D

B) Math Benchmarks

1. MATHVISTA(2024, UCLA, University of Washington, Microsoft Research)

- MathVista is a large-scale visual math benchmark (30k+ problems) that tests a model's ability to solve math using images, diagrams, and charts, not just text. It's widely used to evaluate multimodal LLMs on OCR, reasoning, and visual problem-solving.
- Downstream tasks
  - AI tutors for math & science education.
  - Diagram/geometry understanding for STEM.
  - Data visualization assistants (e.g., reasoning over plots).



Question: Find the missing value in this math puzzle.

Solution:

$(5 - 4)^3 = 1$

$(7 - 3)^3 = 64$

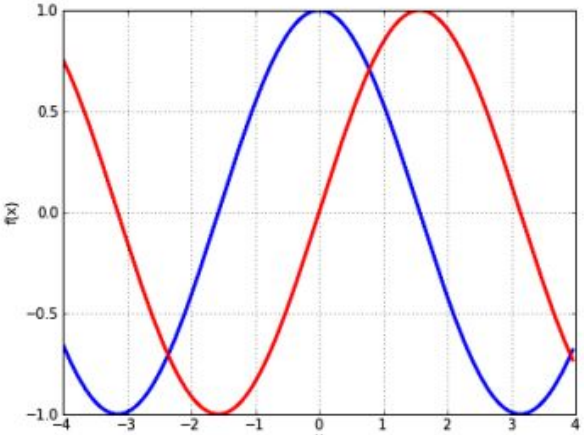
$(8 - 2)^3 = 216$

Similarly, $(11 - 8)^3 = 27$.

So the missing value is 27.

Answer: 27

Category: Math-targeted
Task: Figure question answering
Context: Puzzle test
Grade: Elementary school
Math: Logical reasoning



Question: Which function is monotonic in range [0, pi]?

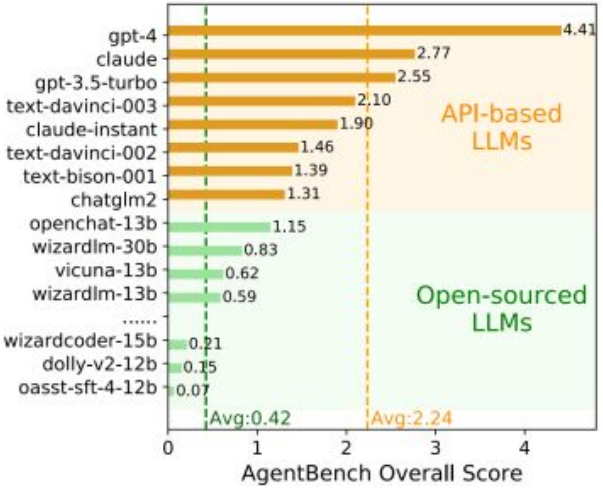
Choices:

(A) the red one (B) the blue one

(C) both (D) none of them

Answer: (B) the blue one

Category: Math-targeted
Task: Textbook question answering
Context: Function plot
Grade: College
Math: Algebraic reasoning



Question: What is the performance gap in the AgentBench Overall Score between the worst API-based LLM and the best open-sourced LLM?

Answer: 0.16

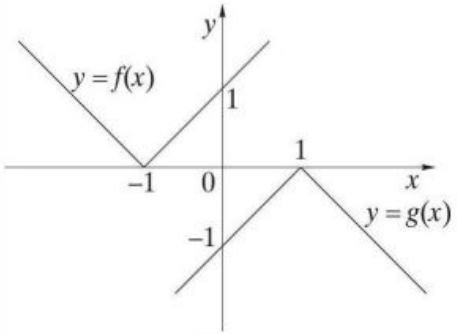
Category: Math-targeted
Task: Figure question answering
Context: Scientific figure
Grade: College
Math: Scientific reasoning

2. MathVision(2024, Chinese University of HK, Sensetime, Shanghai AI)

- MathVision evaluates accuracy and cross-format consistency to test whether models can generalize across different visual styles of the same math problem. Unlike MathVista, which measures broad visual math ability, MathVision is better for checking robustness and fairness across varied representations.
- Downstream tasks
  - Robust AI tutors that don’t fail when diagram style changes.
  - Education technology ensuring fairness across diverse inputs.
  - Cross-representation reasoning in multimodal systems.

▷ *mutual symmetry of functions*

Image:

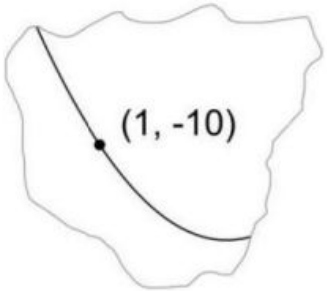


Question: The figure shows graphs of functions f and g defined on real numbers. Each graph consists of two perpendicular halflines. Which is satisfied for every real number x?

(A) $f(x) = -g(x) + 2$
(B) $f(x) = -g(x) - 2$
(C) $f(x) = -g(x + 2)$
(D) $f(x + 2) = -g(x)$
(E) $f(x + 1) = -g(x - 1)$

▷ *quadratic function discriminant*

Image:

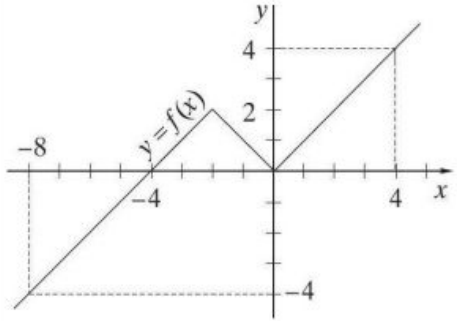


Question: In the (x,y)-plane the coordinate axes are positioned as usual. Point $A(1, -10)$ which is on the parabola $y = ax^2 + bx + c$ was marked. Afterwards the coordinate axis and the majority of the parabola were deleted. Which of the following statements could be false?

(A) $a > 0$ (B) $b < 0$ ...

▷ *find roots of iterative functions*

Image:

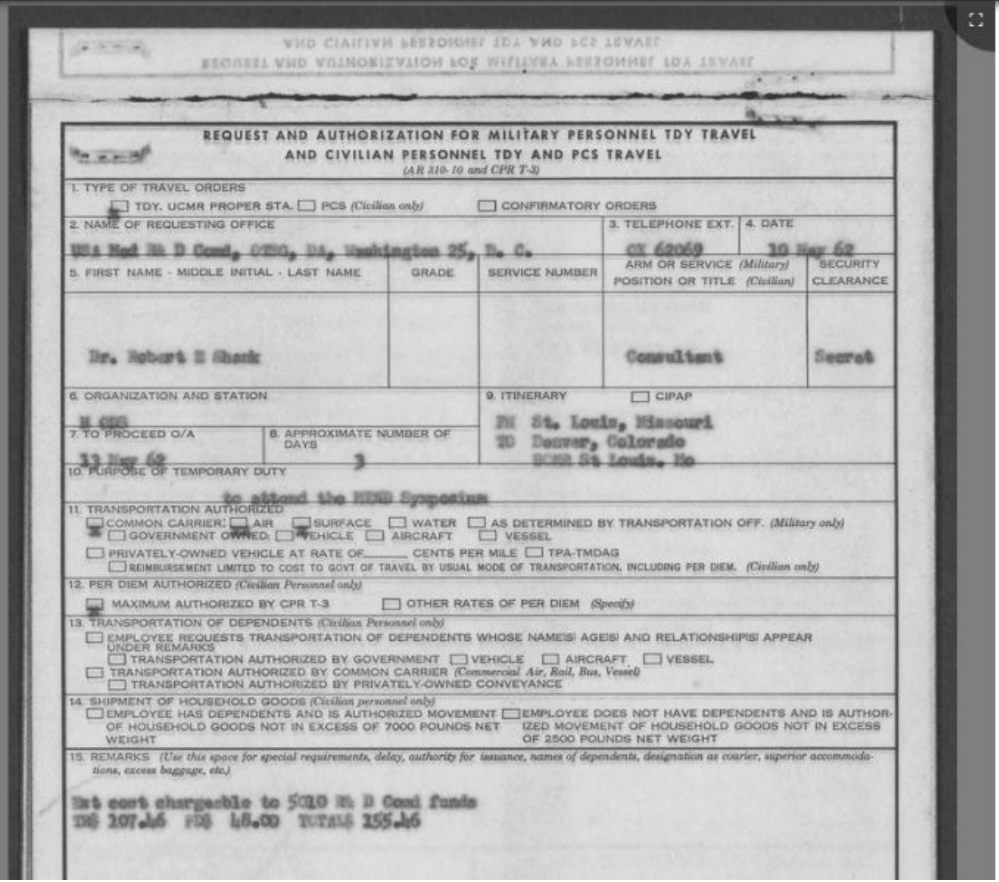


Question: The graph of the function $f(x)$, defined for all real numbers, is formed by two half-lines and one segment, as illustrated in the picture. Clearly, -8 is a solution of the equation $f(f(x)) = 0$, because $f(f(-8)) = f(-4) = 0$. Find all the solutions of the equation $f(f(f(x))) = 0$.

C) OCR Benchmarks

1. **DocVQA(2021, IIIT Hyderabad, UAB Spain)**

- DocVQA evaluates models on document understanding using accuracy, F1, and ANLS as metrics. It focuses on scanned and digital documents (invoices, forms, contracts) where models must combine OCR, layout analysis, and text reasoning. Compared to other OCR datasets like TextVQA (scene text in natural images), DocVQA is better suited for structured document AI applications such as finance, legal, and enterprise automation.
- Downstream tasks
  - Document AI systems (contracts, invoices, forms).
  - Enterprise automation (finance, HR, legal).
  - Assistive tech (screen readers, accessibility tools).



REQUEST AND AUTHORIZATION FOR MILITARY PERSONNEL TDY TRAVEL
AND CIVILIAN PERSONNEL TDY AND PCS TRAVEL
(AR 310-10 and CPR T-3)

1. TYPE OF TRAVEL ORDERS
☒ TDY. UCMR PROPER STA. ☐ PCS (Civilian only) ☐ CONFIRMATORY ORDERS

2. NAME OF REQUESTING OFFICE: USA Med Bd D Comd, OTSG, DA, Washington 25, D. C.
3. TELEPHONE EXT.: CX 62069
4. DATE: 10 May 62

5. FIRST NAME - MIDDLE INITIAL - LAST NAME	GRADE	SERVICE NUMBER	ARM OR SERVICE (Military) POSITION OR TITLE (Civilian)	SECURITY CLEARANCE
Dr. Robert E Shank			Consultant	Secret

6. ORGANIZATION AND STATION: M GHQ
7. TO PROCEED O/A: 13 May 62
8. APPROXIMATE NUMBER OF DAYS: 3
9. ITINERARY ☐ CIPAP: FM St. Louis, Missouri TO Denver, Colorado RTRN St Louis, Mo
10. PURPOSE OF TEMPORARY DUTY: to attend the NIMH Symposium
11. TRANSPORTATION AUTHORIZED
☒ COMMON CARRIER: ☒ AIR ☐ SURFACE ☐ WATER ☐ AS DETERMINED BY TRANSPORTATION OFF. (Military only)
☒ GOVERNMENT OWNED: ☐ VEHICLE ☐ AIRCRAFT ☐ VESSEL
☐ PRIVATELY-OWNED VEHICLE AT RATE OF_____ CENTS PER MILE ☐ TPA-TMDAG
☐ REIMBURSEMENT LIMITED TO COST TO GOVT OF TRAVEL BY USUAL MODE OF TRANSPORTATION, INCLUDING PER DIEM. (Civilian only)
12. PER DIEM AUTHORIZED (Civilian Personnel only)
☒ MAXIMUM AUTHORIZED BY CPR T-3 ☐ OTHER RATES OF PER DIEM (Specify)
13. TRANSPORTATION OF DEPENDENTS (Civilian Personnel only)
☐ EMPLOYEE REQUESTS TRANSPORTATION OF DEPENDENTS WHOSE NAME(S) AGE(S) AND RELATIONSHIP(S) APPEAR UNDER REMARKS
☐ TRANSPORTATION AUTHORIZED BY GOVERNMENT ☐ VEHICLE ☐ AIRCRAFT ☐ VESSEL
☐ TRANSPORTATION AUTHORIZED BY COMMON CARRIER (Commercial Air, Rail, Bus, Vessel)
☐ TRANSPORTATION AUTHORIZED BY PRIVATELY-OWNED CONVEYANCE
14. SHIPMENT OF HOUSEHOLD GOODS (Civilian personnel only)
☐ EMPLOYEE HAS DEPENDENTS AND IS AUTHORIZED MOVEMENT OF HOUSEHOLD GOODS NOT IN EXCESS OF 7000 POUNDS NET WEIGHT ☐ EMPLOYEE DOES NOT HAVE DEPENDENTS AND IS AUTHORIZED MOVEMENT OF HOUSEHOLD GOODS NOT IN EXCESS OF 2500 POUNDS NET WEIGHT
15. REMARKS (Use this space for special requirements, delay, authority for issuance, names of dependents, designation as courier, superior accommodations, excess baggage, etc.)

Est cost chargeable to 5010 M D Cmd funds
TR$ 107.46 PD$ 48.00 TOTAL 155.46

Questions

does it look like an old form?

No answers yet.

yes the form looks like an old form

Cancel Add

What is the value entered in the field "arroximate number of daaaaays"

`3`

Which type of travel order is selected ?

`TDY. UCMR PROPER STA.`

What is the telephone extension number?

`CX 62069`

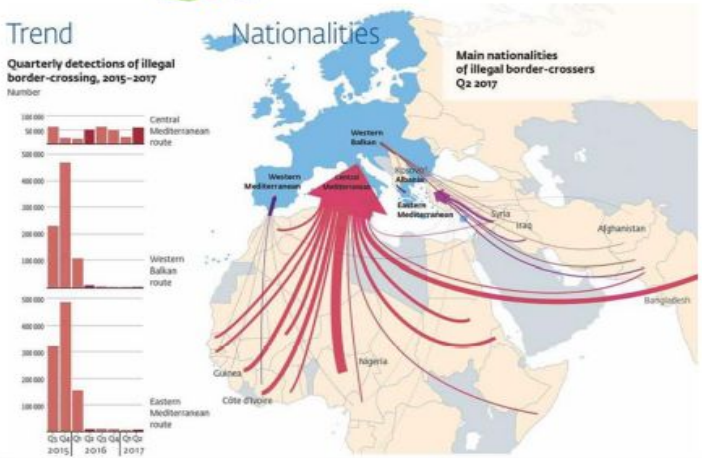
Skip Document

Finish Annotation

2. Seed-Bench-2-Plus(2024, Tencent AI, ARC Lab, CUHK)

- SEED-Bench-2-Plus evaluates VLMs on 200K+ multimodal MCQs with accuracy as the key metric. It covers OCR (scene text, documents, charts) as well as broader perception and reasoning skills, making it a comprehensive benchmark compared to specialized OCR datasets like DocVQA or TextVQA.
- Downstream tasks
  - OCR + multimodal assistants (capable of handling text in images, signs, or docs).
  - Chart/figure understanding for scientific/financial documents.
  - General-purpose VLM evaluation across perception, reasoning, and text-rich scenarios.

Maps



Based on the chart, which Mediterranean route saw the highest number of quarterly illegal border-crossing detections in Q2 of 2017?

A. Western Mediterranean route

B. Central Mediterranean route

C. Eastern Mediterranean route

D. Northern Mediterranean route

Webs

What is the Steam Awards accolade mentioned on the Red Dead Redemption 2 Steam Page?

A. Game of the Year Award

B. Critic's Choice Award

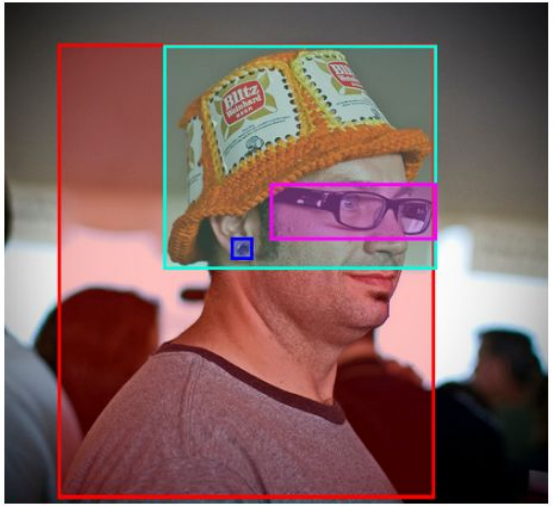
C. Labor of Love Award

D. Outstanding Story Award

D) Visual Grounding/Referring Benchmarks

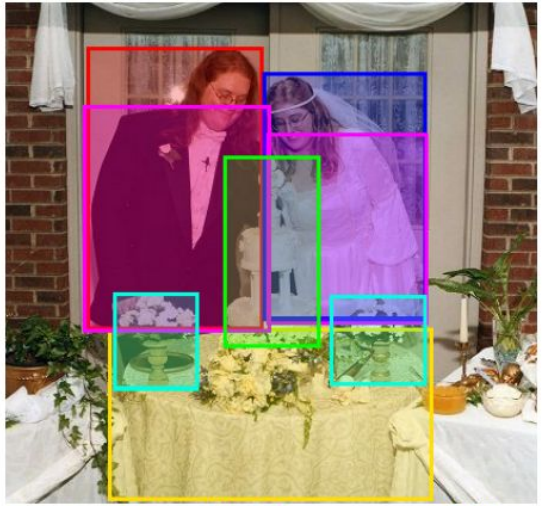
1. **Flicker-30k Entities(2016, Bryan A. Plummer)**

- Flickr30k Entities augments the Flickr30k dataset with 275K phrase-to-region links, enabling evaluation of grounded vision–language understanding. It is typically evaluated with accuracy, Recall@K, or IoU, and is widely used for phrase grounding and referring expression task
- s.
- Downstream tasks
    - Phrase grounding (linking phrases to regions).
    - Referring expression comprehension.
    - Region-to-text retrieval.
    - Pretraining/fine-tuning for VLMs on grounding tasks.



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .



A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

2. RefCOCO(2014, University of North Carolina)

- RefCOCO and its variants are benchmarks for referring expression comprehension (grounding phrases to image regions).
  - RefCOCO: natural, short interactive phrases.
  - RefCOCO+: excludes location words → tests attribute-based grounding.
  - RefCOCOg: longer, descriptive sentences → tests complex language grounding.
- Evaluation is via IoU-based accuracy, making these datasets standard for phrase-to-object grounding in multimodal AI.
- Downstream tasks
  - Referring expression comprehension (find the correct object).
  - Referring expression segmentation (masking the region).
  - Interactive multimodal AI (robots/localization from instructions).
  - Training/fine-tuning VLMs for grounding tasks.

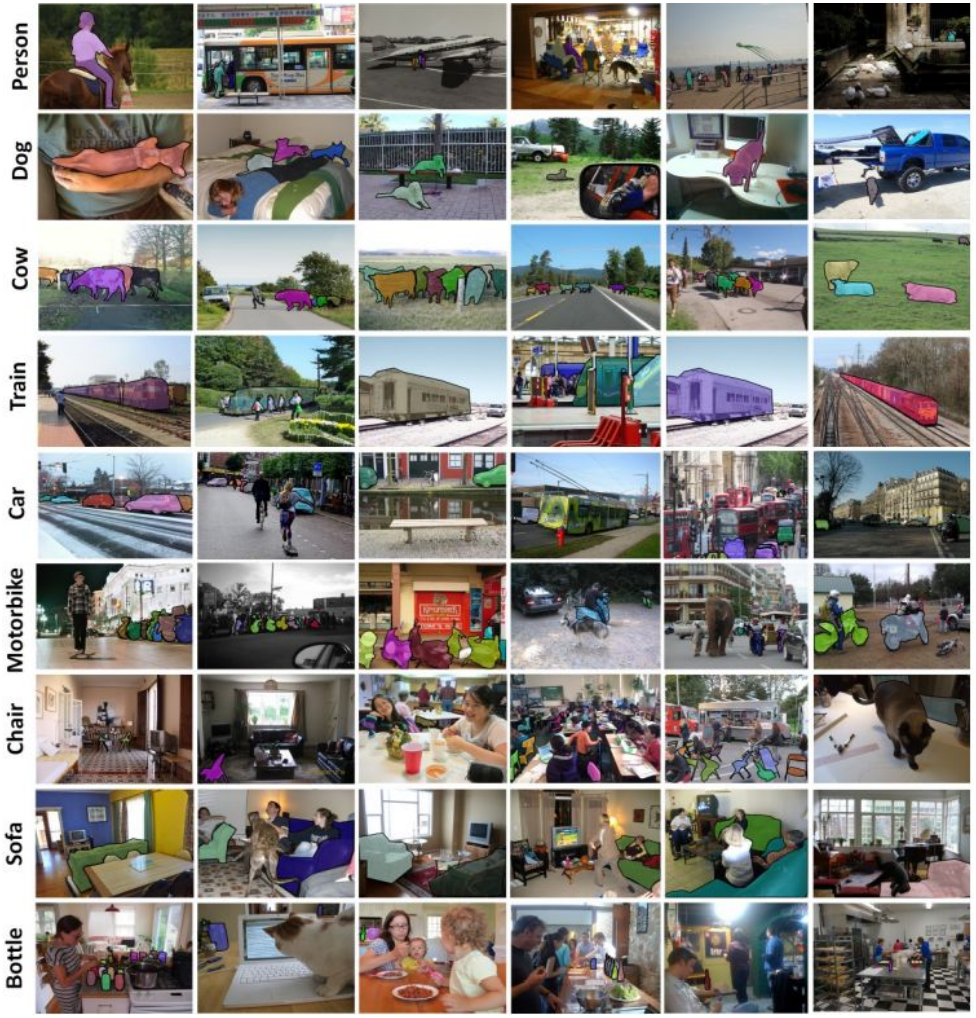
- The blue truck in the bottom right corner
- The light blue truck
- The blue truck on the right

Sample referring expressions for an object in a natural scene.

E) Retrieval and Captioning

MS-COCO(2015, Microsoft)

- MS-COCO is a foundational dataset for image captioning and image–text retrieval, with 330K images and 5 captions each. Captioning models are evaluated with BLEU, CIDEr, SPICE, while retrieval models use Recall@K. Compared to other retrieval/captioning datasets (e.g., Flickr30k, Conceptual Captions), COCO stands out for its rich object-level annotations and everyday scene diversity.


Person
Dog
Cow
Train
Car
Motorbike
Chair
Sofa
Bottle

Remaining Tasks

1. Video Understanding and QA Benchmark
    a. Perception Test
    b. Video-MME
    c. MMBench-Video
    d. EgoTempo
2. Hallucination/ Safety Diagnostics
    a. POPE
    b. HallusionBench
    c. MMHal-Bench
    d. BEAF
3. General Intelligence
    a. MMLU
    b. MMStar
    c. NaturalBench
    d. PHYSBENCH