Topics:

- VLM Datasets and Evaluations
- How to Read a Paper

CS 8803-VLM
ZSOLT KIRA

Administrative

- **Reminders**:
  - Submit reviews night before each session (11:59pm)
    - Grades released for Review 1, soon Review 2
  - Participation is part of the grade!
    - **Please post on Ed and make it lively!**
    - Ask questions and comment during discussions

- **Projects:**
  - Sign up on sheet for teams by 09/10!

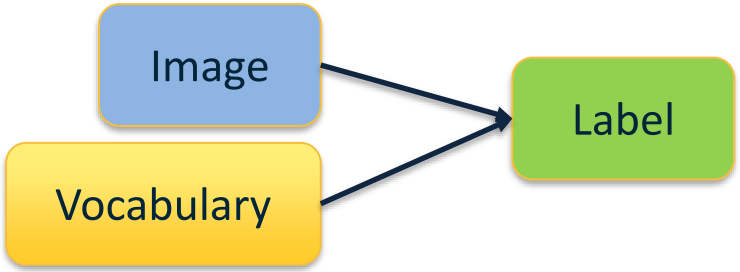
Why does multimodality matter?

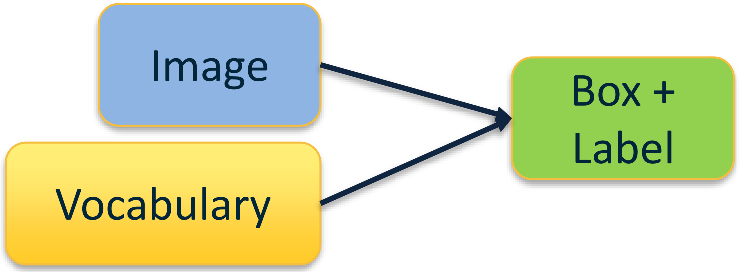
A range of very good reasons:

- Faithfulness: Human experience is multimodal
- Practical: The internet & many applications are multimodal
- Data efficiency and availability:
  - Efficiency: Multimodal data is rich and "high bandwidth" (compared to language; quoting LeCun, "an imperfect, incomplete, and low-bandwidth serialization protocol for the internal data structures we call thoughts"), so better for learning?
  - Scaling: More data is better, and we're running out of high quality text data.

- Cross-modality improvements
- Enables/required for variety of tasks and capabilities!

Open-Vocabulary Classification & Detection





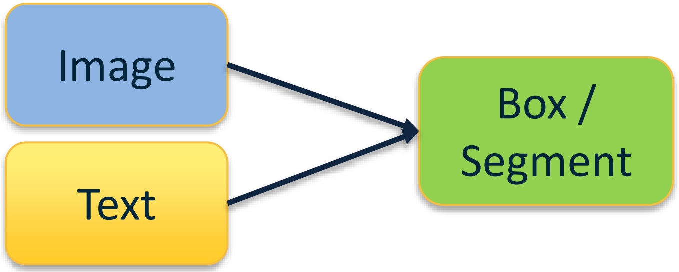
- Language is a universal way to describe what we want
  - Unlike coding, no training (of humans) required

- Improve generalization of vision-based scene understanding via language
  - Last time: Open-vocabulary classification & detection
  - Leverage fixed (but larger!) vocabulary for image tasks

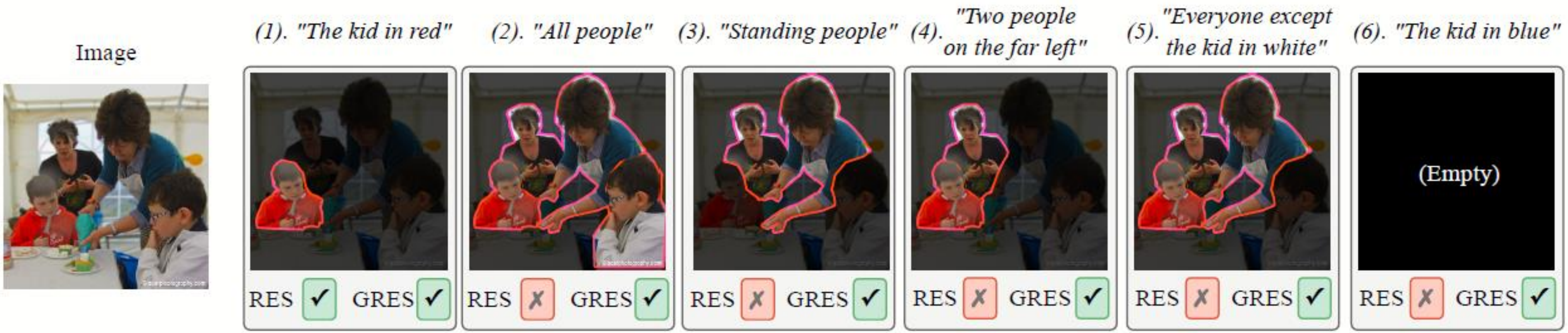
OWL-ST+FT L/14 self-trained on N-grams and fine-tuned on $LVIS_{base}$ (Table 1 row 15)

(Generalized) Referring Expression

- Ideal: Describe anything and have it be detected!
  - "Blue truck with a dog in the back"



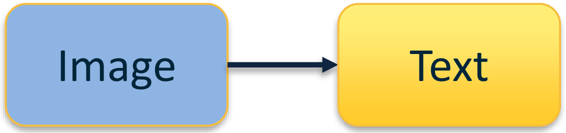
- (Generalized) Referring Expression



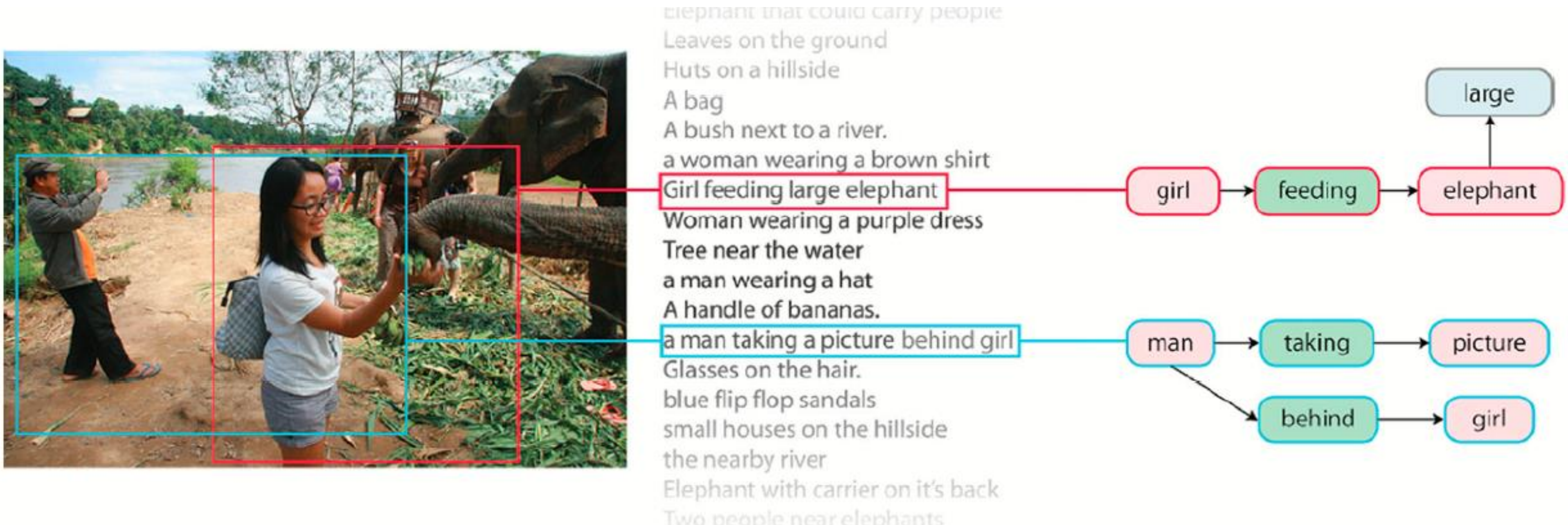
Liu et al., GRES: Generalized Referring Expression Segmentation

Image Captioning

– Image Captioning an early vision-language task



– Captions can vary in detail/how fine-grained it is



Krishna et al., Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations

Image Captioning – Datasets

Dataset	Total Images	Objects/Image	Object Classes	Captions/Image
Visual Genome [63]	108,077	36.17	80,138	5.4m R.D.
MS COCO [76]	330,000	7.57	91	5
Flickr30K Entities [97]	31,783	8.7	44,518	5
OpenImagesV6:V.R. [40]	375,000	8.4	-	1
Flickr30K [139]	31,000	-	-	5
FlickrStyle10K [33]	10,000	-	-	2
OpenImagesV6:L.N. [98]	849,000	-	-	1
SentiCap [86]	3171	-	-	6
TextCaps [110]	28,408	-	-	5
VizWiz-Captions [46]	39,181	-	-	5
nocaps [1]	15,100	-	680	11
Conceptual Captions [106]	3 mil<	-	-	1

Details (R.D. Indicates "Region Descriptions," L.N. Indicates "Localized Narratives," and V.R. Indicates "Visual Relationships").

Ghandi et al., Deep Learning Approaches on Image Captioning: A Review

Image Captioning – Metrics

- BLEU (popular metric used to quantify the quality of machine-generated outputs) -

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right)\left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

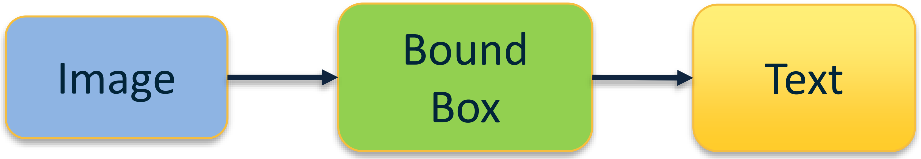
- ROUGE (evaluates text summaries; calculates recall score of generated sentences) - what % of the words or n-grams in the reference occur in the generated output?
- Perpelexity – Confidence of predicting next token

$$\text{Perplexity} = e^{-\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|w_1, w_2, \ldots, w_{i-1})}$$

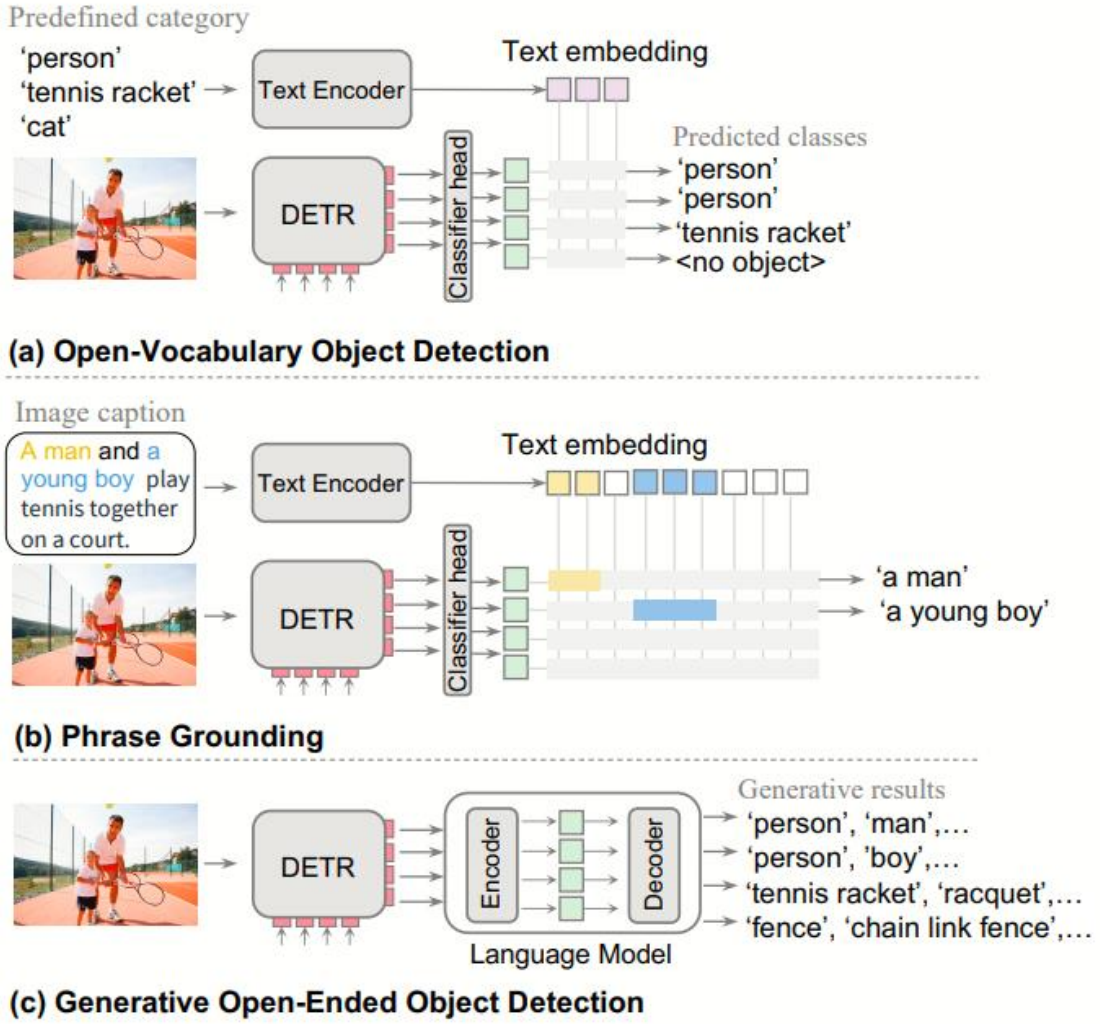
- METEOR (proposed to address the shortcomings of BLEU; introduced semantic matching; score computation is based on how well the generated sentences are aligned)
- CIDEr (recently introduced evaluation metric for image captioning task)
- MRR
- BERTScore
- **Human Evaluation!**

Slide by Emily Park and Mohit Iyyer

Open-Ended Object Detection



- Some new papers attempt to combine text generation and detection



Lin et al., Generative Region-Language Pretraining for Open-Ended Object Detection, CVPR 2024

Image Generation

Text → Image

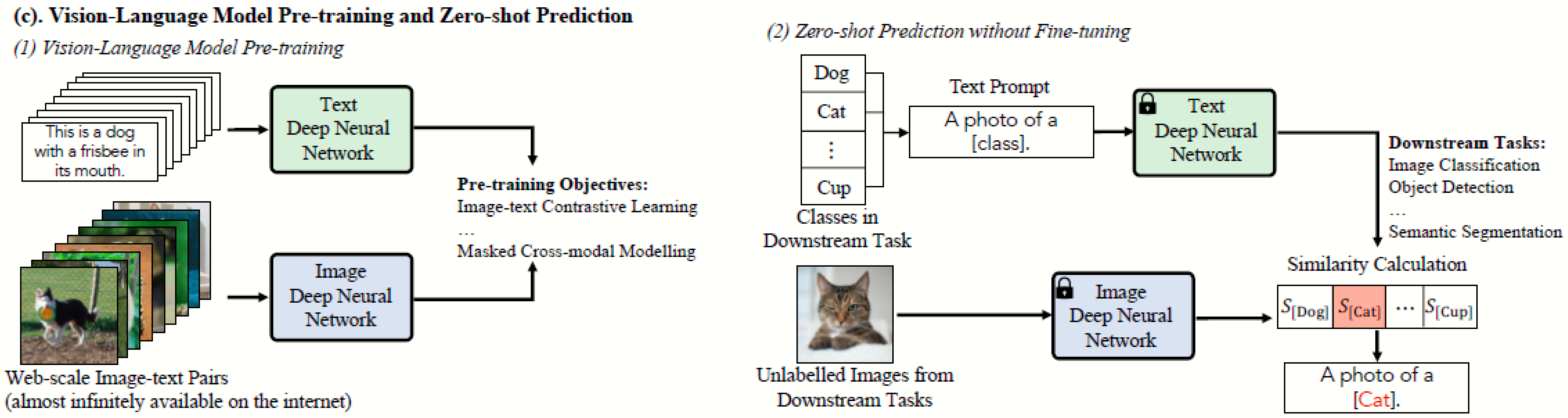
- Language to condition multi-modal generation
  - Images, videos, audio, etc.

Prompt: A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

https://openai.com/index/sora/

Datasets

- Typically various **pre-training** and **finetuning** datasets
- Evaluation done either zero-shot/finetuned on **validation** sets



Zhang et al., Vision-Language Models for Vision Tasks: A Survey

Pre-training Vision & Language Datasets

Dataset	Year	Num. of Image-Text Pairs	Language	Public
SBU Caption [73] [link]	2011	1M	English	✓
COCO Caption [74] [link]	2016	1.5M	English	✓
Yahoo Flickr Creative Commons 100 Million (YFCC100M) [75] [link]	2016	100M	English	✓
Visual Genome (VG) [76] [link]	2017	5.4 M	English	✓
Conceptual Captions (CC3M) [77] [link]	2018	3.3M	English	✓
Localized Narratives (LN) [78] [link]	2020	0.87M	English	✓
Conceptual 12M (CC12M) [79] [link]	2021	12M	English	✓
Wikipedia-based Image Text (WIT) [80] [link]	2021	37.6M	108 Languages	✓
Red Caps (RC) [81] [link]	2021	12M	English	✓
LAION400M [21] [link]	2021	400M	English	✓
LAION5B [20] [link]	2022	5B	Over 100 Languages	✓
WuKong [82] [link]	2022	100M	Chinese	✓
CLIP [10]	2021	400M	English	✗
ALIGN [17]	2021	1.8B	English	✗
FILIP [18]	2021	300M	English	✗
WebLI [83]	2022	12B	109 Languages	✗

Zhang et al., Vision-Language Models for Vision Tasks: A Survey

Fine-tuning/Evaluation **Vision** Datasets

- Often have variants, e.g. Ref/gRefCOCO

Task	Dataset	Year	Classes	Training	Testing	Evaluation Metric
Image Classification	MNIST [88] [link]	1998	10	60,000	10,000	Accuracy
	Caltech-101 [89] [link]	2004	102	3,060	6,085	Mean Per Class
	PASCAL VOC 2007 Classification [90] [link]	2007	20	5,011	4,952	11-point mAP
	Oxford 102 Folwers [91] [link]	2008	102	2,040	6,149	Mean Per Class
	CIFAR-10 [23] [link]	2009	10	50,000	10,000	Accuracy
	CIFAR-100 [23] [link]	2009	100	50,000	10,000	Accuracy
	ImageNet-1k [40] [link]	2009	1000	1,281,167	50,000	Accuracy
	SUN397 [24] [link]	2010	397	19,850	19,850	Accuracy
	SVHN [92] [link]	2011	10	73,257	26,032	Accuracy
	STL-10 [93] [link]	2011	10	1,000	8,000	Accuracy
	GTSRB [94] [link]	2011	43	26,640	12,630	Accuracy
	KITTI Distance [1] [link]	2012	4	6,770	711	Accuracy
	IIIT5k [95] [link]	2012	36	2,000	3,000	Accuracy
	Oxford-IIIT PETS [26] [link]	2012	37	3,680	3,669	Mean Per Class
	Stanford Cars [25] [link]	2013	196	8,144	8,041	Accuracy
	FGVC Aircraft [96] [link]	2013	100	6,667	3,333	Mean Per Class
	Facial Emotion Recognition 2013 [97] [link]	2013	8	32,140	3,574	Accuracy
	Rendered SST2 [98] [link]	2013	2	7,792	1,821	Accuracy
	Describable Textures (DTD) [99] [link]	2014	47	3,760	1,880	Accuracy
	Food-101 [22] [link]	2014	102	75,750	25,250	Accuracy
	Birdsnap [100] [link]	2014	500	42,283	2,149	Accuracy
	RESISC45 [101] [link]	2017	45	3,150	25,200	Accuracy
	CLEVR Counts [102] [link]	2017	8	2,000	500	Accuracy
	PatchCamelyon [103] [link]	2018	2	294,912	32,768	Accuracy
	EuroSAT [104] [link]	2019	10	10,000	5,000	Accuracy
	Hateful Memes [27] [link]	2020	2	8,500	500	ROC AUC
	Country211 [10] [link]	2021	211	43,200	21,100	Accuracy
Image-Text Retrieval	Flickr30k [105] [link]	2014	-	31,783	-	Recall
	COCO Caption [74] [link]	2015	-	82,783	5,000	Recall
Action Recognition	UCF101 [29] [link]	2012	101	9,537	1,794	Accuracy
	Kinetics700 [30] [link]	2019	700	494,801	31,669	Mean(top1, top5)
	RareAct [28] [link]	2020	122	7,607	-	mWAP, mSAP
Object Detection	COCO 2014 Detection [106] [link]	2014	80	83,000	41,000	box mAP
	COCO 2017 Detection [106] [link]	2017	80	118,000	5,000	box mAP
	LVIS [107] [link]	2019	1203	118,000	5,000	box mAP
	ODinW [108] [link]	2022	314	132413	20070	box mAP
Semantic Segmentation	PASCAL VOC 2012 Segmentation [90] [link]	2012	20	1464	1449	mIoU
	PASCAL Content [109] [link]	2014	459	4998	5105	mIoU
	Cityscapes [110] [link]	2016	19	2975	500	mIoU
	ADE20k [111] [link]	2017	150	25574	2000	mIoU

Zhang et al., Vision-Language Models for Vision Tasks: A Survey

What Does “Understanding” Mean?

- Various ways to investigate:
  - Classification, detection, generation
  - Visual Question-answering!
  - Visualization/interpretability

What Does "Understanding" Mean?

- Various ways to investigate:
  - Classification, detection, generation
  - **Visual Question-answering!**
  - Visualization/interpretability

What is VQA?

VQA is a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

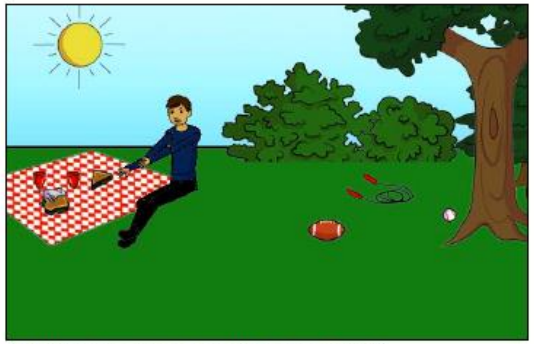
- 265,016 images (COCO and abstract scenes)
- At least 3 questions (5.4 questions on average) per image
- 10 ground truth answers per question
- 3 plausible (but likely incorrect) answers per question
- Automatic evaluation metric

VQA v1

- Various ways to investigate:
  - – Classification, detection, generation
  - – **Visual Question-answering!**
  - – Visualization/interpretability

What color are her eyes?
What is the mustache made of?

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



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Agrawal et al., VQA: Visual Question Answering, ICCV 2015

Issues

- It turns out that for many questions vision is not necessary!
  - How?

The complex compositional structure of language makes problems at the intersection of vision and language challenging. But recent works [6, 47, 49, 16, 18, 1] have pointed out that language also provides a strong prior that can result in good superficial performance, without the underlying models truly understanding the visual content.

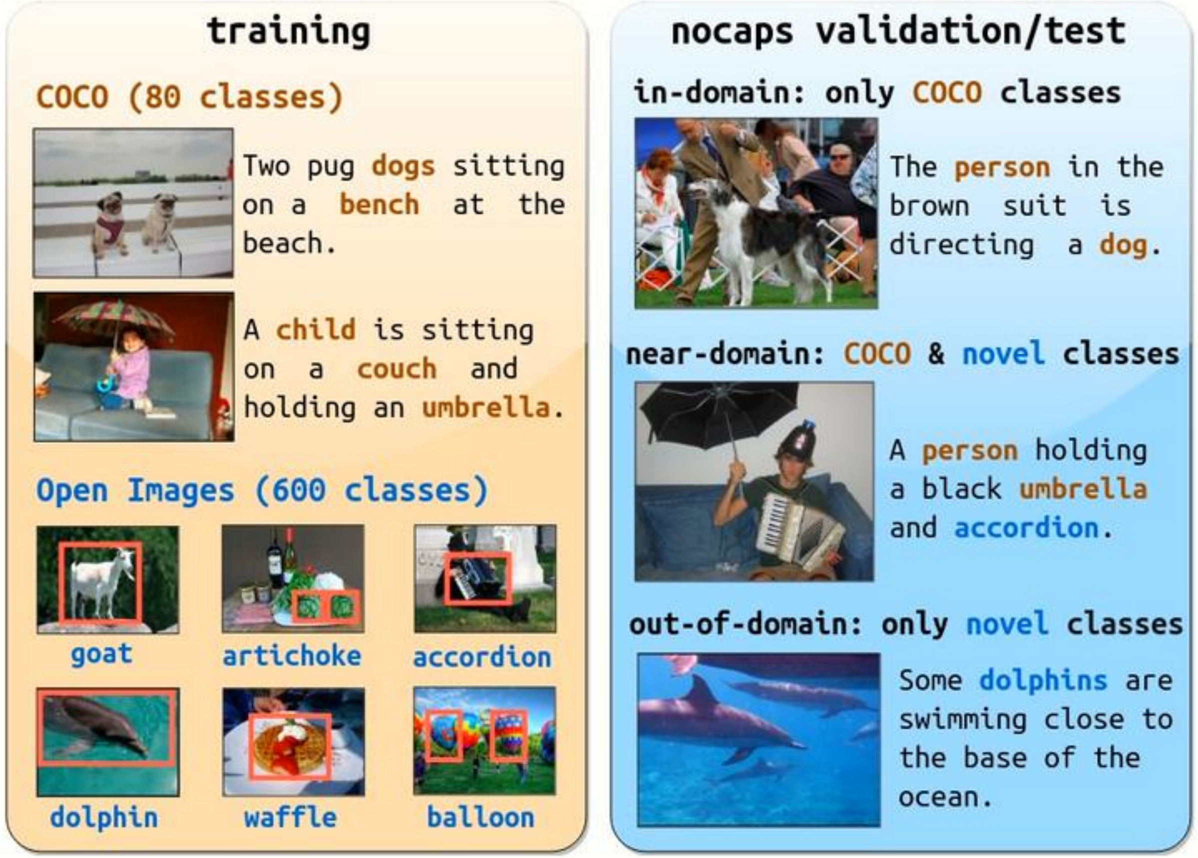
This phenomenon has been observed in image captioning [6] as well as visual question answering [47, 49, 16, 18, 1]. For instance, in the VQA [3] dataset, the most common sport answer "tennis" is the correct answer for 41% of the questions starting with "What sport is", and "2" is the correct answer for 39% of the questions starting with "How many". Moreover, Zhang *et al.* [47] points out a particular 'visual priming bias' in the VQA dataset – specifically, subjects saw an image while asking questions about it. Thus, people only ask the question "Is there a clock tower in the picture?" on images actually containing clock towers. As one particularly perverse example – for questions in the VQA dataset starting with the n-gram "Do you see a . . . ", blindly answering "yes" without reading the rest of the question or looking at the associated image results in a VQA accuracy of 87%!

Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, CVPR 2017

VQAv2


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
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

Testing Generalization



The **nocaps** benchmark for novel object captioning (at scale).

Image captioning models have achieved impressive results on datasets containing limited visual concepts and large amounts of paired image-caption training data. However, if these models are to ever function in the wild, a much larger variety of visual concepts must be learned, ideally from less supervision. To encourage the development of image captioning models that can learn visual concepts from alternative data sources, such as object detection datasets, we present the first large-scale benchmark for this task. Dubbed **nocaps**, for novel object captioning at scale, our benchmark consists of 166,100 human-generated captions describing 15,100 images from the Open Images validation and test sets. The associated training data consists of COCO image-caption pairs, plus Open Images image-level labels and object bounding boxes. Since Open Images contains many more classes than COCO, nearly 400 object classes seen in test images have no or very few associated training captions (hence, **nocaps**). We extend existing novel object captioning models to establish strong baselines for this benchmark and provide analysis to guide future work.

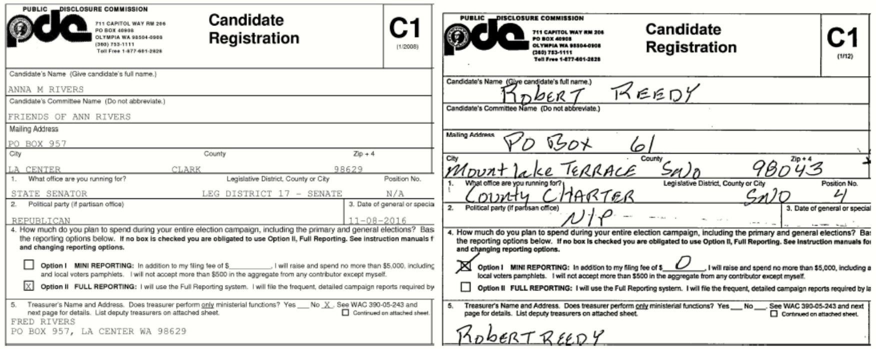
Agrawal et al., nocaps: novel object captioning at scale

Out-of-Distribution Variants

- Distribution Shifts to Images
  - IV-VQA
  - CV-VQA
- Distribution Shifts to Questions
  - VQA-Rephrasings
  - VQA-LOL
- Distribution Shifts to Answers
  - VQA-CP
- Distribution Shifts to Multi-modalities.
  - VQA-GEN
  - VQA-CE
  - VQA-VS Adversarial Distribution Shifts
- AdVQA
- AVQA

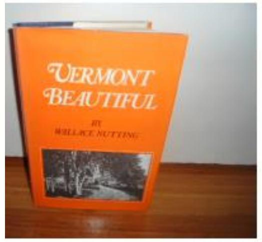
Other Forms of Image Understanding

- Lots of applications beyond natural images
  - OCR
  - Document/Infographic understanding
  - Keypoint detection
  - Video / Action Recognition
  - Cross-image alignment



Q: In which years did Anna M. Rivers run for the State senator office?
A: [2016, 2020]
E: [454, 10901]

Tito et al., DocQA



Q. What is the title of this book?
A. Vermont Beautiful
Q. Who is the author of this book?
A. Wallace Nutting
Q. What type of book is this?
A. Travel

Mishra et al., OCRQA

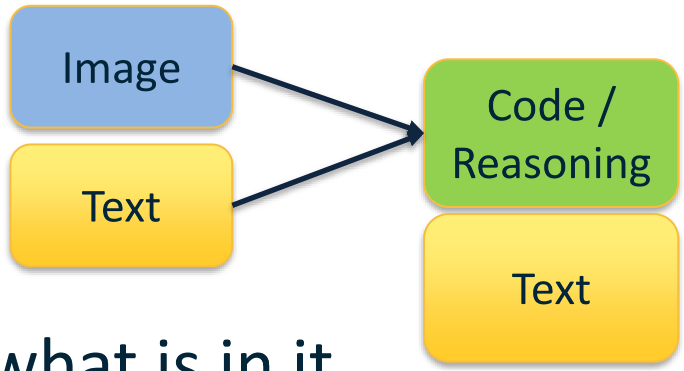
How many companies have more than 10K delivery workers?
Answer: 2 Evidence: Figure
Answer-source: Non-extractive Operation: Counting Sorting

Who has better coverage in Toronto - Canada post or Amazon?
Answer: canada post Evidence: Text
Answer-source: Question-span Image-span Operation: *none*

In which cities did Canada Post get maximum media coverage?
Answer: vancouver, montreal Evidence: Text Map
Answer-source: Multi-span Operation: *none*

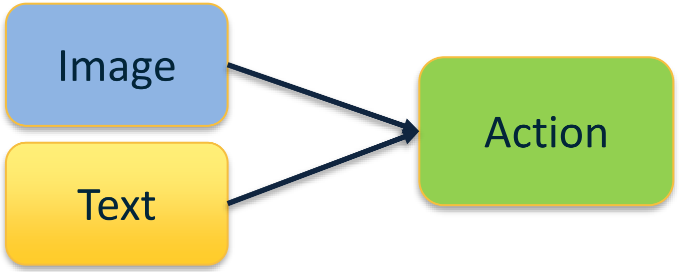
Mathew et al., InfographVQA

What Does “Reasoning” Mean?



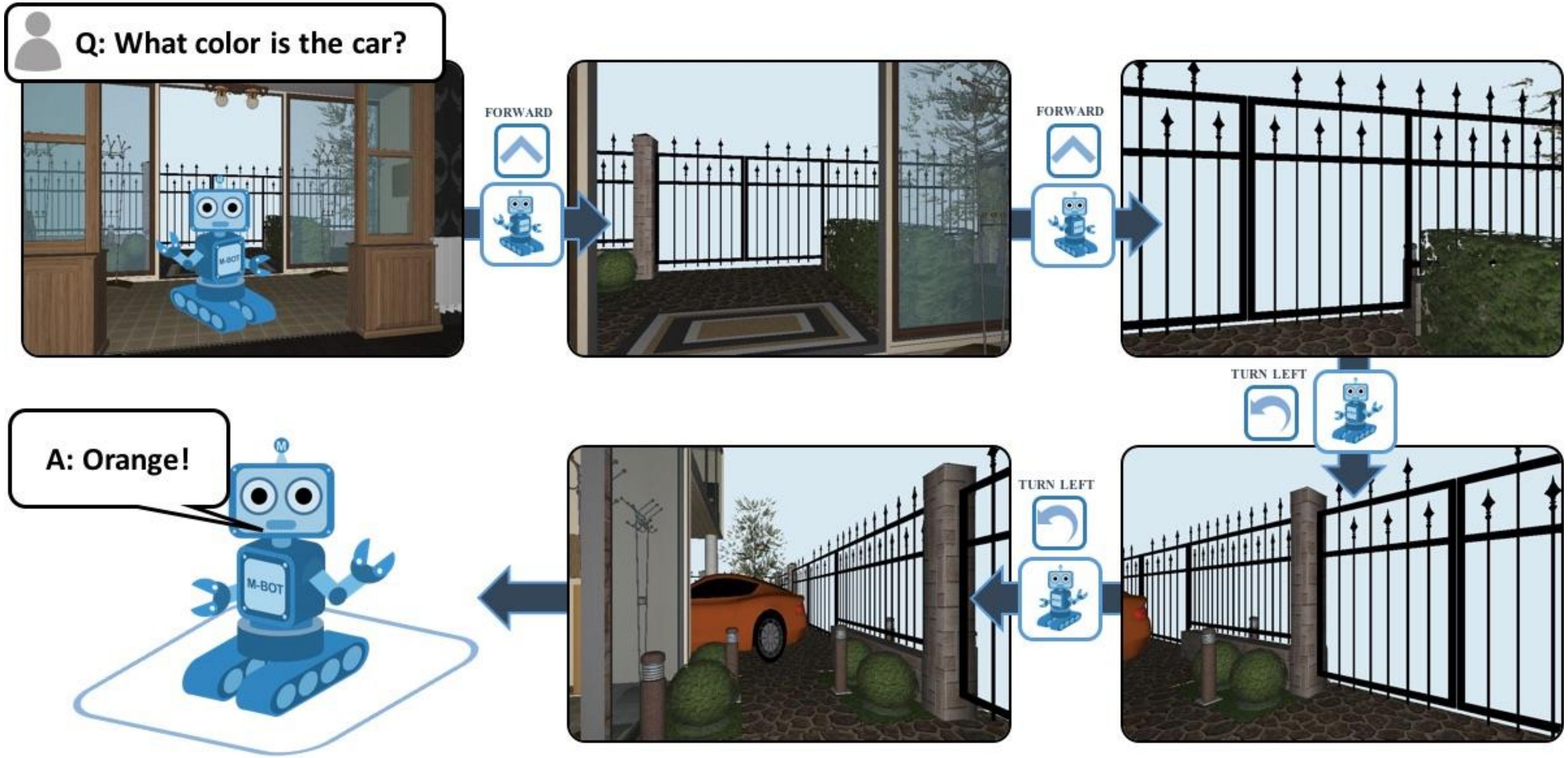
- Want to leverage image and reason/plan about what is in it
  - More complex QA
  - Image -> Math reasoning
  - Image -> Code

Decision-Making



- Want to leverage image and reason/plan about what is in it
  - More complex QA
  - Image -> Math reasoning
  - Image -> Code
  - **Image -> Action (Vision-Language-Action or VLA models)**

Embodied QA

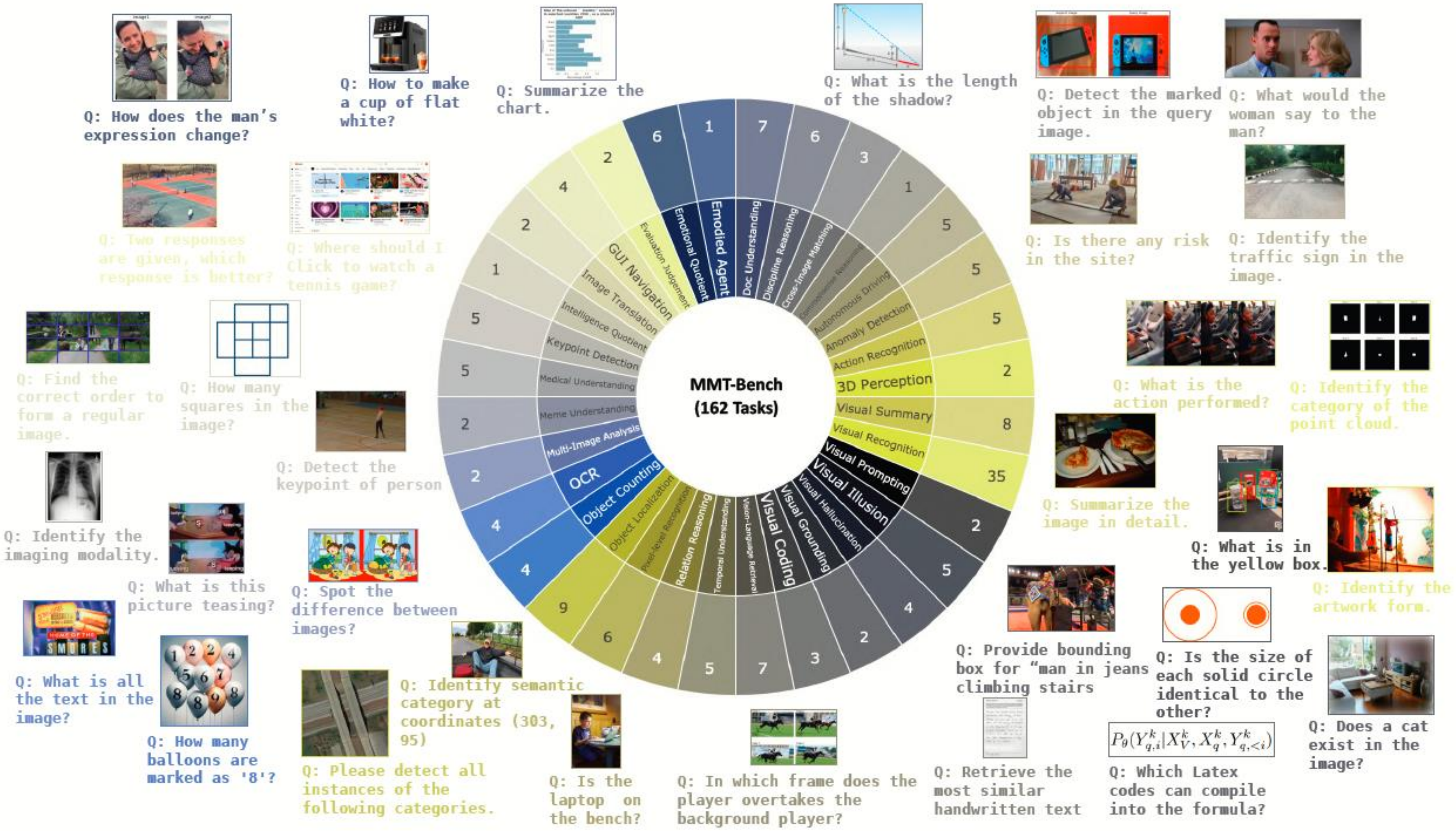


Source: https://embodiedqa.org/

Combining Everything!

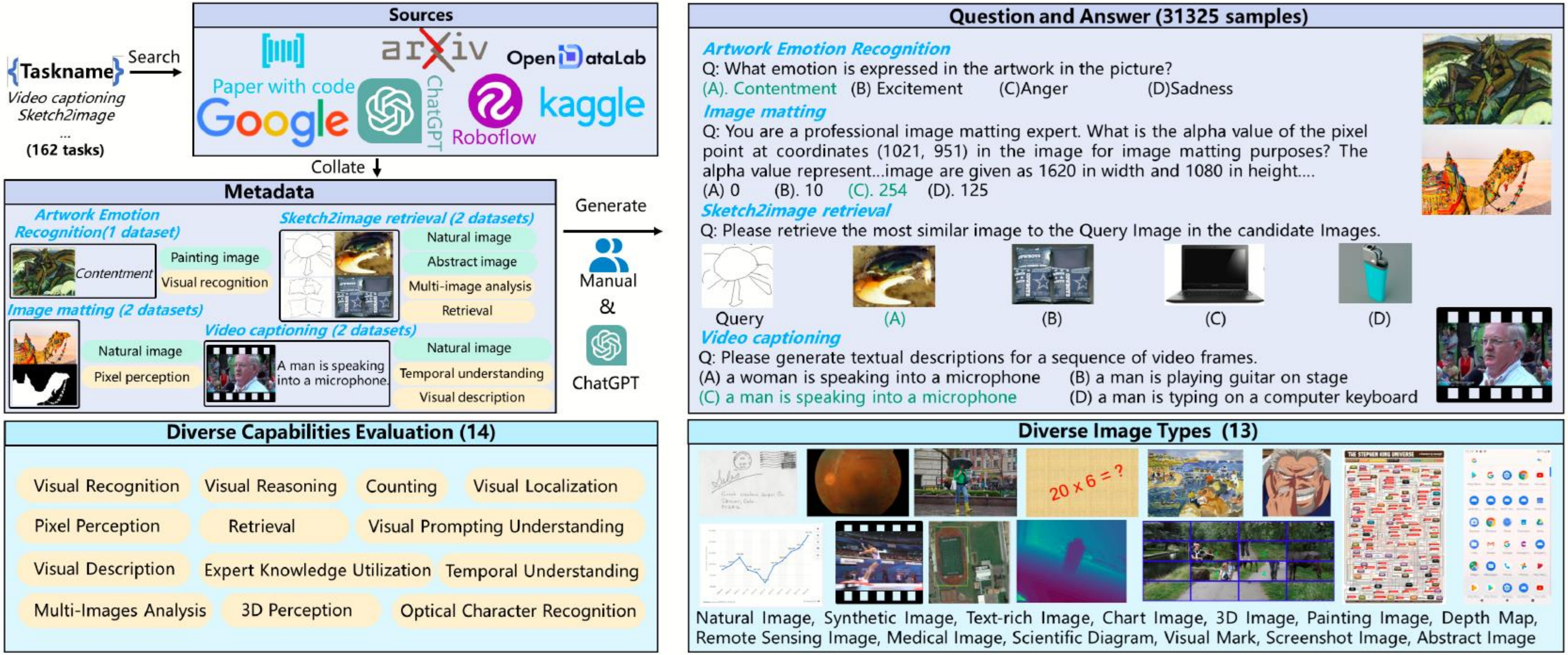
- Many large datasets and evaluations combine all of these tasks!
  - VQA
  - Document understanding
  - OCR
  - Embodied
  - …

MMT-Bench



MMT-Bench
(162 Tasks)
Emodied Agent
Emotional Quotient
Evaluation Judgement
GUI Navigation
Image Translation
Intelligence Quotient
Keypoint Detection
Medical Understanding
Meme Understanding
Multi-Image Analysis
OCR
Object Counting
Object Localization
Pixel-level Recognition
Relation Reasoning
Temporal Understanding
Vision-Language Retrieval
Visual Coding
Visual Grounding
Visual Hallucination
Visual Illusion
Visual Prompting
Visual Recognition
Visual Summary
3D Perception
Action Recognition
Anomaly Detection
Autonomous Driving
Commonsense Reasoning
Cross-Image Matching
Discipline Reasoning
Doc Understanding
7 6 3 1 5 5 5 2 8 35 2 5 4 2 3 7 5 4 6 9 4 4 2 2 5 5 1 2 4 2 6 1
Q: How does the man's expression change?
Q: How to make a cup of flat white?
Q: Summarize the chart.
Q: What is the length of the shadow?
Q: Detect the marked object in the query image.
Q: What would the woman say to the man?
Q: Two responses are given, which response is better?
Q: Where should I click to watch a tennis game?
Q: Is there any risk in the site?
Q: Identify the traffic sign in the image.
Q: Find the correct order to form a regular image.
Q: How many squares in the image?
Q: What is the action performed?
Q: Identify the category of the point cloud.
Q: Detect the keypoint of person
Q: Summarize the image in detail.
Q: Identify the imaging modality.
Q: What is this picture teasing?
Q: Spot the difference between images?
Q: What is in the yellow box.
Q: Identify the artwork form.
Q: What is all the text in the image?
Q: How many balloons are marked as '8'?
Q: Identify semantic category at coordinates (303, 95)
Q: Please detect all instances of the following categories.
Q: Is the laptop on the bench?
Q: In which frame does the player overtakes the background player?
Q: Provide bounding box for "man in jeans climbing stairs
Q: Retrieve the most similar handwritten text
Q: Is the size of each solid circle identical to the other?
Q: Which Latex codes can compile into the formula?
$P_\theta(Y^k_{q,i}|X^k_V, X^k_q, Y^k_{q,<i})$
Q: Does a cat exist in the image?

MMT-Bench



Sources
{Taskname} Search
Video captioning
Sketch2image
...
(162 tasks)
arXiv
OpenDataLab
Paper with code
Google
ChatGPT
Roboflow
kaggle
Collate
Metadata
Artwork Emotion Recognition(1 dataset)
Contentment
Painting image
Visual recognition
Sketch2image retrieval (2 datasets)
Natural image
Abstract image
Multi-image analysis
Retrieval
Image matting (2 datasets)
Natural image
Pixel perception
Video captioning (2 datasets)
A man is speaking into a microphone.
Natural image
Temporal understanding
Visual description
Generate
Manual
&
ChatGPT
Question and Answer (31325 samples)
Artwork Emotion Recognition
Q: What emotion is expressed in the artwork in the picture?
(A). Contentment (B) Excitement (C)Anger (D)Sadness
Image matting
Q: You are a professional image matting expert. What is the alpha value of the pixel point at coordinates (1021, 951) in the image for image matting purposes? The alpha value represent...image are given as 1620 in width and 1080 in height....
(A) 0 (B). 10 (C). 254 (D). 125
Sketch2image retrieval
Q: Please retrieve the most similar image to the Query Image in the candidate Images.
Query (A) (B) (C) (D)
Video captioning
Q: Please generate textual descriptions for a sequence of video frames.
(A) a woman is speaking into a microphone (B) a man is playing guitar on stage
(C) a man is speaking into a microphone (D) a man is typing on a computer keyboard
Diverse Capabilities Evaluation (14)
Visual Recognition
Visual Reasoning
Counting
Visual Localization
Pixel Perception
Retrieval
Visual Prompting Understanding
Visual Description
Expert Knowledge Utilization
Temporal Understanding
Multi-Images Analysis
3D Perception
Optical Character Recognition
Diverse Image Types (13)
20 x 6 = ?
Natural Image, Synthetic Image, Text-rich Image, Chart Image, 3D Image, Painting Image, Depth Map, Remote Sensing Image, Medical Image, Scientific Diagram, Visual Mark, Screenshot Image, Abstract Image

More Complex Tasks/Datasets

Visual Math and Text Reasoning

What is “a” in the first equation?
What is y in the last equation?

5 + a = 7	40 + y = 45
10 + y = 15	39 + a = 41
20 + a = 22	24 + y = 29
48 + a = 50	15 + a =17
10 + a = 12	34 + y = 39

The value of “a” in the first equation “5 + a = 7” is 2.

The value of “y” in the last equation “34 + y = 39” is 5.

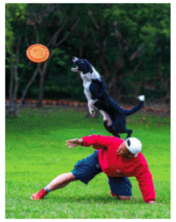
Visual-Conditioned Joke/Meme

What is funny about it?

The humor in the comic may come from the situation, the characters, or the dialogue. The person is feeling unwell and is asked if they are pregnant or have the coronavirus, which is a current and relevant topic. The humor may come from the absurdity of the situation or the unexpectedness of the question.

Spatial/Coordinate Understanding

Where is the frisbee?
What is the relative position of the frisbee and the man?



The frisbee is located at coordinates (63,119,133,177).

The frisbee is located above the man in the image. The man is falling on the ground while the frisbee is in the air. The frisbee is also closer to the top of the image than the man..

Visual Planning and Prediction

What should I do after kneading the dough, before dividing it into pieces? For how long?

After kneading the dough, you should put it in a warm place before dividing it into pieces, according to the recipe in the image.

You should put the dough in a warm place for 1 hour.

Multi-Image Reasoning

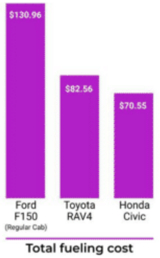
How much in total did I spend on travel?
How much in total did I pay for taxes?

You spent $1299.90 on airfare and $43.83 on an Uber ride for a total of $1343.73 on travel.

You paid $323.23 in total for taxes. This includes $321.90 for taxes and fees for the flight purchase and $1.33 for sales tax for the Windmill Family Restaurant.

Multi-Hop Document Understanding

Excluding the regular cab, what is the average total fueling cost?



The average total fueling cost excluding the Ford F150 (Regular Cab) is $76.55. This is calculated by taking the sum of the total fueling cost of Toyota RAV4 and Honda Civic and dividing it by 2. ($82.56 + $70.55) / 2 = $76.55.

Open-World Concept Understanding

What is the brand of these drink? Which one is unsweetened?

The brand of these drinks is “BETTER THAN MILK”. The unsweetened one is “BETTER THAN MILK organic OAT DRINK”.

Who are the celebrities in this image?

Keira Knightley and Natalie Portman are the celebrities in this image.

Is it safe to eat this mushroom?

Morel mushrooms are generally considered safe to eat when properly cooked.

Video Summarization/Event Localization

Can you breakdown this video tutorial to step-by-step instructions, with start and end time?

Here's a step-by-step breakdown of the tutorial with start and end time.

… Add Lettuce Add Tomato Add Bacon …

Yang et al., MM-ReAct MM-ReAct: Prompting ChatGPT for Multimodal Reasoning and Action

Discussion

- What stands out when looking across this space?

- Some interesting characteristics/caveats
  - Bias can exist across question types/answers and modalities
  - Sometimes, it is discovered that datasets can be solved without using vision
  - Emphasis on different modalities driven by datasets, architecture, loss, etc.

- Some trends:
  - Combination of datasets and evaluation across **many** of these tasks

Example: CogVLM2 Pre-Training Data

3.1 Pre-training Data

The aim of visual language pre-training is to endow models with the capability to comprehend visual input and align with language space based on large-scale image-text pairs. While there are several open-source large-scale image-text pair datasets, such as LAION [68] and DataComp [15], they generally contain significant noise and obtaining high-quality image-text pairs is challenging. Additionally, these datasets focus on coarse-grained natural language descriptions of real images, resulting in limited distribution. To address this, we employs two main techniques to obtain and process the pre-training dataset:

Hong et al., CogVLM2: Visual Language Models for Image and Video Understanding

Example: CogVLM2 Pre-Training Data

Iterative Refinement. While large-scale image-text datasets provide with massive visual language knowledge, they are often noisy or weakly related. Therefore, we use iterative refinement to enhance the data quality. To begin with, the initial model is trained on publicly available datasets, and then used to re-annotate a new batch of data. The annotations generated by the model undergo meticulous manual correction to ensure their accuracy. The corrected data is subsequently used to iteratively refine and enhance future versions of the model. This iterative process fosters continuous improvement in the quality of the training data and, consequently, the model's performance.

Synthetic Data Generation. The large-scale image-text datasets often focus on coarse-grained natural language descriptions of real images, resulting in limited distribution. For example, they commonly lack data for Chinese text recognition and GUI image understanding. To endow models with a more diverse range of fundamental visual capabilities, we create part of the datasets by synthesizing data according to specific rules or utilizing advanced tools to generate high-quality image-text pairs.

Utilizing these two techniques, the construction of pre-training data for CogVLM family is progressive and incremental. Here we presents the datasets and their usage in chronological order:

LAION-2B and COYO-700M [9] are two extensive, publicly available datasets comprising numerous images paired with corresponding captions. These datasets form the foundational base for the pre-training stages of all models in CogVLM family, offering a diverse collection of image-text pairs essential for effective model training.

LAION-40M-grounding is an in-house grounding dataset developed using LAION-400M [69] and GLIPv2 [91]. This specialized dataset is designed to enhance the model's grounding capabilities, making it particularly suitable for use in models such as CogVLM-grounding and CogAgent, which require precise and accurate grounding annotations.

The **Digital World Grounding Dataset** consists of 7 million English and 5 million Chinese entries. This dataset is created by crawling web pages with a web browser, capturing screenshots along with all visible DOM elements and their corresponding rendered boxes using Playwright [1]. This comprehensive approach allows for the creation of REC (Referring Expression Comprehension) and REG (Referring Expression Generation) question-answer pairs, significantly enhancing the model's ability to understand and generate natural language descriptions for visual elements.

The **Synthetic OCR Dataset** is another vital component of the pre-training data. This dataset includes 120 million English and 150 million Chinese entries, focusing on four specific OCR scenarios: (1) fully generated OCR images with source text printed on the images using Python; (2) real-world images with extracted text obtained using PaddleOCR [32]; (3) academic papers with extracted LaTeX code by Nougat [8]; and (4) HTML or LaTeX code of tables and formulae rendered to images using various tools. This extensive dataset is utilized in models such as CogAgent, CogVLM2, and GLM-4V to enhance their OCR capabilities.

Finally, **CLAY-1B** is an in-house recaption dataset built upon LAION-2B and COYO-700M. This dataset is developed with the aid of a fine-tuned CogVLM model specifically designed to generate long, detailed captions for images. The Chinese captions in this dataset are translated by a fine-tuned ChatGLM. CLAY-1B is used in models like CogVLM2 and GLM-4V to improve their captioning abilities.

Hong et al., CogVLM2: Visual Language Models for Image and Video Understanding

Post-Training

4.2 Post-training Settings

Image Supervised Fine-tuning. In CogVLM2 and GLM-4V, we employed a two-stage SFT training approach. In the first stage, we utilized all VQA training datasets and the 300K alignment corpora to enhance the model's foundational capabilities, addressing the limitations of pre-training on image captioning tasks. In the second stage, we selected a subset of VQA datasets and the 50K preference alignment data to optimize the model's output style, closely aligning with human preferences.

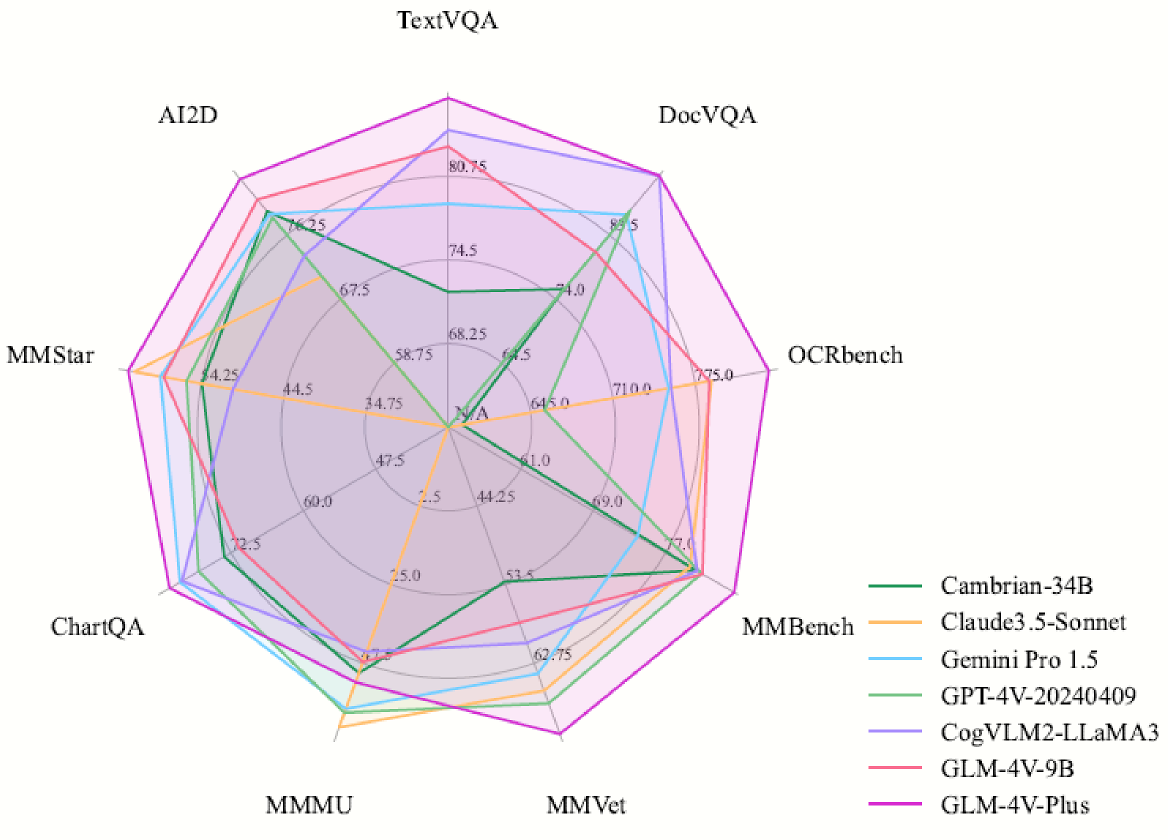
In the first stage, the model underwent 3000 iterations with a learning rate of 1e-5 and a global batch size of 2340. Subsequently, in the second stage, we reduced the global batch size to 1150 for 750 steps. We performed the image SFT process by fine-tuning all parameters. To enhance and ensure the stability of the training, we activated the visual encoder's parameters and adjusted its learning rate to be one-tenth of that used for the remaining training parameters.

Table 2: VQA datasets used in image understanding models. The "Type" column signifies the format of the answers provided. "0" corresponds to concise responses, such as multiple-choice, Y/N, etc. "1" denotes comprehensive answers that incorporate a chain of thought processes.

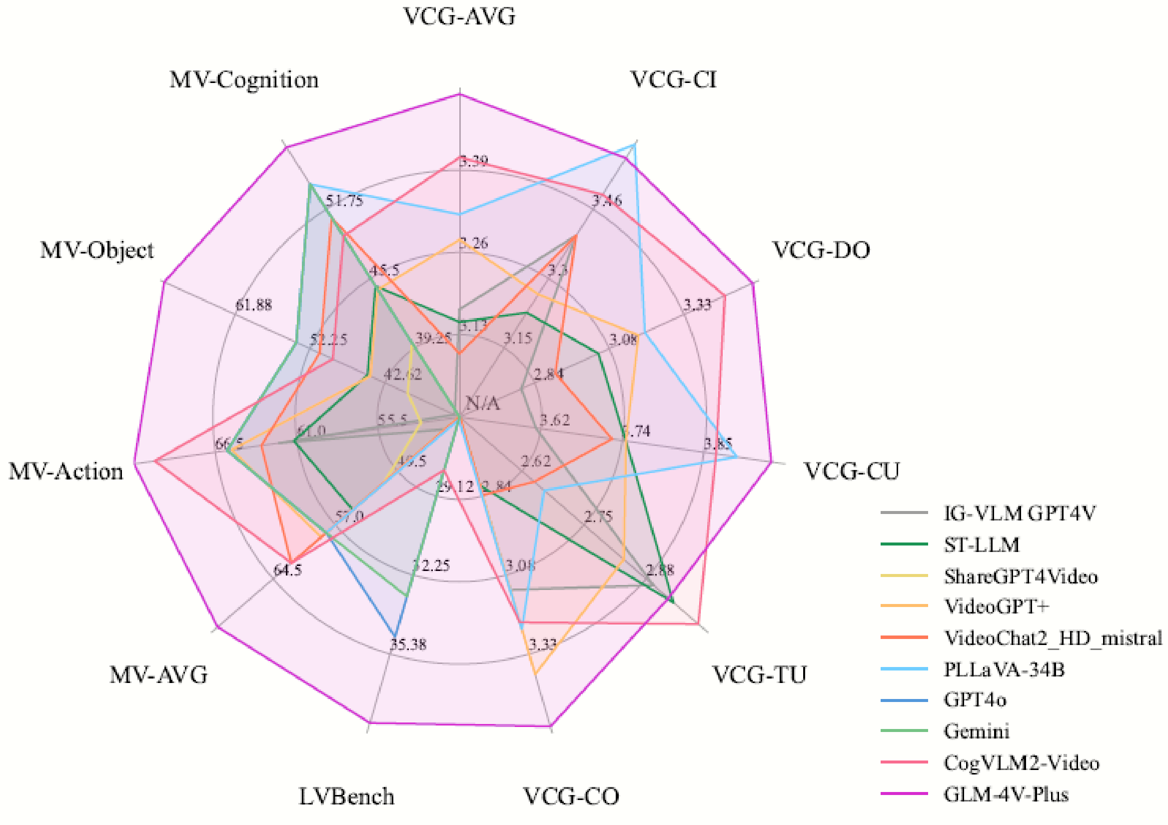
Categories	Datasets	Type	CogVLM	CogVLM2	GLM4V-9B
General QA	OKVQA [55]	0	✓	✓	✓
	STVQA [7]	0		✓	✓
	VGQA [30]	0		✓	✓
	VQAV2 [4]	0	✓	✓	✓
	A-OKVQA [70]	0			✓
	TQA [28]	0			✓
OCR	IAM [56]	1			✓
	DocVQA [62]	0		✓	✓
	OCRVQA [64]	0	✓	✓	✓
	TextVQA [73]	0	✓	✓	✓
	Rendered_text [2]	0			✓
Math & Science	GeoMetry3K [47]	0		✓	✓
	Geo170K [16]	1		✓	✓
	GeoQA [10]	0		✓	✓
	Geomverse [25]	1			✓
	Raven [89]	1			✓
	InterGPS [47]	0			✓
	Ai2D [26]	0			✓
	ScienceQA [48]	1	✓	✓	✓
Chart Analysis	ChartQA [57]	0		✓	✓
	FigureVQA [24]	0		✓	✓
	InfoVQA [60]	0		✓	✓
	DVQA [22]	0		✓	✓
	ArxivQA [37]	1		✓	✓
	TabMWP [49]	1			✓
	VQARAD [31]	0			✓
Other	VSR [40]	0			✓
	TDIUC [23]	0		✓	✓
	TallyQA [1]	0		✓	✓
	IconQA [50]	0			✓
	VisText [75]	0			✓
	Diagram_image_to_text [3]	1			✓

Hong et al., CogVLM2: Visual Language Models for Image and Video Understanding

CogVLM2 Results



(a) Evaluation results on image tasks



(b) Evaluation results on video tasks.

Hong et al., CogVLM2: Visual Language Models for Image and Video Understanding

VLM Leaderboards!

<4B | 4B-10B | 10B-20B | 20B-40B | >40B | Unknown | API | OpenSource | Proprietary

Rank	Method	Param (B)	Language Model	Vision Model	Avg Score	Avg Rank	MMBench_V11	MMStar	MMMU_VAL	MathVista	OC
1	GPT-4o (0806, detail-high)				71.5	4.12	80.5	64.7	69.9	62.7	80
2	InternVL2-Llama3-76B	76	Llama-3-70B-Instruct	InternViT-6B	71	3.62	85.5	67.1	58.3	65.6	84
3	GPT-4o (0513, detail-high)				69.9	6	82.2	63.9	69.2	61.3	73
4	InternVL2-40B	40	Nous-Hermes-2-Yi-34B	InternViT-6B	69.7	4.75	85	64.7	55.2	64	83
5	Step-1.5V		Step-1.5	stepencoder	68.4	7.62	82.5	63.3	59.3	67.8	72
6	Claude3.5-Sonnet				67.9	9.75	78.5	62.2	65.9	61.6	78
7	InternVL2-26B	26	InternLM2-20B	InternViT-6B	66.4	9.88	81.2	61	50.7	59.4	82
8	GPT-4o (0513, detail-low)				65.7	13.88	82.8	61.6	62.8	56.5	66
9	CongRong				65.5	12.25	80.7	60.6	48.3	61	82
10	MiniCPM-V-2.6	8	Qwen2-7B	SigLIP-400M	65.2	13.62	78	57.5	49.8	60.6	85
11	Gemini-1.5-Pro				64.4	16.75	73.9	59.1	60.6	57.7	75
12	GPT-4o-mini (0718, detail-high)				64.1	18.75	76	54.8	60	52.4	78

https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Vision Arena

Spaces WildVision / vision-arena like 457 Running

App Files Community 15

Rank	V-L Model	WV-Arena Elo	95% CI	Battles	MMMU
1	gpt-4o	1217	+14/-21	1497	OpenAI
2	claude-3-5-sonnet-20240620	1169	+22/-23	540	Anthropic
3	gpt-4o-mini-2024-07-18	1134	+99/-94	40	OpenAI
4	gpt-4-turbo	1127	+66/-59	83	OpenAI
5	gemini-1.5-pro-latest	1122	+29/-41	283	Google
6	gpt-4-vision-preview	1103	+16/-12	2950	OpenAI
7	gemini-1.5-flash-latest	1085	+28/-21	764	Google
8	Reka-Flash	1072	+25/-19	750	Reka AI
9	claude-3-opus	1070	+20/-18	1388	Anthropic
10	qwen-vl-max	1054	+44/-43	131	Alibaba
11	yi-vl-plus	1053	+26/-22	792	01 AI
12	phi-3-vision-128k-instruct	1048	+156/-125	12	Microsoft
13	gemini-pro-vision	1037	+17/-11	2630	Google
14	Reka-Core	1030	+32/-33	259	Reka AI
15	llava-v1.6-34b	1029	+15/-14	2095	UW Madison

https://huggingface.co/spaces/WildVision/vision-arena

Summary

- Large number of tasks and datasets, both for pre-training and evaluation!

- Moving towards more "generalist" models
  - This gets more difficult to evaluate!

- Specialist models (documents, figures, etc.) can still do better for now
  - Leads to a number of questions such as finetuning of generalist models to specialize, without losing generalization!

Reading Research Papers

Slides originally by Judy Hoffman, modified by Zsolt Kira

Where to start?

- How did you read OWLv2?
- What background did you already have?
  - Object Detection
  - CLIP, etc.
  - Open-vocabulary detectors
- When you want to read a new paper
  - What do you read at first?
  - What questions do you ask yourself?
  - What information do you look for?

Abstract Example

- Open-vocabulary object detection has benefited greatly from pretrained vision-language models, but is still limited by the amount of available detection training data. While detection training data can be expanded by using Web image-text pairs as weak supervision, this has not been done at scales comparable to image-level pretraining. Here, we scale up detection data with self-training, which uses an existing detector to generate pseudo-box annotations on image-text pairs. Major challenges in scaling self-training are the choice of label space, pseudo-annotation filtering, and training efficiency. We present the OWLv2 model and OWL-ST self-training recipe, which address these challenges. OWLv2 surpasses the performance of previous state-of-the-art open-vocabulary detectors already at comparable training scales (~10M examples). However, with OWL-ST, we can scale to over 1B examples, yielding further large improvement: With an L/14 architecture, OWL-ST improves AP on LVIS rare classes, for which the model has seen no human box annotations, from 31.2% to 44.6% (43% relative improvement). OWL-ST unlocks Web-scale training for open-world localization, similar to what has been seen for image classification and language modelling.

Georgia Tech

Reading Exercise

- What problem does this paper focus on?
  - Is this new or already explored?
  - Is this important?
  - What key applications this is relevant for?
  - What assumptions does this paper make about

Abstract Example

- Open-vocabulary object detection has benefited greatly from pretrained vision-language models, but is still limited by the amount of available detection training data. While detection training data can be expanded by using Web image-text pairs as weak supervision, this has not been done at scales comparable to image-level pretraining. Here, we scale up detection data with self-training, which uses an existing detector to generate pseudo-box annotations on image-text pairs. Major challenges in scaling self-training are the choice of label space, pseudo-annotation filtering, and training efficiency. We present the OWLv2 model and OWL-ST self-training recipe, which address these challenges. OWLv2 surpasses the performance of previous state-of-the-art open-vocabulary detectors already at comparable training scales (~10M examples). However, with OWL-ST, we can scale to over 1B examples, yielding further large improvement: With an L/14 architecture, OWL-ST improves AP on LVIS rare classes, for which the model has seen no human box annotations, from 31.2% to 44.6% (43% relative improvement). OWL-ST unlocks Web-scale training for open-world localization, similar to what has been seen for image classification and language modelling.

Reading Exercise

- What problem does this paper focus on?
  - Is this new or already explored?
  - Is this important?
  - What key applications this is relevant for?
  - What assumptions does this paper make? Are these similar to what has been done before? Extra restrictive? Less restrictive?

Reading Exercise

- What problem does this paper focus on?
- What is the key “golden nugget” – intuition, idea, etc. that leads to approach

Abstract Example

- Open-vocabulary object detection has benefited greatly from pretrained vision-language models, but is still limited by the amount of available detection training data. While detection training data can be expanded by using Web image-text pairs as weak supervision, this has not been done at scales comparable to image-level pretraining. Here, we scale up detection data with self-training, which uses an existing detector to generate pseudo-box annotations on image-text pairs. Major challenges in scaling self-training are the choice of label space, pseudo-annotation filtering, and training efficiency. We present the OWLv2 model and OWL-ST self-training recipe, which address these challenges. OWLv2 surpasses the performance of previous state-of-the-art open-vocabulary detectors already at comparable training scales (~10M examples). However, with OWL-ST, we can scale to over 1B examples, yielding further large improvement: With an L/14 architecture, OWL-ST improves AP on LVIS rare classes, for which the model has seen no human box annotations, from 31.2% to 44.6% (43% relative improvement). OWL-ST unlocks Web-scale training for open-world localization, similar to what has been seen for image classification and language modelling.

Georgia Tech

Reading Exercise

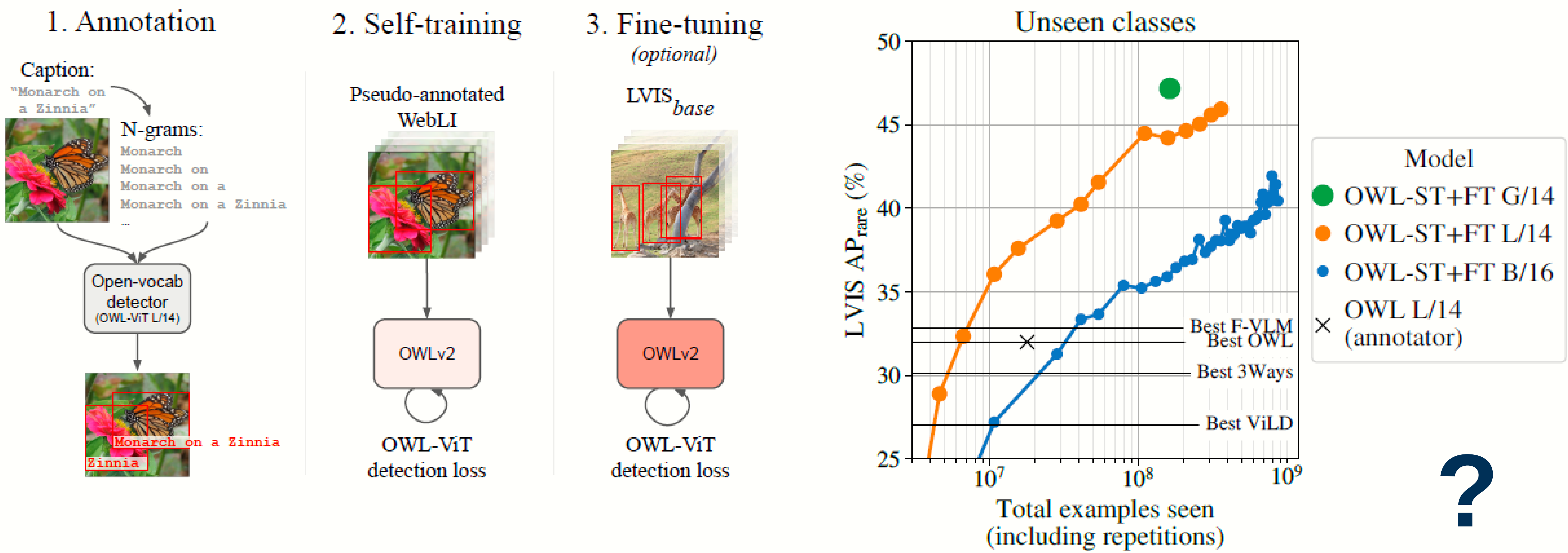
- What problem does this paper focus on?
- What is the key “golden nugget” – intuition, idea, etc. that leads to approach
- What approach does this paper take?

Abstract Example

- Open-vocabulary object detection has benefited greatly from pretrained vision-language models, but is still limited by the amount of available detection training data. While detection training data can be expanded by using Web image-text pairs as weak supervision, this has not been done at scales comparable to image-level pretraining. Here, we scale up detection data with self-training, which uses an existing detector to generate pseudo-box annotations on image-text pairs. Major challenges in scaling self-training are the choice of label space, pseudo-annotation filtering, and training efficiency. We present the OWLv2 model and OWL-ST self-training recipe, which address these challenges. OWLv2 surpasses the performance of previous state-of-the-art open-vocabulary detectors already at comparable training scales (~10M examples). However, with OWL-ST, we can scale to over 1B examples, yielding further large improvement: With an L/14 architecture, OWL-ST improves AP on LVIS rare classes, for which the model has seen no human box annotations, from 31.2% to 44.6% (43% relative improvement). OWL-ST unlocks Web-scale training for open-world localization, similar to what has been seen for image classification and language modelling.

?

Figure 1



1. Annotation
Caption:
"Monarch on a Zinnia"
N-grams:
Monarch
Monarch on
Monarch on a
Monarch on a Zinnia
...
Open-vocab detector (OWL-ViT L/14)
Monarch on a Zinnia
Zinnia
2. Self-training
Pseudo-annotated WebLI
OWLv2
OWL-ViT detection loss
3. Fine-tuning
(optional)
LVIS base
OWLv2
OWL-ViT detection loss
Unseen classes
LVIS AP rare (%)
Total examples seen (including repetitions)
Best F-VLM
Best OWL
Best 3Ways
Best ViLD
Model
OWL-ST+FT G/14
OWL-ST+FT L/14
OWL-ST+FT B/16
OWL L/14 (annotator)

?

Reading Exercise

- What problem does this paper focus on?
- What is the key “golden nugget” – intuition, idea, etc. that leads to approach
- What approach does this paper take?
  - Abstract -> Intro and/or Figure 1 -> Method Section -> Code
  - Not uncommon to have math section --(leap--> Algorithm. Look at algorithm first.

Reading Exercise

- What problem does this paper focus on?
- What approach does this paper take?
- What is the key “golden nugget” – intuition, idea, etc. that leads to approach
- What prior approaches exist to solve this problem?
  - Will need to explore related work to answer this
- How does this work validate their approach?

Reading Exercise

- What problem does this paper focus on?
- What approach does this paper take?
- What prior approaches exist to solve this problem?
  - Will need to explore related work to answer this
- How do they validate their approach?
  - What data do they use?
  - What baselines do they compare against?

Georgia Tech

Abstract Example

- Open-vocabulary object detection has benefited greatly from pretrained vision-language models, but is still limited by the amount of available detection training data. While detection training data can be expanded by using Web image-text pairs as weak supervision, this has not been done at scales comparable to image-level pretraining. Here, we scale up detection data with self-training, which uses an existing detector to generate pseudo-box annotations on image-text pairs. Major challenges in scaling self-training are the choice of label space, pseudo-annotation filtering, and training efficiency. We present the OWLv2 model and OWL-ST self-training recipe, which address these challenges. OWLv2 surpasses the performance of previous state-of-the-art open-vocabulary detectors already at comparable training scales (~10M examples). However, with OWL-ST, we can scale to over 1B examples, yielding further large improvement: With an L/14 architecture, OWL-ST improves AP on LVIS rare classes, for which the model has seen no human box annotations, from 31.2% to 44.6% (43% relative improvement). OWL-ST unlocks Web-scale training for open-world localization, similar to what has been seen for image classification and language modelling.

Paper reading advice

- **First pass** – Key Concepts
  - Try to answer the key questions about the paper
  - Read abstract / intro / teaser figure / key result table(s)
- **Second pass** – More Insight / Understanding
  - Read approach section in more detail
  - Study equations / algorithm boxes / figures
  - Look at ablation studies
- **Third pass** – Think critically
  - Did they validate all claims? Are claims significant? How does this paper do things differently than what came before?

Georgia Tech