Carnegie Mellon University

Universidade de Macau

NEURAL INFORMATION PROCESSING SYSTEMS

Revisiting Few Shot Object Detection with Vision-Language Models

Anish Madan

Neehar Peri

Shu Kong

Deva Ramanan

Remarkable Zero Shot Open-Vocab Detection Performance

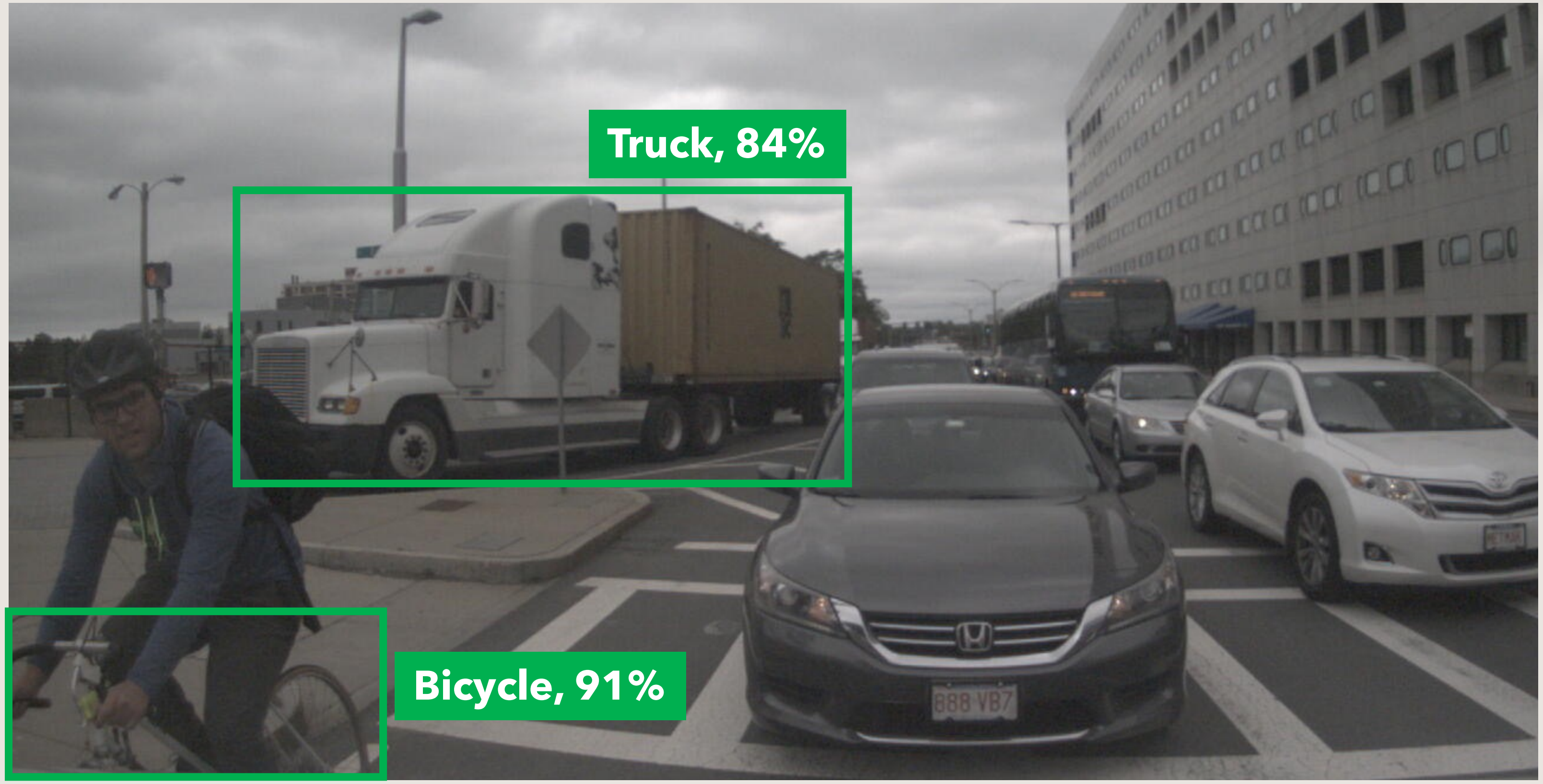
1. laptop
2. keyboard
3. newspaper
4. mug

…

N. bottle

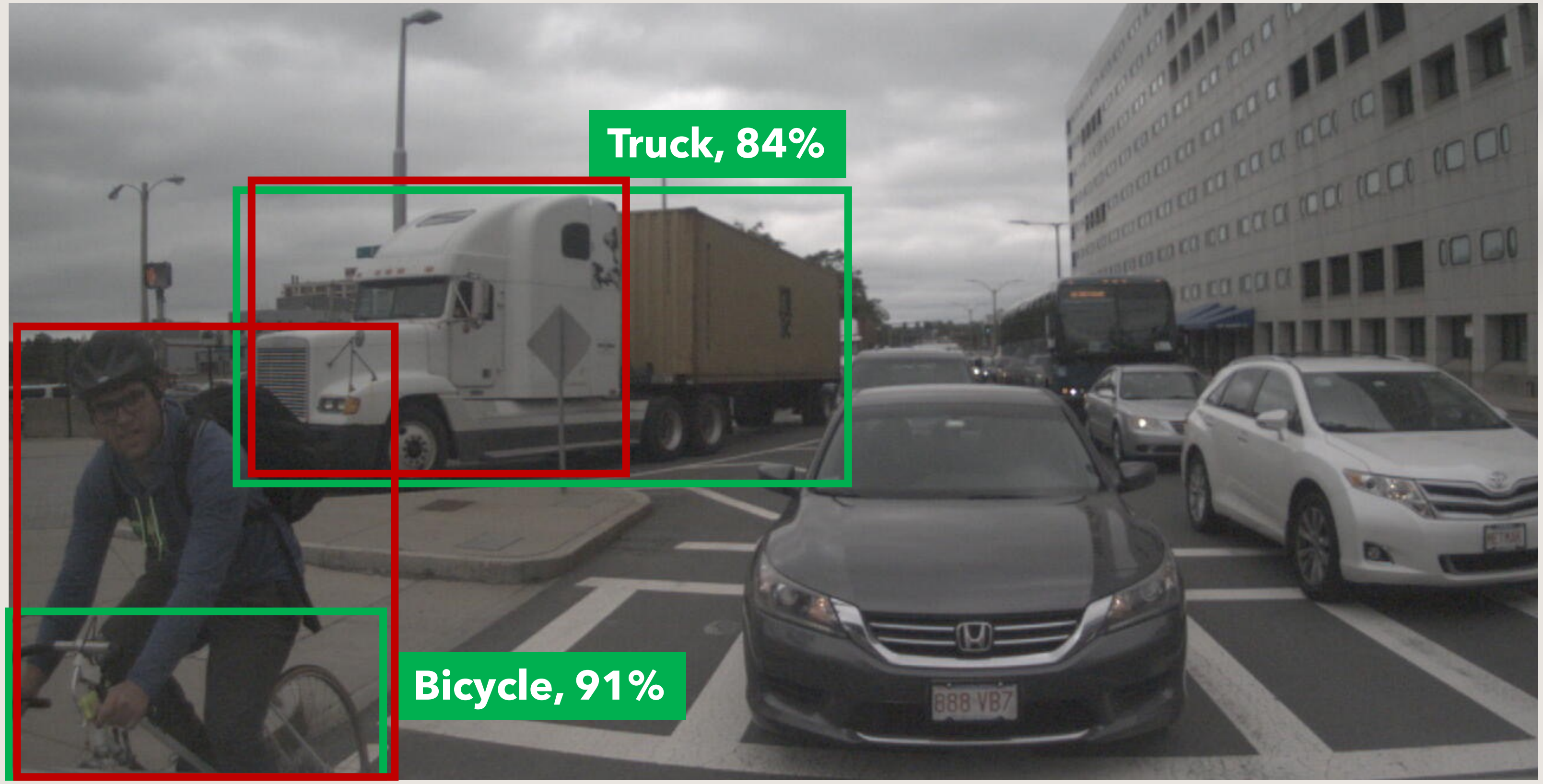
Zero-Shot Foundational Models are all we need?

Running an Open-Vocabulary Detector (GLIP) on AV Data



Zero-Shot Foundational Models are all we need? Not Quite!

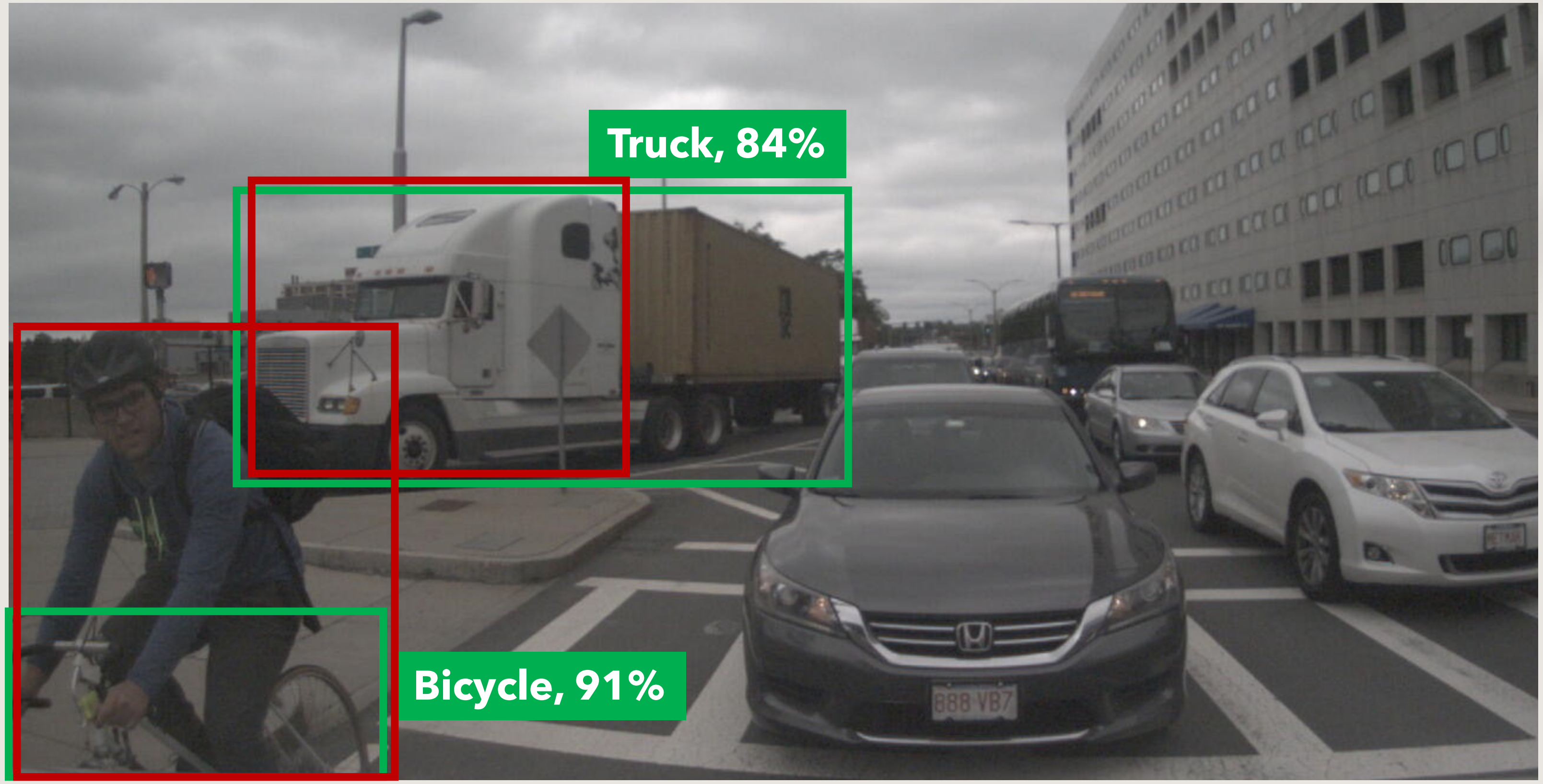
Running an Open-Vocabulary Detector (GLIP) on AV Data



Off-the-shelf VLM predictions don't match the ground-truth annotations!

Zero-Shot Foundational Models are all we need? Not Quite!

Running an Open-Vocabulary Detector (GLIP) on AV Data



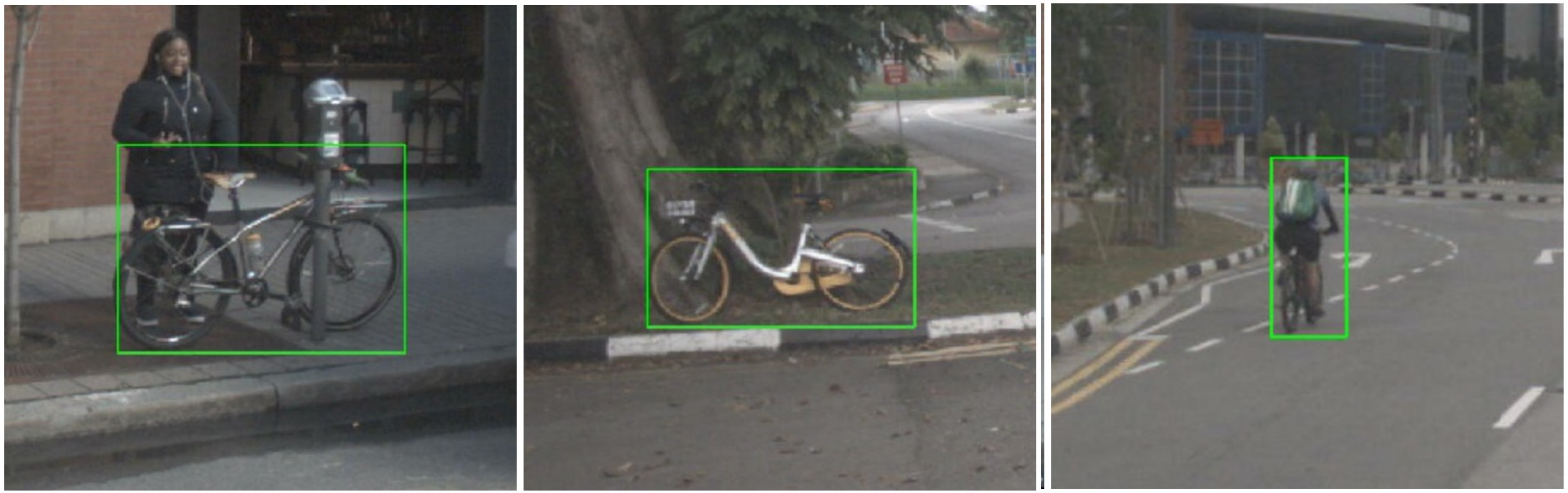
Off-the-shelf VLM predictions don't match the ground-truth annotations!

Why does this concept gap exist?

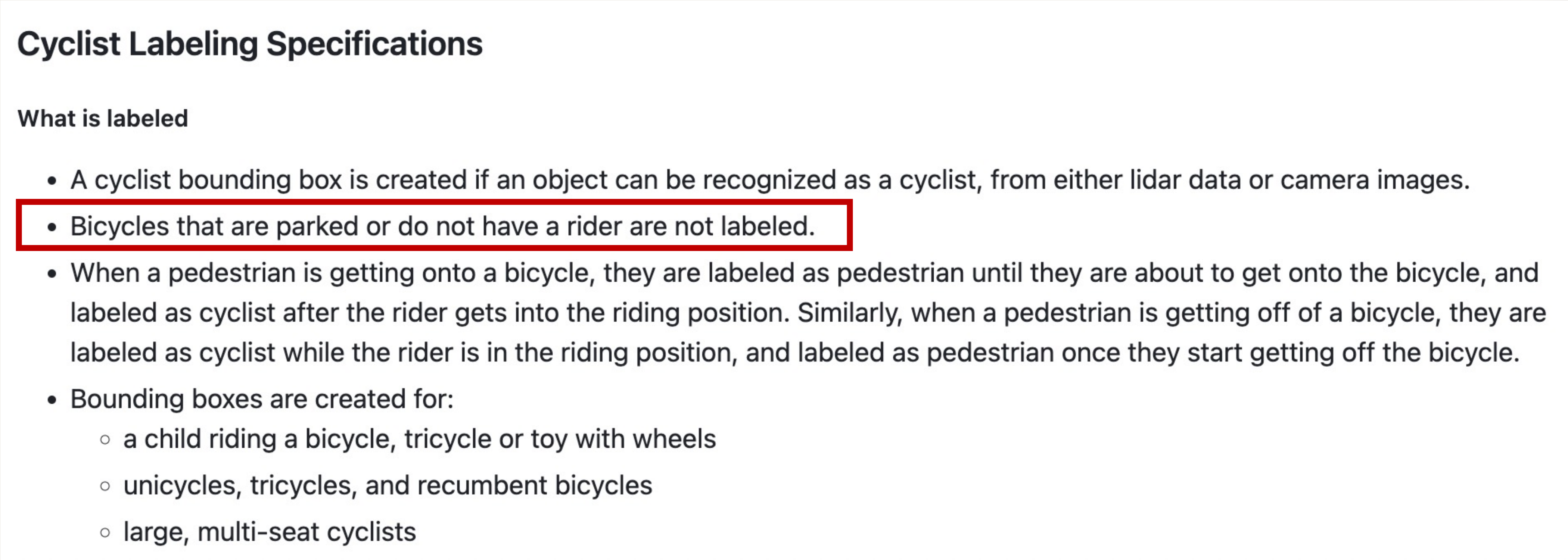
NuImages Labelling Instructions …

Bicycle

- Human or electric powered 2-wheeled vehicle designed to travel at lower speeds either on road surface, sidewalks or bicycle paths.
    - If there is a rider, include the rider in the box
    - If there is a passenger, include the passenger in the box
    - If there is a pedestrian standing next to the bicycle, do NOT include in the annotation



… differs from Waymo's Labelling Instructions



Cyclist Labeling Specifications

What is labeled

- A cyclist bounding box is created if an object can be recognized as a cyclist, from either lidar data or camera images.
- Bicycles that are parked or do not have a rider are not labeled.
- When a pedestrian is getting onto a bicycle, they are labeled as pedestrian until they are about to get onto the bicycle, and labeled as cyclist after the rider gets into the riding position. Similarly, when a pedestrian is getting off of a bicycle, they are labeled as cyclist while the rider is in the riding position, and labeled as pedestrian once they start getting off the bicycle.
- Bounding boxes are created for:
  - a child riding a bicycle, tricycle or toy with wheels
  - unicycles, tricycles, and recumbent bicycles
  - large, multi-seat cyclists

Labelling Instructions are key multi-modal cues

Debris

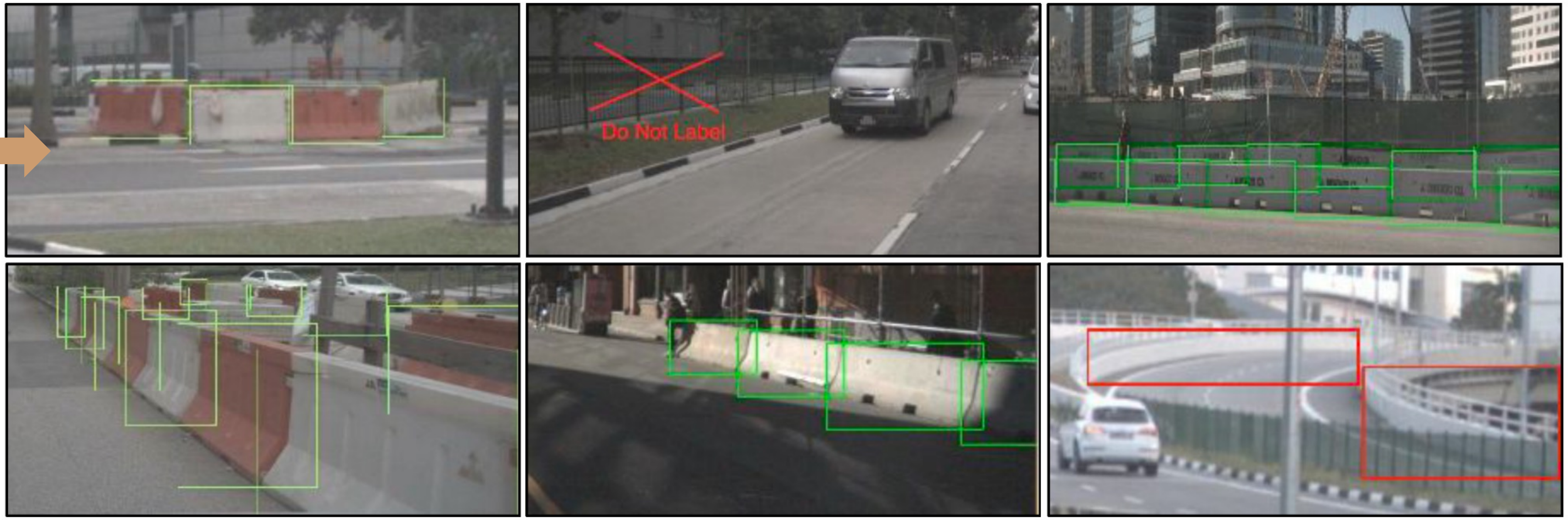
- Debris or movable object that is left **on the driveable surface** that is too large to be driven over safely, e.g tree branch, full trash bag etc.

Pushable Pullable Object

- Objects that a pedestrian may push or pull. For example dolleys, wheel barrows, garbage-bins with wheels, or shopping carts. Typically not designed to carry humans.

Human Annotators Need Multi-Modal Concept Alignment too!

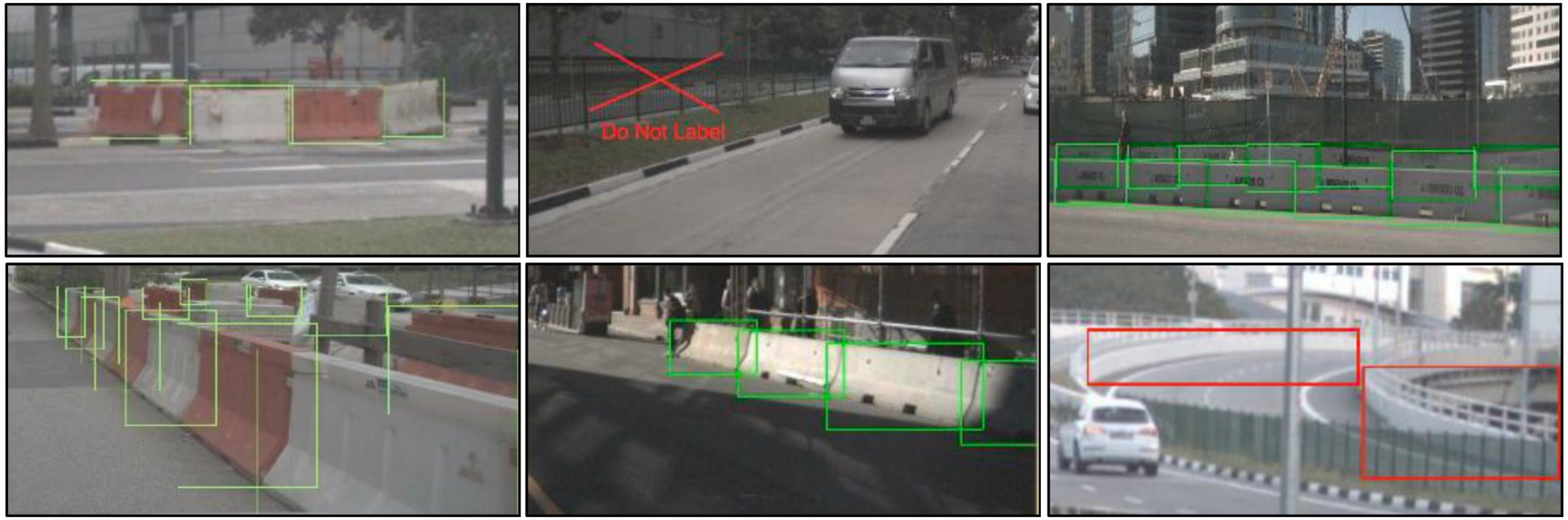
Visual Examples



Barrier

- ➔ *Any metal, concrete or water barrier temporarily placed in the scene in order to re-direct vehicle or pedestrian traffic. In particular, includes barriers used at construction zones.*
- ➔ *If there are multiple barriers either connected or just placed next to each other, they should be annotated separately.*
- ➔ *If barriers are installed permanently, then do **NOT** include them.*

Human Annotators Need Multi-Modal Concept Alignment too!

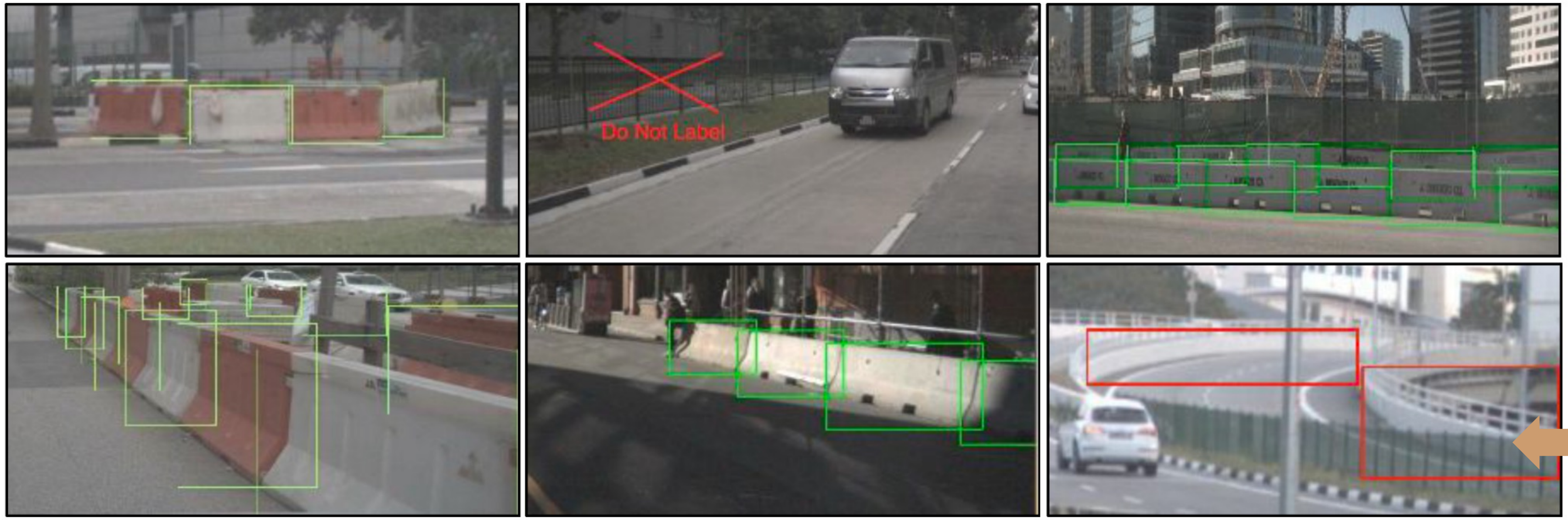


Rich Text

Barrier

- ➔ *Any metal, concrete or water barrier temporarily placed in the scene in order to re-direct vehicle or pedestrian traffic. In particular, includes barriers used at construction zones.*
- ➔ *If there are multiple barriers either connected or just placed next to each other, they should be annotated separately.*
- ➔ *If barriers are installed permanently, then do **NOT** include them.*

Human Annotators Need Multi-Modal Concept Alignment too!



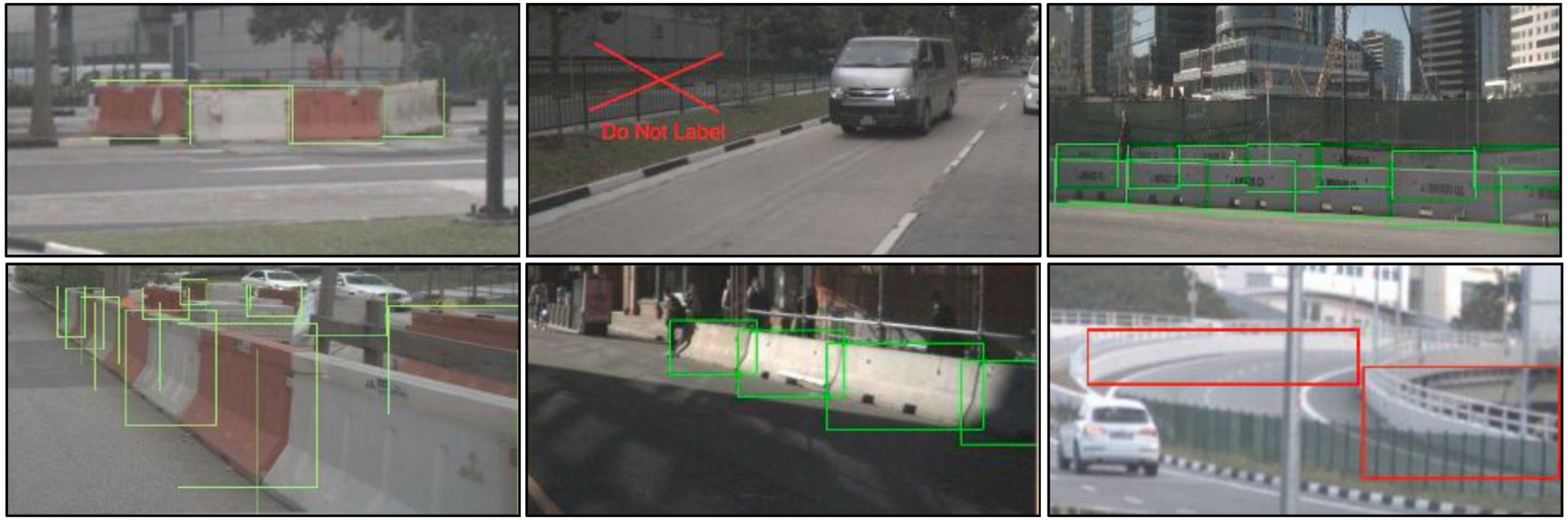
Negatives

Barrier

- ➔ *Any metal, concrete or water barrier temporarily placed in the scene in order to re-direct vehicle or pedestrian traffic. In particular, includes barriers used at construction zones.*
- ➔ *If there are multiple barriers either connected or just placed next to each other, they should be annotated separately.*
- ➔ *If barriers are installed permanently, then do* ***NOT*** *include them.*

Negatives

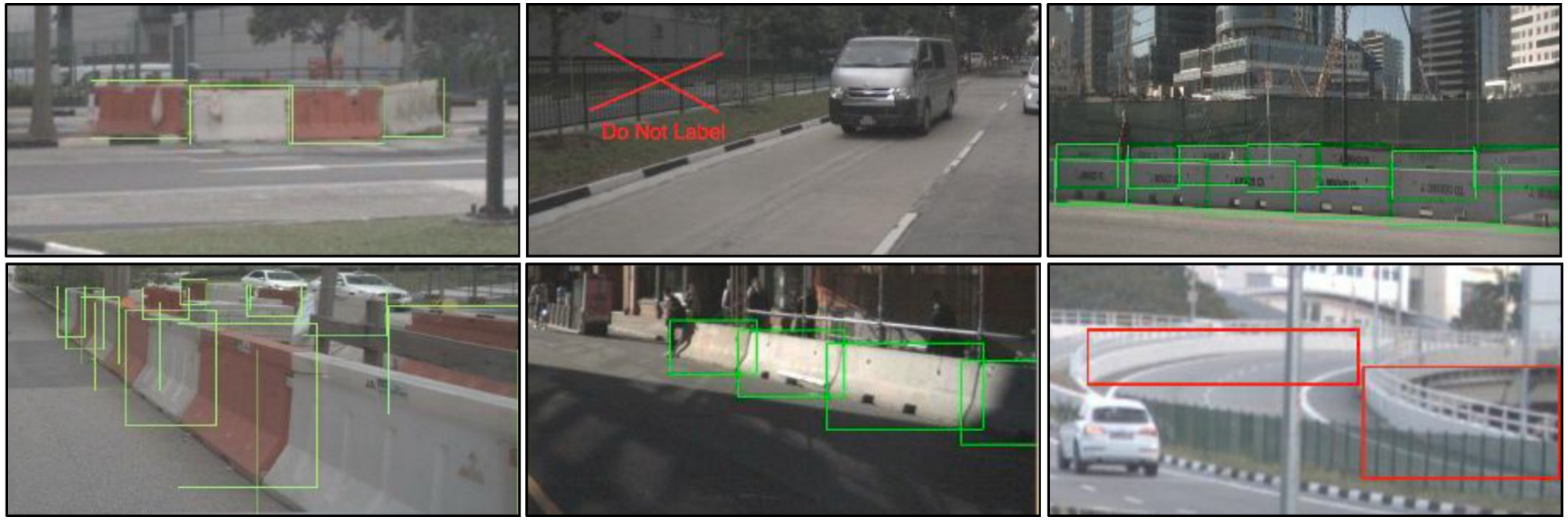
Can we **Align** Foundation Models like Human Annotators?



Barrier

- ➔ *Any metal, concrete or water barrier temporarily placed in the scene in order to re-direct vehicle or pedestrian traffic. In particular, includes barriers used at construction zones.*
- ➔ *If there are multiple barriers either connected or just placed next to each other, they should be annotated separately.*
- ➔ *If barriers are installed permanently, then do* ***NOT*** *include them.*

Can we **Align** Foundation Models like Human Annotators?

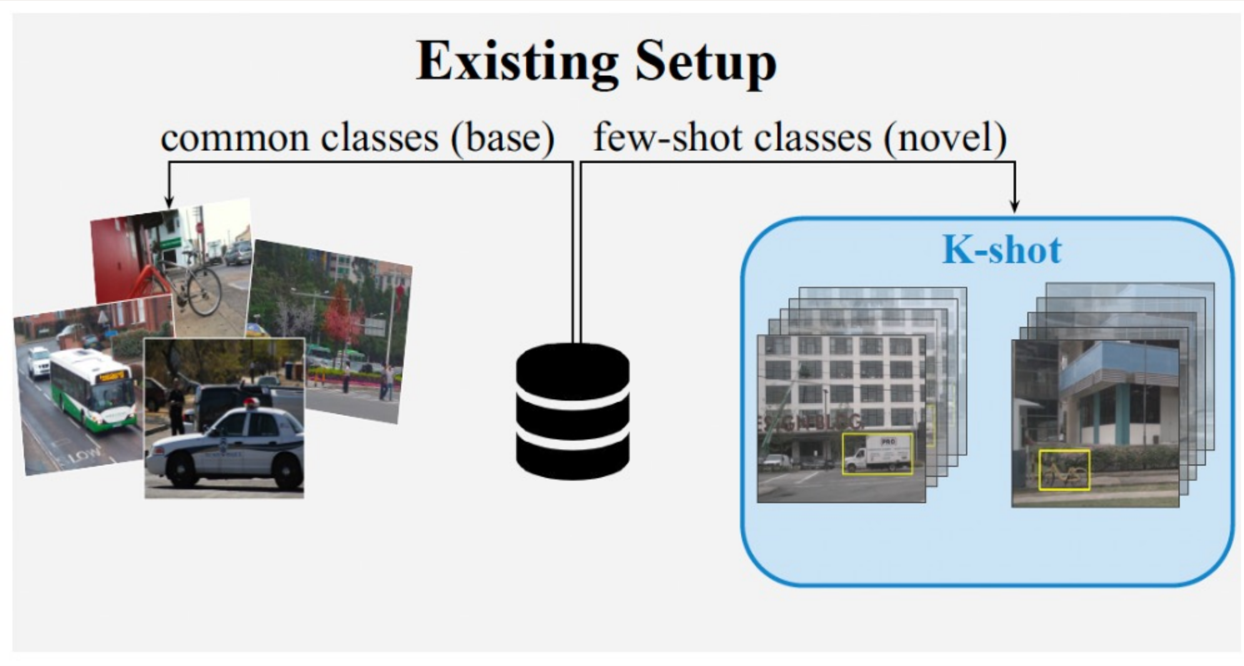


Barrier

- ➔ *Any metal, concrete or water barrier temporarily placed in the scene in order to re-direct vehicle or pedestrian traffic. In particular, includes barriers used at construction zones.*
- ➔ *If there are multiple barriers either connected or just placed next to each other, they should be annotated separately.*
- ➔ *If barriers are installed permanently, then do* ***NOT*** *include them.*

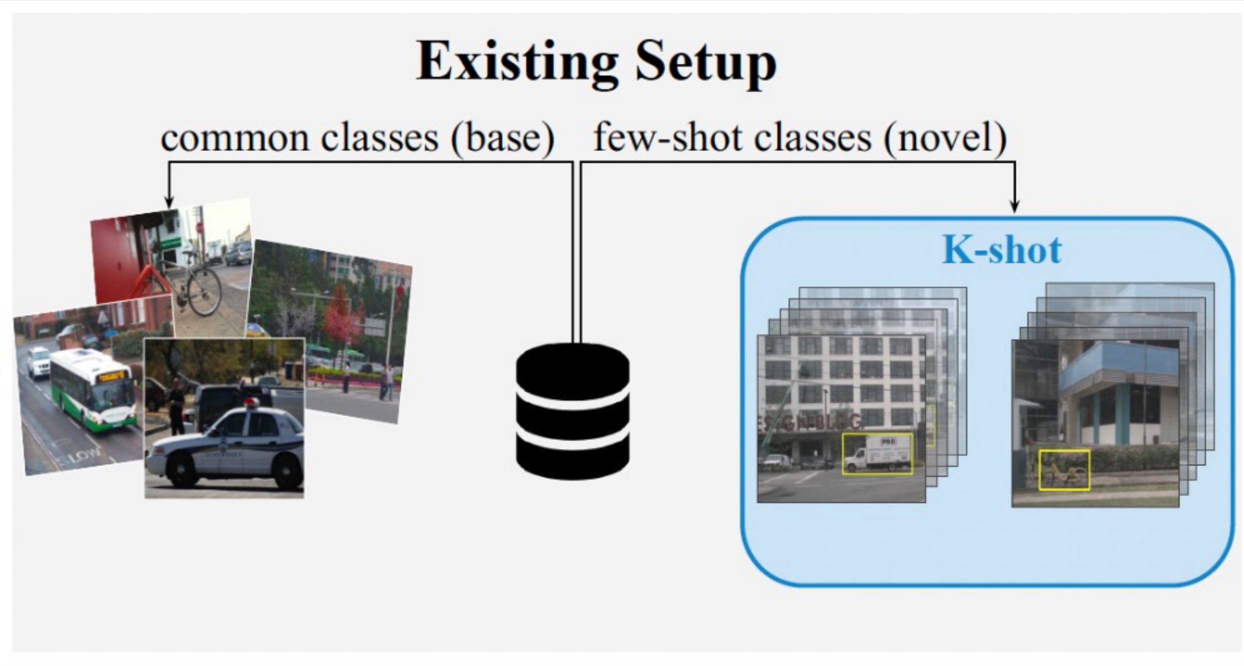
Yes! By adapting to few multi-modal examples via *fine-tuning, prompt tuning, in-context learning*

Isn't this like Few-Shot Object Detection?



Foundational VLMs violate current FSOD benchmarking protocols

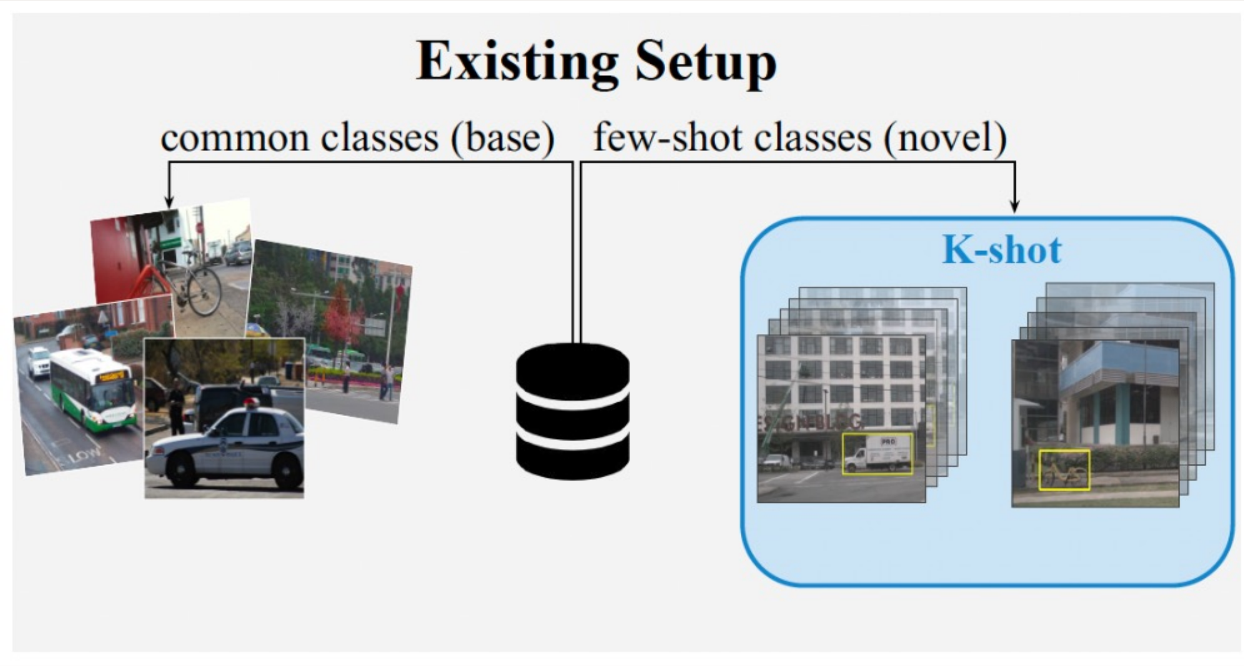
Isn’t this like Few-Shot Object Detection?



No Base vs. Novel Splits: VLMs like CLIP are trained on private datasets, so we can’t define novel

Foundational VLMs violate current FSOD benchmarking protocols

Isn't this like Few-Shot Object Detection?

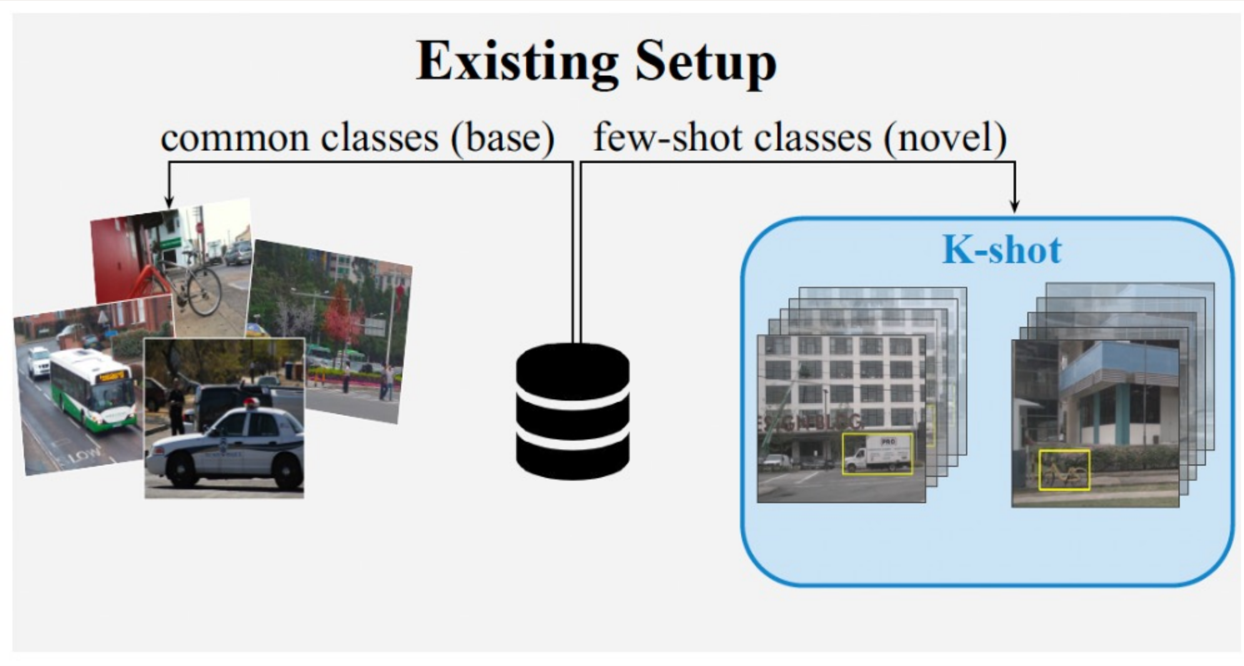


No Base vs. Novel Splits: VLMs like CLIP are trained on private datasets, so we can't define novel

Concept Leakage: Foundation models have been trained on diverse data, so they have seen car, cat, and person (considered novel in COCO)

Foundational VLMs violate current FSOD benchmarking protocols

Isn't this like Few-Shot Object Detection?



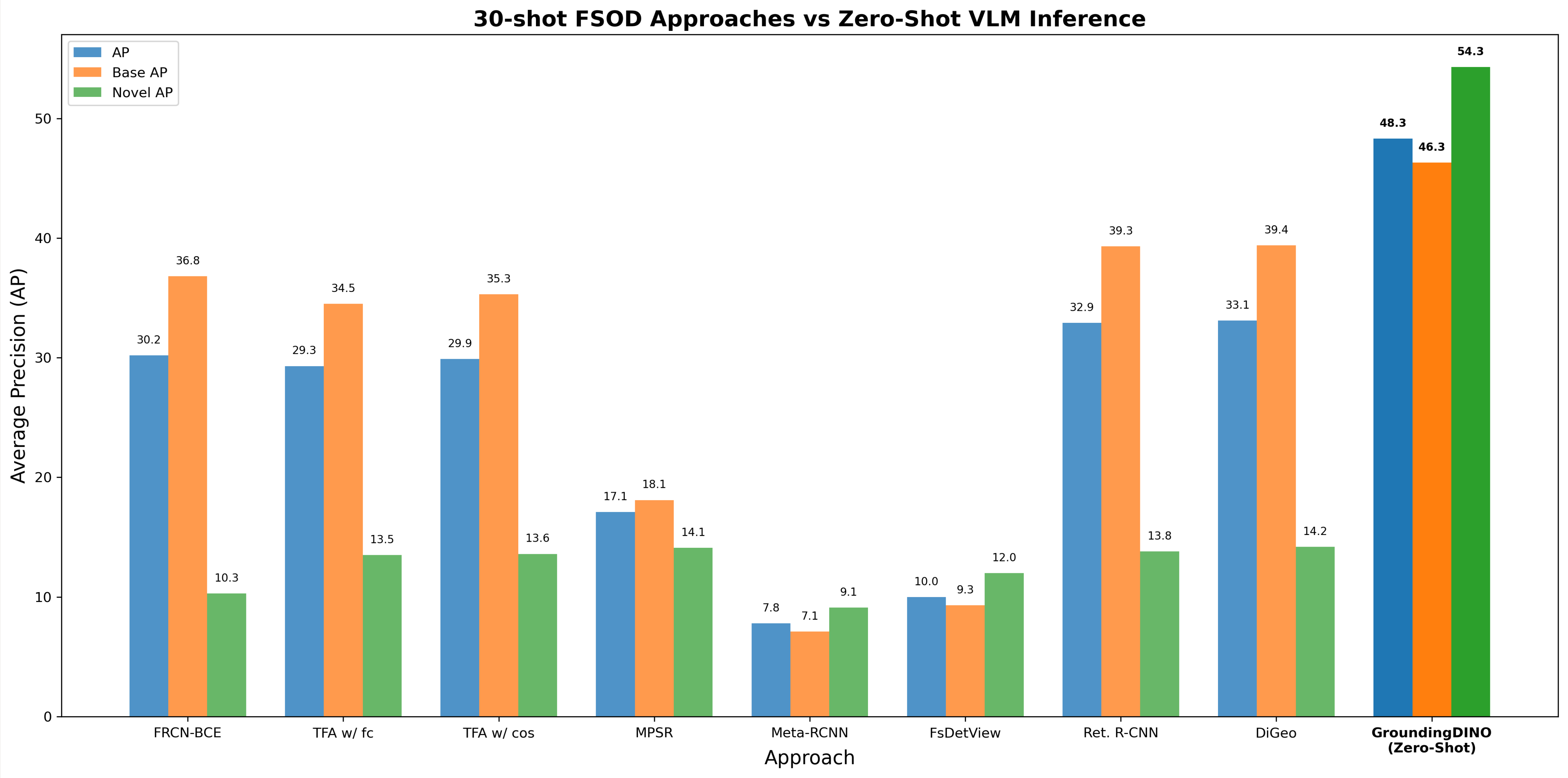
No Base vs. Novel Splits: VLMs like CLIP are trained on private datasets, so we can't define novel

Concept Leakage: Foundation models have been trained on diverse data, so they have seen car, cat, and person (considered novel in COCO)

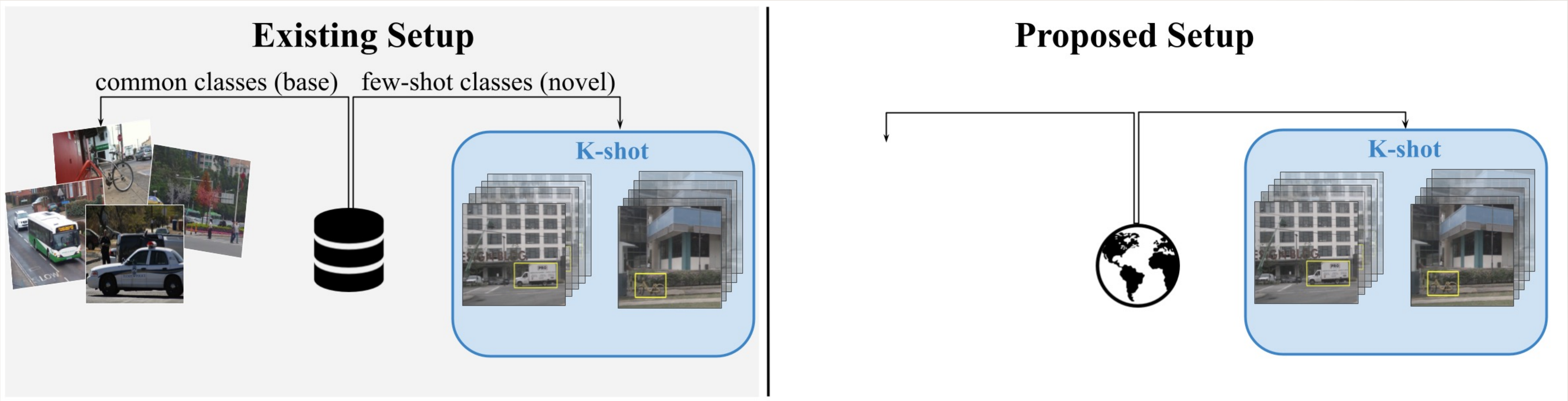
Role of Language: Existing setup ignores language cues

Foundational VLMs violate current FSOD benchmarking protocols

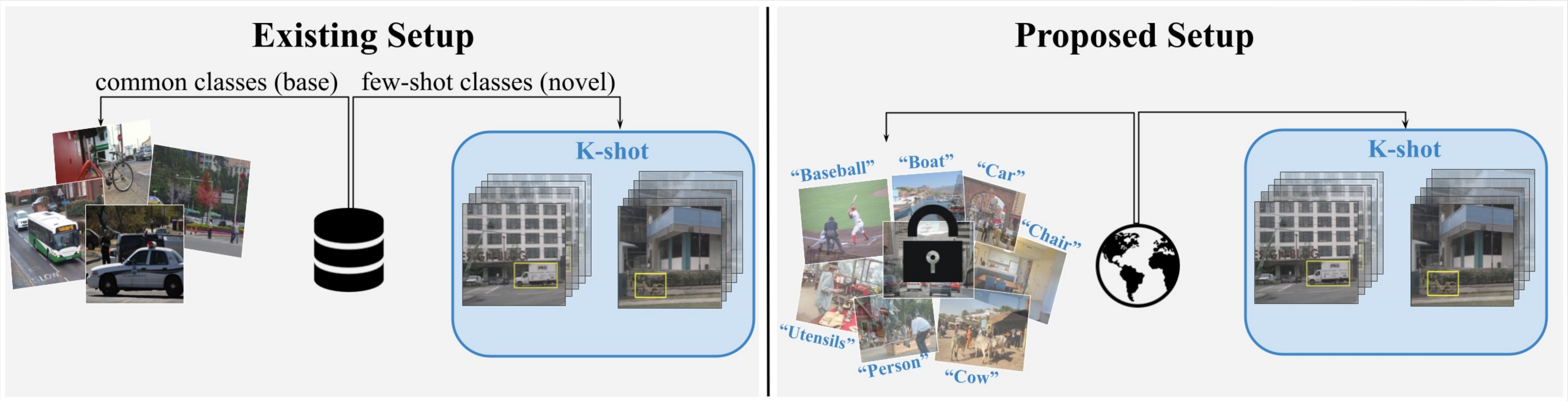
Zero-Shot VLMs beat SOTA FSOD Methods



Foundational VLMs should "Enter the Conversation"



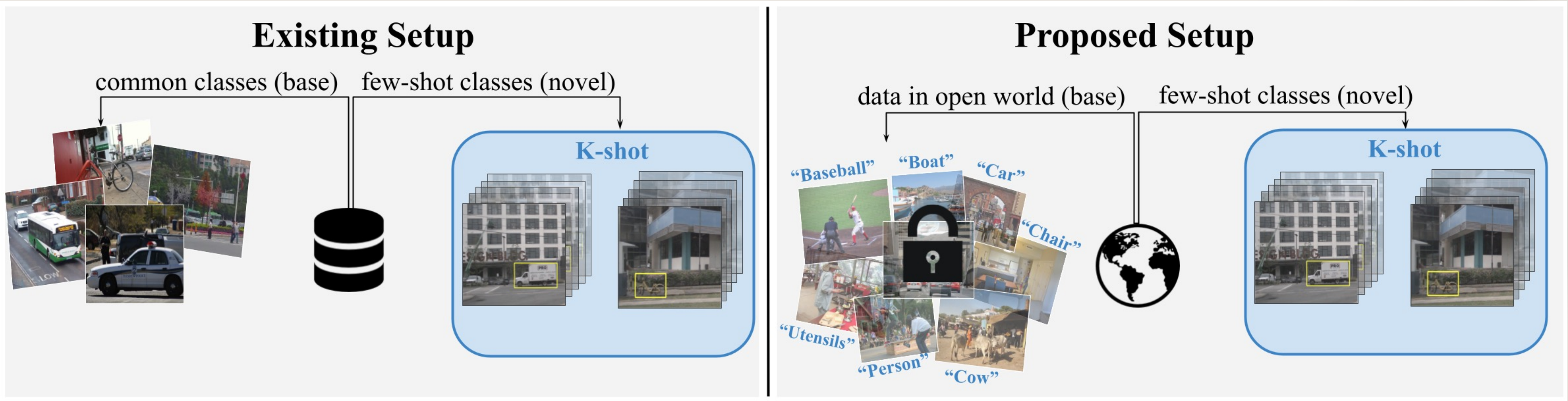
Foundational VLMs should "Enter the Conversation"



We should embrace

- ➢ web-scale pre-training

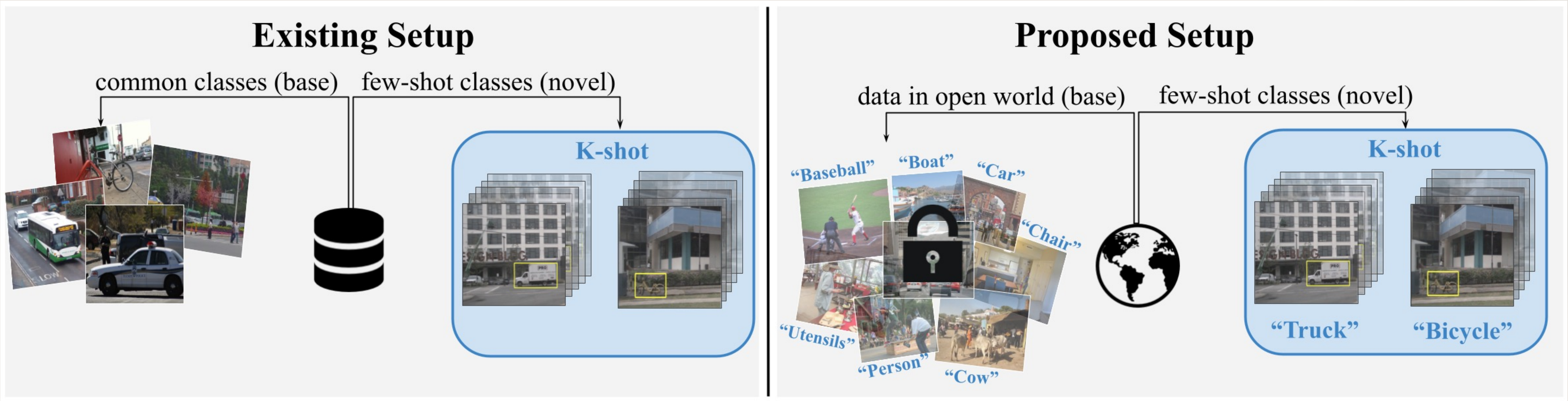
Foundational VLMs should "Enter the Conversation"



We should embrace

- ➢ web-scale pre-training
- ➢ concept-leakage by re-framing `base` vs. `novel`

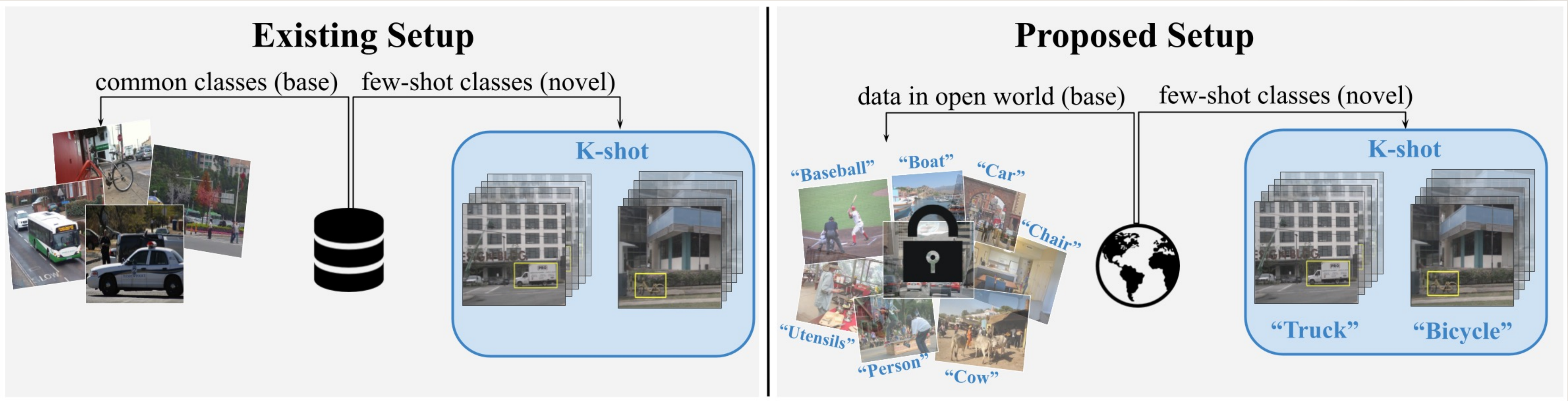
Foundational VLMs should "Enter the Conversation"



We should embrace

- ➢ web-scale pre-training
- ➢ concept-leakage by re-framing `base` vs. `novel`
- ➢ language cues as additional signal

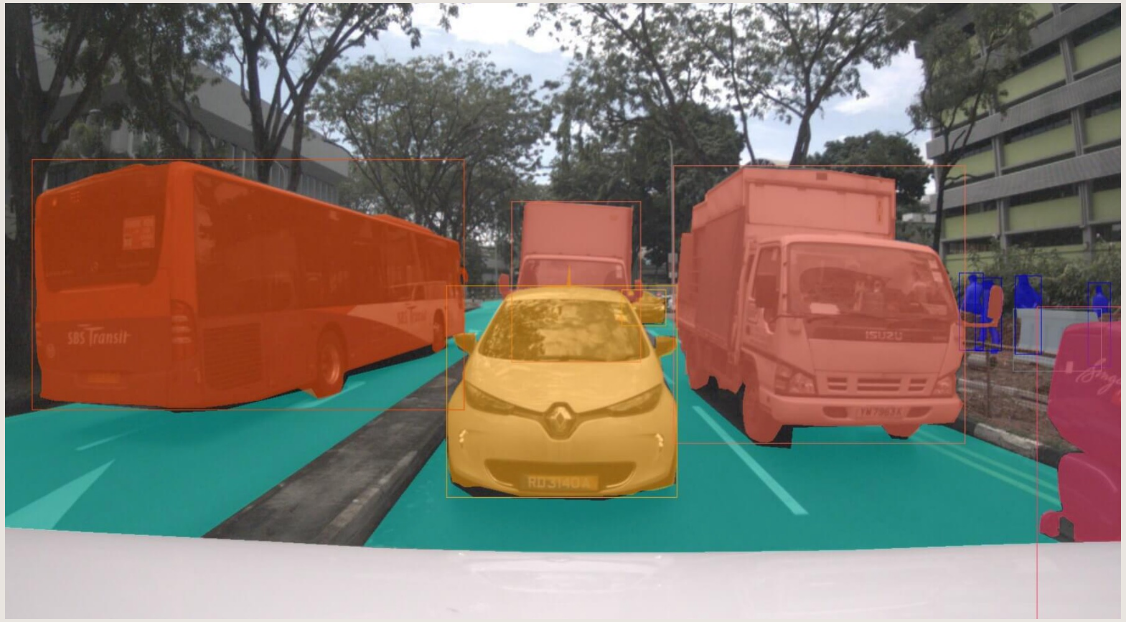
Foundational FSOD Benchmark

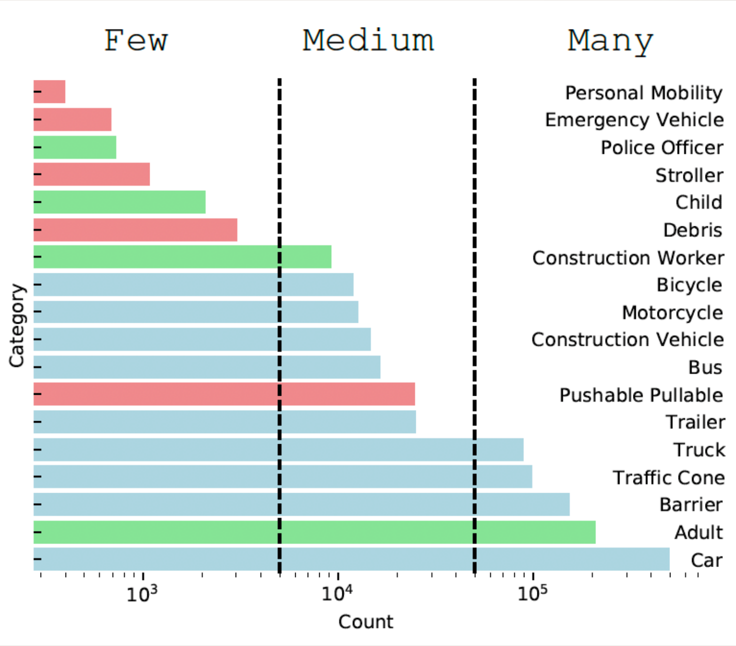


We should embrace

- web-scale pre-training
- concept-leakage by re-framing `base` vs. `novel`
- language cues as additional signal

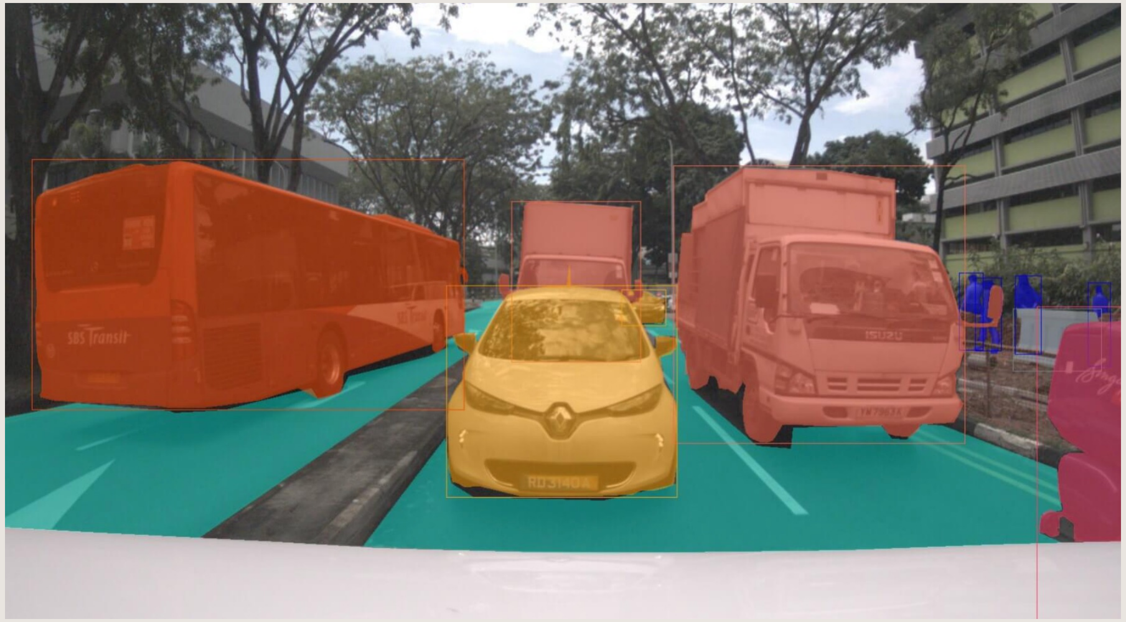
Repurposing nuImages for Foundational FSOD

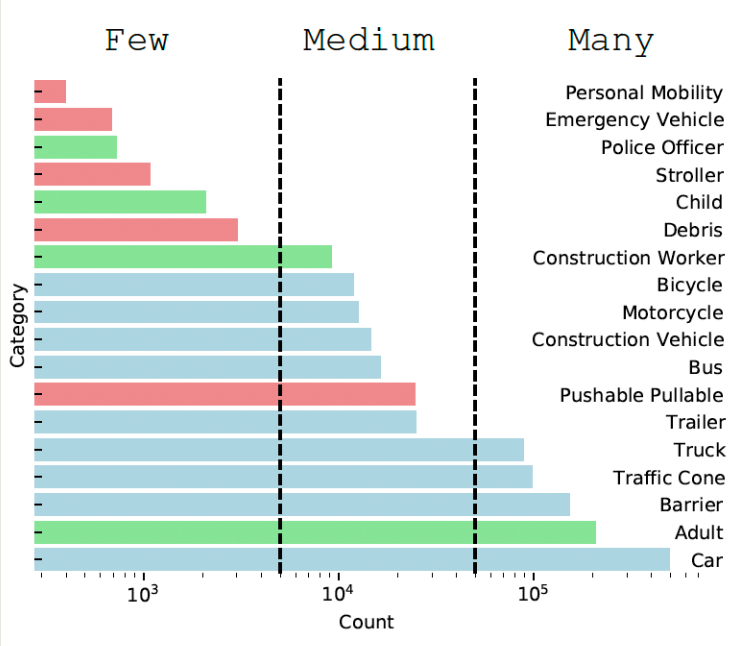




➢ 2D AV dataset (not typically used for FSOD) with **challenging open-world categories** like `pushable-pullable` and `debris`

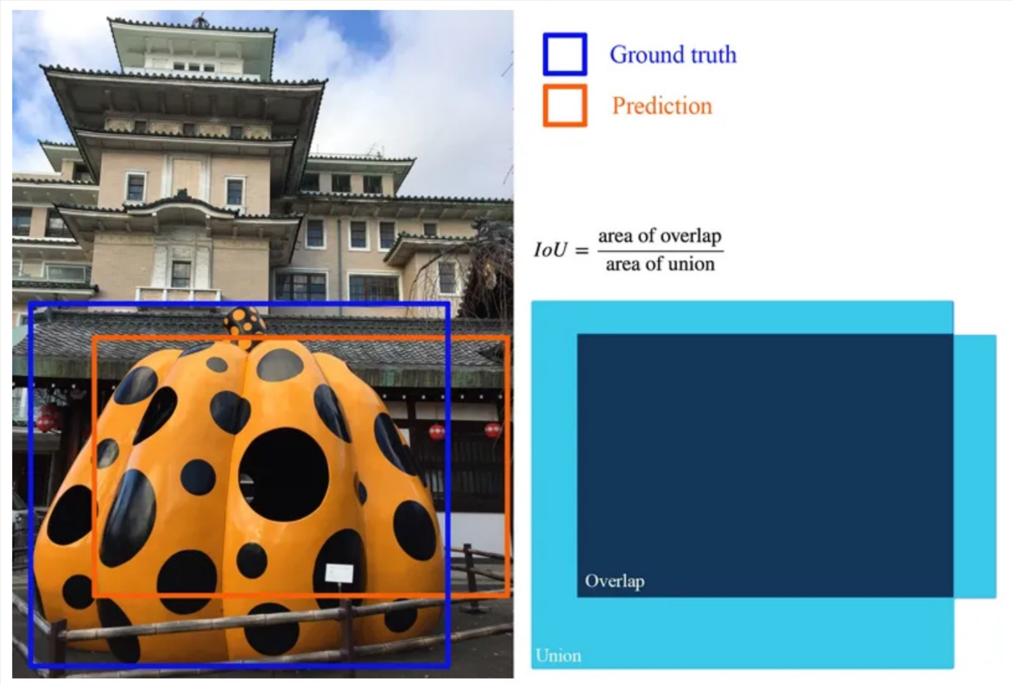
Repurposing nuImages for Foundational FSOD





- 2D AV dataset (not typically used for FSOD) with **challenging open-world categories** like `pushable-pullable` and `debris`
- Contains **publicly available multi-modal annotator instructions**

Evaluation Metric: Mean Average Precision



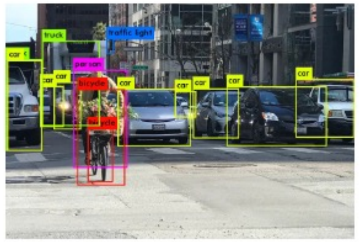
Confidence	Correct?	Precision	Recall
0.9	✅	1.0	0.33
0.7	❌	0.5	0.33
0.5	✅	0.67	0.67
0.3	❌	0.5	0.67
0.2	❌	0.4	0.67

- COCO-style evaluation for **18 classes**
- Classes are grouped by frequency: **Many**, **Medium** and **Few**

10-shot Foundational FSOD

Approach	Backbone	Pre-Train Data	Average Precision (AP)			
			All	Many	Med	Few
Zero-Shot Detection						
Detic [61]	SWIN-B	LVIS, COCO, IN-21K	14.40	25.83	16.59	2.32
GLIP [29]	SWIN-L	FourODs,GoldG,Cap24M	17.01	23.36	19.86	8.40
MQ-GLIP-Text [53]	SWIN-L	Objects365, FourODs, GoldG, Cap24M	17.01	23.36	19.85	8.41
Prompt Engineering						
Detic [61]	SWIN-B	LVIS, COCO, IN-21K	14.92	26.48	17.29	2.53
GLIP [29]	SWIN-L	FourODs, GoldG, Cap24M	17.15	23.82	19.36	9.02
Standard Fine-Tuning						
RegionCLIP [58]	RN50	CC3M	3.86	6.08	5.13	0.54
Detic [61]	SWIN-B	LVIS, COCO, IN-21K	16.09	25.46	20	3.73
Federated Fine-Tuning (Ours)						
Detic [61]	SWIN-B	LVIS, COCO, IN-21K	17.24	28.07	20.71	4.18
Detic [61] w/ Prompt Engineering	SWIN-B	LVIS, COCO, IN-21K	17.71	28.46	21.14	4.75
Language Prompt Tuning						
GLIP [29]	SWIN-L	FourODs,GoldG,Cap24M	19.41	22.18	**25.16**	**10.39**
Visual Prompting						
MQ-GLIP-Image [53]	SWIN-L	Objects365,FourODs,GoldG,Cap24M	14.07	24.39	15.89	3.34
Multi-Modal Prompting						
MQ-GLIP [53]	SWIN-L	Objects365,FourODs,GoldG,Cap24M	21.42	32.19	23.29	10.26
Multi-Modal Chat Assistants						
GPT-4o Zero-Shot Classification [1]	*Private*	*Private*	9.95	16.81	12.11	1.71
Iterative Prompting: MQ-GLIP	Private	Private	**22.03**	**33.42**	24.72	9.41

1st CVPR Foundational FSOD Challenge



Foundational Few-Shot Object Detection Challenge

Organized by: foundational_fsod

Published

Starts on: Apr 10, 2024 8:00:00 PM EST (GMT - 5:00)

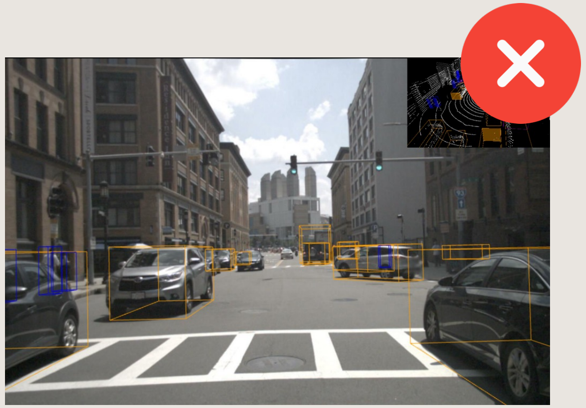
Ends on: Jun 7, 2099 7:59:59 PM EST (GMT - 5:00)

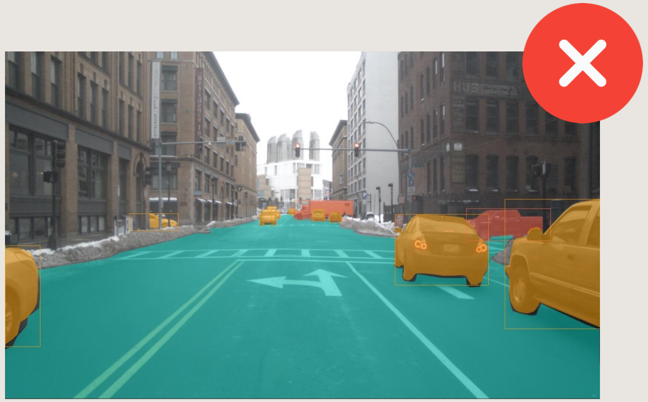
Multi-Modal | VLMs | FSOD | Detection | Computer Vision

Challenge Results presented at the **Visual Perception via Learning in an Open World Workshop,** CVPR 2024

Foundational FSOD Challenge: Setup

- 10-shot nuImages split
- 2-month timeline for submissions
- **Constraint**: Can pre-train/fine-tune on anything except nuImages and nuScenes





Foundational FSOD Challenge: Results

8 Teams

50+ Submissions

Foundational FSOD Challenge: Results

8 Teams

50+ Submissions

Approach	Backbone	Average Precision (AP) All	Many	Med	Few
Multi-Modal Prompting					
MQ-GLIP	SWIN-L	21.42	32.19	23.29	10.26
CVPR 2024 Competition Results					
PHP_hhh	Private	**45.35**	**64.25**	**53.43**	**20.19**
NJUST KMG	SWIN-L	32.56	50.21	34.87	15.16
zjyd_cxy_vision	SWIN-L	31.57	46.59	33.32	17.03

Leaderboard

Foundational FSOD Challenge: Results

8 Teams

50+ Submissions

Approach	Backbone	Average Precision (AP) All	Many	Med	Few
Multi-Modal Prompting					
MQ-GLIP	SWIN-L	21.42	32.19	23.29	10.26
CVPR 2024 Competition Results					
PHP_hhh	Private	**45.35**	**64.25**	**53.43**	**20.19**
NJUST KMG	SWIN-L	32.56	50.21	34.87	15.16
zjyd_cxy_vision	SWIN-L	31.57	46.59	33.32	17.03

Leaderboard

Top submission outperforms our best baseline by over **2x**!

Project Page