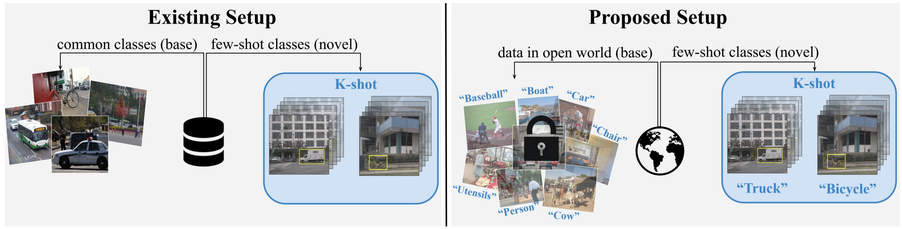
# Revisiting Few-Shot Object Detection with Vision-Language Models

Anish Madan*, Neehar Peri*, Shu Kong†, Deva Ramanan†

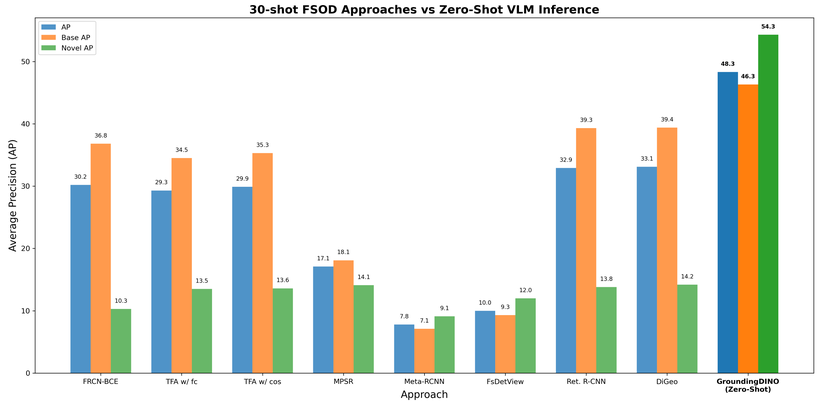



## Standard FSOD

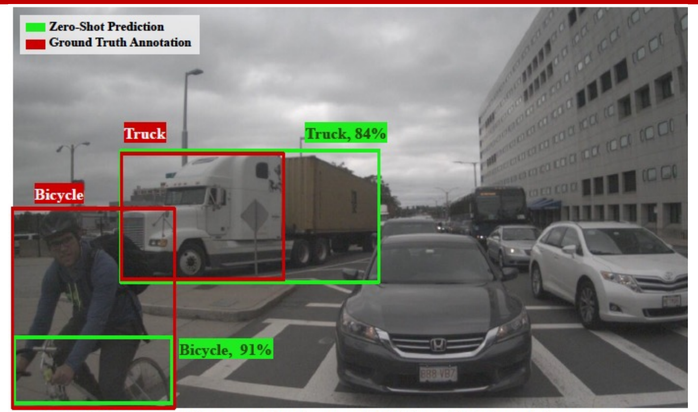
- Pre-train on `base` classes
- Fine-tune on K-shots of `novel` classes
- Disjoint `novel` and `base` classes

## Foundational FSOD

- Pre-train on `web-scale` datasets
- Fine-tune on K-shots of `target` classes
- Foundation models can now "enter the conversation"




Zero-shot VLMs beat all prior FSOD approaches. Is FSOD trivially solved with foundation models? We argue **no**!




Poor Concept Alignment between VLM and Dataset Annotations

**nuImages Bicycle**

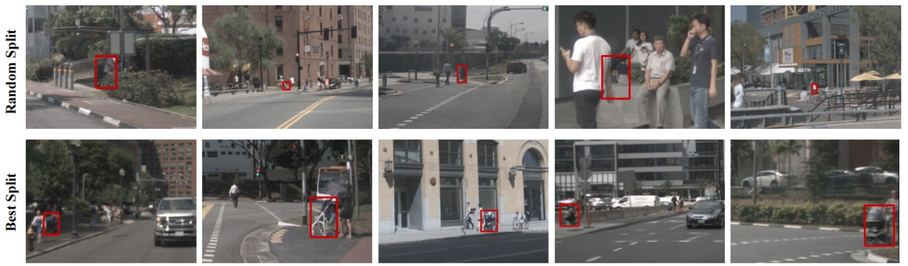
- *Human or electric powered 2-wheeled vehicle designed to travel at lower speeds either on road surface, sidewalks or bicycle paths.*
- *If there is a rider, include the rider in the box*
- *If there is a pedestrian standing next to the bicycle, do NOT include in the annotation*

**nuImages Trucks**

- *Vehicles primarily designed to haul cargo including pick-ups, lorries, trucks and semi-tractors. Trailers hauled after a semi-tractor should be labeled as trailer.*
- *A pickup truck is a light duty truck with an enclosed cab and an open or closed cargo area.*

Multimodal Annotation Instructions

**Insight:** VLMs must be ***aligned*** (like human annotators) with a few multi-modal examples!




***Which few-shot examples should we use to fine-tune***? Pick examples from clear images that are large and unoccluded!

| Approach | Pre-Train Data | Average Precision (AP) | | | |
|---|---|---|---|---|---|
| | | All | Many | Med | Few |
| **Zero-Shot Detection** | | | | | |
| RegionCLIP | CC3M | 2.50 | 3.20 | 3.80 | 0.40 |
| Detic | LVIS, COCO, IN-21K | 14.40 | 25.83 | 16.59 | 2.32 |
| GLIP | FourODs, GoldG, Cap24M | 17.01 | 23.36 | 19.86 | 8.40 |
| MQ-GLIP-Text | Objects365, FourODs, GoldG, Cap24M | 17.01 | 23.36 | 19.85 | 8.41 |
| **Prompt Engineering** | | | | | |
| Detic | LVIS, COCO, IN-21K | 14.92 | 26.48 | 17.29 | 2.53 |
| GLIP | FourODs, GoldG, Cap24M | 17.15 | 23.82 | 19.36 | 9.02 |
| **Standard Fine-Tuning** | | | | | |
| RegionCLIP | CC3M | 3.86 | 6.08 | 5.13 | 0.54 |
| Detic | LVIS, COCO, IN-21K | 16.09 | 25.46 | 20.00 | 3.73 |
| **Federated Fine-Tuning (Ours)** | | | | | |
| Detic | LVIS, COCO, IN-21K | 17.24 | 28.07 | 20.71 | 4.18 |
| Detic w/ Prompt Engineering | LVIS, COCO, IN-21K | 17.71 | 28.46 | 21.14 | 4.75 |
| **Language Prompt Tuning** | | | | | |
| GLIP | FourODs, GoldG, Cap24M | 19.41 | 22.18 | **25.16** | **10.39** |
| **Visual Prompting** | | | | | |
| MQ-GLIP-Image | O365, FourODs, GoldG, Cap24M | 14.07 | 24.39 | 15.89 | 3.34 |
| **Multi-Modal Prompting** | | | | | |
| MQ-GLIP | O365, FourODs, GoldG, Cap24M | 21.42 | 32.19 | 23.29 | 10.26 |
| **Multi-Modal Chat Assistants** | | | | | |
| GPT-4o Zero-Shot Classification | *Private* | 9.95 | 16.81 | 12.11 | 1.71 |
| MQ-GLIP Iterative Prompting | *Private* | **22.03** | **33.42** | 24.72 | 9.41 |
| **CVPR 2024 Competition Results** | | | | | |
| PHP_hhh | *Private* | **45.35** | **64.25** | **53.43** | **20.19** |
| NJUST KMG | O365V2, OIV6, GoldG, V3Det, COCO, LVIS, GRIT, RefCOCO* ... | 32.56 | 50.21 | 34.87 | 15.16 |
| zjyd_cxy_vision | O365V2, OIV6, GoldG, V3Det, COCO, LVIS, GRIT, RefCOCO* ... | 31.57 | 46.59 | 33.32 | 17.03 |