# VPLOW@CVPR'25

# The 5th Workshop of Visual Perception and Learning in an Open World


Shu Kong

University of Macau


June 11, 2025

# Welcome

- Hope everyone is *safe* and *healthy*!

- Goals of the workshop
  - connect people and exchange ideas *about Open-World Vision*
  - discuss new opportunities and challenges *about Open-World Vision*

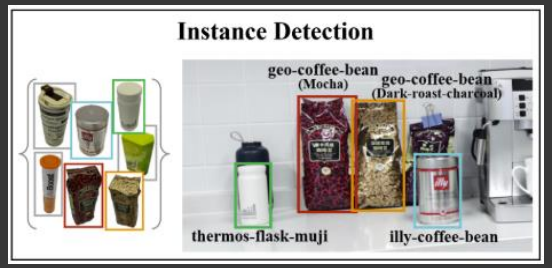
- Hybrid workshop
  - on-site: enjoy and involve by asking questions
  - online via zoom (provided by CVPR'25)

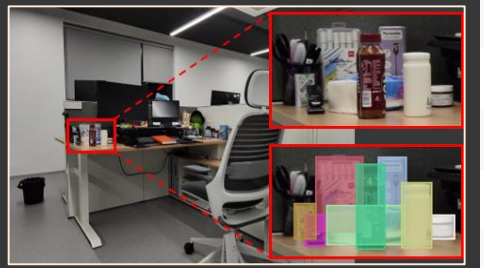
# Schedule

https://vplow.github.io/vplow_5th.html

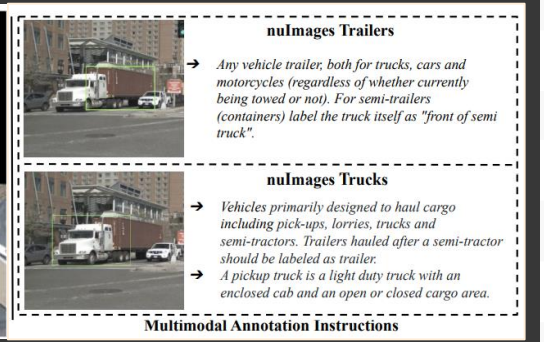
Two challenges

1. InsDet: Object Instance Detection



2. Foundational Few-Shot Object Detection


| CDT | Event | Presenter / Title |
|---|---|---|
| 09:00 - 09:20 | Opening remarks | **Shu Kong** University of Macau<br>**Visual Perception via Learning in an Open World** |
| 09:20 - 10:00 | Invited talk #1 | **Kristen Grauman**, UT Austin<br>**Human activity in the open world** |
| 10:00 - 10:40 | Invited talk #2 | **Gunshi Gupta**, **Yarin Gal**, University of Oxford<br>**tba** |
| 10:40 - 11:20 | Invited talk #3 | **Grant Van Horn**, UMass-Amherst<br>**tba** |
| 11:20 - 12:00 | Invited talk #4 | **Abhinav Gupta**, CMU<br>**tba** |
| 12:00 - 13:00 | **Lunch** | |
| 13:00 - 13:40 | Invited talk #5 | **Yuxiong Wang**, UIUC<br>**tba** |
| 13:40 - 14:20 | Invited talk #6 | **Deepak Pathak**, CMU<br>**tba** |
| 14:20 - 15:00 | Invited talk #7 | **Liangyan Gui**, UIUC<br>**tba** |
| 15:00 - 15:05 | **Coffee break** | |
| 15:05 - 15:45 | Invited talk #7 | **Georgia Gkioxari**, Caltech<br>**tba** |
| 15:45 - 16:20 | Challenge-1 | **Challenge 1: InsDet**<br>**Object Instance Detection Challenge** |
| 16:20 - 16:55 | Challenge-2 | **Challenge-2: Foundational FSOD**<br>**Foundational Few-Shot Object Detection Challenge v2** |
| 16:55 - 17:00 | Closing remarks | **Neehar Peri** CMU |

# Together we serve

## Speakers


Shu Kong
UMacau


Georgia Gkioxari
Caltech


Grant Van Horn
UMass, Amherst


Deepak Pathak
CMU


Yarin Gal
University of Oxford


Gunshi Gupta
University of Oxford


Abhinav Gupta
CMU


Kristen Grauman
UT Austin


Liangyan Gui
UIUC


Yu-Xiong Wang
UIUC

## Advisory Board


Deva Ramanan
Carnegie Mellon University


Terrance Boult
University of Colorado
Colorado Springs


Walter J. Scheirer
University of Notre Dame

## Organizers

Please contact Shu Kong with any questions: aimerykong [at] gmail [dot] com


Shu Kong
UMacau


Neehar Peri
CMU


Yu-Xiong Wang
University of Illinois at
Urbana-Champaign


Andrew Owens
University of Michigan


Abhinav Shrivastava
University of Maryland

## Challenge Organizers


Shu Kong
UMacau


Qianqian Shen
Zhejiang University


Yunhan Zhao
Google / UCI


Anish Madan
CMU


Neehar Peri
CMU


Deva Ramanan
CMU

roboflow

## Coordinators


Tian Liu
Texas A&M

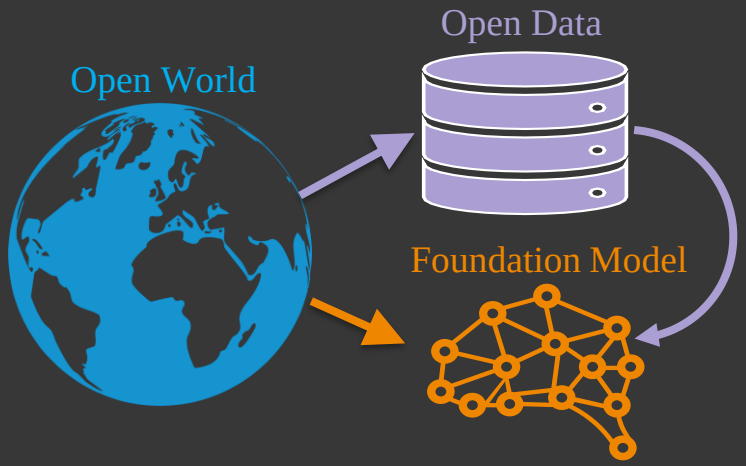

Yunhan Zhao
Google / UCI


Neehar Peri
CMU

# A brief introduction to the open world

Nowadays, we train Foundation Models (FMs) in the open world, on open data

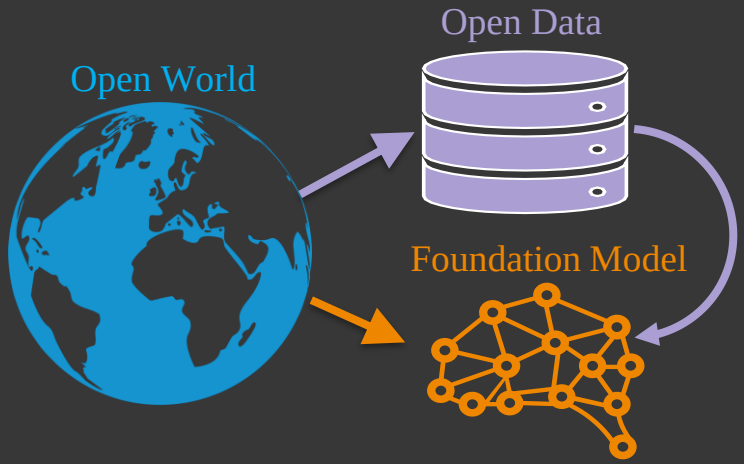
- open-vocabulary recognition
- zero-shot recognition
- ….

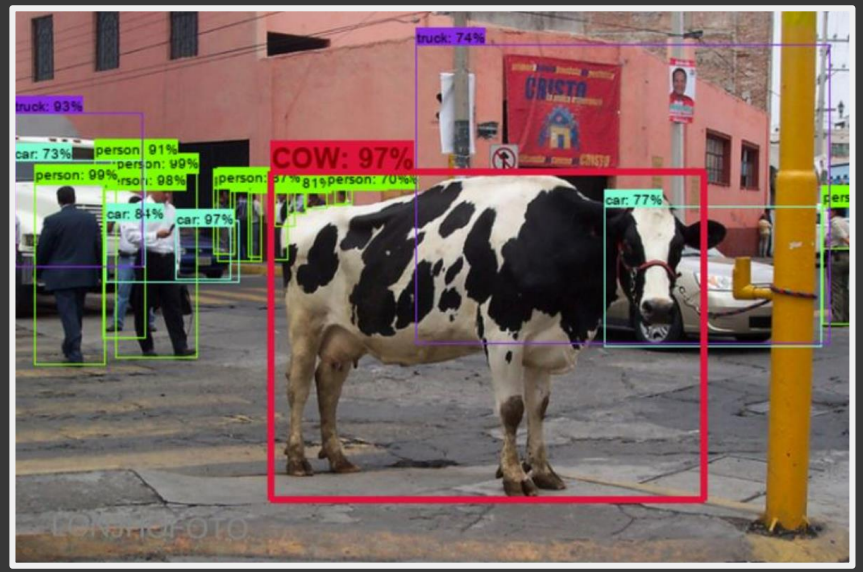


# Open-World Foundation Models (FMs) for Open-Vocabulary Detection

Nowadays, we train Foundation Models (FMs) in the open world, on open data

- open-vocabulary recognition
- zero-shot recognition
- ….



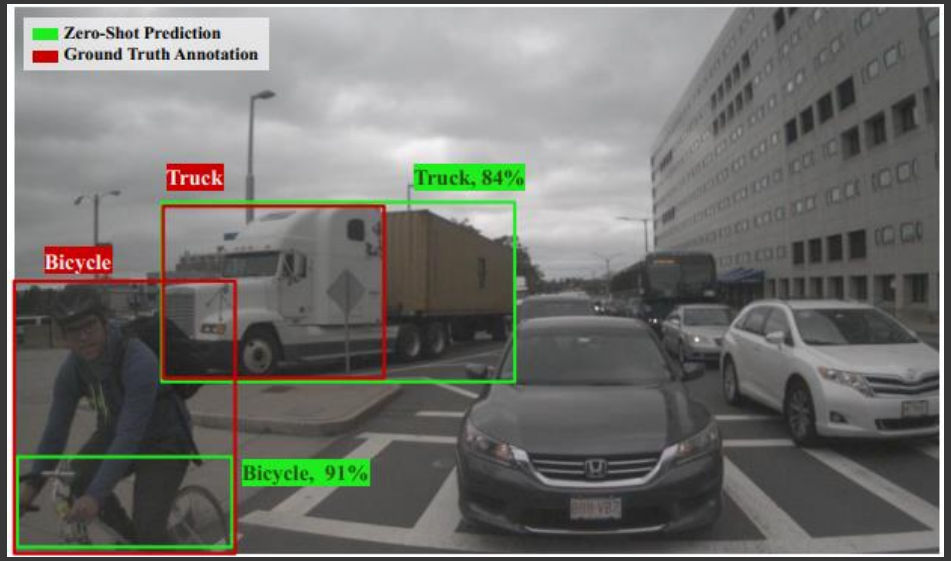



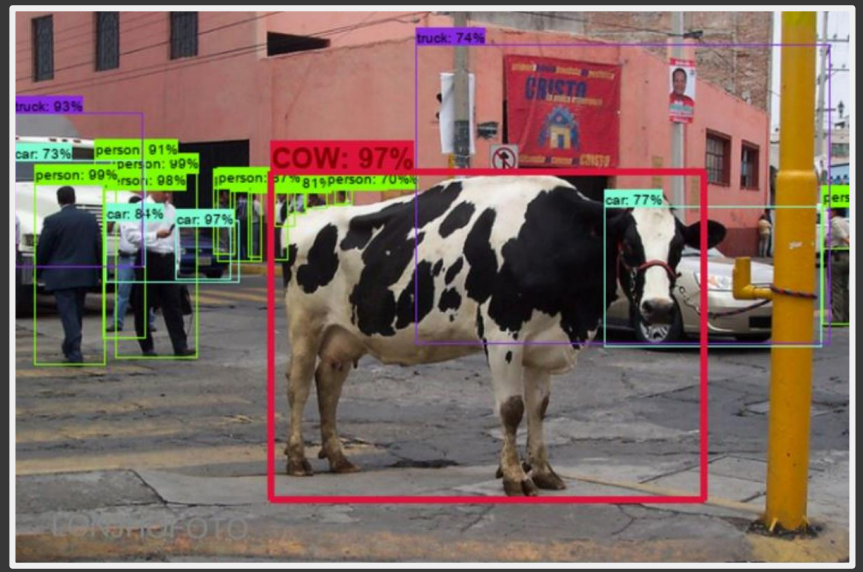

Cool!!!

Liu, et al., “Grounding DINO: Marrying dino with grounded pre-training for open-set object detection”, ECCV, 2024
Cheng, et al., "YOLO-World: Real-Time Open-Vocabulary Object Detection", CVPR, 2024

# FMs struggles!?

nuImages dataset



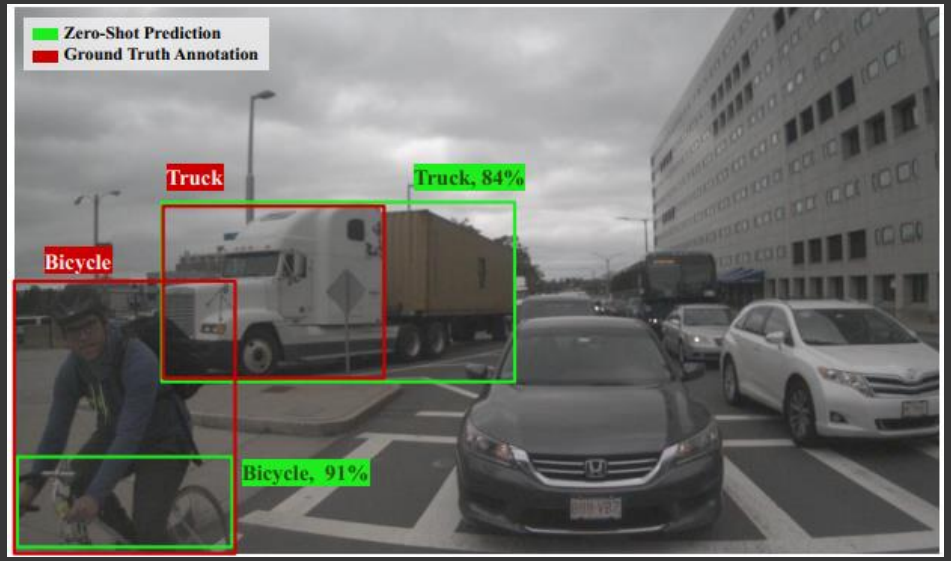

Poor alignments between foundational detector and ground-truth annotations in nuImages dataset. Why?

Cool!!!

Liu, et al., “Grounding DINO: Marrying dino with grounded pre-training for open-set object detection”, ECCV, 2024
Madan, et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

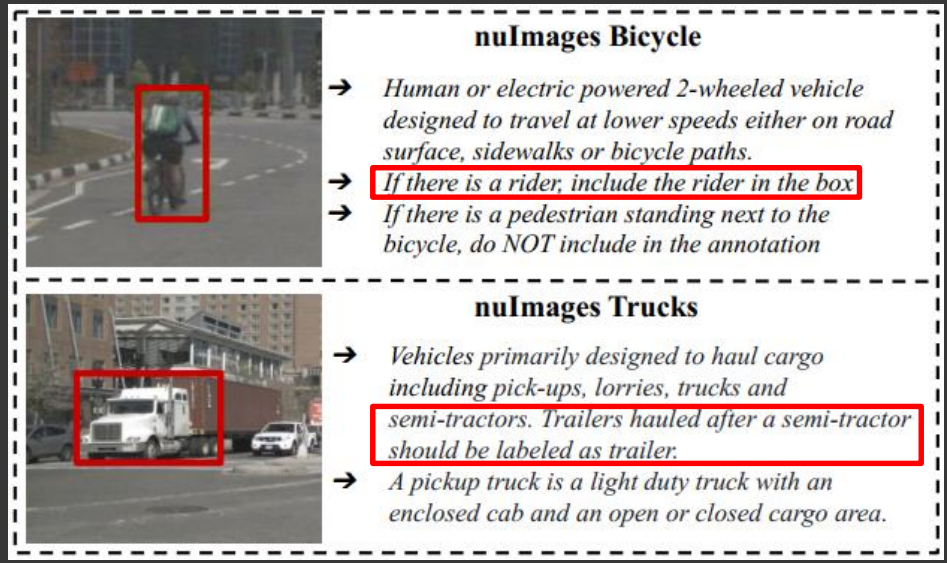
# FM struggles on specific downstream tasks!

nuImages dataset




Poor alignments between foundational detector and ground-truth annotations in nuImages dataset. Why?

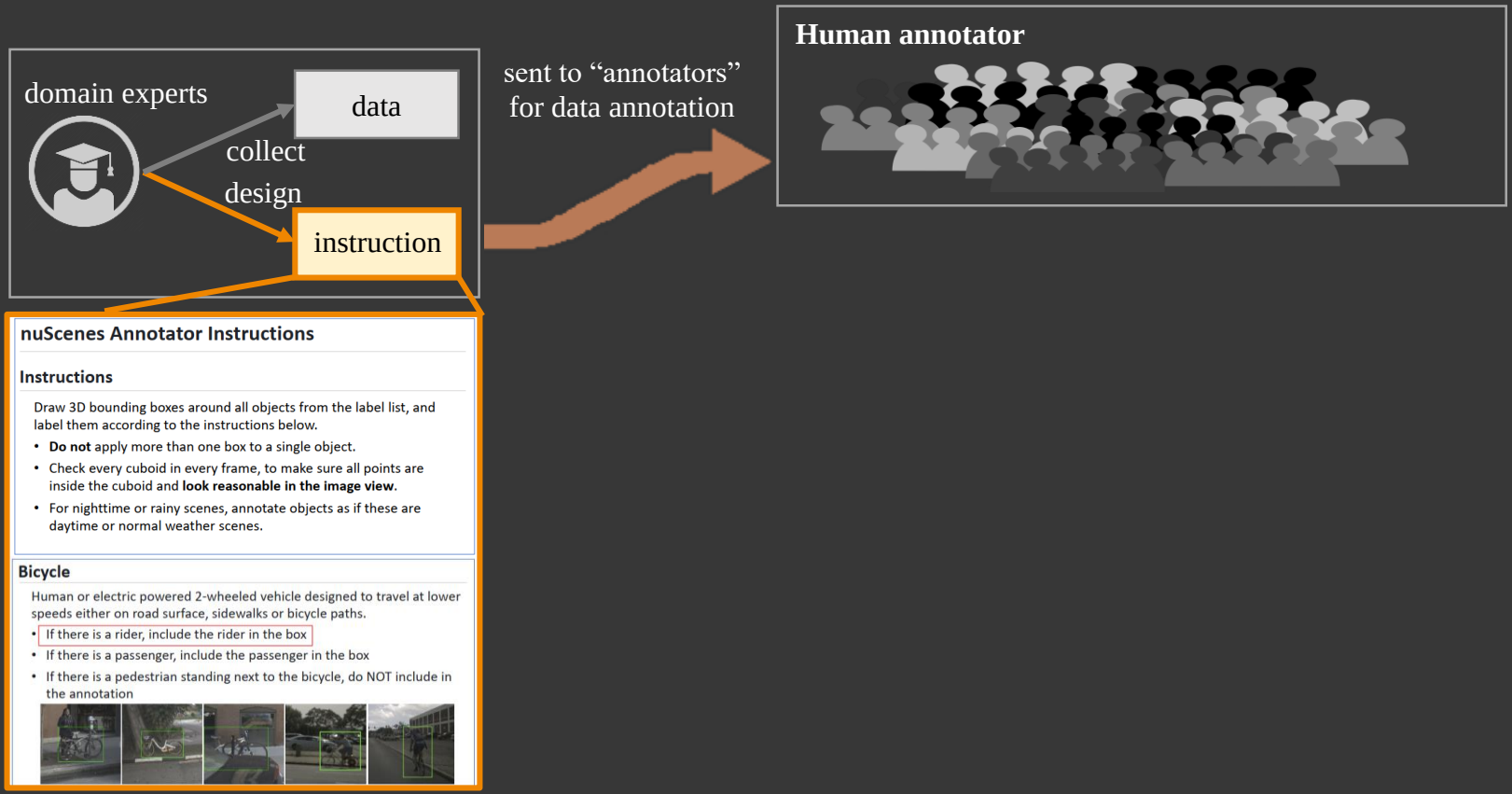
A snippet of annotation guidelines from nuImages




Annotation instructions designed by autonomous driving experts with special considerations.

Liu, et al., “Grounding DINO: Marrying dino with grounded pre-training for open-set object detection”, ECCV, 2024
Madan, et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

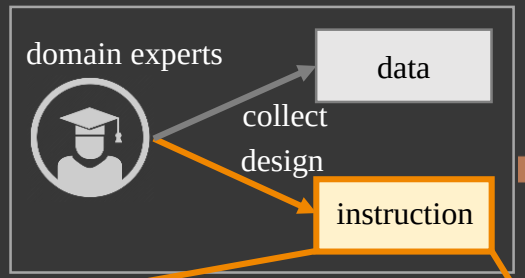
# Speaking of data labeling…






Madan et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

# How about exploit the open world for (automating) data labeling?

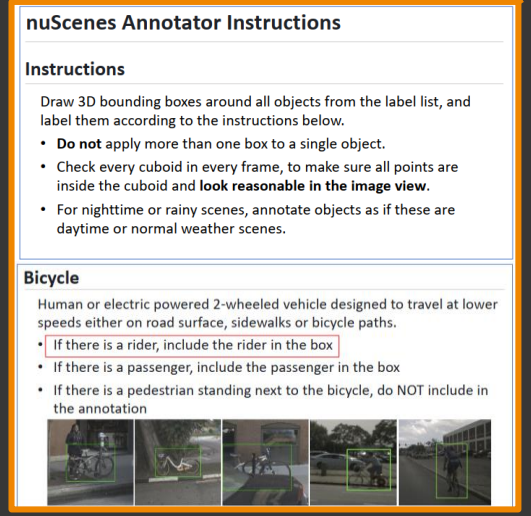



sent to “annotators” for data annotation

**Human annotator**

**AI annotator / Foundation Models**

- Large Language Models (LLMs)
- Vision-Language Models (VLMs)
- Foundation Vision Models (FVMs)
- Large Multi-Modal Models (LMMs)



**nuScenes Annotator Instructions**

**Instructions**

Draw 3D bounding boxes around all objects from the label list, and label them according to the instructions below.

- **Do not** apply more than one box to a single object.
- Check every cuboid in every frame, to make sure all points are inside the cuboid and **look reasonable in the image view.**
- For nighttime or rainy scenes, annotate objects as if these are daytime or normal weather scenes.

**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

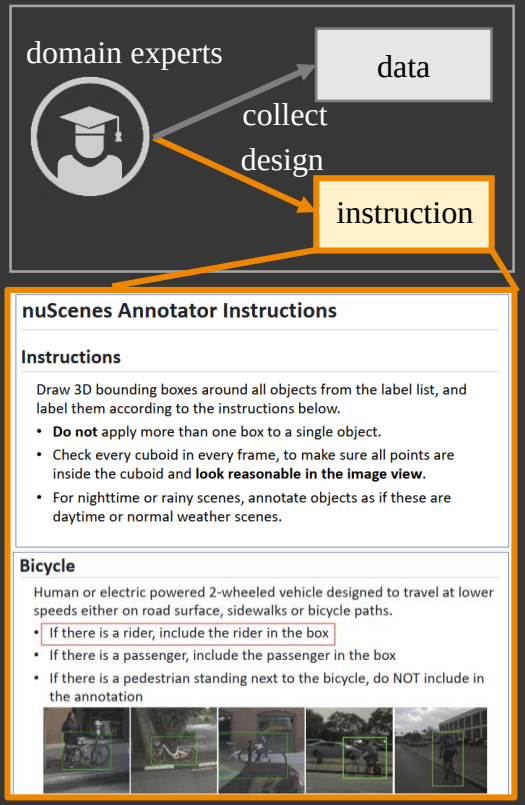


- Can we replace human annotators with FMs?
- How to adapt FMs to align with experts?
- This is a multimodal few-shot learning problem.

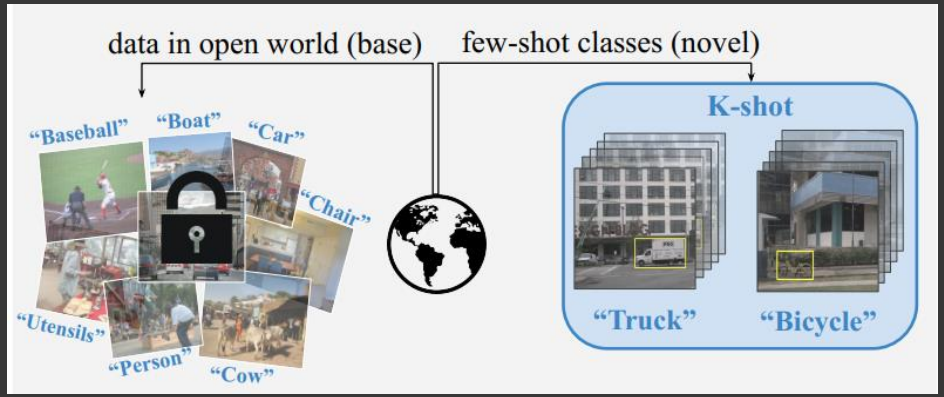


Madan et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

# How about exploit the open world for (automating) data labeling?




Realistically embracing the open world, leveraging FMs to learn from few-shot visuals and texts



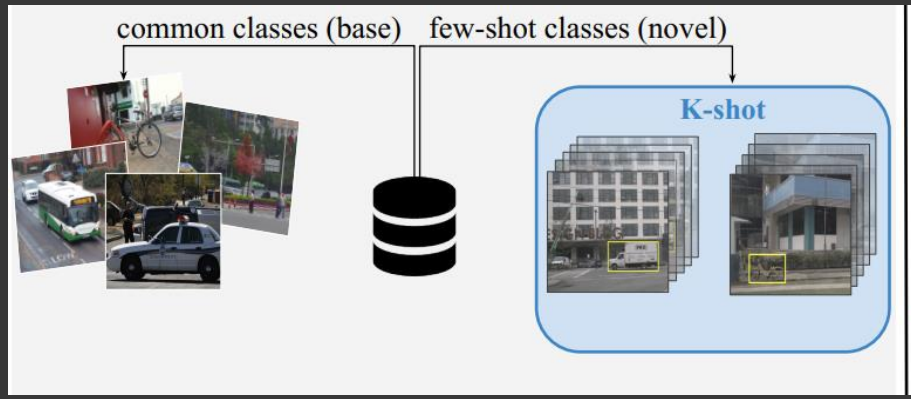

multimodal few-shot learning



Madan et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

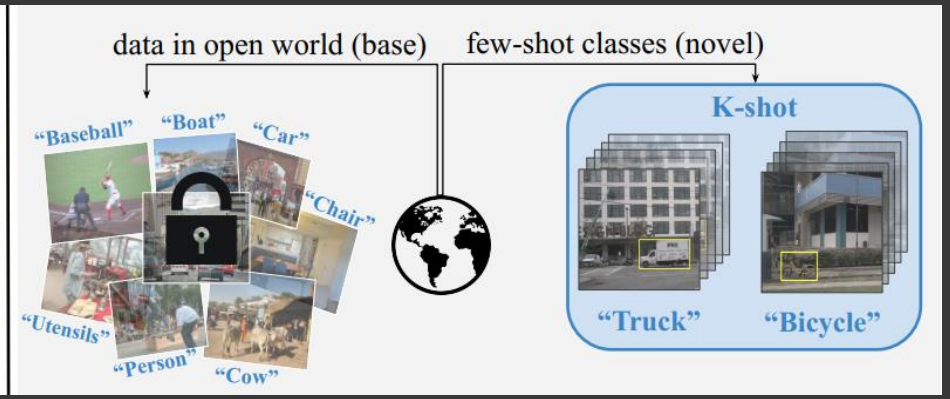
# Multimodal Few-Shot Learning

e.g., artificially splitting 80 classes of COCO into base set (60 classes) and novel set (20 classes)




Existing few-shot learning setup

Realistically embracing the open world, leveraging FMs to learn from few-shot visuals and texts




multimodal few-shot learning



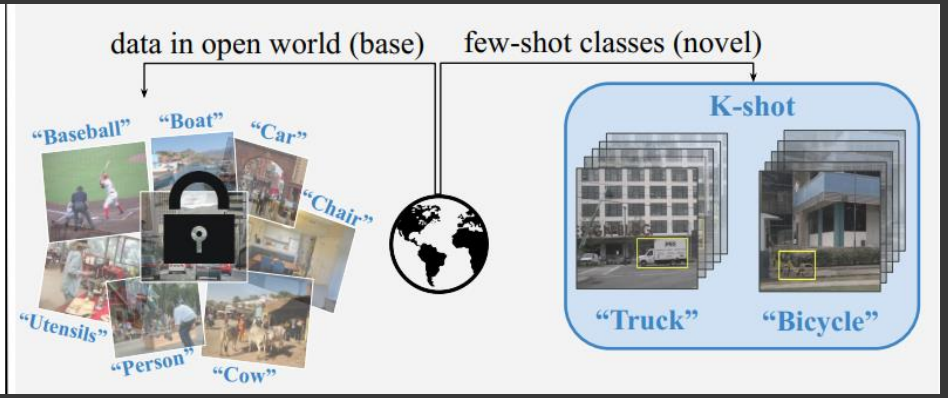
Madan et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

# Multimodal Few-Shot Learning

Validating various methods, collecting effective approaches, summarizing useful techniques


16:20 – 16:55 at our workshop!

Realistically embracing the open world, leveraging FMs to learn from few-shot visuals and texts



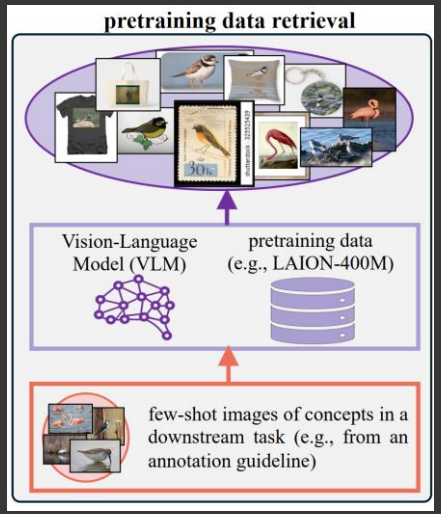

multimodal few-shot learning

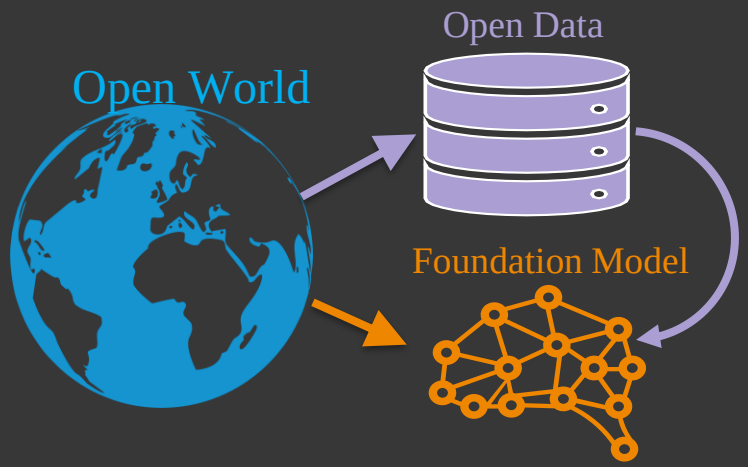


Madan et al., "Revisiting Few-Shot Object Detection with Vision-Language Models", NeurIPS, 2024

# Augment Few-Shot Data with Open Data

Retrieval-based Data Augmentation: retrieval task-relevant data from open data, e.g., open-source FMs' pretraining data



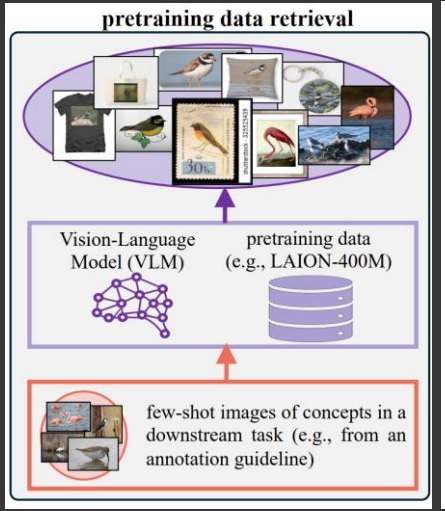


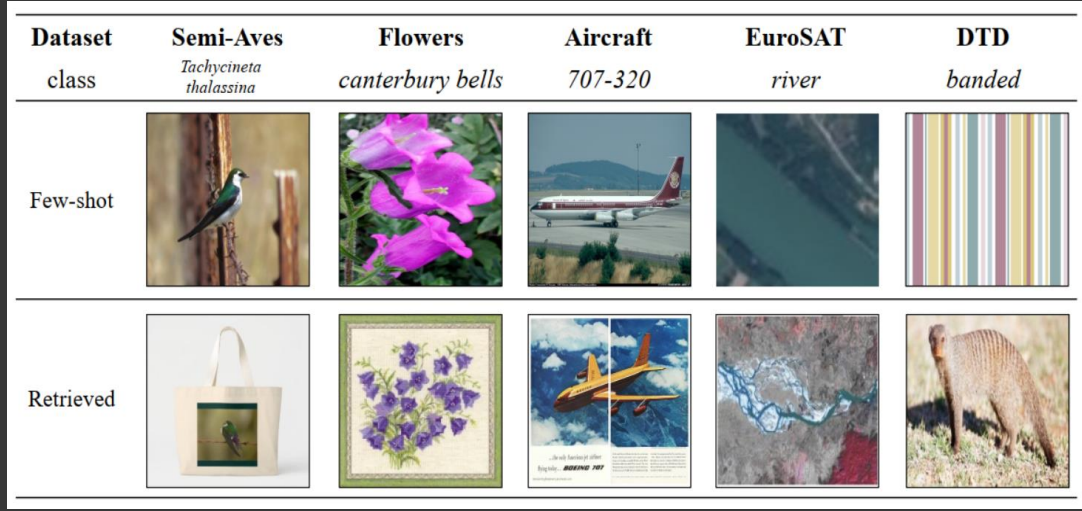
Liu, et al., "Few-Shot Recognition via Stage-Wise Retrieval-Augmented Finetuning", CVPR 2025 --- ExHall D Poster #425, 10:30 am CDT, 14 June

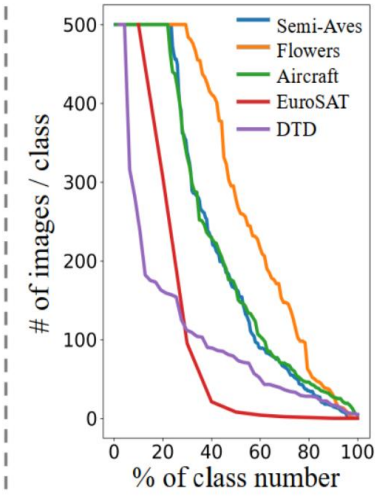
# Augment Few-Shot Data with Open Data

Retrieval-based Data Augmentation: retrieval task-relevant data from open data, e.g., open-source FMs' pretraining data

Despite being intuitive, it has two challenges: (1) domain gaps, and (2) imbalanced distribution



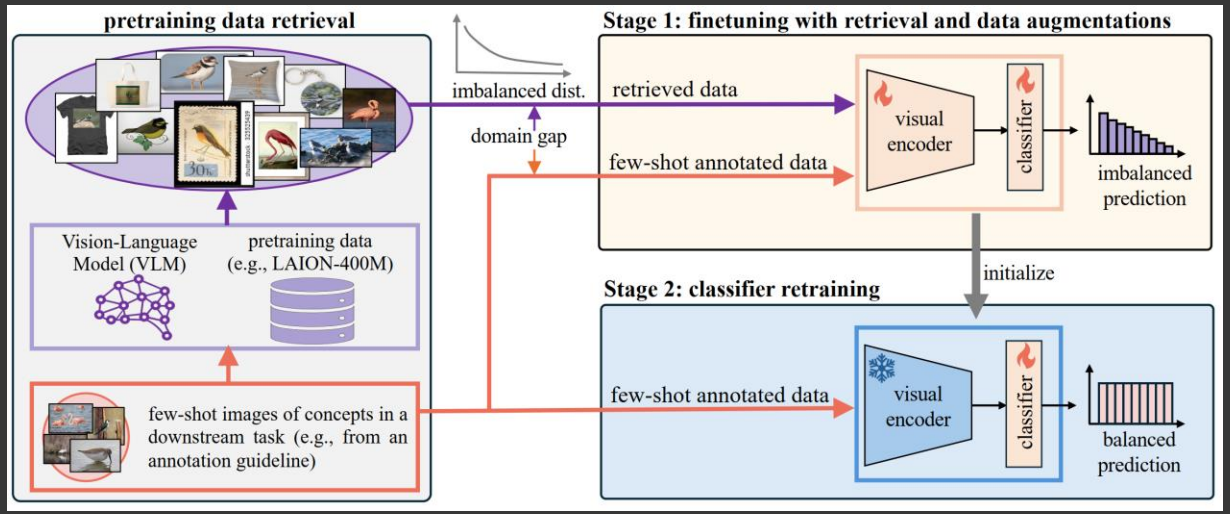

Liu, et al., "Few-Shot Recognition via Stage-Wise Retrieval-Augmented Finetuning", CVPR 2025 --- ExHall D Poster #425, 10:30 am CDT, 14 June
Wang, et al., "Robust Few-Shot Vision-Language Model Adaptation", arXiv:2506.04713, 2025

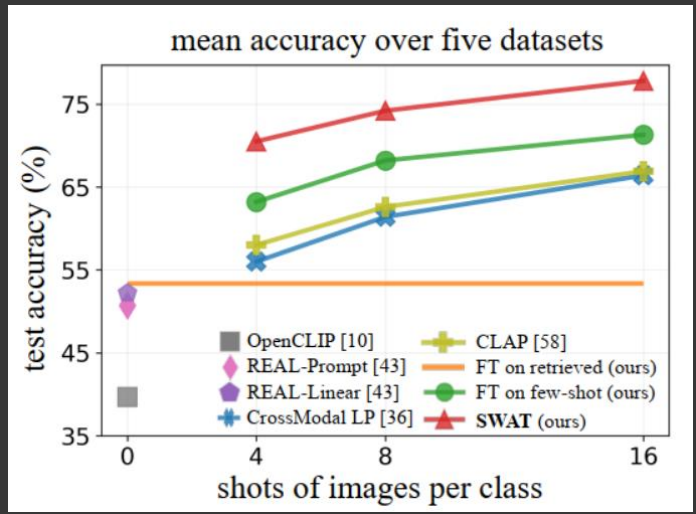
# Augment Few-Shot Data with Open Data

Retrieval-based Data Augmentation: retrieval task-relevant data from open data, e.g., open-source FMs' pretraining data

Despite being intuitive, it has two challenges: (1) domain gaps, and (2) imbalanced distribution

Stage-wise finetuning as a simple yet effective method, related to transfer learning and long-tailed learning.




Liu, et al., "Few-Shot Recognition via Stage-Wise Retrieval-Augmented Finetuning", CVPR 2025 --- ExHall D Poster #425, 10:30 am CDT, 14 June
Wang, et al., "Robust Few-Shot Vision-Language Model Adaptation", arXiv:2506.04713, 2025
Parashar, et al., "The Neglected Tails of Vision-Language Models", CVPR, 2024

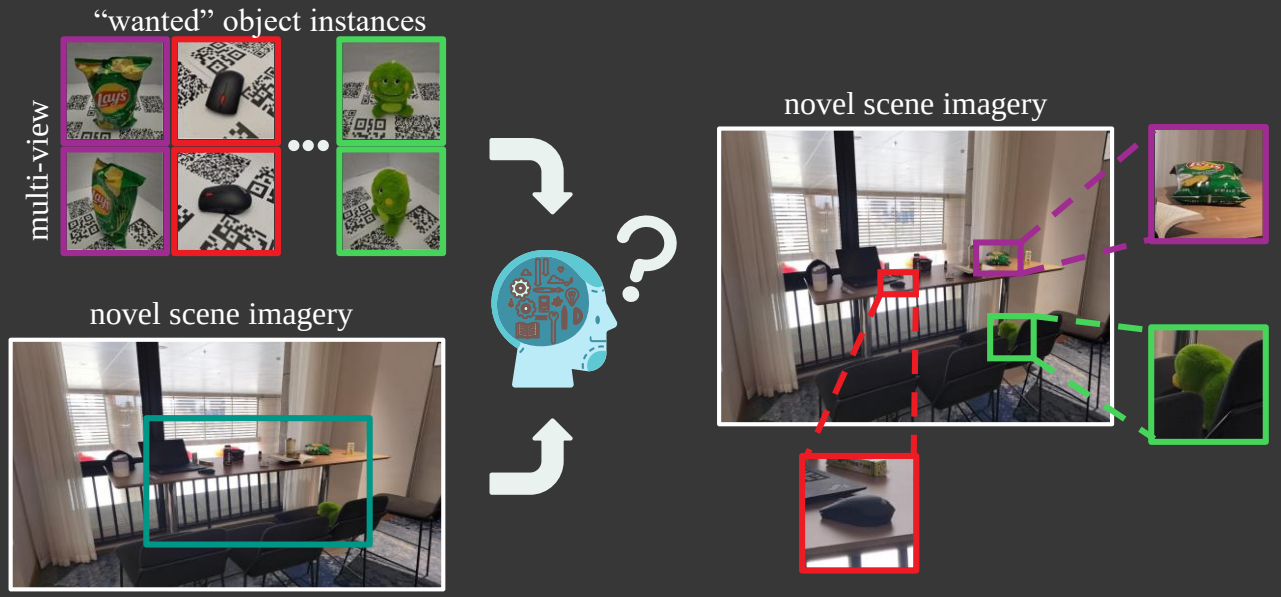
# Exploit Open Data to Solve Instance Detection (InsDet)

- InsDet aims to localize the “wanted” object in distance.
- It is a prerequisite step in many applications such as robotics and AR/VR.


Shen, et al., “Solving Instance Detection from an Open-World Perspective”, CVPR 2025 --- ExHall D Poster #431, 4pm CDT, 13 June
Zhao, et al., “Instance Tracking in 3D Scenes from Egocentric Videos”, CVPR 2024

# Exploit Open Data to Solve Instance Detection (InsDet)

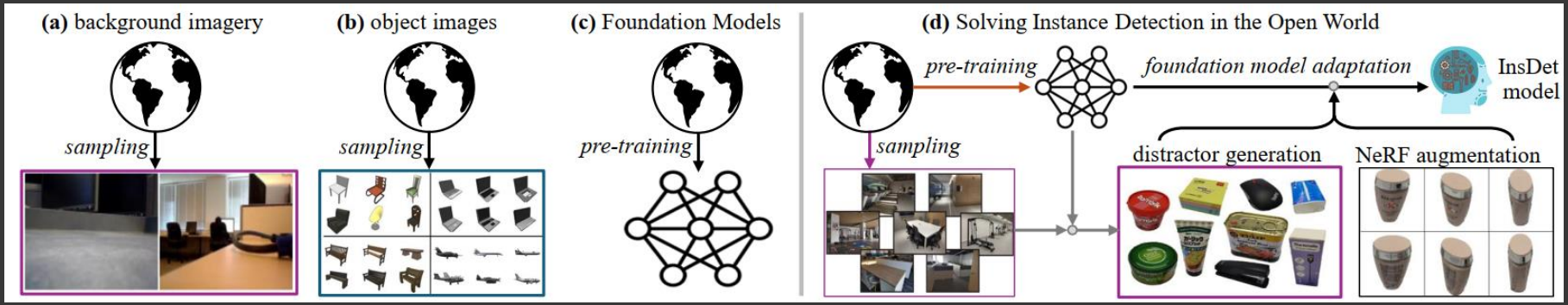
- Open-set testing imagery is never-before-seen and hence unknown to an instance detector.
- Domain gaps exist between visual references and instance proposals (due to occlusions, lighting variations, etc.).
- Robustness and generalization are desperately needed to detect diverse instances.




Shen, et al., "Solving Instance Detection from an Open-World Perspective", CVPR 2025 --- ExHall D Poster #431, 4pm CDT, 13 June
Zhao, et al., "Instance Tracking in 3D Scenes from Egocentric Videos", CVPR 2024
Shen, et al., "A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture", NeurIPS 2024

# Exploit Open Data to Solve Instance Detection (InsDet)

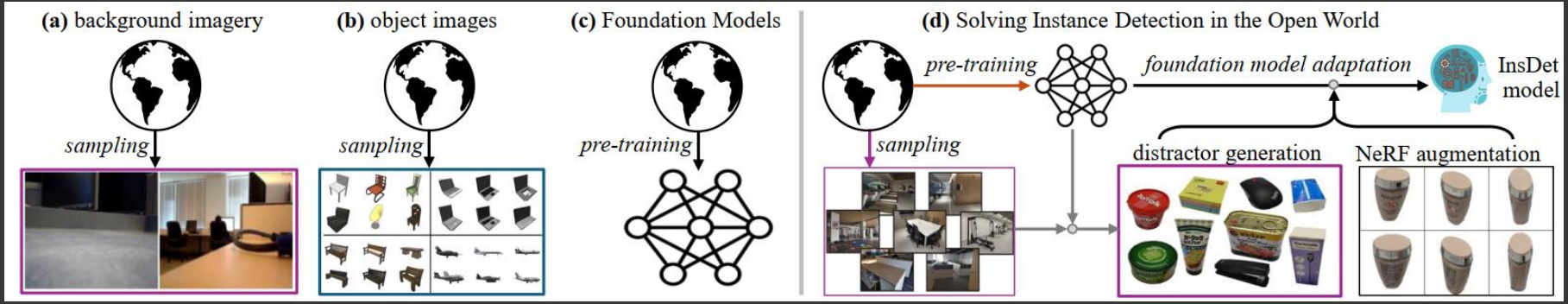
- Previous methods also exploit the open world, but insufficiently.
- Why not make more use of the open world?

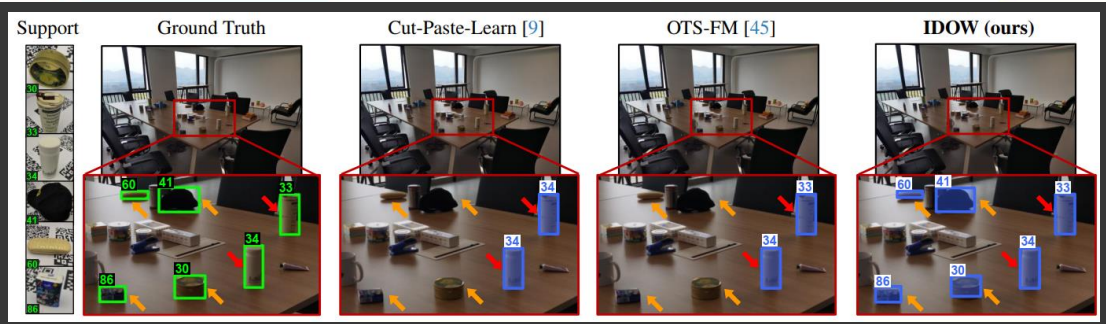



Shen, et al., "Solving Instance Detection from an Open-World Perspective", CVPR 2025 --- ExHall D Poster #431, 4pm CDT, 13 June
Shen, et al., "A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture", NeurIPS 2024
Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017
Li et al. "VoxDet: Voxel Learning for Novel Instance Detection", NeurIPS, 2023

# Exploit Open Data to Solve Instance Detection (InsDet)

- Previous methods also exploit the open world, but insufficiently.
- Why not make more use of the open world?



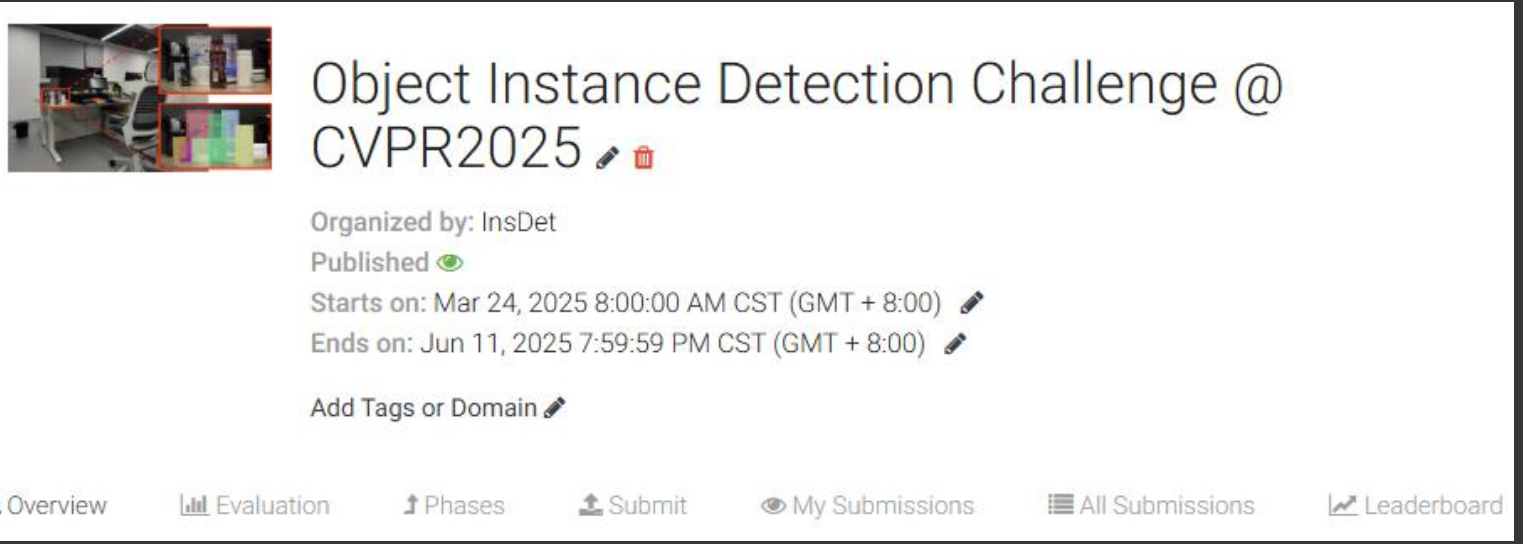




Shen, et al., "Solving Instance Detection from an Open-World Perspective", CVPR 2025 --- ExHall D Poster #431, 4pm CDT, 13 June
Shen, et al., "A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture", NeurIPS 2024
Dwibed & Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection", ICCV, 2017
Li et al. "VoxDet: Voxel Learning for Novel Instance Detection", NeurIPS, 2023

# Exploit Open Data to Solve Instance Detection (InsDet)

- Previous methods also exploit the open world, but insufficiently.
- Why not make more use of the open world?

15:45 - 16:20 at our workshop!




Shen, et al., “Solving Instance Detection from an Open-World Perspective”, CVPR 2025 --- ExHall D Poster #431, 4pm CDT, 13 June
Shen, et al., “A High-Resolution Dataset for Instance Detection with Multi-View Instance Capture”, NeurIPS 2024
Dwibed & Hebert, “Cut, paste and learn: Surprisingly easy synthesis for instance detection”, ICCV, 2017
Li et al. “VoxDet: Voxel Learning for Novel Instance Detection”, NeurIPS, 2023

# Thank you for joining!

- Embrace the open world – the foundation models and open data!
- Watch out for misalignment between AI and experts (like you)!
- Challenges and opportunities are coupled in the open world!