

The Concept Misalignment between Experts and AI

from Data Labeling to Data Versioning

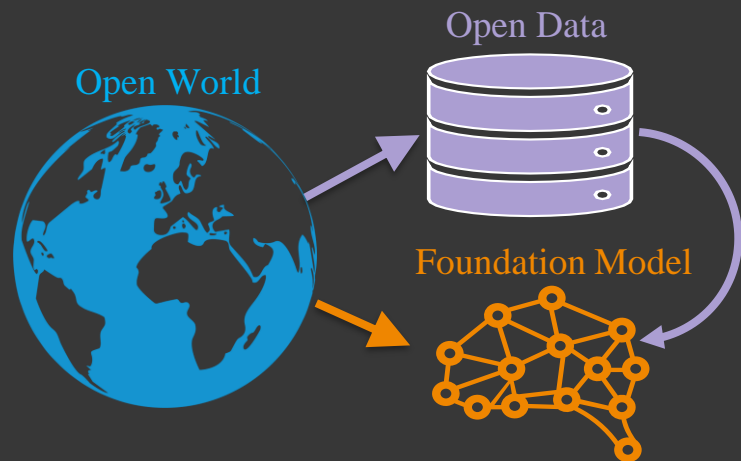


Shu Kong

University of Macau

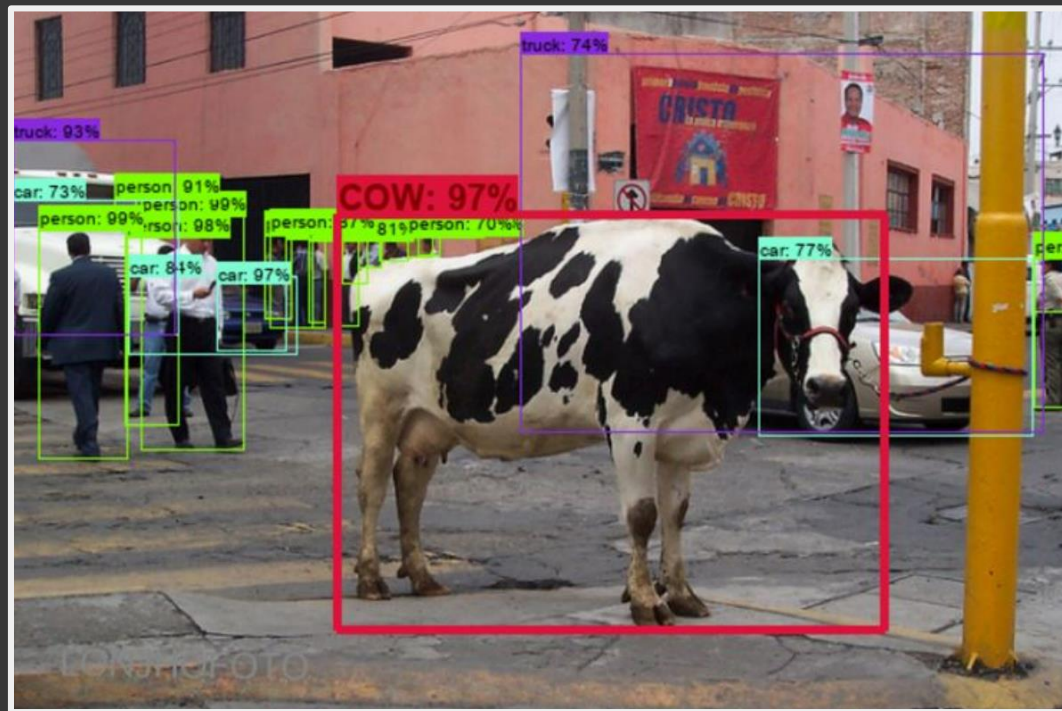
March 4, 2025

Representing the Open World: Foundation Model and Open Data



Foundation Model for Open-Vocabulary Detection

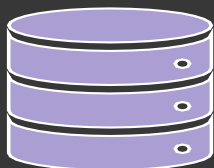
Cool!!!



Open World



Open Data

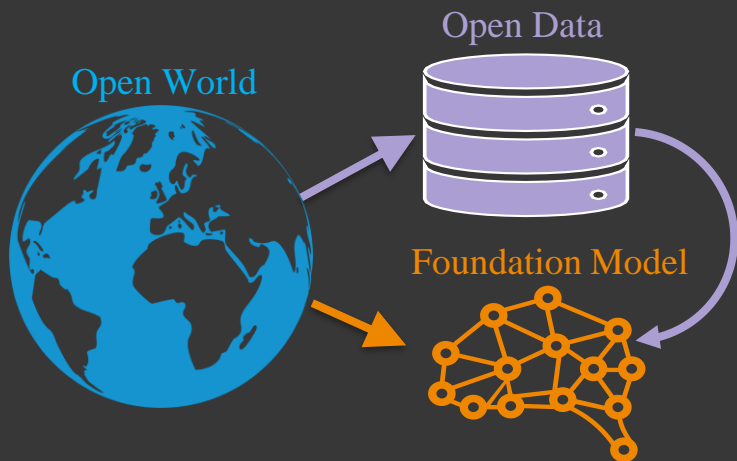


Foundation Model



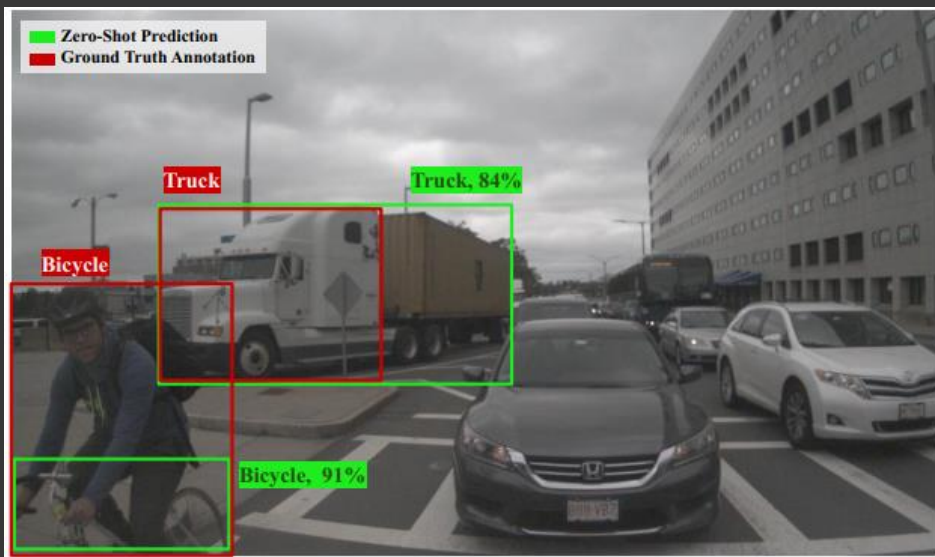
Foundation Model for Open-Vocabulary Detection

Cool!!!



Data Labeling where a foundation model struggles!

nuImages dataset

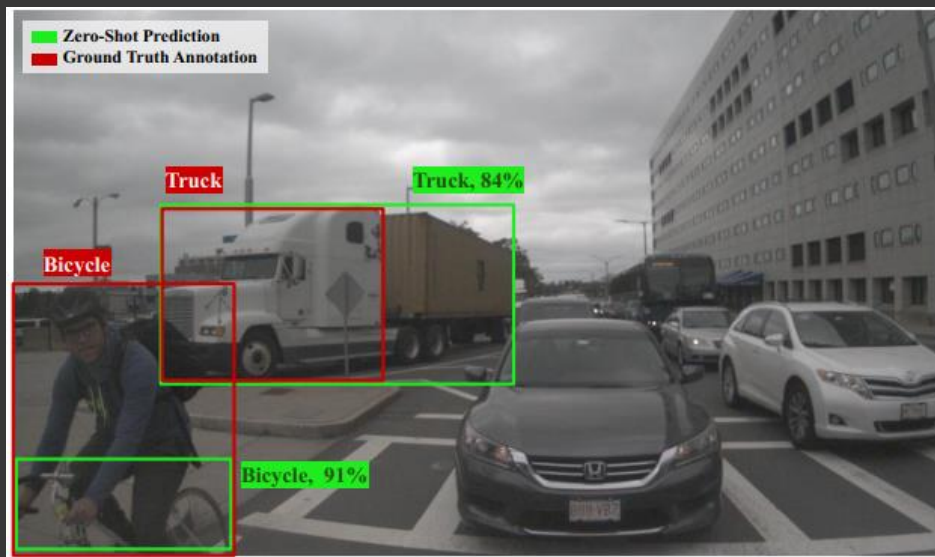


Poor alignments between foundational detector and ground-truth annotations in nuImages dataset.

Why?

Data Labeling where a foundation model struggles!

nuImages dataset



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

Why?

A snippet of annotation guidelines from nuImages

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

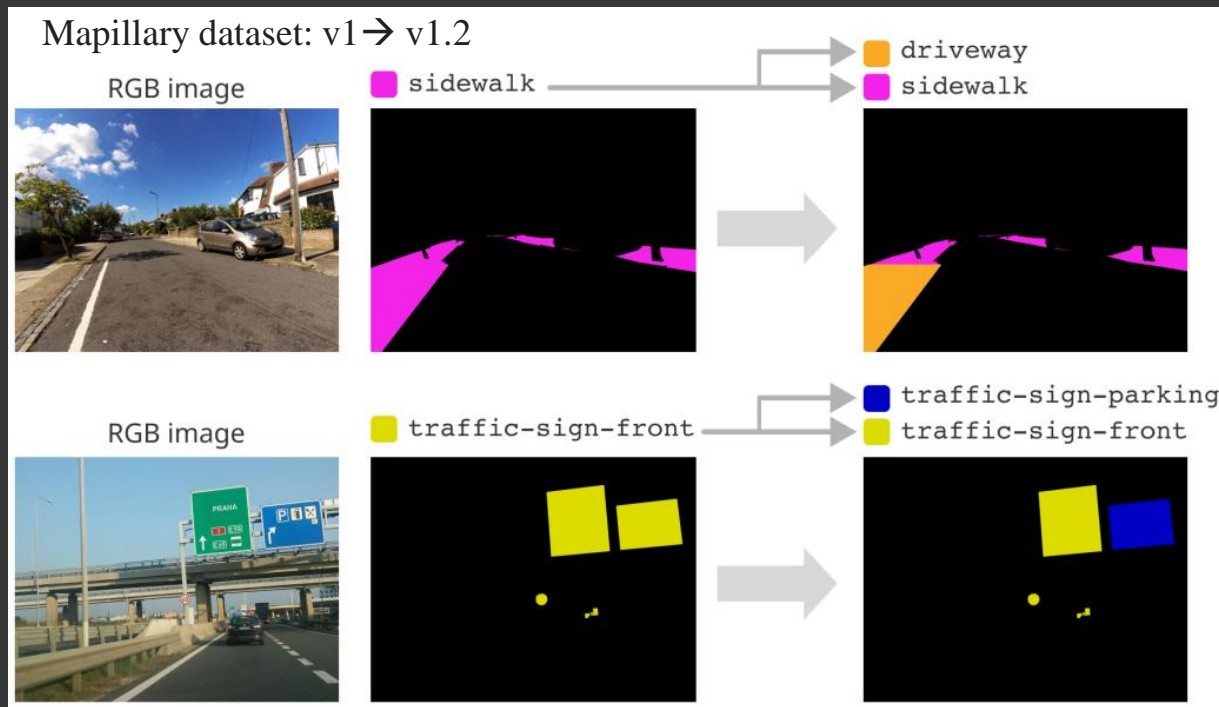
- Vehicles primarily designed to haul cargo including pick-ups, lorries, trucks and semi-tractors. Trailers hauled after a semi-tractor should be labeled as trailer.
- A pickup truck is a light duty truck with an enclosed cab and an open or closed cargo area.

Annotation instructions designed by autonomous driving experts.

Due to practical considerations!

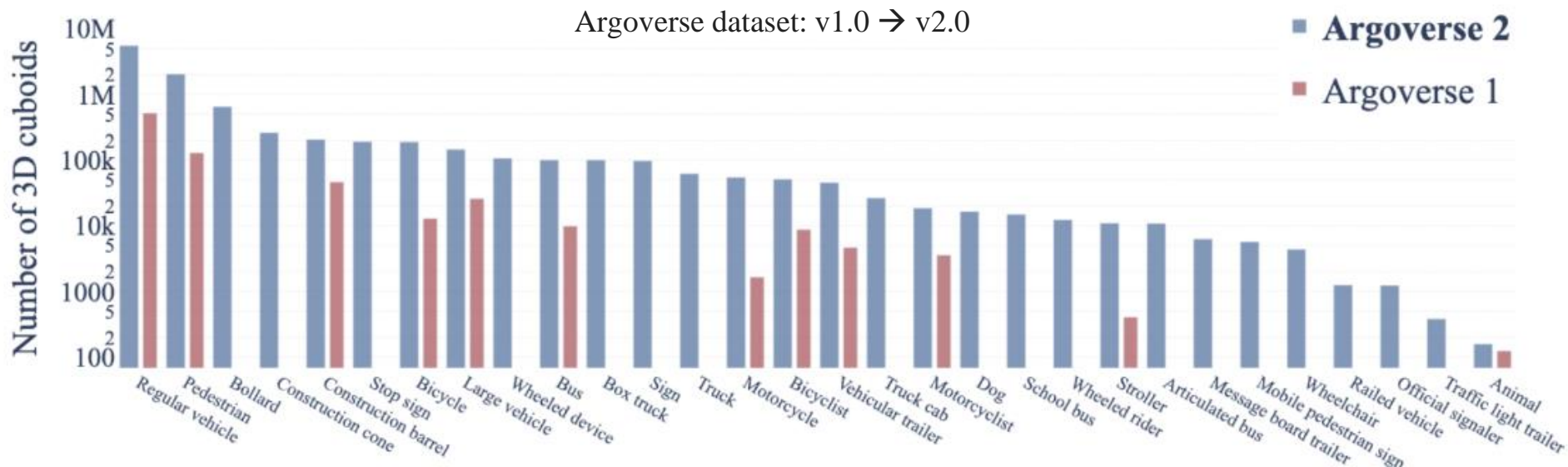
Data Versioning where a foundation model can continue to struggle!

Class ontologies evolve over time to meet needs in the open world.



Data Versioning where a foundation model can continue to struggle!

Class ontologies evolve over time to meet needs in the open world.



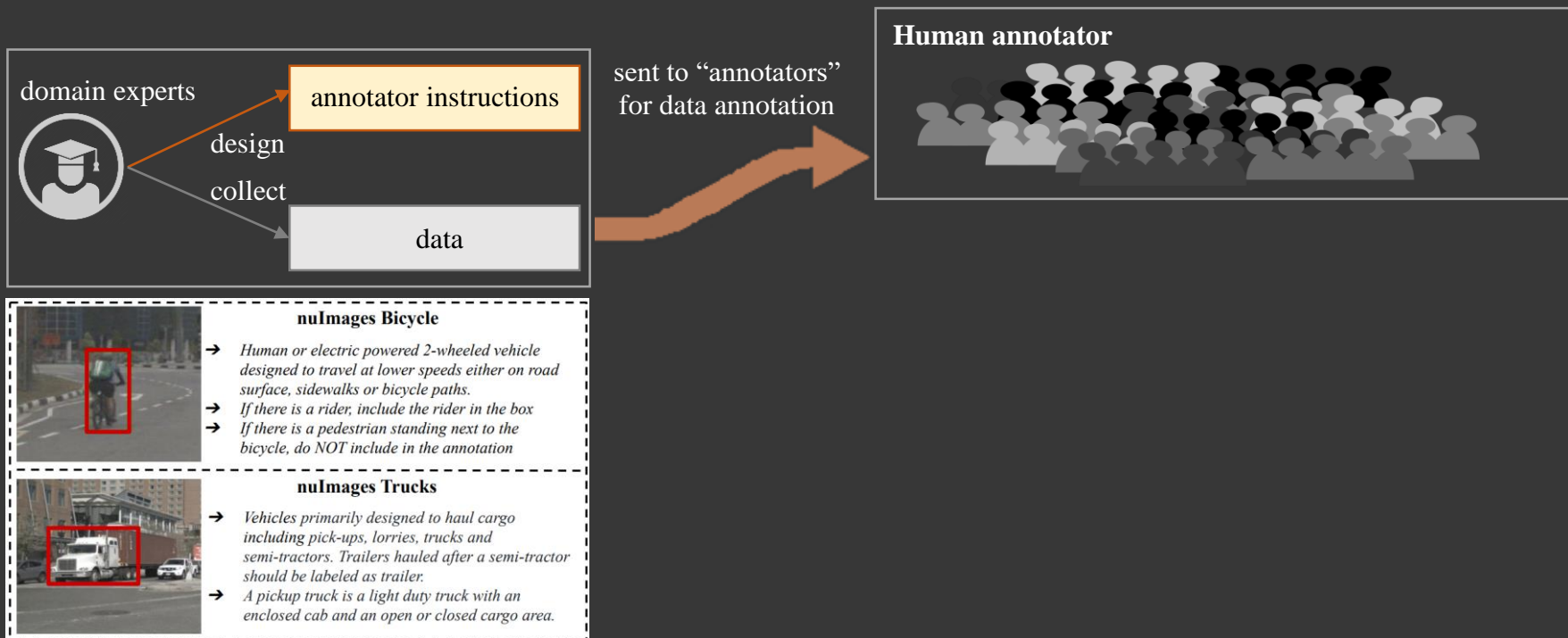
Chang, et al. "Argoverse: 3d tracking and forecasting with rich maps." CVPR, 2019.

Wilson, et al. "Argoverse 2: Next generation datasets for self-driving perception and forecasting." NeurIPS, 2021.

Lin, et al. "Continual learning with evolving class ontologies", NeurIPS, 2022

Let's formulate an interesting problem!

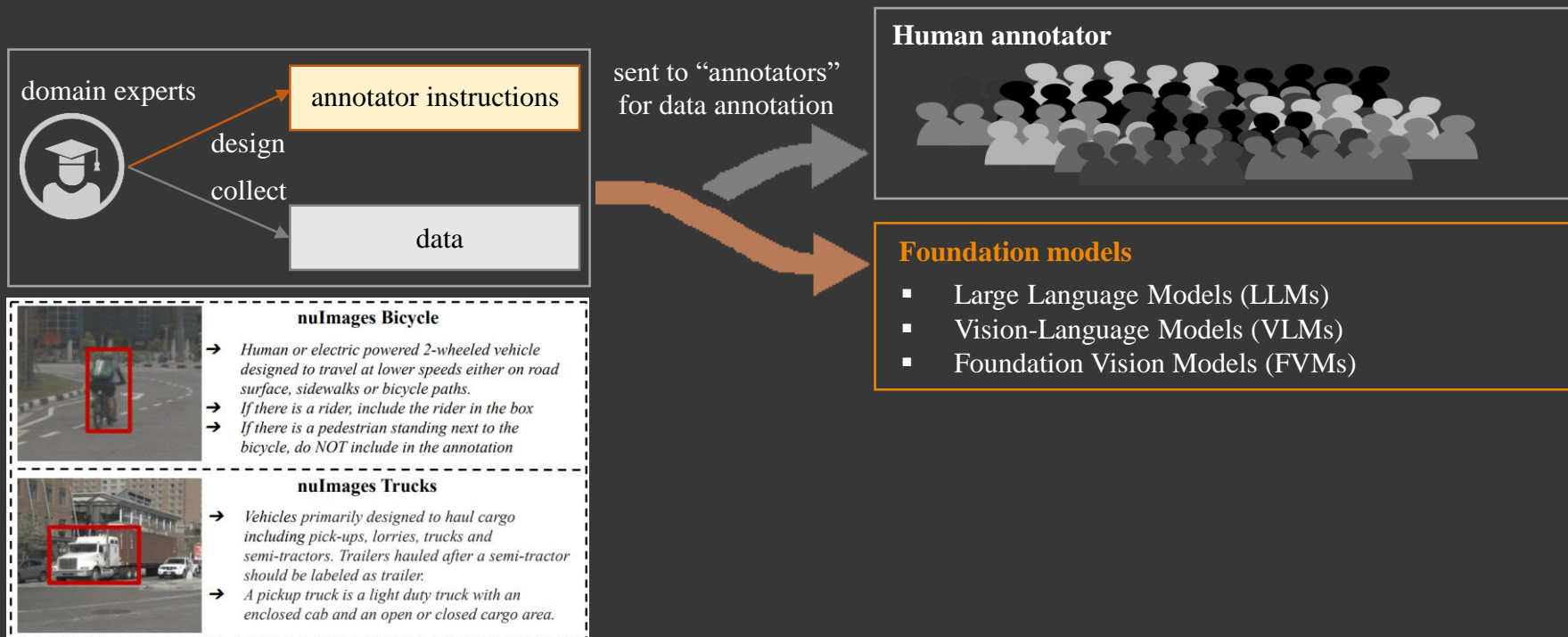
How to adapt foundation models to align with experts?



Let's formulate an interesting problem!

How to adapt foundation models to align with experts?

Can we replace human annotators with foundation models for data annotation?

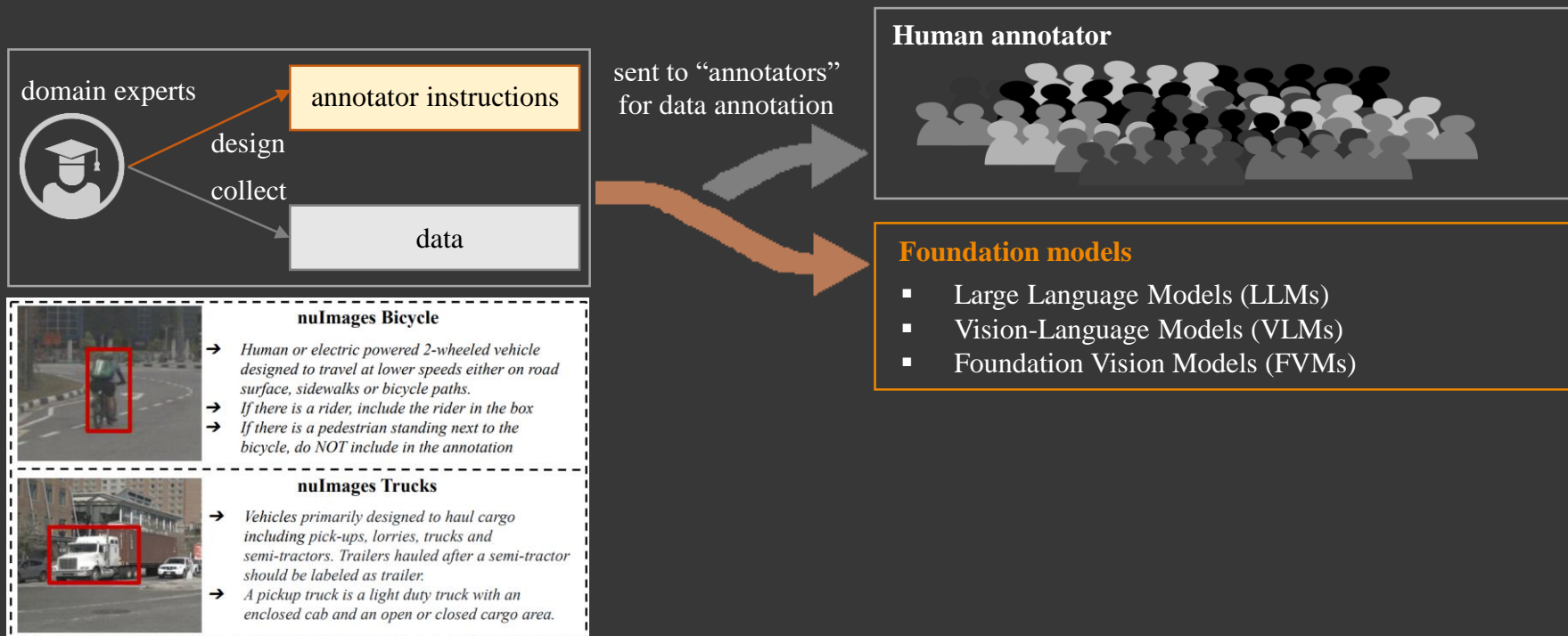


Let's formulate an interesting problem!

How to adapt foundation models to align with experts?

Can we replace human annotators with foundation models for data annotation?

Can we adapt foundation models w.r.t annotation guidelines?



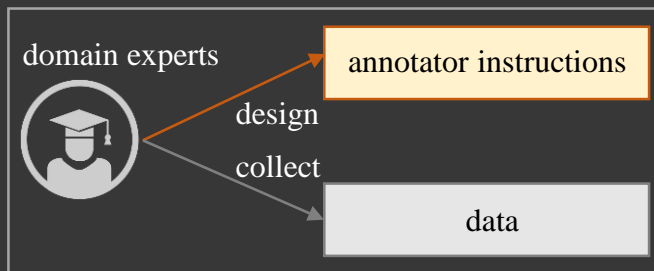
Let's formulate an interesting problem!

How to adapt foundation models to align with experts?


Can we replace human annotators with foundation models for data annotation?

Can we adapt foundation models w.r.t annotation guidelines?

Technically, this is a **multimodal few-shot learning** problem.




nulImages Bicycle



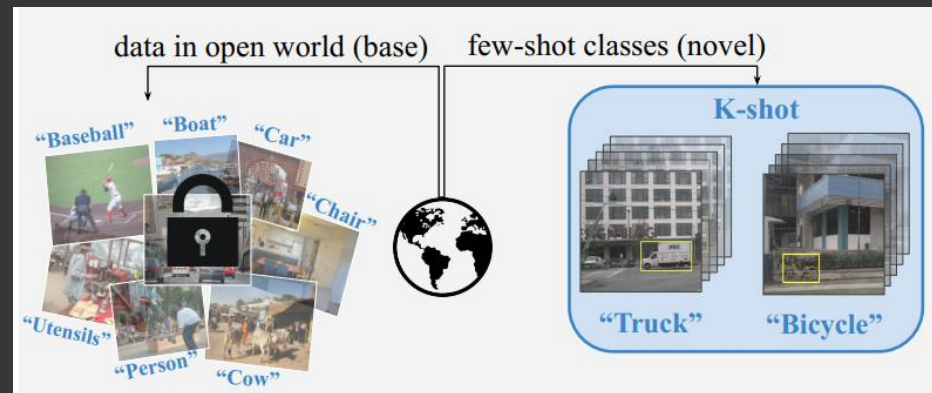
- Human or electric powered 2-wheeled vehicle designed to travel at lower speeds either on road surface, sidewalks or bicycle paths.
- If there is a rider, include the rider in the box
- If there is a pedestrian standing next to the bicycle, do NOT include in the annotation

nulImages 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.

The proposed **multimodal few-shot learning** setup



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

Let's formulate an interesting problem!

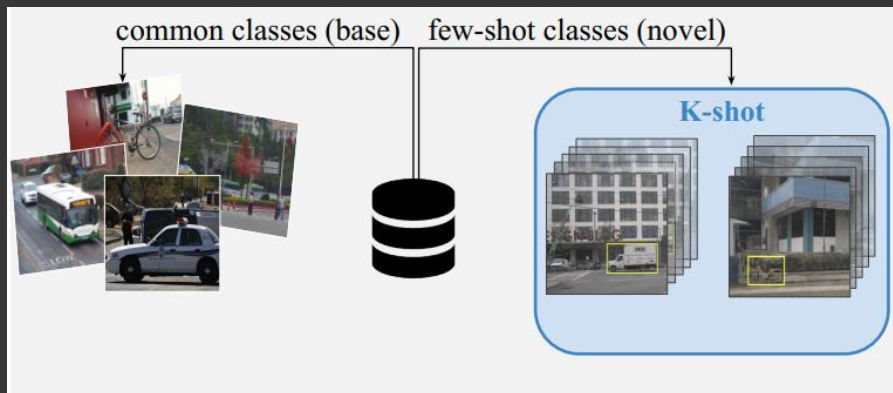
How to adapt foundation models to align with experts?

Can we replace human annotators with foundation models for data annotation?

Can we adapt foundation models w.r.t annotation guidelines?

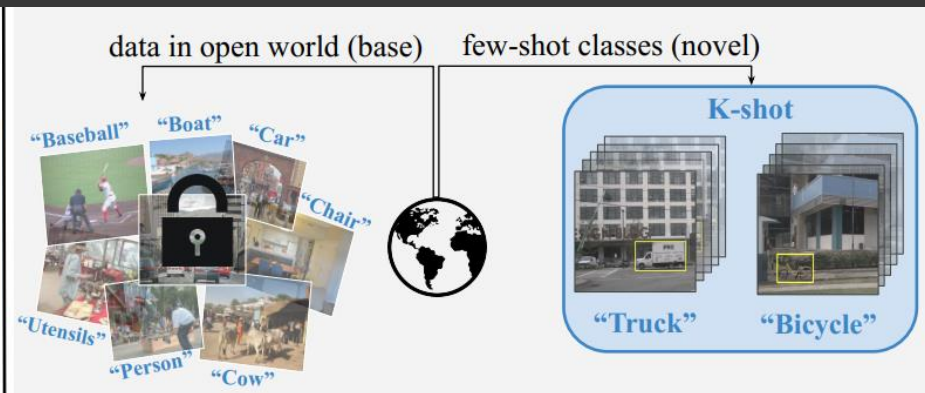
Technically, this is a **multimodal few-shot learning** problem.

Existing few-shot learning setup



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

The proposed **multimodal few-shot learning** setup



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

Multimodal few-shot learning

How to adapt foundation models to align with experts?

Can we replace human annotators with foundation models for data annotation?

Can we adapt foundation models w.r.t annotation guidelines?

Technically, this is a **multimodal few-shot learning** problem.

Challenge at CVPR'24 and CVPR'25



Foundational Few-Shot Object Detection Challenge

Organized by: foundational_fsod

Starts on: Apr 11, 2024 8:00:00 AM CST (GMT + 8:00)

Ends on: Jun 8, 2024 7:59:59 AM CST (GMT + 8:00)

Multi-Modal

VLMS

FSOD

Detection

Computer Vision

Overview

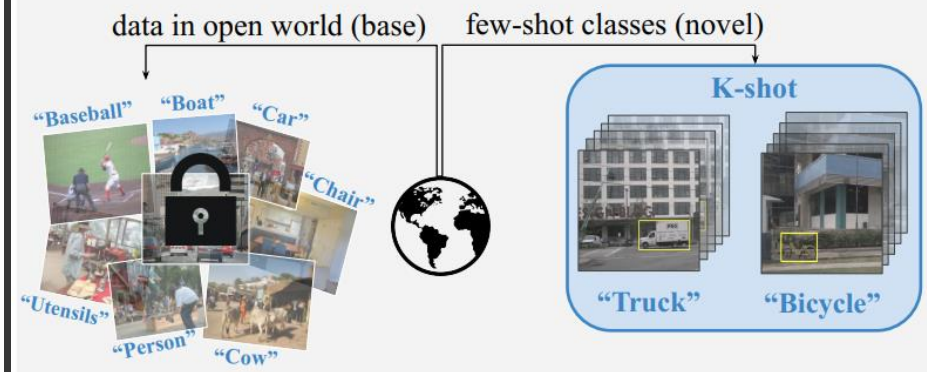
Evaluation

Phases

Participate

Leaderboard

The proposed **multimodal few-shot learning** setup



Validating various methods, collecting effective approaches, summarizing useful techniques

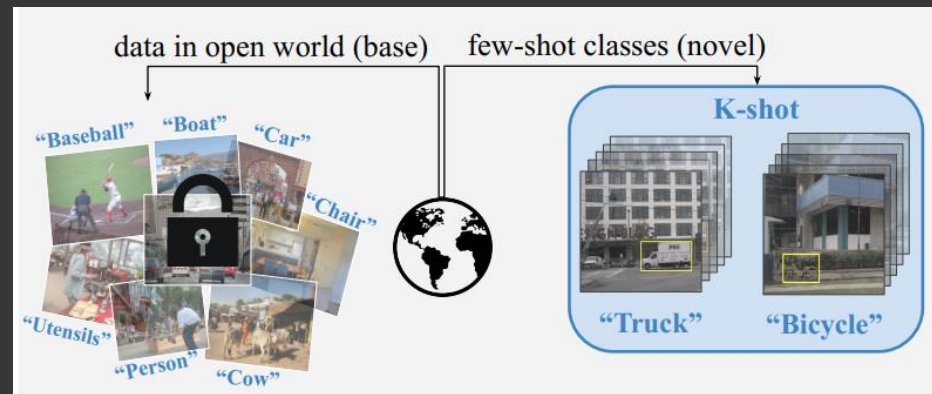
Realistically embracing the open world, leveraging foundation models to learn from few-shot visuals and texts

Approaches

Embracing the open world, esp. foundation models, we compare various approaches:

1. Prompt engineering
2. Standard finetuning
3. Language prompt tuning
4. Visual prompting
5. Multimodal prompting
6. Multimodal chat assistants

The proposed **multimodal few-shot learning** setup



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

Approaches

Embracing the open world, esp. foundation models, we compare various approaches:

1. Prompt engineering
2. Standard finetuning
3. Language prompt tuning
4. Visual prompting
5. Multimodal prompting
6. **Multimodal chat assistants**

Approach	Backbone	Pre-Train Data	Average Precision (AP)			
			All	Many	Med	Few
Zero-Shot Detection						
RegionCLIP [64]	RN50	CC3M	2.50	3.20	3.80	0.40
Detic [67]	SWIN-B	LVIS, COCO, IN-21K	14.40	25.83	16.59	2.32
GroundingDINO [33]	SWIN-T	Objects365, GoldG, Cap4M	12.05	17.29	15.45	3.72
GLIP [30]	SWIN-L	FourODs, GoldG, Cap24M	17.01	23.36	19.86	8.40
MQ-GLIP-Text [59]	SWIN-L	Objects365, FourODs, GoldG, Cap24M	17.01	23.36	19.85	8.41
Prompt Engineering						
Detic [67]	SWIN-B	LVIS, COCO, IN-21K	14.92	26.48	17.29	2.53
GLIP [30]	SWIN-L	FourODs, GoldG, Cap24M	17.15	23.82	19.36	9.02
Standard Fine-Tuning						
RegionCLIP [64]	RN50	CC3M	3.86	6.08	5.13	0.54
Detic [67]	SWIN-B	LVIS, COCO, IN-21K	16.09	25.46	20	3.73
Federated Fine-Tuning (Ours)						
Detic [67]	SWIN-B	LVIS, COCO, IN-21K	17.24	28.07	20.71	4.18
Detic [67] w/ Prompt Engineering	SWIN-B	LVIS, COCO, IN-21K	17.71	28.46	21.14	4.75
Language Prompt Tuning						
GLIP [30]	SWIN-L	FourODs, GoldG, Cap24M	19.41	22.18	25.16	10.39
Visual Prompting						
MQ-GLIP-Image [59]	SWIN-L	Objects365, FourODs, GoldG, Cap24M	14.07	24.39	15.89	3.34
Multi-Modal Prompting						
MQ-GLIP [59]	SWIN-L	Objects365, FourODs, GoldG, Cap24M	21.42	32.19	23.29	10.26
Multi-Modal Chat Assistants						
GPT-4o Zero-Shot Classification [1]	<i>Private</i>	<i>Private</i>	9.95	16.81	12.11	1.71
MQ-GLIP Iterative Prompting	<i>Private</i>	<i>Private</i>	22.03	33.42	24.72	9.41
CVPR 2024 Competition Results						
NJUST KMG	SWIN-L	Objects365V2, OpenImageV6, GoldG, V3Det, COCO2014, COCO2017, LVISV1, GRIT, RefCOCO, RefCOCO+, RefCOCOg, gRef-COCO	32.56	50.21	34.87	15.16
zjyd_cxy_vision	SWIN-L	Objects365V2, COCO2017, LVIS, GoldG, VG, OpenImagesV6, V3Det, PhraseCut, RefCOCO, RefCOCO+, RefCOCOg, gRef-COCO	31.57	46.59	33.32	17.03

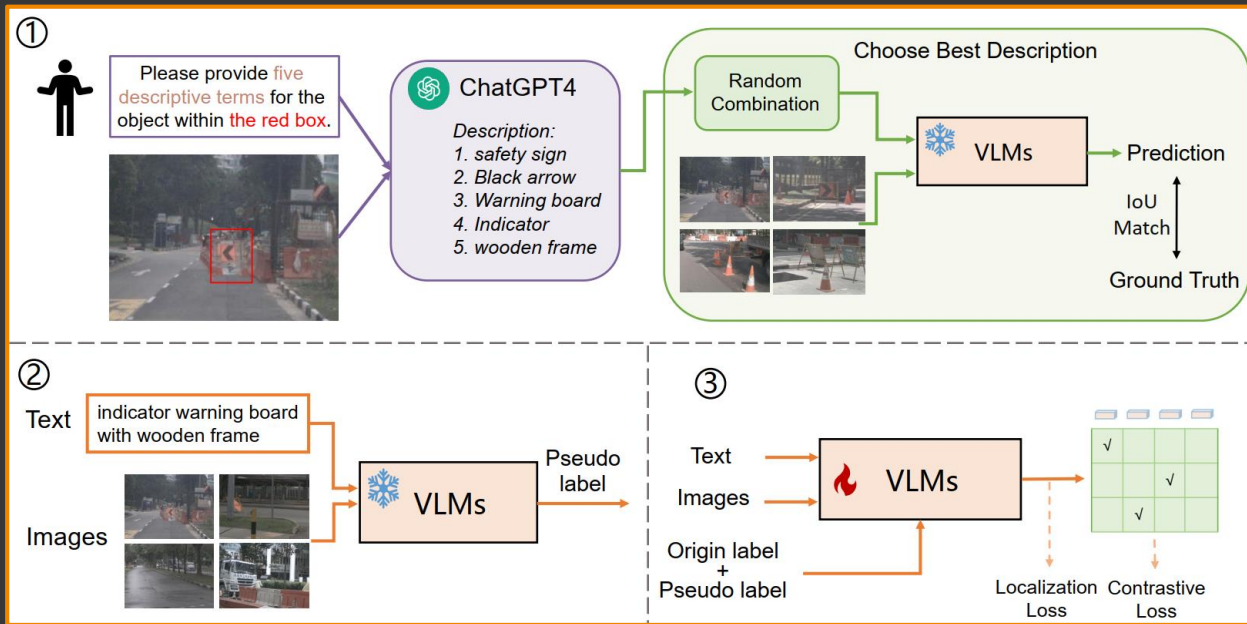
Approaches



It's cool to embrace foundation models!

Embracing the open world, esp. foundation models, we compare various approaches:

1. Prompt engineering
2. Standard finetuning
3. Language prompt tuning
4. Visual prompting
5. Multimodal prompting
6. **Multimodal chat assistants**



CVPR 2024 Competition Results

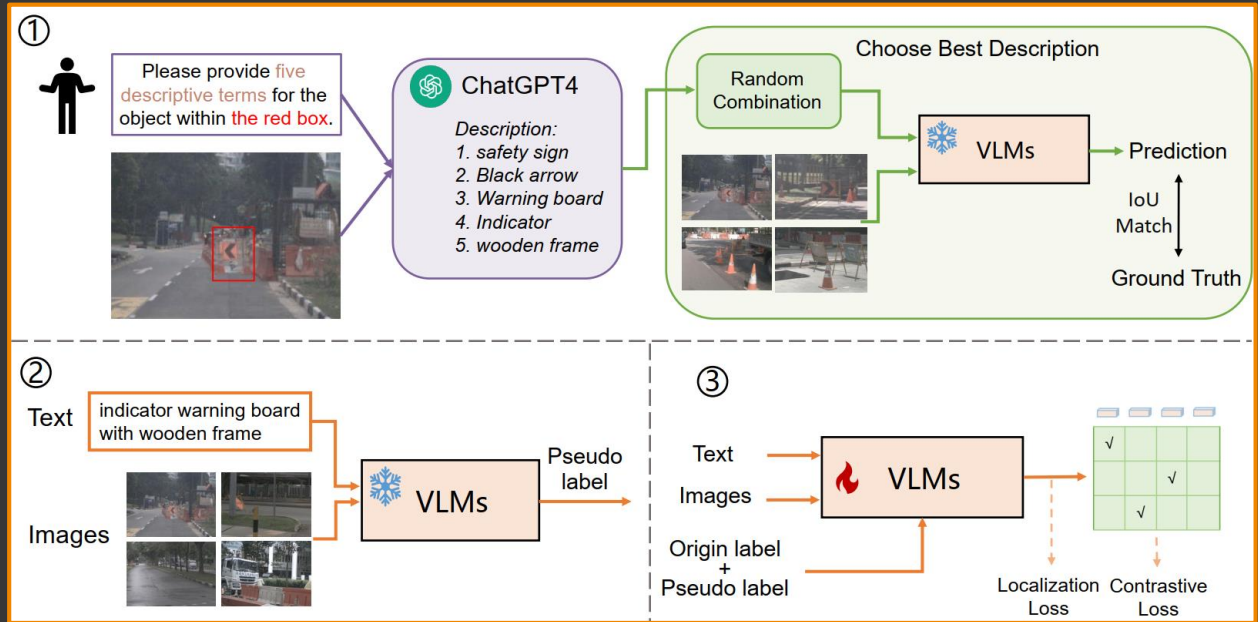
NJUST KMG	SWIN-L	Objects365V2, OpenImageV6, GoldG, V3Det, COCO2014, COCO2017, LVISV1, GRIT, RefCOCO, RefCOCO+, RefCOCOg, gRef-COCO	32.56	50.21	34.87	15.16
-----------	--------	---	-------	-------	-------	-------

Approaches

 It's cool to embrace foundation models! Really?

Embracing the open world, esp. foundation models, we compare various approaches:

1. Prompt engineering
2. Standard finetuning
3. Language prompt tuning
4. Visual prompting
5. Multimodal prompting
6. **Multimodal chat assistants**




CVPR 2024 Competition Results

NJUST KMG	SWIN-L	Objects365V2, OpenImageV6, GoldG, V3Det, COCO2014, COCO2017, LVISV1, GRIT, RefCOCO, RefCOCO+, RefCOCOg, gRef-COCO	32.56	50.21	34.87	15.16
-----------	--------	---	-------	-------	-------	-------



Chat assistant can fail too!

60	59: 'vine snake',	ImageNet
61	60: 'night snake, Hypsiglena torquata',	
62	61: 'boa constrictor, Constrictor constrictor',	
63	62: 'rock python, rock snake, Python sebae',	

GPT-4 misclassifies “night snake” as “European Adder”



What is the animal in this image?

 *European Adder* 

Chat assistant can fail too!

```
60 59: 'vine snake',
61 60: 'night snake, Hypsiglena torquata',
62 61: 'boa constrictor, Constrictor constrictor',
63 62: 'rock python, rock snake, Python sebae',
```

ImageNet

GPT-4 misclassifies “*night snake*” as “*European Adder*”

Do



What is the animal in this image?



European Adder



A photo of *European Adder*

Chat assistant can fail too!

It has seen the whole internet data, right?



Concept: **night snake**

Definition: a small light brown or beige colored snake.

What is the species name of the animal in the photo?

GPT4-V: European adder ❌

LLaVA1.5: garter snake ❌

Generate a photo of a night snake

DALL-E 3: ❌

SD-XL: ❌

The neglected long tails in VLMs

Hypothesis: certain concepts are insufficiently presented in the open world.

“night snake” is one of rare concepts in the open world

Concept: night snake

Definition: a small light brown or beige colored snake.

What is the species name of the animal in the photo?

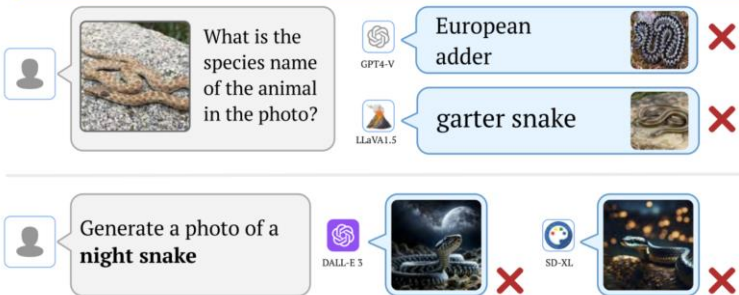
GPT4-V: European adder ❌

LLaVA1.5: garter snake ❌

Generate a photo of a night snake

DALL-E 3: ❌

SD-XL: ❌



The neglected long tails in VLMs












Hypothesis: certain concepts are insufficiently presented in the open world.

Justification: We *count* the occurrence of pretraining texts related to the concept of interest.

“night snake” is one of rare concepts in the open world

Concept: night snake

Definition: a small light brown or beige colored snake.

		What is the species name of the animal in the photo?	 European adder  ❌
			 garter snake  ❌
	Generate a photo of a night snake	  ❌	  ❌



The neglected long tails in VLMs

Hypothesis: certain concepts are insufficiently presented in the open world.

Justification: We *count* the occurrence of pretraining texts related to the concept of interest.

Challenge: billions of training examples (e.g., LAION-2B).



Measure concept frequency

Hypothesis: certain concepts are insufficiently presented in the open world.

Justification: We *count* the occurrence of pretraining texts related to the concept of interest.

Challenge: billions of training examples (e.g., LAION-2B). We use string matching!



Measure concept frequency

Hypothesis: certain concepts are insufficiently presented in the open world.

Justification: We *count* the occurrence of pretraining texts related to the concept of interest.

Challenge: billions of training examples (e.g., LAION-2B). We use string matching!

sneakers



Los Angeles Times
collectible Nike sneakers cost ...

running shoes



New York Magazine
The Best Running Shoes for Me...

trainer shoes



Adidas
adidas LA Trainer Shoes - ...

tiger



tiger shark



Tiger Woods



Lexical variation, e.g., synonyms

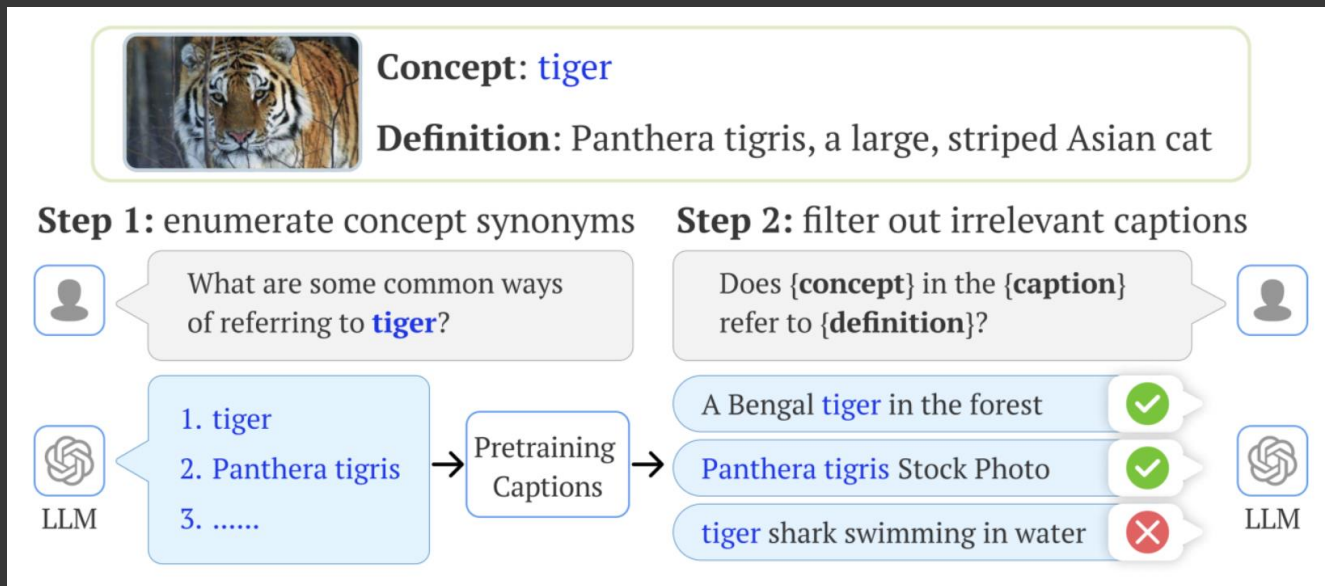
Linguistic ambiguity

Measure concept frequency

Hypothesis: certain concepts are insufficiently presented in the open world.

Justification: We *count* the occurrence of pretraining texts related to the concept of interest.

Challenge: billions of training examples (e.g., LAION-2B). We use string matching!



The neglected long tails in VLMs

Hypothesis: certain concepts are insufficiently presented in the open world.

Justification: We *count* the occurrence of pretraining texts related to the concept of interest.

Evidence: a strong correlation between *concept frequency* and per-concept accuracy.

“night snake” is one of rare concepts in the open world

Concept: night snake

Definition: a small light brown or beige colored snake.

What is the species name of the animal in the photo?

GPT4-V: European adder ❌

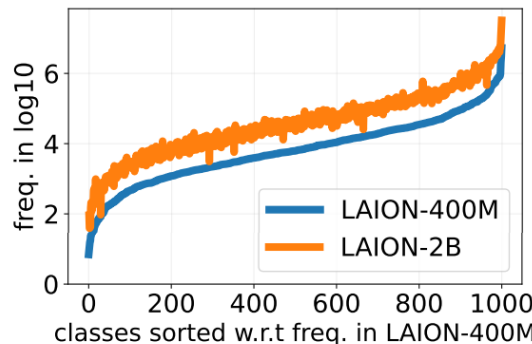
LLaVA1.5: garter snake ❌

Generate a photo of a night snake

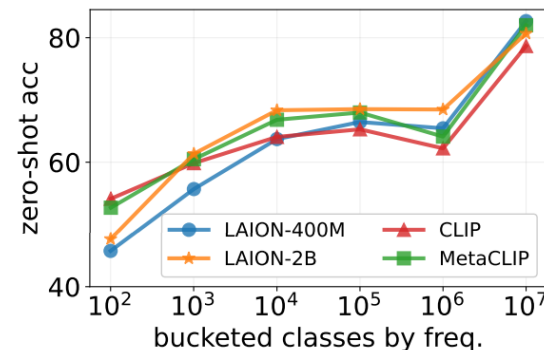
DALL-E 3: ❌

SD-XL: ❌

(a) freq. of ImageNet concepts

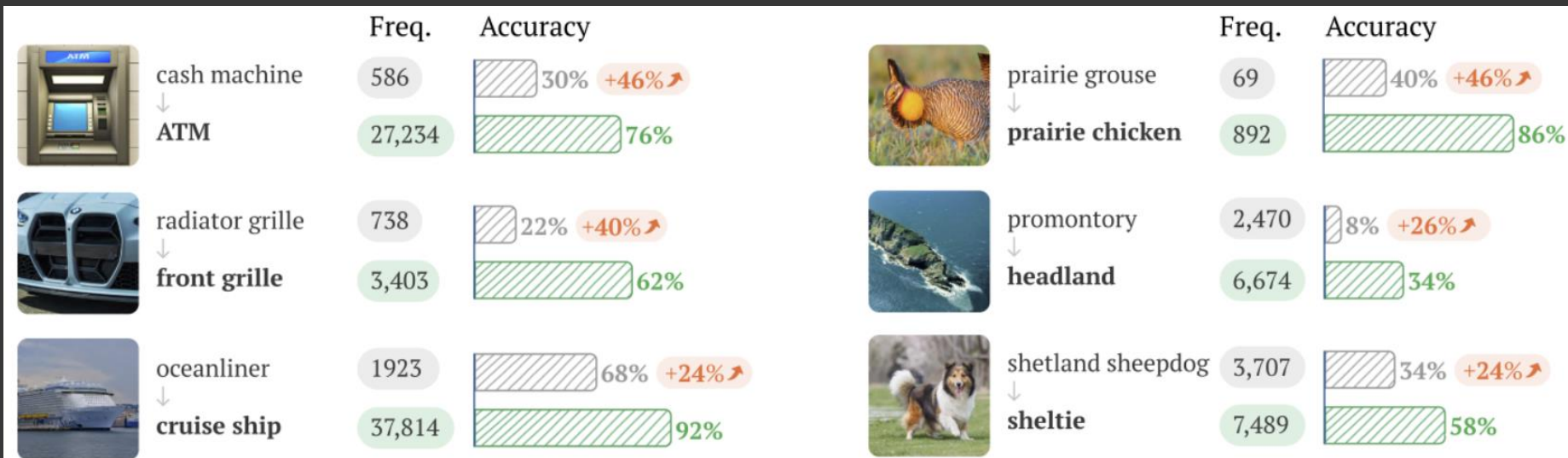


(b) freq. vs. zero-shot accuracy


















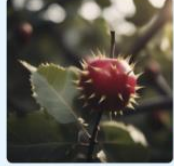


Insight 1: prompt VLM using the most frequent synonym

This simple change significantly boosts zero-shot accuracy!



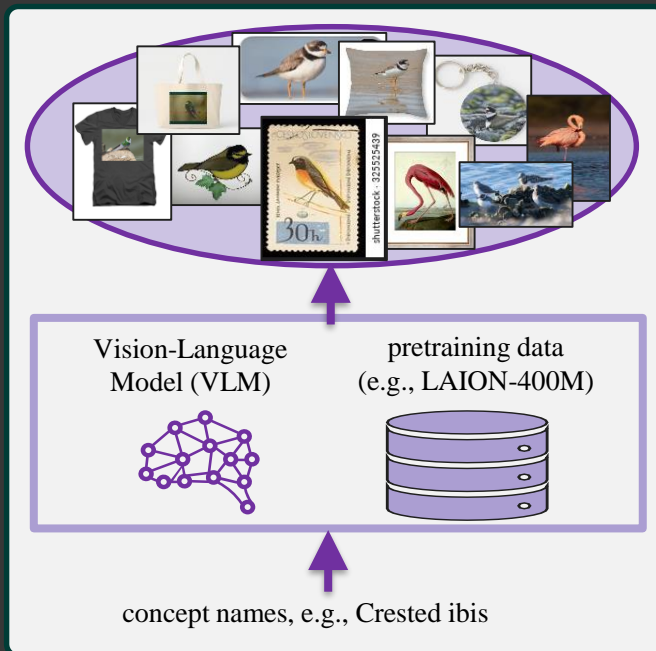
Insight 2: use the most frequent synonym in image generation

replace the original query with its most frequent synonym in prompts

<p>A photo of a bank swallow</p> 	<p>Generate a photo of a bank swallow.</p>  <p>DALL-E 3</p>  <p>✗</p>	<p>Generate a photo of a sand martin.</p>  <p>DALL-E 3</p>  <p>✓</p>	<p>Generate a photo of a bank swallow.</p>  <p>SD-XL</p>  <p>✗</p>	<p>Generate a photo of a sand martin.</p>  <p>SD-XL</p>  <p>✓</p>
<p>A photo of a thorn apple</p> 	<p>Generate a photo of a thorn apple.</p>  <p>DALL-E 3</p>  <p>✗</p>	<p>Generate a photo of a datura.</p>  <p>DALL-E 3</p>  <p>✓</p>	<p>Generate a photo of a thorn apple.</p>  <p>SD-XL</p>  <p>✗</p>	<p>Generate a photo of a datura.</p>  <p>SD-XL</p>  <p>✓</p>

Insight 3: use all synonyms for Retrieval Augmented Learning (RAL)

[REACT] is the state-of-the-art RAL method for zero-shot recognition

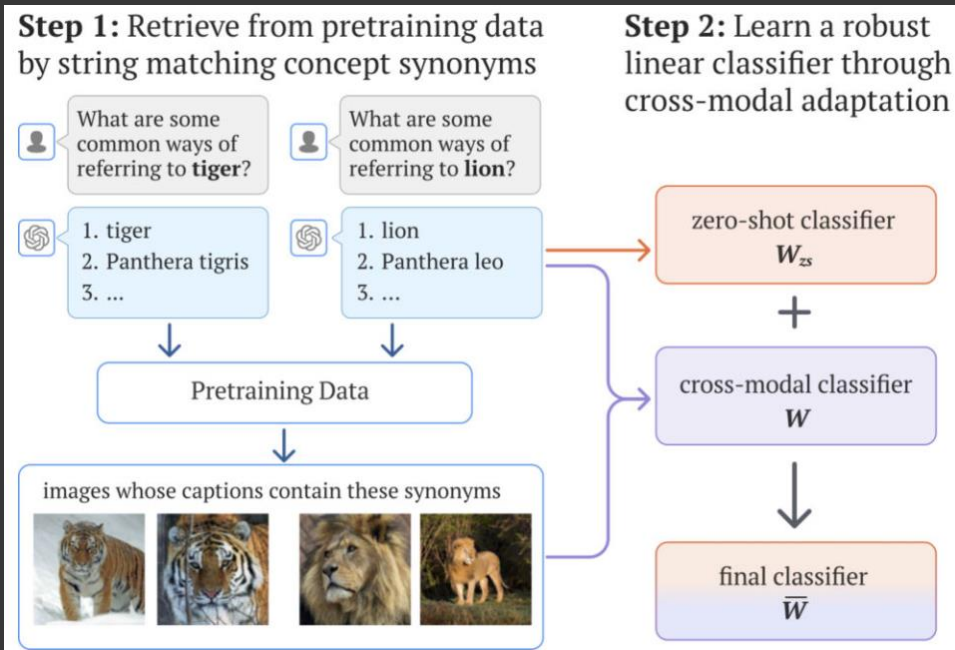
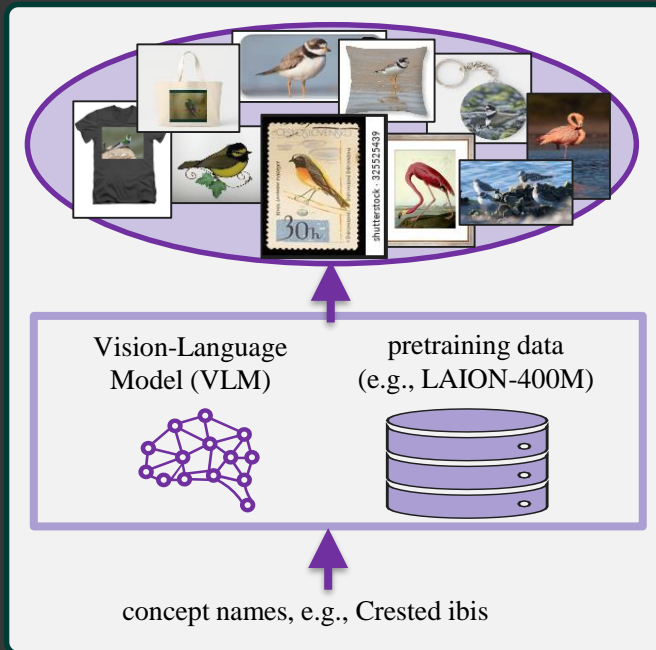


[REACT] Liu, et al., "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023

Insight 3: use all synonyms for Retrieval Augmented Learning (RAL)

[REACT] is the state-of-the-art RAL method for zero-shot recognition

[Ours] exploits all synonyms to retrieve data using string matching



[REACT] Liu, et al., “Learning customized visual models with retrieval-augmented knowledge”, CVPR, 2023

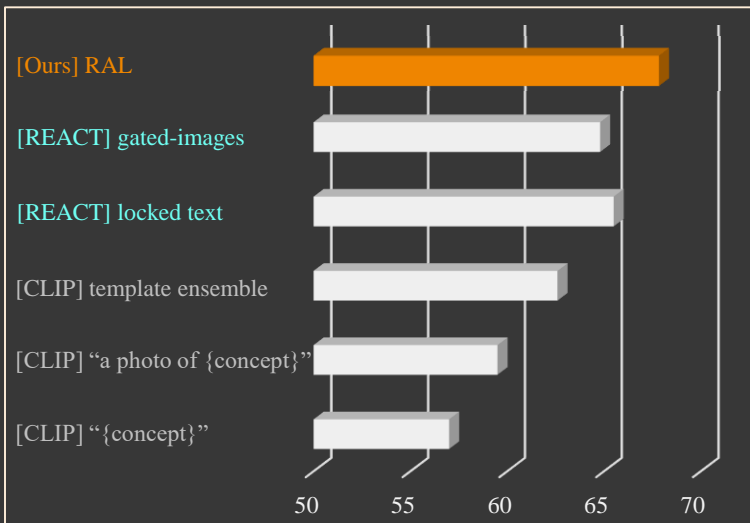
[Ours] Parashar, et al., “The Neglected Tails of Vision-Language Models”, CVPR, 2024

Insight 3: use all synonyms for Retrieval Augmented Learning (RAL)

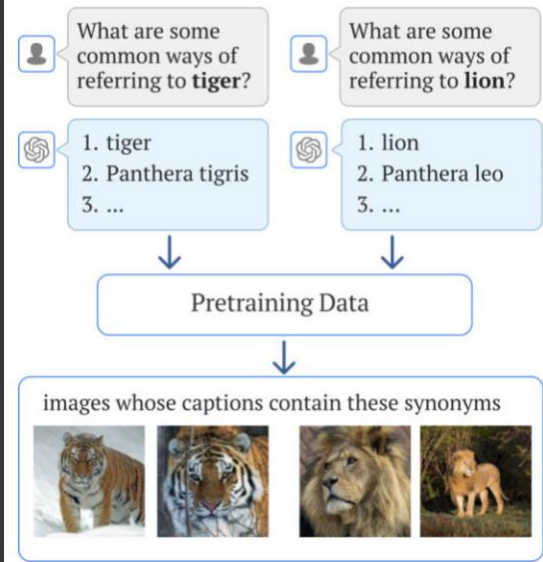
[REACT] is the state-of-the-art RAL method for zero-shot recognition

[Ours] exploits all synonyms to retrieve data using string matching

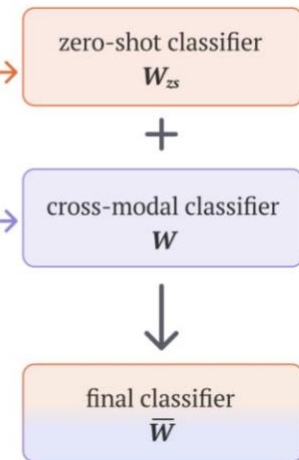
Accuracy averaged over eight datasets such as ImageNet, Food...



Step 1: Retrieve from pretraining data by string matching concept synonyms



Step 2: Learn a robust linear classifier through cross-modal adaptation



[REACT] Liu, et al., "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023

[Ours] Parashar, et al., "The Neglected Tails of Vision-Language Models", CVPR, 2024

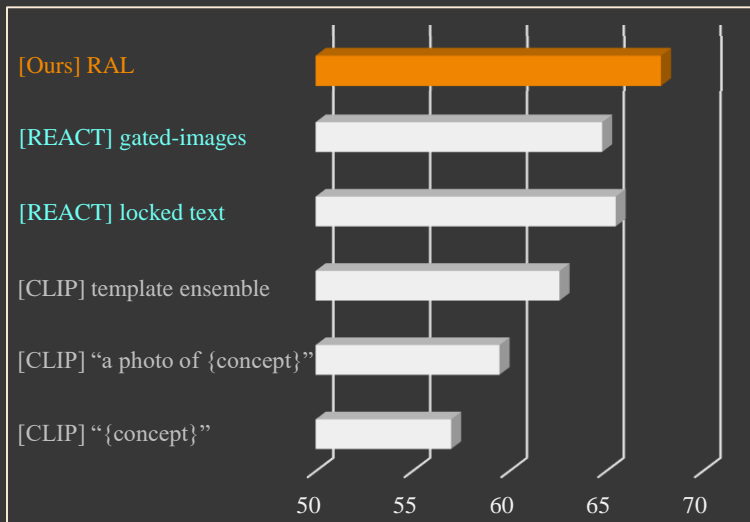
[CLIP] Radford, et al. "Learning transferable visual models from natural language supervision." ICML, 2021

Insight 3: use all synonyms for Retrieval Augmented Learning (RAL)

[REACT] is the state-of-the-art RAL method for zero-shot recognition

[Ours] exploits all synonyms to retrieve data using string matching

Accuracy averaged over eight datasets such as ImageNet, Food...



Stage	Resource	REACT	[Our] RAL	Relative Cost
Retrieval	retrieved examples	400M	0.5M	0.1%
	time	200 hrs	6 hrs	3%
	storage	10 TB	25 GB	0.25%
Learning	training images	10M	0.5M	5%
	time	256 hrs	2 mins	0.01%
	# of learned parameters	87M	0.5M	0.6%
	GPU memory	256 GB	2 GB	0.8%

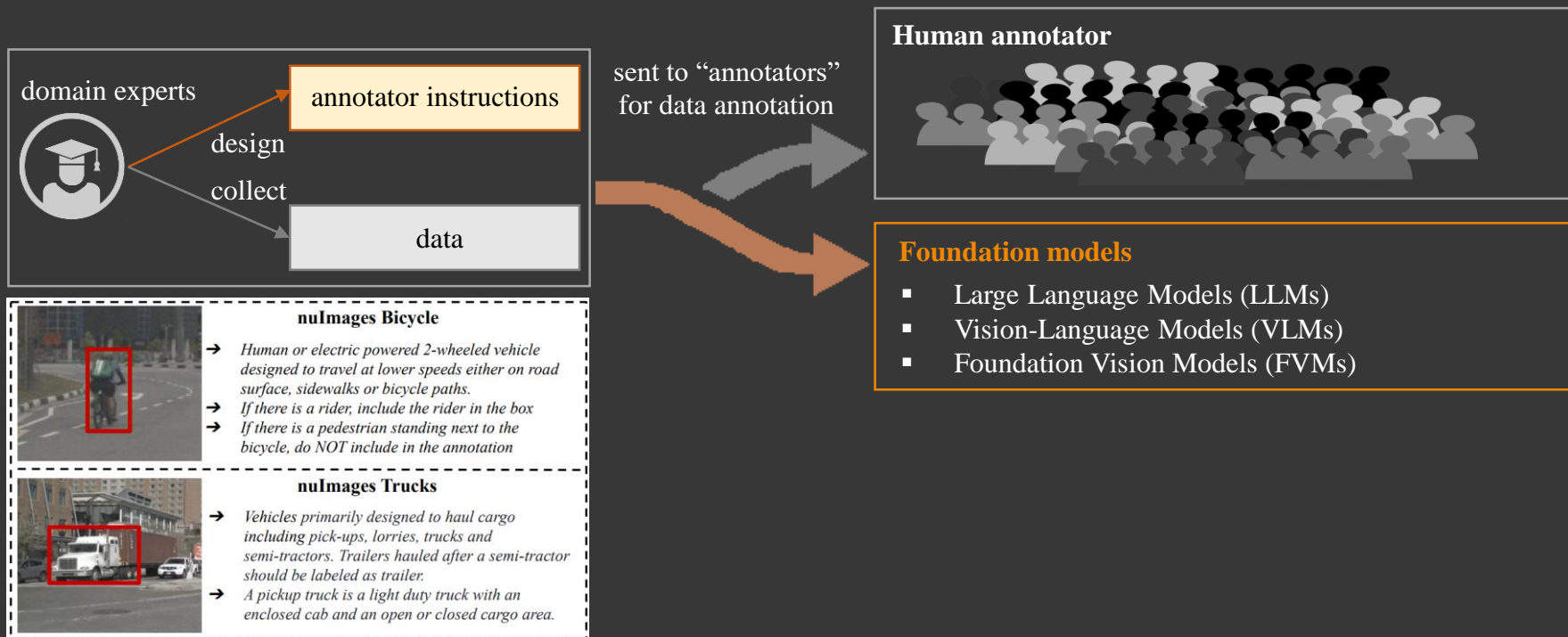
[REACT] Liu, et al., "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023

[Our] Parashar, et al., "The Neglected Tails of Vision-Language Models", CVPR, 2024

[CLIP] Radford, et al. "Learning transferable visual models from natural language supervision." ICML, 2021

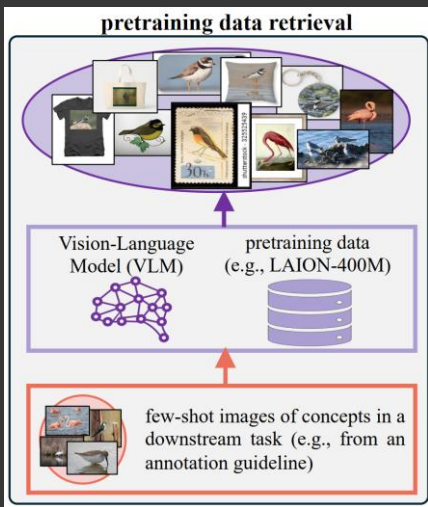
Exploit the open world for auto-annotation

We study few-shot recognition by adapting a Vision-Language Model (VLM)



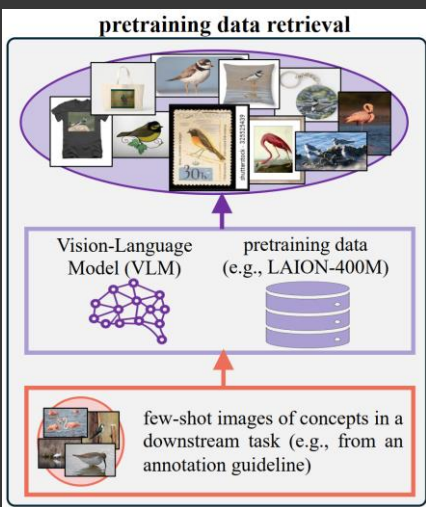
Exploit the open world for auto-annotation











- Retrieve data

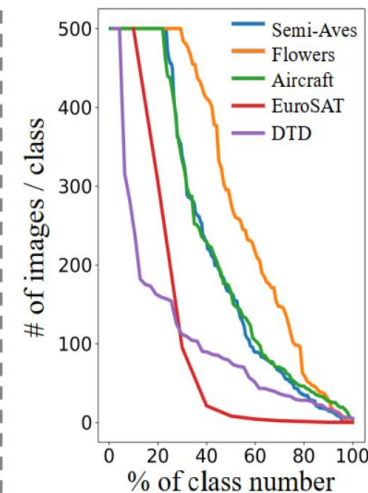


Exploit the open world for auto-annotation

- Retrieve data, which has (1) domain gaps, and (2) imbalanced distributions.

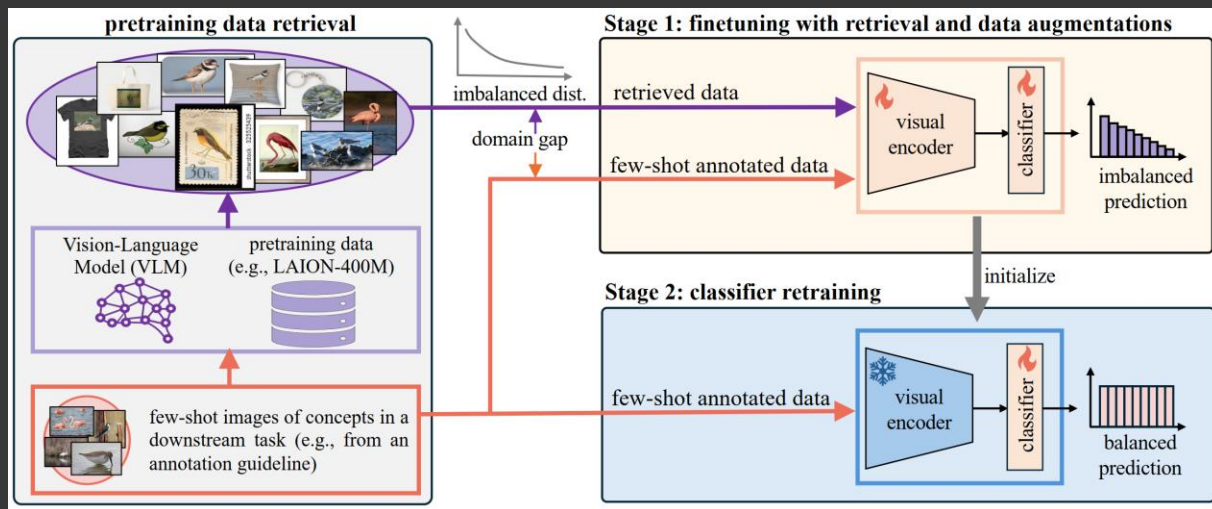


Dataset	Semi-Aves	Flowers	Aircraft	EuroSAT	DTD
class	<i>Tachycineta thalassina</i>	<i>canterbury bells</i>	707-320	river	banded
Few-shot					
Retrieved					



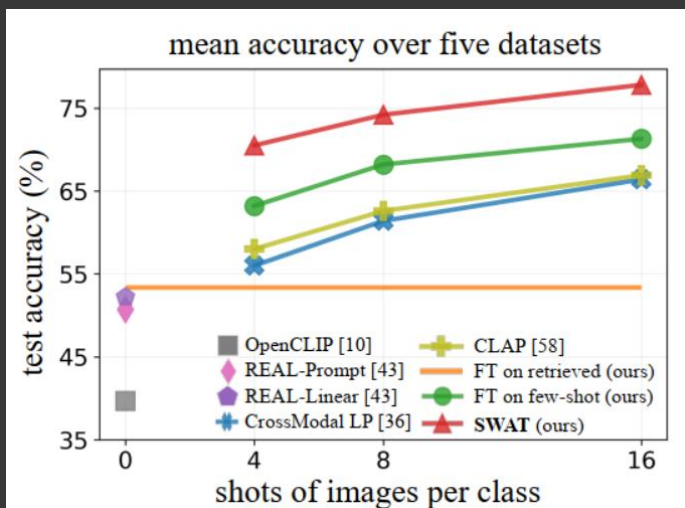
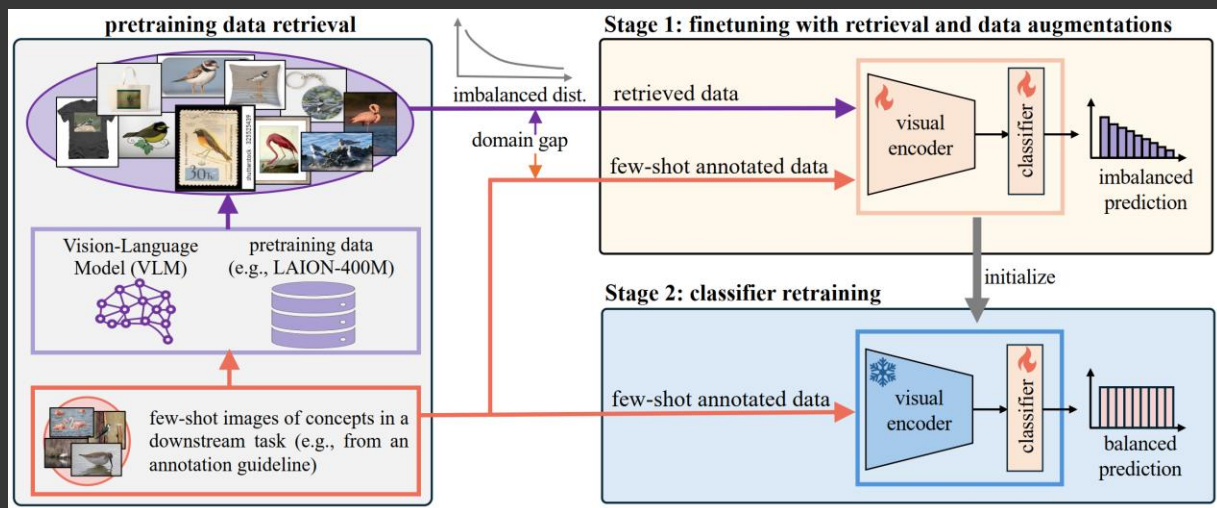
Exploit the open world for auto-annotation

- Retrieve data, which has (1) domain gaps, and (2) imbalanced distributions.
- We solve the above issues via **Stage-Wise retrieval Augmented fine-Tuning (SWAT)**, cf. decoupled feature and classifier for long-tailed recognition, and transfer learning for domain adaptation.



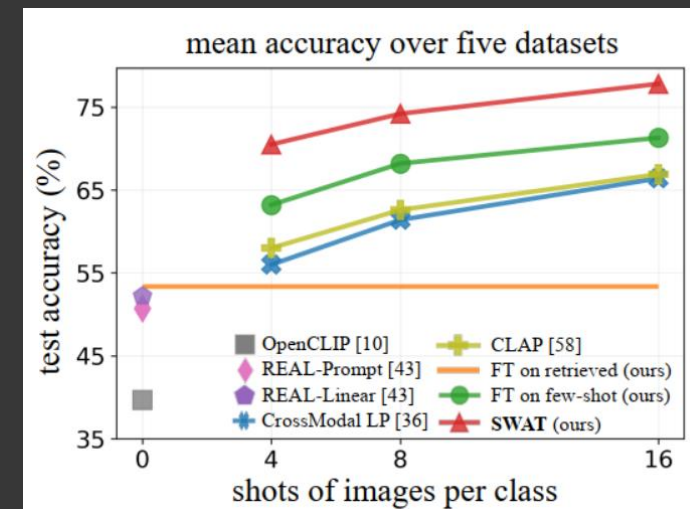
Exploit the open world for auto-annotation

- Retrieve data, which has (1) domain gaps, and (2) imbalanced distributions.
- We solve the above issues via **Stage-Wise retrieval Augmented fine-Tuning (SWAT)**, cf. decoupled feature and classifier for long-tailed recognition, and transfer learning for domain adaptation.
- SWAT** performs the best.



Exploit the open world for auto-annotation

- Retrieve data, which has (1) domain gaps, and (2) imbalanced distributions.
- We solve the above issues via **Stage-Wise retrieval Augmented fine-Tuning (SWAT)**, cf. decoupled feature and classifier for long-tailed recognition, and transfer learning for domain adaptation.
- **SWAT** performs the best.
- **Few-shot finetuning** outperforms existing few-shot learning methods!
- **Finetuning on retrieved data** underperforms **zero-shot method (REAL-Linear)** due to domain gaps & imbalanced distributions.



Conclusions & Thank you!

- Embrace the open world – the foundation models and open data!
- Watch out for misalignment between AI and experts (like you)!
- Be aware of the imbalance of the open world!

