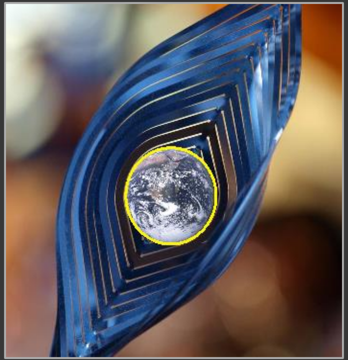
VPLOW@CVPR’24

The 4th Workshop of Visual Perception and Learning in an Open World



Shu Kong

Texas A&M University
University of Macau

June 18, 2024

Welcome

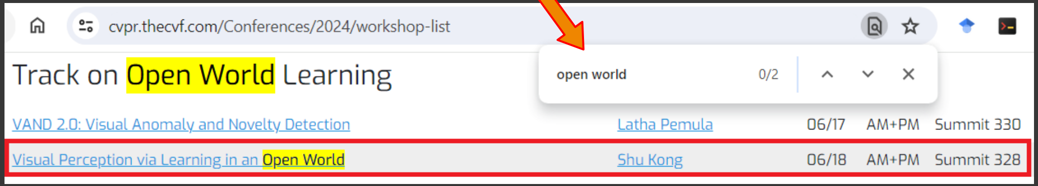
- Hope everyone is *safe* and *healthy*!

- 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’24)

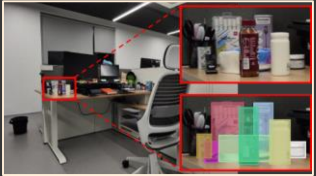
Schedule

https://vplow.github.io/vplow_4th.html

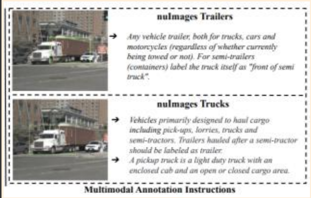


Four challenges

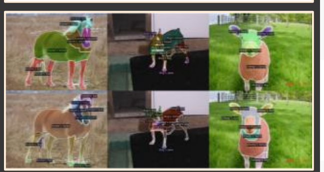
1. InsDet: Object Instance Detection



2. Foundational Few-Shot Object Detection



3. OV-PARTS: Open-Vocabulary Part Segmentation

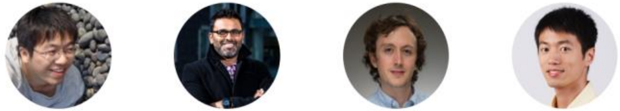


4. V3Det: Vast Vocabulary Visual Detection

PDT / Time in Vancouver	Event	Title/Presenter
08:30 - 08:50	Opening remarks	Shu Kong Texas A&M, University of Macau **Visual Perception via Learning in an Open World**
08:50 - 09:30	Invited talk #1	Walter Scheirer, University of Notre Dame **Open Issues in Open World Learning**
09:30 - 10:10	Invited talk #2	Deva Ramanan CMU **Open World Learning in the Era of MultiModal Foundation Models**
10:10 - 10:15	**Coffee break**	
10:15 - 10:55	Invited talk #3	Andrew Owens, UMich **title tba**
10:55 - 11:35	Challenge-1	Challenge 1: InsDet **Object Instance Detection Challenge**
11:35 - 13:30	**Lunch**	
13:30 - 14:10	Invited talk #4	Xiaolong Wang UCSD **Spatial Perception and Control in the Wild**
14:10 - 14:50	Invited talk #5	Ziwei Liu NTU **Building Open-World Multimodal AI Assistant**
14:50 - 15:30	Invited talk #6	Yu-Xiong Wang UIUC **All-in-One: Bridging Generative and Discriminative Learning in the Open World**
15:30 - 15:35	**Coffee break**	
15:35 - 16:15	Challenge-2	Challenge-2: Foundational FSOD **Foundational Few-Shot Object Detection Challenge**
16:15 - 16:55	Challenge-3	Challenge-3: OV-PARTS **Challenge of Open-Vocabulary Part Segmentation**
16:55 - 17:35	Challenge-4	Challenge-4: V3Det **Challenge of Vast Vocabulary Visual Detection**
17:35 - 17:40	Closing remarks	Neehar Peri CMU

Together we serve

Speakers



Shu Kong
UMacau, Texas A&M

Deva Ramanan
Carnegie Mellon University

Walter J. Scheirer
University of Notre Dame

Ziwei Liu
Nanyang Technological University

Xiaolong Wang
UC San Diego

Andrew Owens
University of Michigan

Yu-Xiong Wang
University of Illinois at Urbana-Champaign

Advisory Board

Deva Ramanan
Carnegie Mellon University

Terrance Boult
University of Colorado Colorado Springs

Walter J. Scheirer
University of Notre Dame

Organizers

Shu Kong
UMacau, Texas A&M

Yanan Li
Zhejiang Lab

Neehar Peri
CMU

Yu-Xiong Wang
University of Illinois at Urbana-Champaign

Andrew Owens
University of Michigan

Deepak Pathak
Carnegie Mellon University

Carl Vondrick
Columbia University

Abhinav Shrivastava
University of Maryland

Coordinators

Tian Liu
Texas A&M

Shubham Parashar
Texas A&M

Yunhan Zhao
UC Irvine

Challenge Organizers

Shu Kong
UMacau, Texas A&M

Yanan Li
Zhejiang Lab

Qianqian Shen
Zhejiang University

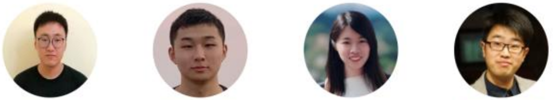
Yunhan Zhao
UC Irvine

Anish Madan
CMU

Neehar Peri
CMU

Deva Ramanan
CMU

Meng Wei
HKU



Xiaoyu Yue
University of Sydney

Wenwei Zhang
Shanghai AI Lab

Xihui Liu
HKU

Jiangmiao Pang
Shanghai AI Lab

Pan Zhang
Shanghai AI Lab

Tao Chu
SCUT

Yuhang Cao
CUHK

Zeyi Sun
SJTU

Ziyu Liu
SJTU

Xiaoyi Dong
Shanghai AI Lab

Yuhang Zang
Shanghai AI Lab

Tong Wu
CUHK

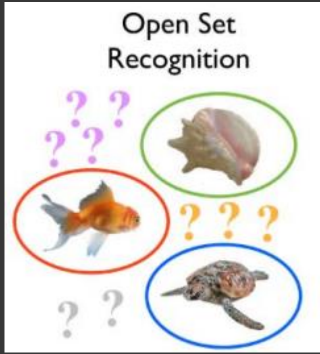
Jiaqi Wang
Shanghai AI Lab

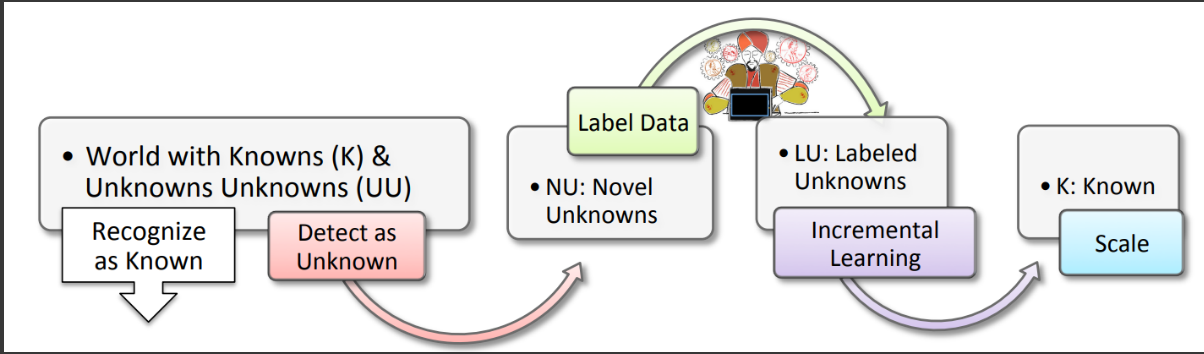
Dahua Lin
CUHK

A brief introduction to the open world

Previously, we emphasize *testing in the open world* *(while training in a closed world)*

- Open-set recognition (OSR)
- Open-world recognition: OSR + continual learning for new concepts w/ human annotation





W. Scheirer, A. de Rezende Rocha, A. Sapkota, T. Boult, "Toward Open Set Recognition", PAMI, 2013
A. Bendale, T. Boult, "Towards open world recognition", CVPR, 2015

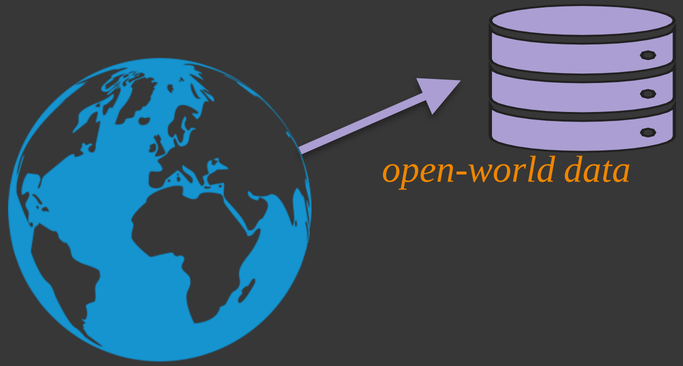
A brief introduction to the open world

Previously, we emphasize *testing in the open world (while training in a closed world)*

- Open-set recognition (OSR)
- Open-world recognition: OSR + continual learning for new concepts w/ human annotation

Nowadays, we also *train in the open world*

- *data* sampling from the open world, e.g., sampling outlier data for better OSR.

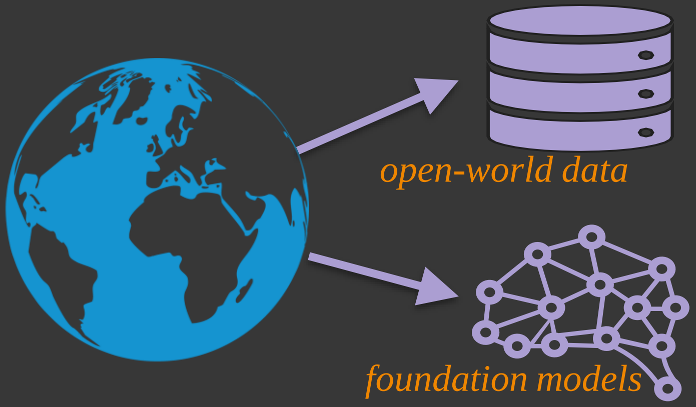


D. Hendrycks, M. Mazeika, T. Dietterich, "Deep anomaly detection with outlier exposure", ICLR, 2019
S. Kong, D. Ramanan, "OpenGAN: Open-set recognition via open data generation", ICCV, 2021

A brief introduction to the open world

Previously, we emphasize *testing in the open world (while training in a closed world)*
- Open-set recognition (OSR)
- Open-world recognition: OSR + continual learning for new concepts w/ human annotation

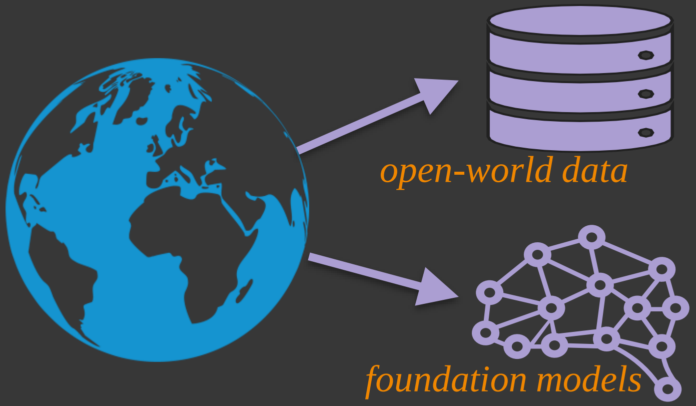
Nowadays, we also *train in the open world*
- *data* sampling from the open world, e.g., sampling outlier data for better OSR.
- *foundation models* pretrained in the open world, e.g., CLIP.



Foundation models are open-world models, enabling:
- open-vocabulary recognition
- zero-shot recognition
- ….

A. Radford, et al. "Learning transferable visual models from natural language supervision." ICML, 2021
A. Kirillov, et al. "Segment anything." ICCV, 2023

Can we predict testing performance with distribution shift?

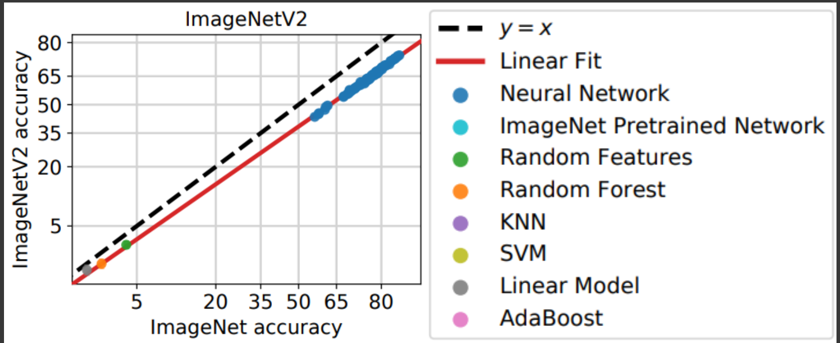


Can we predict testing performance with distribution shift?

ImageNet / in-distribution

mongoose

ImageNetV2 / out-of-distribution

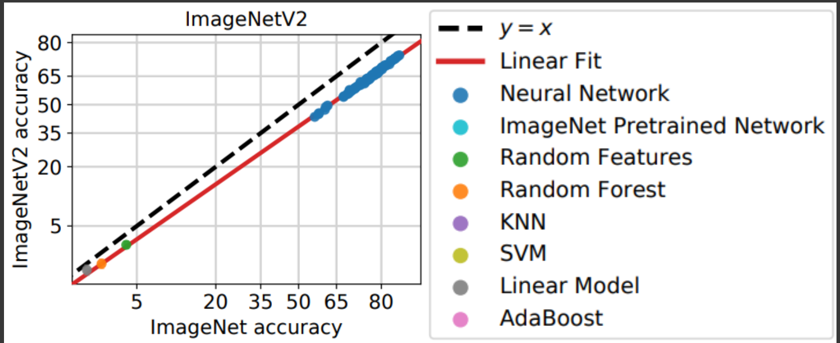


Accuracy-on-the-line: empirically, OOD performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts.

J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021
C. Baek, et al., "Agreement-on-the-line: Predicting the performance of neural networks under distribution shift", NeurIPS, 2022

Can we predict testing performance with distribution shift?

ImageNet / in-distribution   ImageNetV2 / out-of-distribution



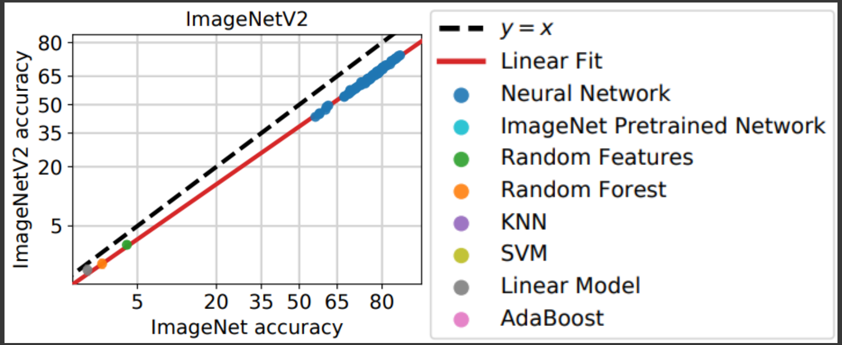
Accuracy-on-the-line: empirically, OOD performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts.

But the models are trained in the closed world! ☹

J. Miller, et al., "Accuracy on the Line: On the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization", ICML, 2021
C. Baek, et al., "Agreement-on-the-line: Predicting the performance of neural networks under distribution shift", NeurIPS, 2022

Can we predict testing performance with distribution shift?

ImageNet / in-distribution   ImageNetV2 / out-of-distribution

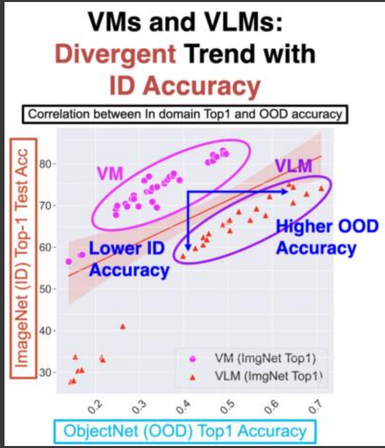


Accuracy-on-the-line: empirically, OOD performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts.

But the models are trained in the closed world! ☹

We test 75 models including Vision Models (VMs) and Vision-Language Models (VLMs), trained in either the closed world (ImageNet) or the *open world (data from the Internet)*.

Accuracy-on-the-line does *NOT* hold anymore! ☹

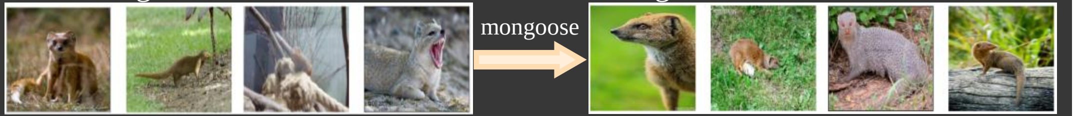


foundation models

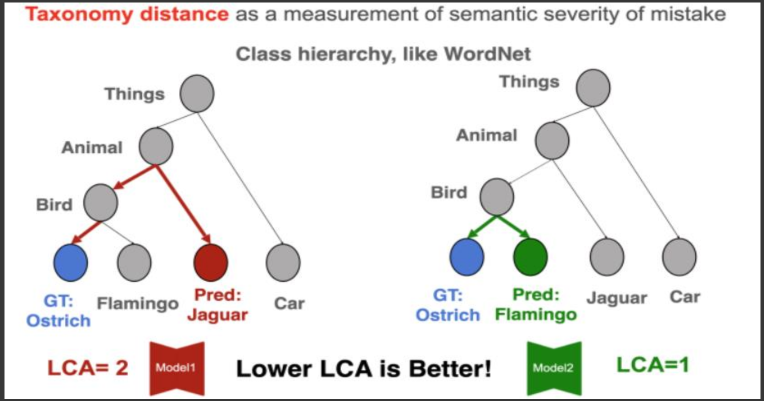
Shi, Gare, Tian, Chai, Lin, Vasudevan, Feng, Ferroni, Kong, “LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies”, ICML, 2024

Can we predict testing performance with distribution shift?

ImageNet / in-distribution | ImageNetV2 / out-of-distribution

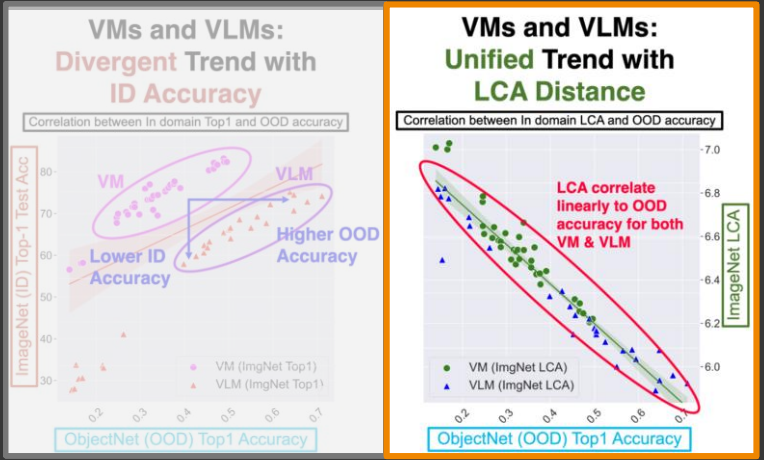


LCA-on-the-line: using least common ancestor (LCA) to predict OOD performance. It is a better metric!



We test 75 models including Vision Models (VMs) and Vision-Language Models (VLMs), trained in either the closed world (ImageNet) or the *open world (data from the Internet)*.

Accuracy-on-the-line does *NOT* hold anymore!

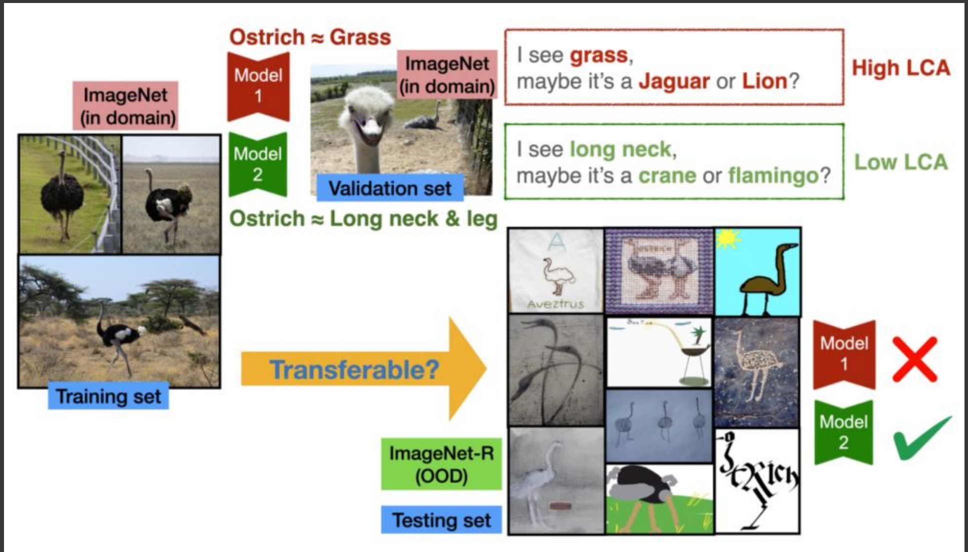


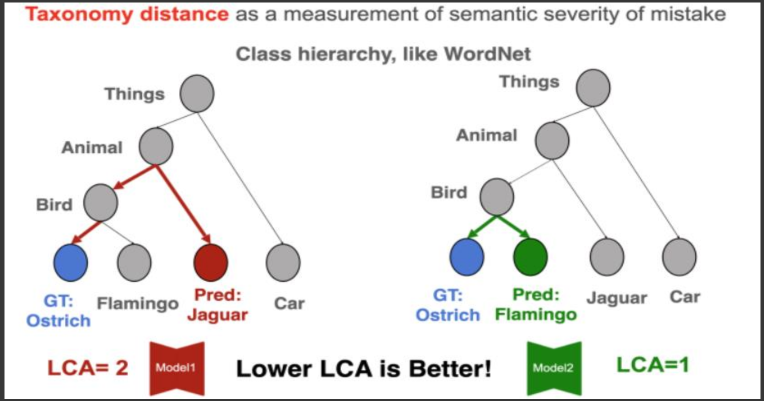
Shi, Gare, Tian, Chai, Lin, Vasudevan, Feng, Ferroni, Kong, “LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies”, ICML, 2024

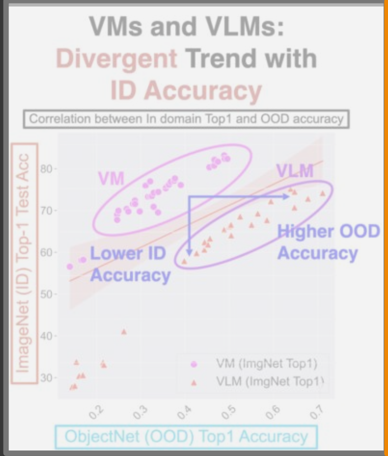
Can we predict testing performance with distribution shift?

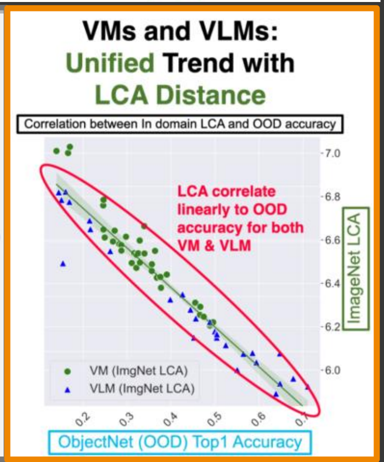
Intuition: A model that makes better mistakes (measured by LCA) can mitigate spurious correlation, leading to better generalization.

LCA-on-the-line: using least common ancestor (LCA) to predict OOD performance. It is a better metric!





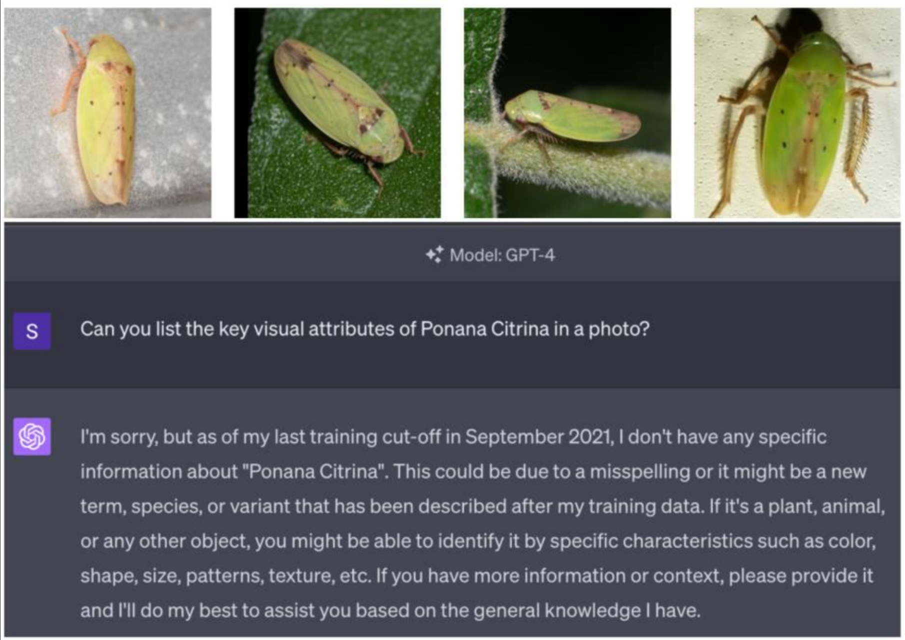




Shi, Gare, Tian, Chai, Lin, Vasudevan, Feng, Ferroni, Kong, "LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies", ICML, 2024

An interesting failure case

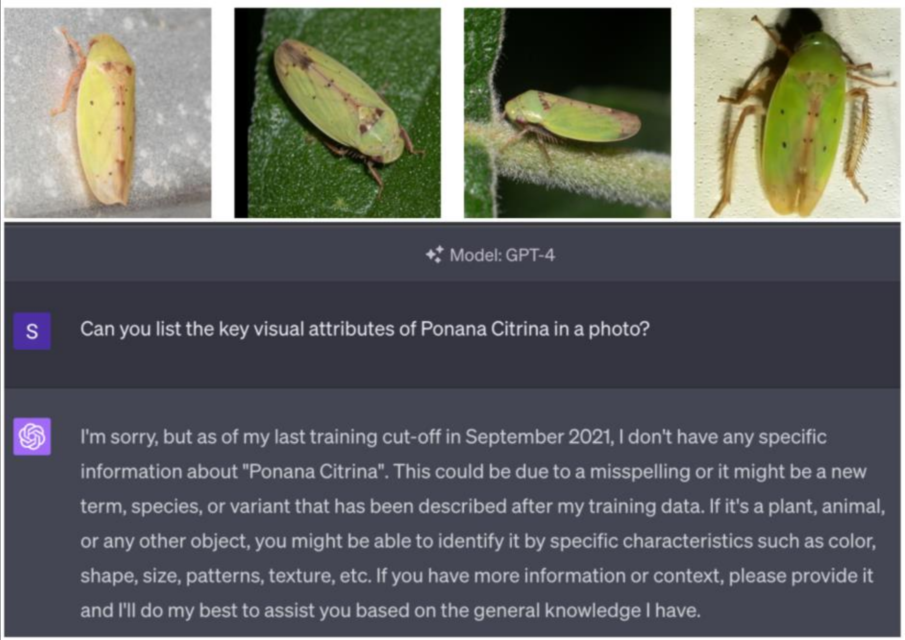
Why does GPT-4 fail as it is trained on internet-data in the open world!?

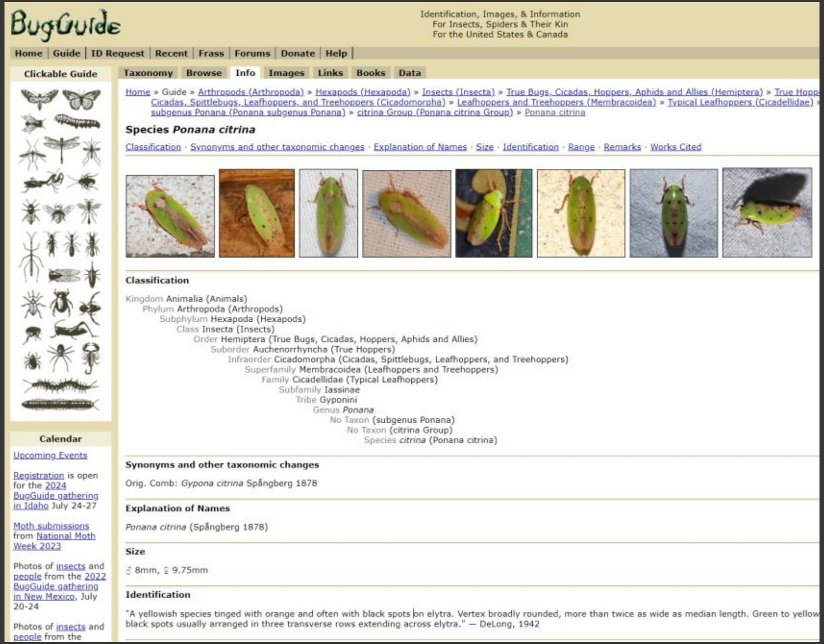


S. Parashar, Z. Lin, Y. Li, S. Kong, "Prompting Scientific Names for Zero-Shot Species Recognition", EMNLP, 2023

An interesting failure case

Why does GPT-4 fail as it is trained on internet-data in the open world!?



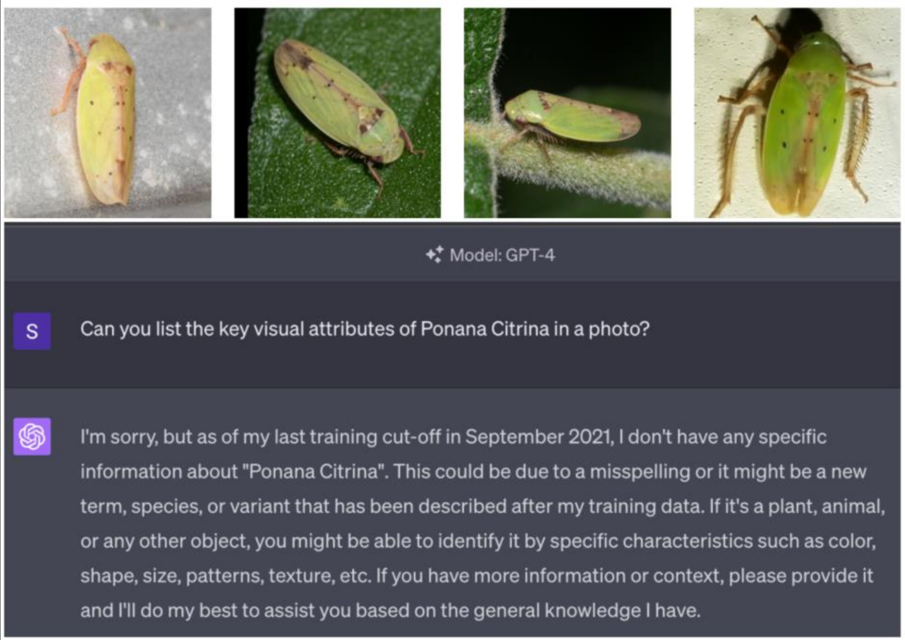


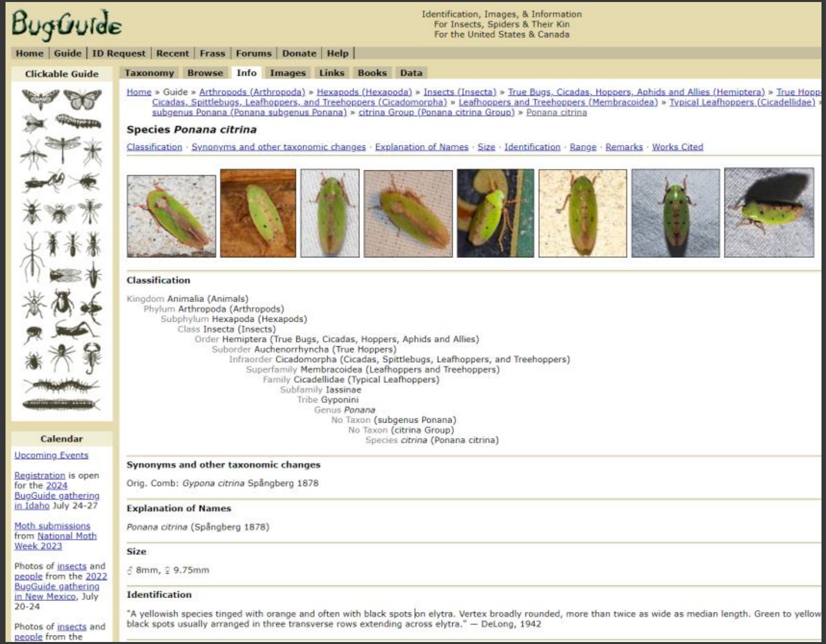
S. Parashar, Z. Lin, Y. Li, S. Kong, “Prompting Scientific Names for Zero-Shot Species Recognition”, EMNLP, 2023

An interesting failure case

Why does GPT-4 fail as it is trained on internet-data in the open world!?

Hypothesis: Some concepts (esp. scientific names in Latin) are too few in the open world to train models.





S. Parashar, Z. Lin, Y. Li, S. Kong, “Prompting Scientific Names for Zero-Shot Species Recognition”, EMNLP, 2023

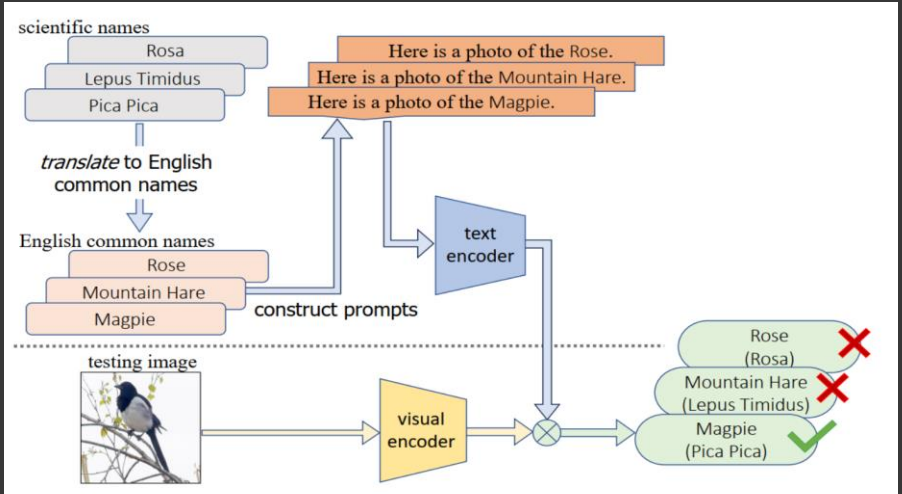
An interesting failure case

Why does GPT-4 fail as it is trained on internet-data in the open world!?
Hypothesis: Some concepts (esp. scientific names in Latin) are too few in the open world to train models.
Justification: We count!

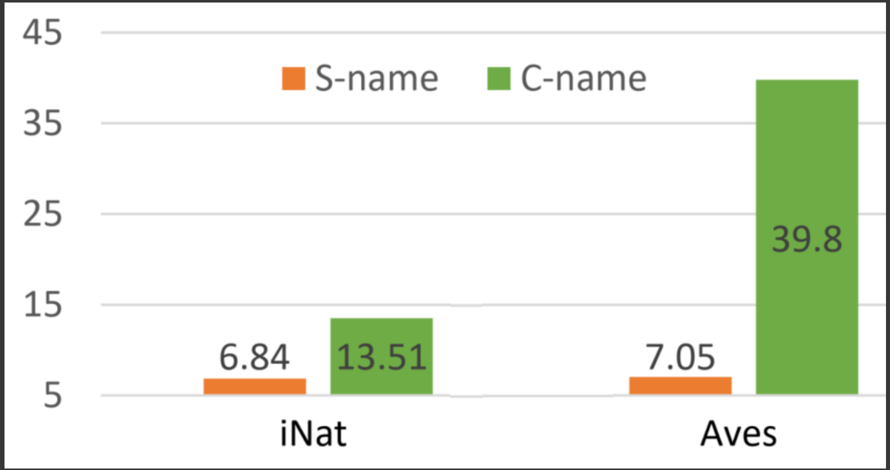
C. Schuhmann, et al. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs." arXiv:2111.02114, 2021
C. Schuhmann, et al. "Laion-5b: An open large-scale dataset for training next generation image-text models", NeurIPS 2022
S. Parashar, Z. Lin, Y. Li, S. Kong, “Prompting Scientific Names for Zero-Shot Species Recognition”, EMNLP, 2023

An interesting failure case

Why does GPT-4 fail as it is trained on internet-data in the open world!?
Hypothesis: Some concepts (esp. scientific names in Latin) are too few in the open world to train models.
Justification: We count!
Remedy: Translating Latin scientific names to English common names.



zero-shot accuracy by prompting OpenCLIP



S. Parashar, Z. Lin, Y. Li, S. Kong, “Prompting Scientific Names for Zero-Shot Species Recognition”, EMNLP, 2023

An interesting failure case

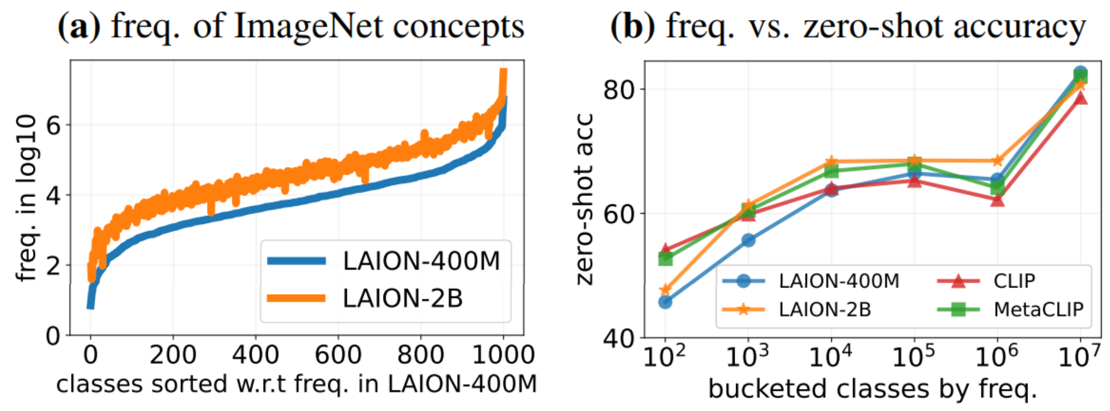
Why do foundation models fail to handle some concepts?
Hypothesis: certain concepts are insufficiently presented in the open world.

Arch 4A-E Poster #324
Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, “The Neglected Tails of Vision-Language Models”, CVPR, 2024

An interesting failure case

Why do foundation models fail to handle some concepts?
Hypothesis: certain concepts are insufficiently presented in the open world.
Evidence: a strong correlation between **concept frequency** and per-concept accuracy.



Arch 4A-E Poster #324
Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

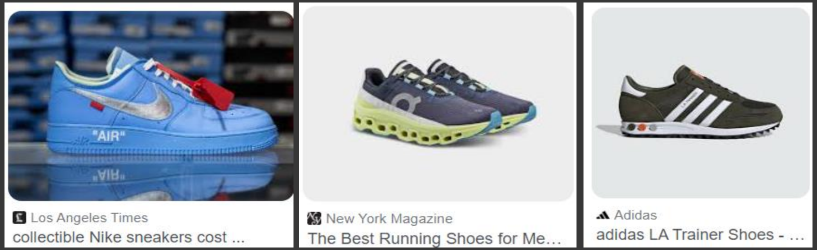
S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

Measure concept frequency

Intuitively, 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!

Lexical variation, e.g., synonyms



Linguistic ambiguity

C. Schuhmann, et al. “Laion-400m: Open dataset of clip-filtered 400 million image-text pairs”, arXiv, 2021
SY Gadre, et al. “Datacomp: In search of the next generation of multimodal datasets”, NeurIPS, 2024

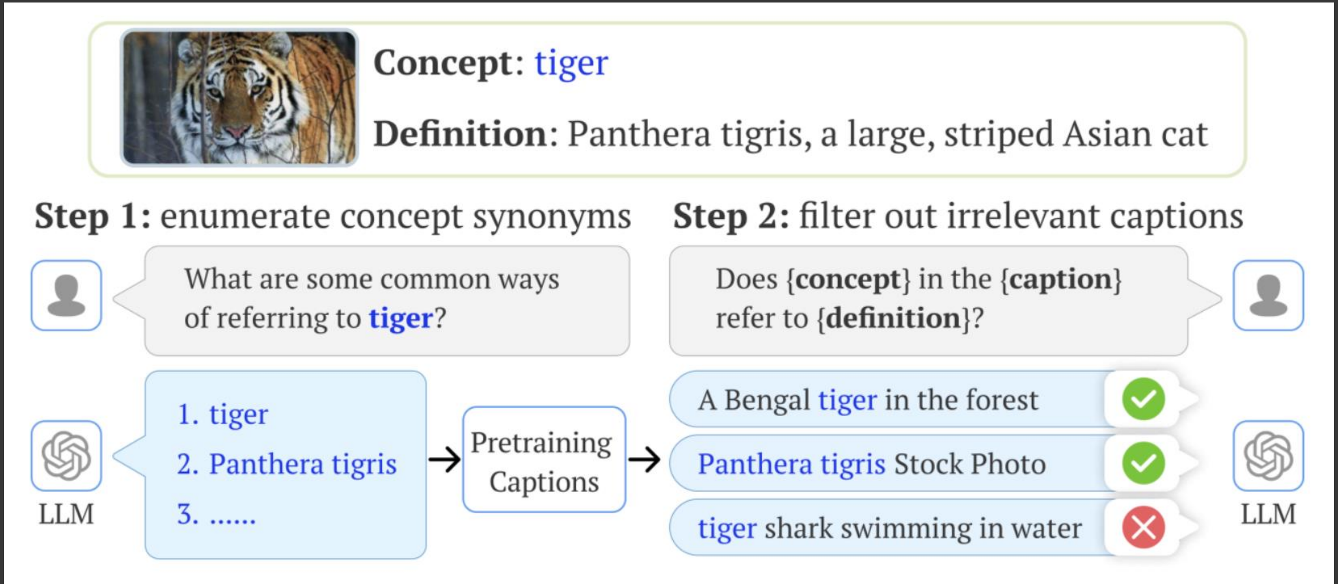
Measure concept frequency

Intuitively, 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!

Lexical variation, e.g., synonyms

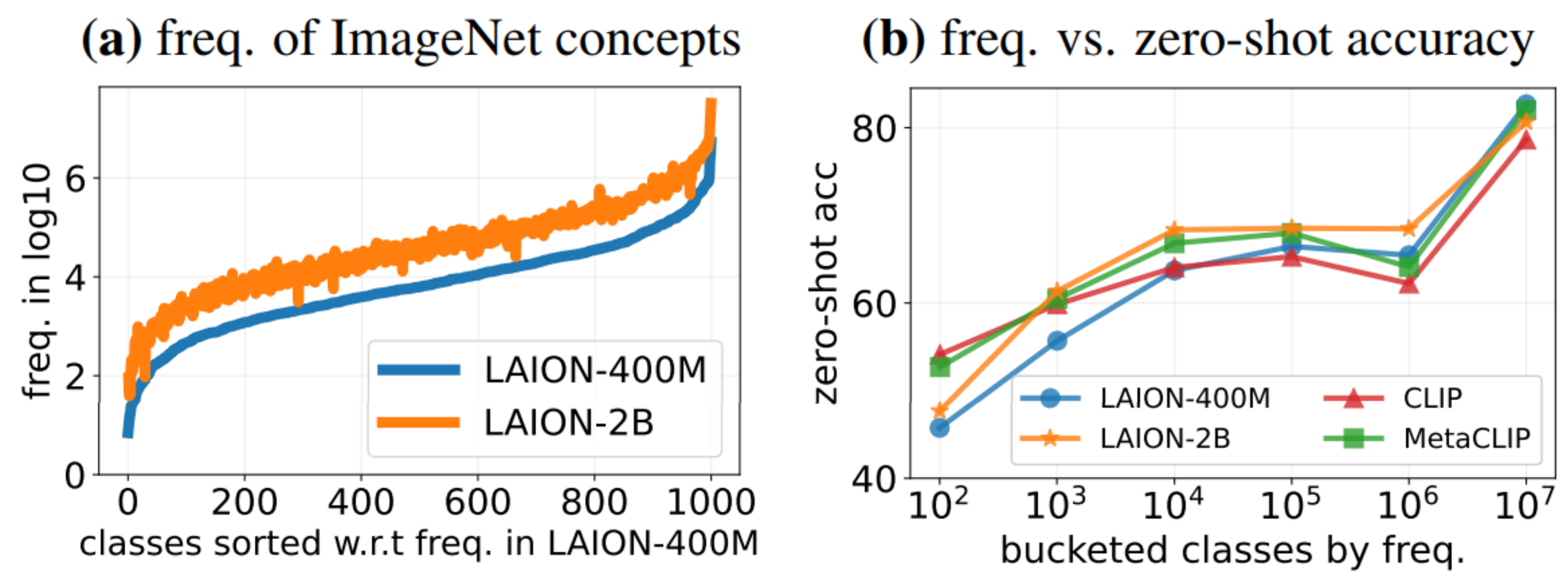
Linguistic ambiguity



S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

Measure concept frequency

Reliably measuring concept frequency reveals its strong correlation with per-concept accuracy!

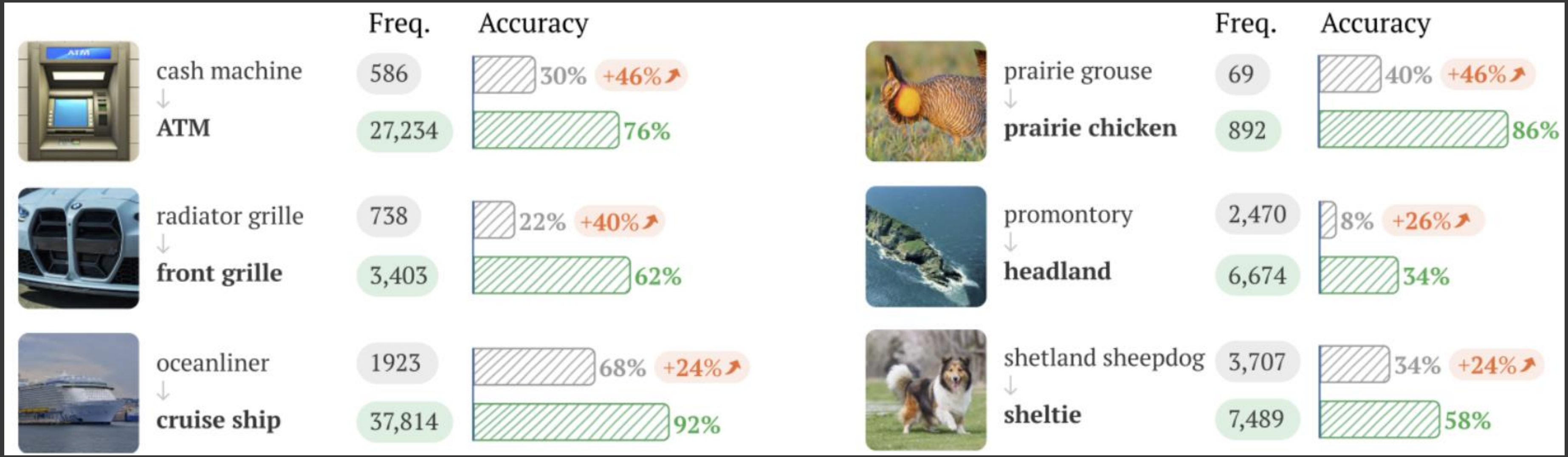


Arch 4A-E Poster #324
Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

Insight 1: prompt VLM using the most frequent synonym

This simple change significantly boosts zero-shot accuracy!

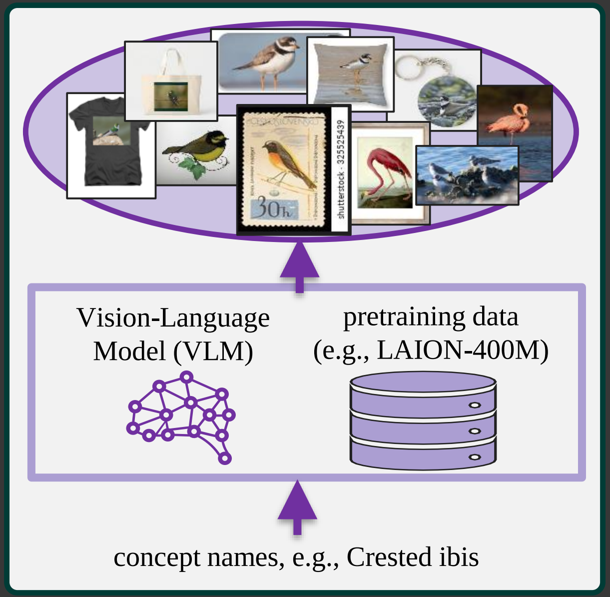


Arch 4A-E Poster #324
Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, “The Neglected Tails of Vision-Language Models”, CVPR, 2024

Insight 2: use all synonyms for Retrieval Augmented Learning (RAL)

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

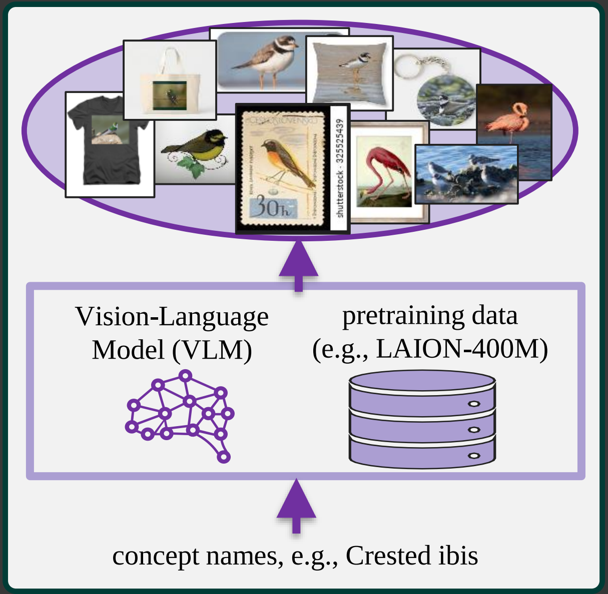


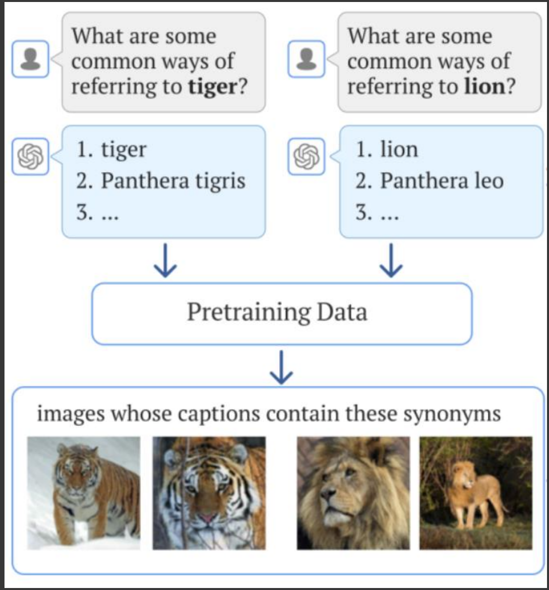
[REACT] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, YJ Lee, C. Li. "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023

Insight 2: use all synonyms for Retrieval Augmented Learning (RAL)

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

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





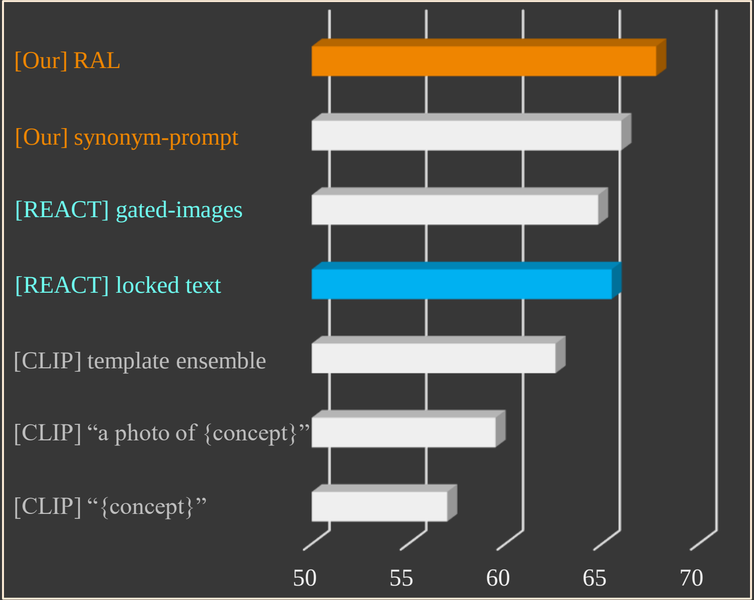
[REACT] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, YJ Lee, C. Li. "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023
[Our] S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

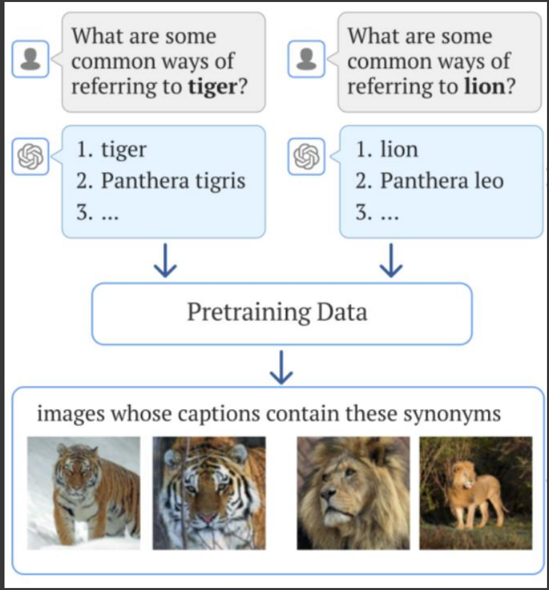
Insight 2: use all synonyms for Retrieval Augmented Learning (RAL)

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

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

Accuracy averaged over eight datasets such as ImageNet, Food101, DTD, EuroSAT, etc.





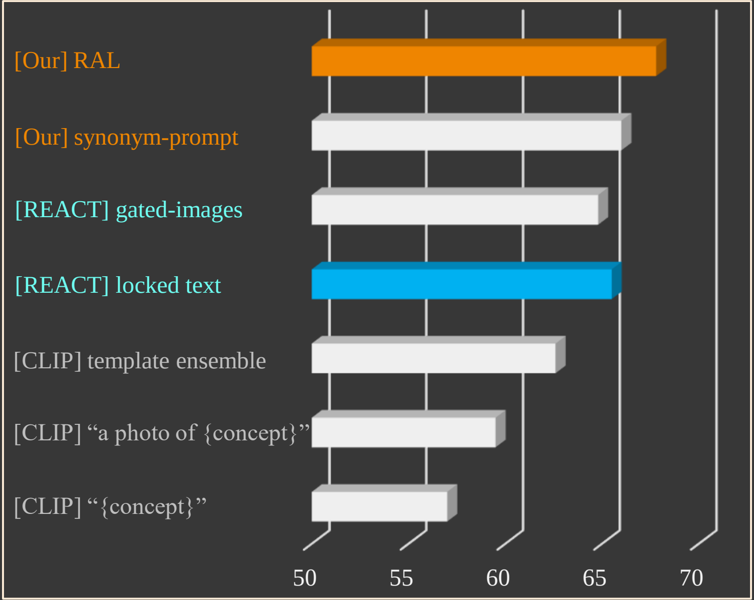
[REACT] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, YJ Lee, C. Li. "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023
[Our] S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024
[CLIP] A. Radford, et al. "Learning transferable visual models from natural language supervision." ICML, 2021

Insight 2: use all synonyms for Retrieval Augmented Learning (RAL)

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

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

Accuracy averaged over eight datasets such as ImageNet, Food101, DTD, EuroSAT, etc.



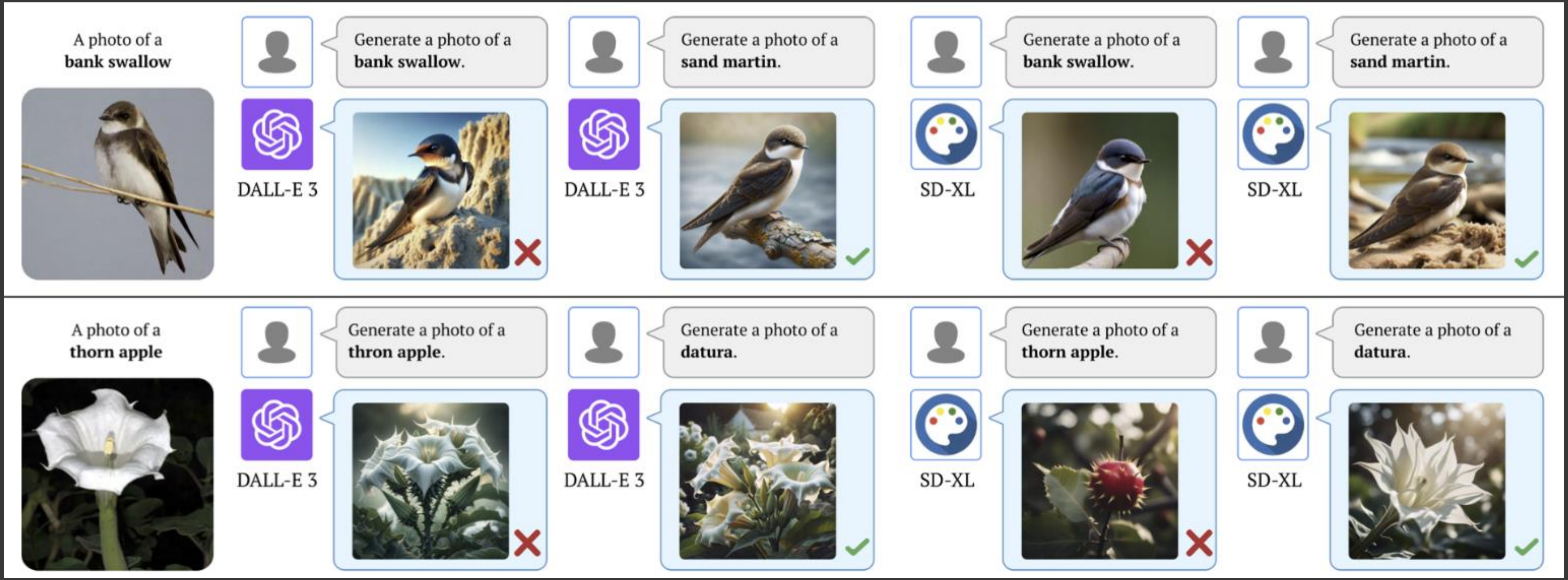
Stage	Resource	REACT	[Our] RAL	Relative Cost
Retrieval	retrieved examples	400M	0.5M	0.1%
	time	200 hrs	6 hrs	3%
	storage	10 TB	25 GB	0.25%
Learning	training images	10M	0.5M	5%
	time	256 hrs	2 mins	0.01%
	# of learned parameters	87M	0.5M	0.6%
	GPU memory	256 GB	2 GB	0.8%

Arch 4A-E Poster #324
Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

Insight 3: use the most frequent synonym in image generation

Recipe: replacing the original query with its most frequent synonym in prompts

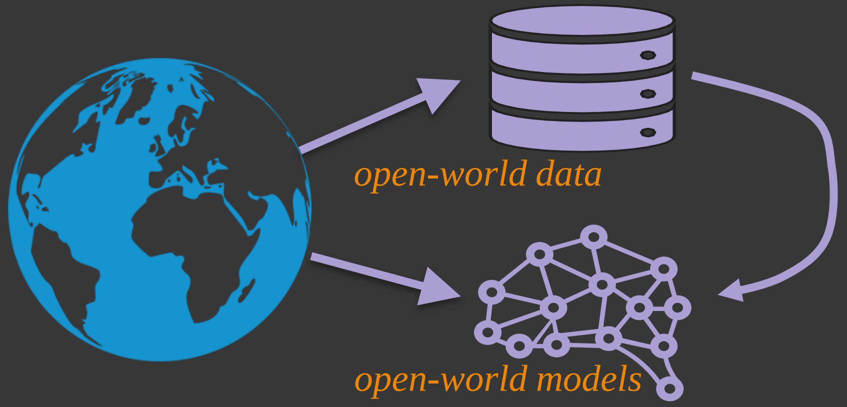


Arch 4A-E Poster #324
Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

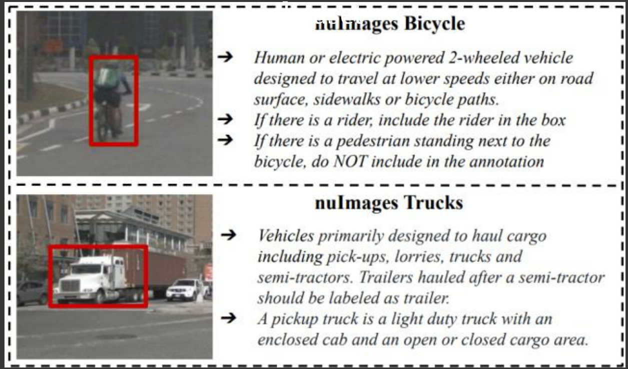
Remarks

- Foundation models are open-world models!
- Open-world models have open issues!
- Attention to open-world data!



Exploit open-world foundation models for data annotation

A snippet of annotation guidelines from nuScenes


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.*

domain experts

annotator instructions

design

collect

data

sent to “annotators” for data annotation

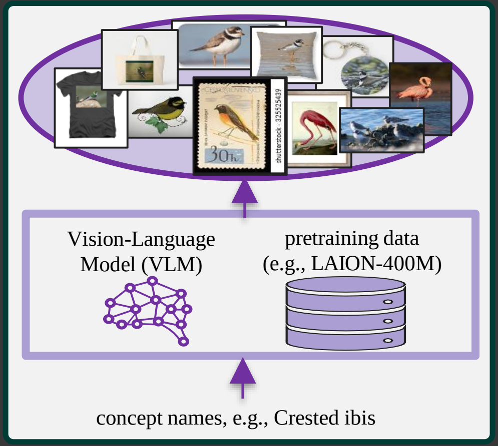
Human annotator

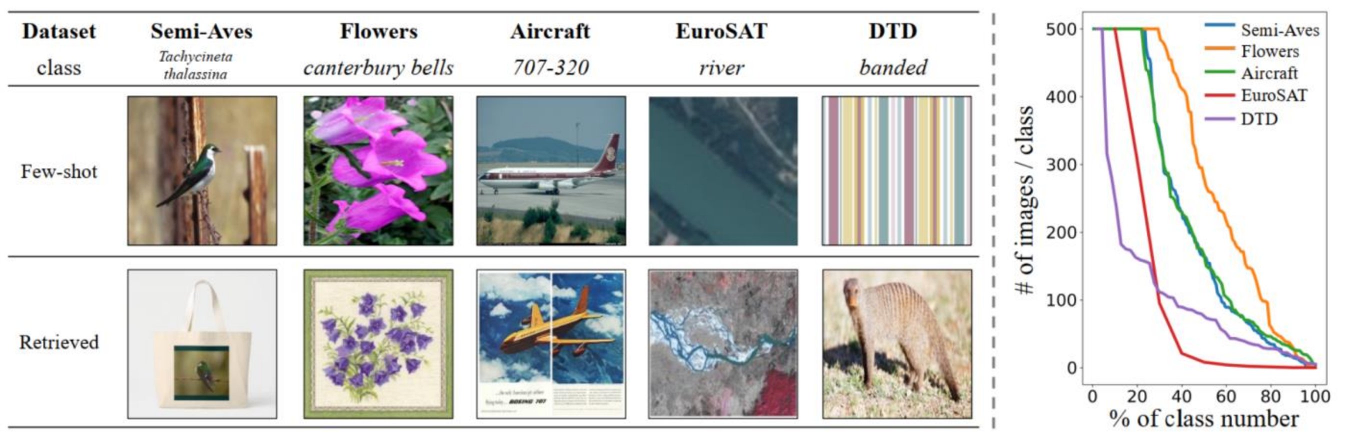
Foundation models

- Large Language Models (LLMs)
- Vision-Language Models (VLMs)
- Foundation Vision Models (FVMs)

T. Liu, H. Zhang, S. Parashar, S. Kong, ‘Few-Shot Recognition via Stage-Wise Augmented Finetuning’, ArXiv:2406.11148, 2024

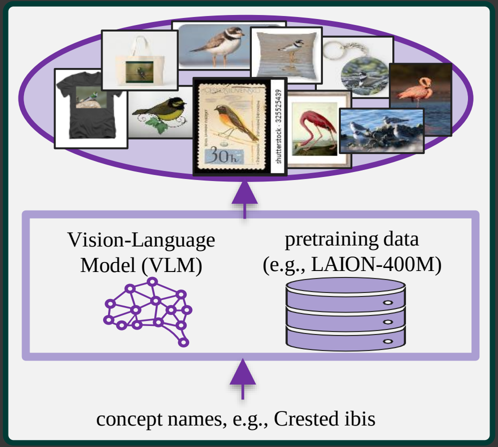
Exploit open-world foundation models for data annotation

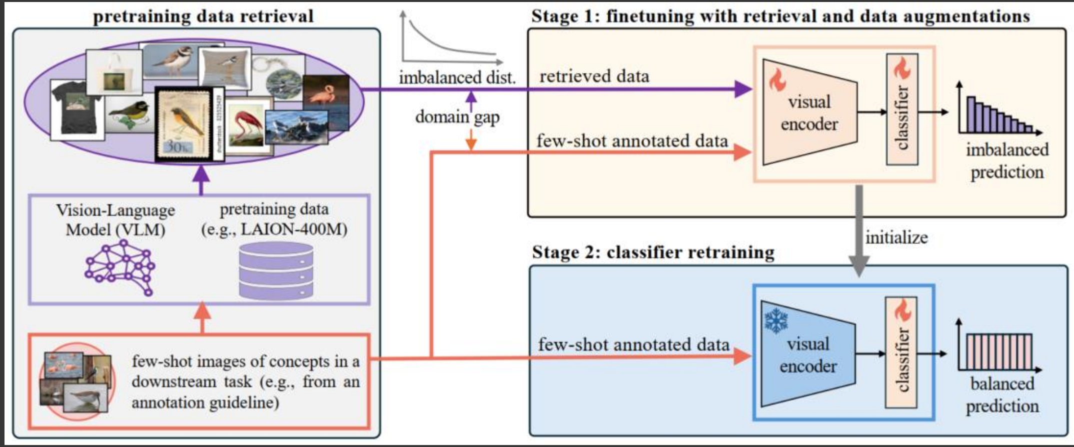




T. Liu, H. Zhang, S. Parashar, S. Kong, 'Few-Shot Recognition via Stage-Wise Augmented Finetuning', ArXiv:2406.11148, 2024

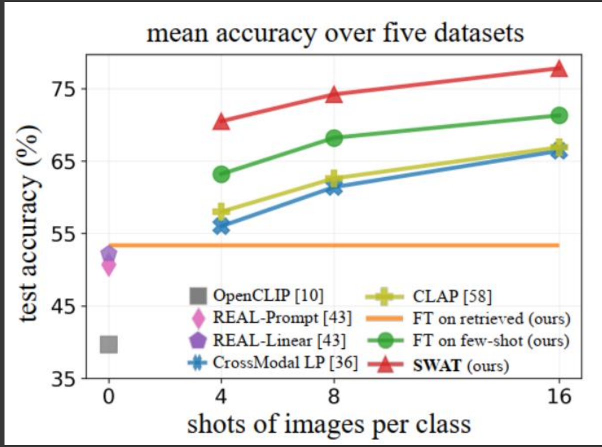
Exploit open-world foundation models for data annotation

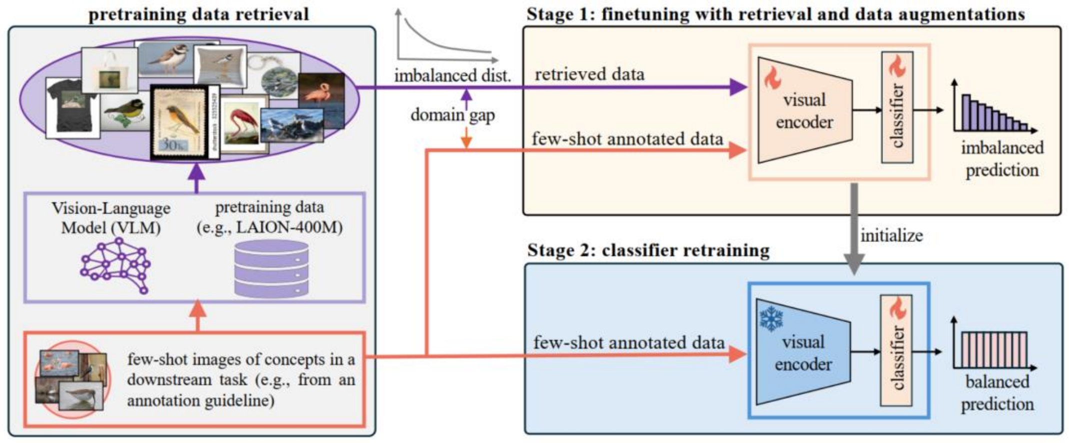




T. Liu, H. Zhang, S. Parashar, S. Kong, 'Few-Shot Recognition via Stage-Wise Augmented Finetuning', ArXiv:2406.11148, 2024

Exploit open-world foundation models for data annotation





T. Liu, H. Zhang, S. Parashar, S. Kong, ‘Few-Shot Recognition via Stage-Wise Augmented Finetuning’, ArXiv:2406.11148, 2024

Remarks

- Foundation models are open-world models!
- Open-world models have open issues, which are yet to be discovered!
- Attention to open-world data!