

VPLOW@CVPR'24

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



Shu Kong

Texas A&M University

University of Macau

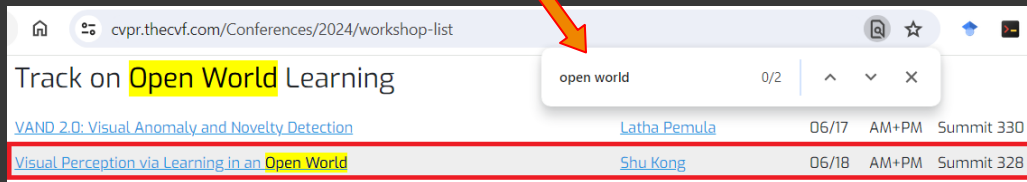
June 18, 2024

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://vpflow.github.io/vpflow_4th.html



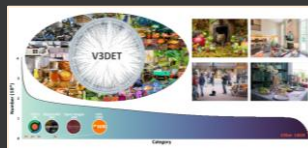
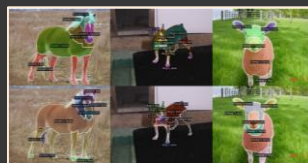
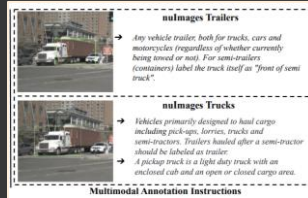
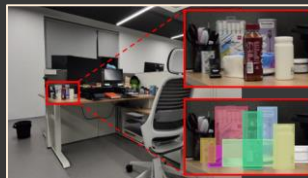
Track on **Open World** Learning




[VAND 2.0: Visual Anomaly and Novelty Detection](#) [Latha Pemula](#) 06/17 AM+PM Summit 330

[Visual Perception via Learning in an Open World](#) [Shu Kong](#) 06/18 AM+PM Summit 328

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 Boulton
University of Colorado
Colorado Springs



Walter J. Scheirer
University of Notre Dame

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

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

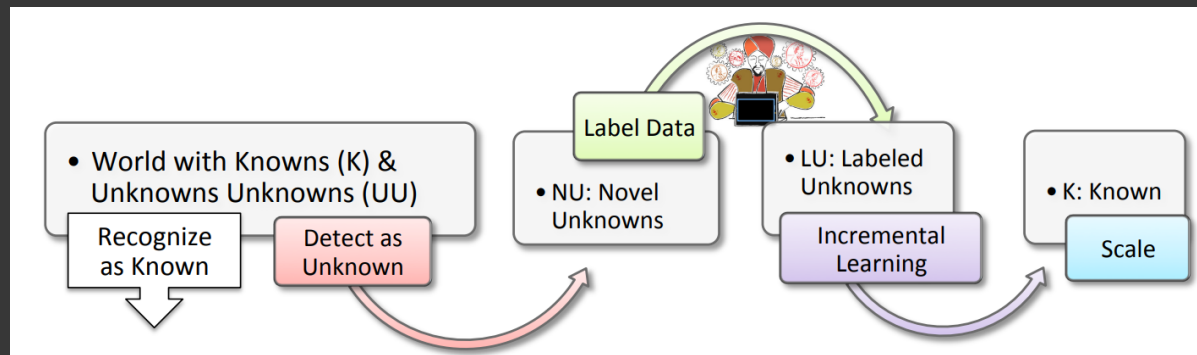


Yunhan Zhao
UC Irvine

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



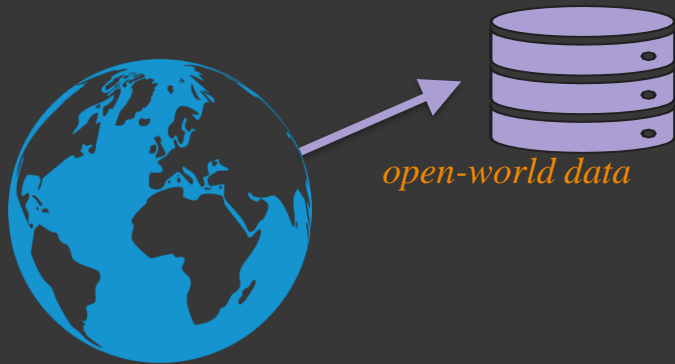
A brief introduction to the open world

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.



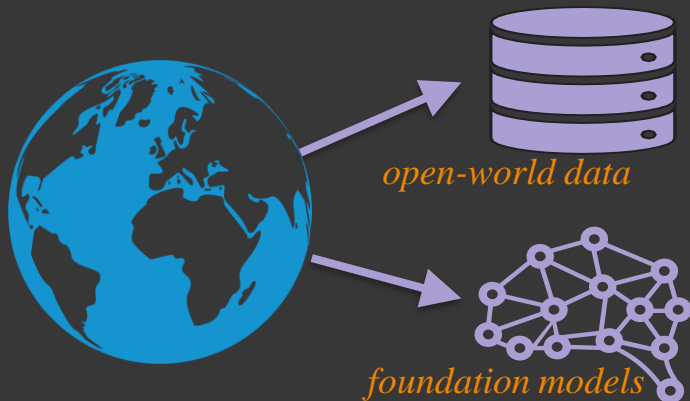
A brief introduction to the open world

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*

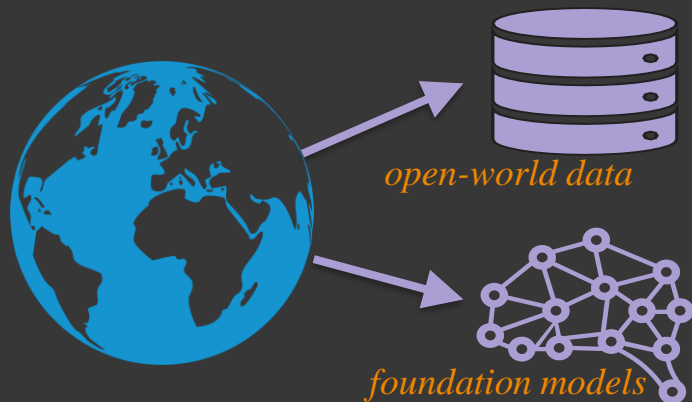
- *data* sampling from the open world, e.g., sampling outlier data for better OSR.
- *foundation models* pretrained in the open world, e.g., CLIP.



Foundation models are open-world models, enabling:

- open-vocabulary recognition
- zero-shot recognition
-

Can we predict testing performance with distribution shift?



Can we predict testing performance with distribution shift?

ImageNet / in-distribution

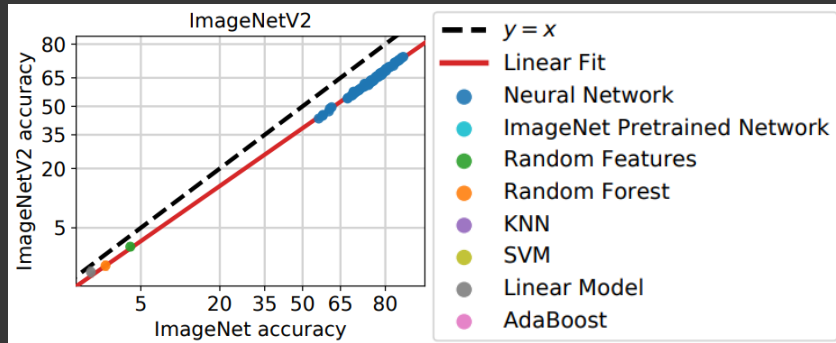


mongoose

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.



Can we predict testing performance with distribution shift?

ImageNet / in-distribution



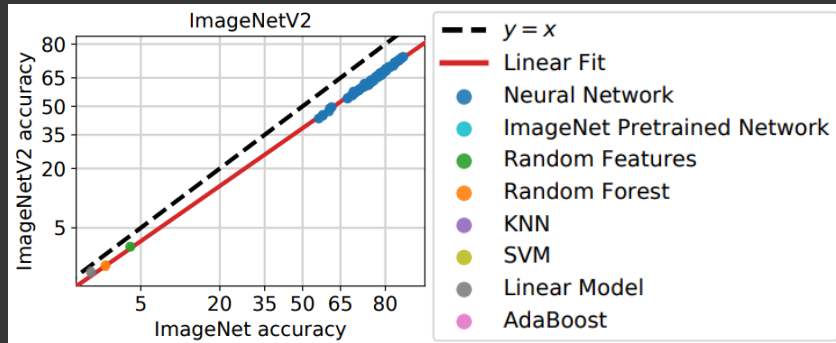
mongoose

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.

But the models are trained in the closed world! ☹️



Can we predict testing performance with distribution shift?

ImageNet / in-distribution

ImageNetV2 / out-of-distribution

mongoose

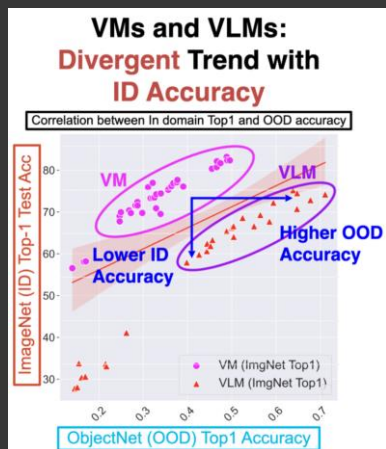
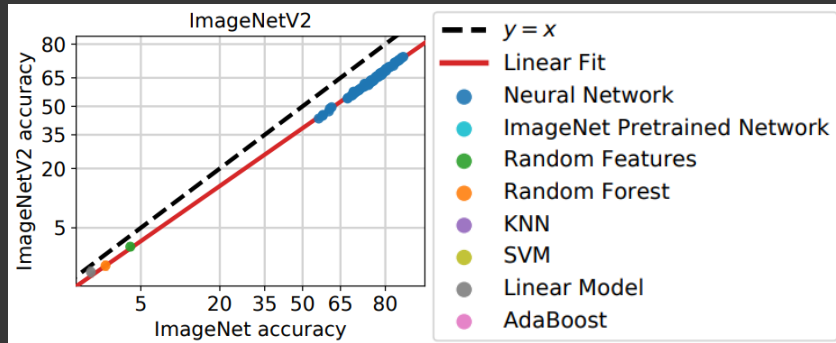


Accuracy-on-the-line: empirically, OOD performance is strongly correlated with in-distribution performance for a wide range of models and distribution shifts.

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! 😞



Can we predict testing performance with distribution shift?

ImageNet / in-distribution



mongoose

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)*.

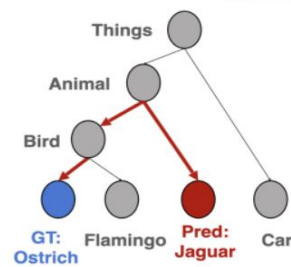
Accuracy-on-the-line does NOT hold anymore! 😞



foundation models

Taxonomy distance as a measurement of semantic severity of mistake

Class hierarchy, like WordNet



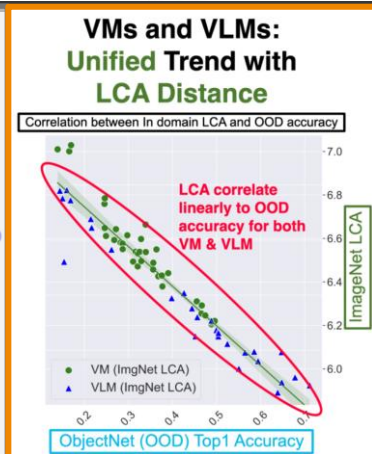
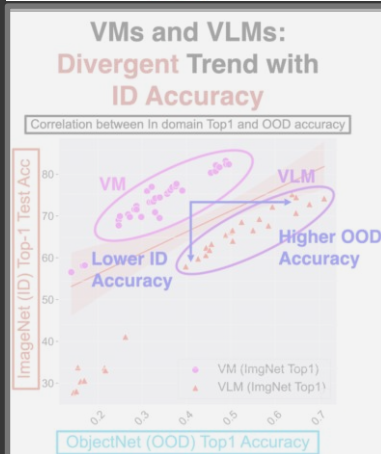
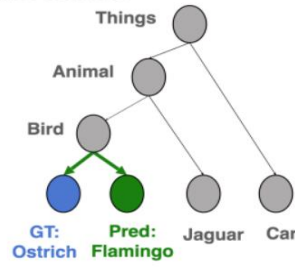
LCA=2



Lower LCA is Better!



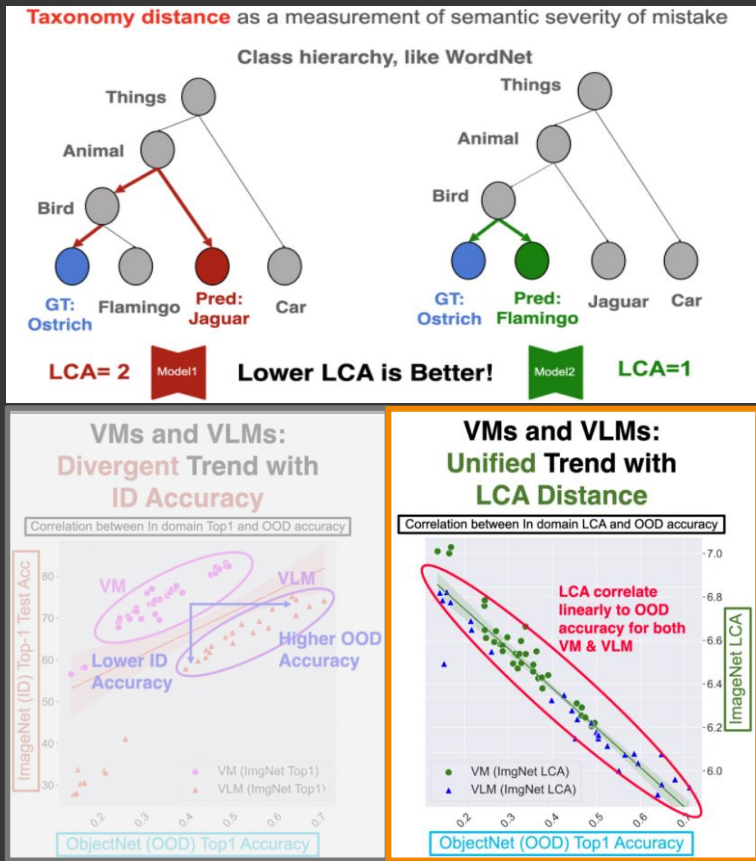
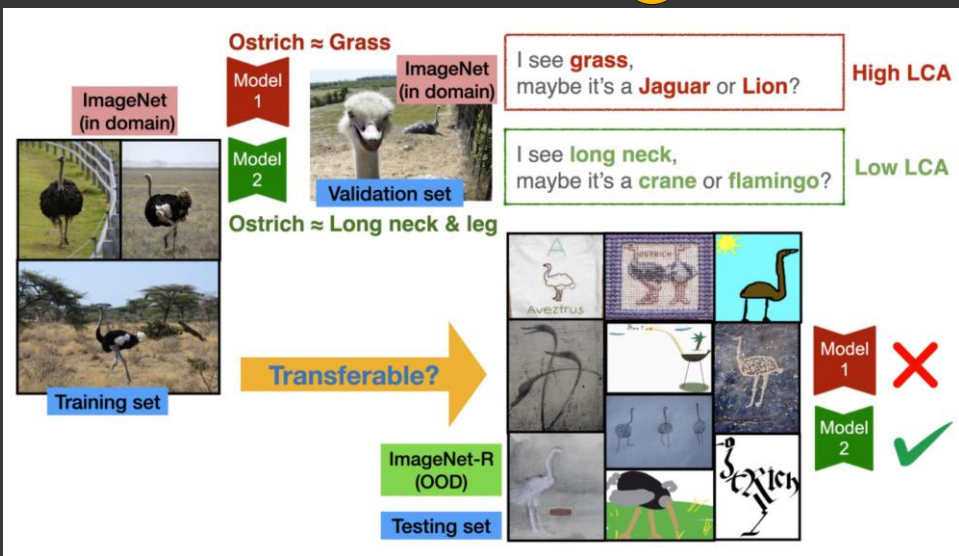
LCA=1



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! 😊



An interesting failure case

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



Model: GPT-4

S

Can you list the key visual attributes of Ponana Citrina in a photo?



I'm sorry, but as of my last training cut-off in September 2021, I don't have any specific information about "Ponana Citrina". This could be due to a misspelling or it might be a new term, species, or variant that has been described after my training data. If it's a plant, animal, or any other object, you might be able to identify it by specific characteristics such as color, shape, size, patterns, texture, etc. If you have more information or context, please provide it and I'll do my best to assist you based on the general knowledge I have.

An interesting failure case

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



Model: GPT-4

S Can you list the key visual attributes of Ponana Citrina in a photo?



I'm sorry, but as of my last training cut-off in September 2021, I don't have any specific information about "Ponana Citrina". This could be due to a misspelling or it might be a new term, species, or variant that has been described after my training data. If it's a plant, animal, or any other object, you might be able to identify it by specific characteristics such as color, shape, size, patterns, texture, etc. If you have more information or context, please provide it and I'll do my best to assist you based on the general knowledge I have.

BugGuide Identification, Images, & Information For Insects, Spiders & Their Kin For the United States & Canada

Home | Guide | ID Request | Recent | Frass | Forums | Donate | Help

Clickable Guide

Taxonomy | Browse | Info | Images | Links | Books | Data

Home » Guide » Arthropods (Arthropoda) » Hexapoda (Hexapoda) » Insects (Insecta) » True Bugs, Cicadas, Hoppers, Aphids and Allies (Hemiptera) » True Hoppers (Cicadelloidea) » Leafhoppers and Treehoppers (Cicadomorpha) » Leafhoppers and Treehoppers (Membracidae) » Typical Leafhoppers (Cicadellidae) » subgenus Ponana (Ponana subgenus Ponana) » citrina Group (Ponana citrina Group) » Ponana citrina

Species Ponana citrina

Classification · Synonyms and other taxonomic changes · Explanation of Names · Size · Identification · Range · Remarks · Works Cited

Classification

Kingdom Animalia (Animals)
Phylum Arthropoda (Arthropods)
Subphylum Hexapoda (Hexapoda)
Class Insecta (Insects)
Order Hemiptera (True Bugs, Cicadas, Hoppers, Aphids and Allies)
Suborder Auchenorrhyncha (True Hoppers)
Infraclass Cicadomorpha (Cicadas, Spittlebugs, Leafhoppers, and Treehoppers)
Superfamily Membracodea (Leafhoppers and Treehoppers)
Family Cicadellidae (Typical Leafhoppers)
Subfamily Tassiniae
Tribe Gypsonini
Genus Ponana
No Taxon (subgenus Ponana)
No Taxon (citrina Group)
Species citrina (Ponana citrina)

Synonyms and other taxonomic changes

Orig. Comb: *Gypona citrina* Spångberg 1878

Explanation of Names

Ponana citrina (Spångberg 1878)

Size

♂ 8mm, ♀ 9.75mm

Identification

"A yellowish species tinged with orange and often with black spots on elytra. Vertex broadly rounded, more than twice as wide as median length. Green to yellow black spots usually arranged in three transverse rows extending across elytra." — DeLong, 1942

Calendar

Upcoming Events

Registration is open for the 2024 BugGuide gathering in Idaho July 24-27

Moth submissions

from National Moth Week 2023

Photos of insects and

speciee from the

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.



Model: GPT-4

S

Can you list the key visual attributes of Ponana Citrina in a photo?



I'm sorry, but as of my last training cut-off in September 2021, I don't have any specific information about "Ponana Citrina". This could be due to a misspelling or it might be a new term, species, or variant that has been described after my training data. If it's a plant, animal, or any other object, you might be able to identify it by specific characteristics such as color, shape, size, patterns, texture, etc. If you have more information or context, please provide it and I'll do my best to assist you based on the general knowledge I have.

BugGuide Identification, Images, & Information For Insects, Spiders & Their Kin For the United States & Canada

Home | Guide | ID Request | Recent | Frass | Forums | Donate | Help

Clickable Guide

Taxonomy | Browse | Info | Images | Links | Books | Data

Home » Guide » Arthropods (Arthropoda) » Hexapoda (Hexapoda) » Insects (Insecta) » True Bugs, Cicadas, Hoppers, Aphids and Allies (Hemiptera) » True Hoppers (Cicadas, Spittlebugs, Leafhoppers, and Treehoppers (Cicadomorpha) » Leafhoppers and Treehoppers (Membracidae) » Typical Leafhoppers (Cicadellidae) » subgenus Ponana (Ponana subgenus Ponana) » citrina Group (Ponana citrina Group) » Ponana citrina

Species Ponana citrina

Classification · Synonyms and other taxonomic changes · Explanation of Names · Size · Identification · Range · Remarks · Works Cited

Classification

Kingdom Animalia (Animals)
Phylum Arthropoda (Arthropods)
Subphylum Hexapoda (Hexapoda)
Class Insecta (Insects)
Order Hemiptera (True Bugs, Cicadas, Hoppers, Aphids and Allies)
Suborder Auchenorrhyncha (True Hoppers)
Infracorder Cicadomorpha (Cicadas, Spittlebugs, Leafhoppers, and Treehoppers)
Superfamily Membracodea (Leafhoppers and Treehoppers)
Family Cicadellidae (Typical Leafhoppers)
Subfamily Tassiniae
Tribe Gypsonini
Genus Ponana
No Taxon (subgenus Ponana)
No Taxon (citrina Group)
Species citrina (Ponana citrina)

Synonyms and other taxonomic changes

Orig. Comb: *Gypona citrina* Spångberg 1878

Explanation of Names

Ponana citrina (Spångberg 1878)

Size

♂ 8mm, ♀ 9.75mm

Identification

"A yellowish species tinged with orange and often with black spots on elytra. Vertex broadly rounded, more than twice as wide as median length. Green to yellow black spots usually arranged in three transverse rows extending across elytra." — DeLong, 1942

Calendar

Upcoming Events

Registration is open for the 2024 BugGuide gathering in Idaho July 24-27

Moth submissions from National Moth Week 2023

Photos of insects and people from the 2022 BugGuide gathering in New Mexico, July 20-24

Photos of insects and people from the

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

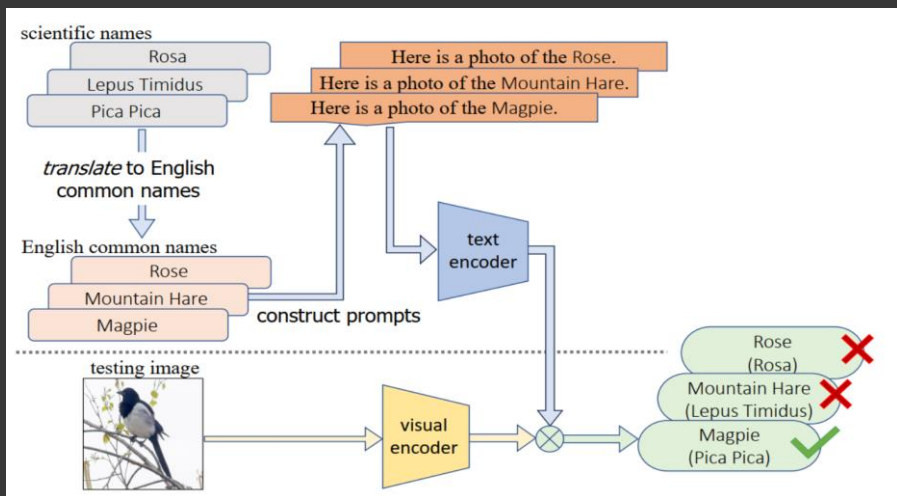
An interesting failure case

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

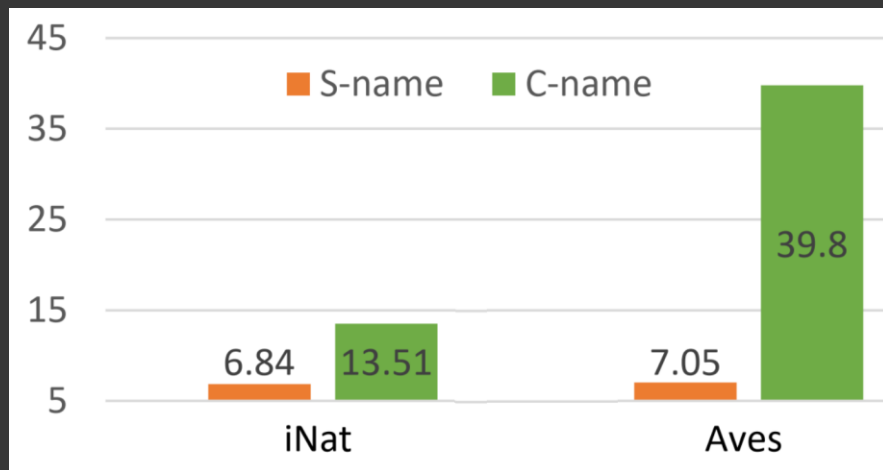
Hypothesis: Some concepts (esp. scientific names in Latin) are too few in the open world to train models.

Justification: We count!

Remedy: Translating Latin scientific names to English common names.



zero-shot accuracy by prompting OpenCLIP





An interesting failure case

Why do foundation models fail to handle some concepts?


Hypothesis: certain concepts are insufficiently presented 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**

Arch 4A-E Poster #324

Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

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.


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**

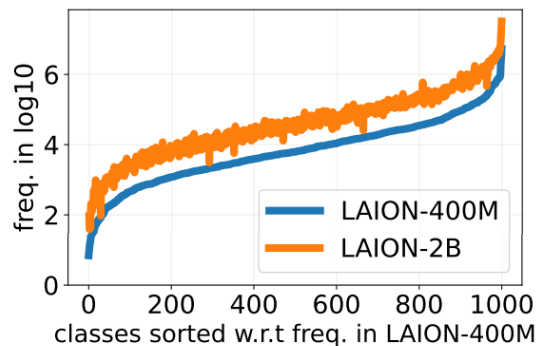



✗

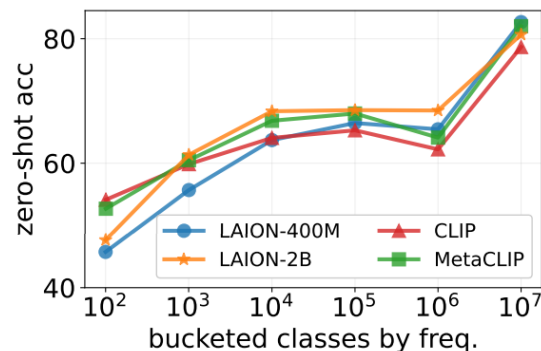



✗

(a) freq. of ImageNet concepts



(b) freq. vs. zero-shot accuracy



Arch 4A-E Poster #324

Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

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



Los Angeles Times
collectible Nike sneakers cost ...



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



Adidas
adidas LA Trainer Shoes - ...

Linguistic ambiguity



tiger



tiger shark



Tiger Woods

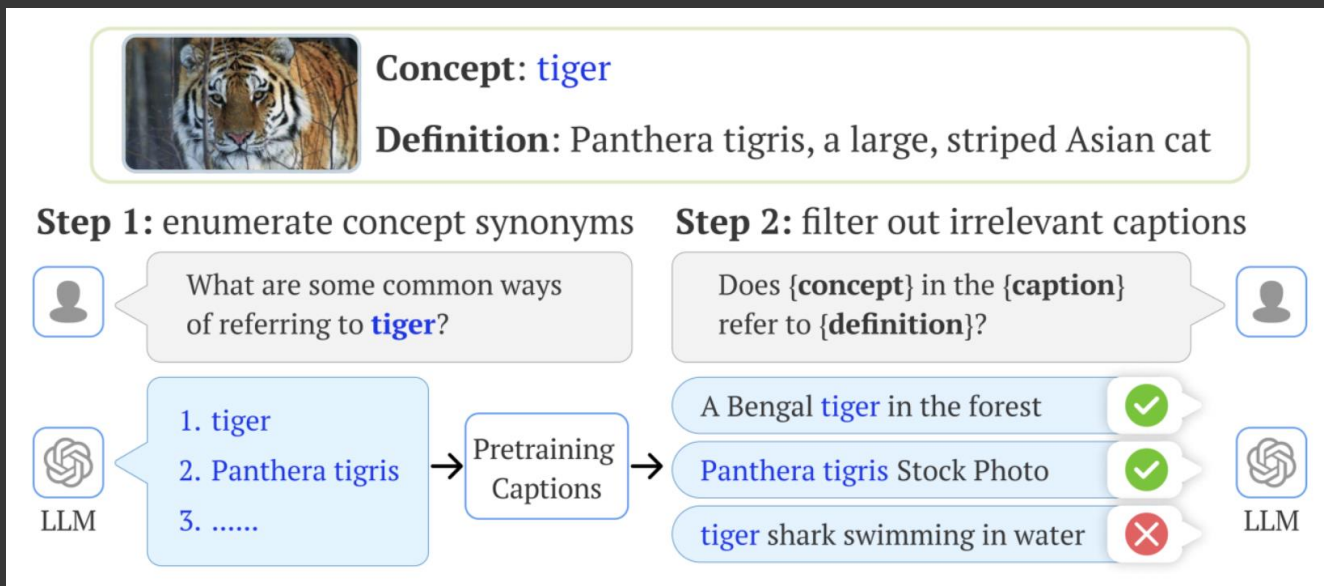
Measure concept frequency

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!*

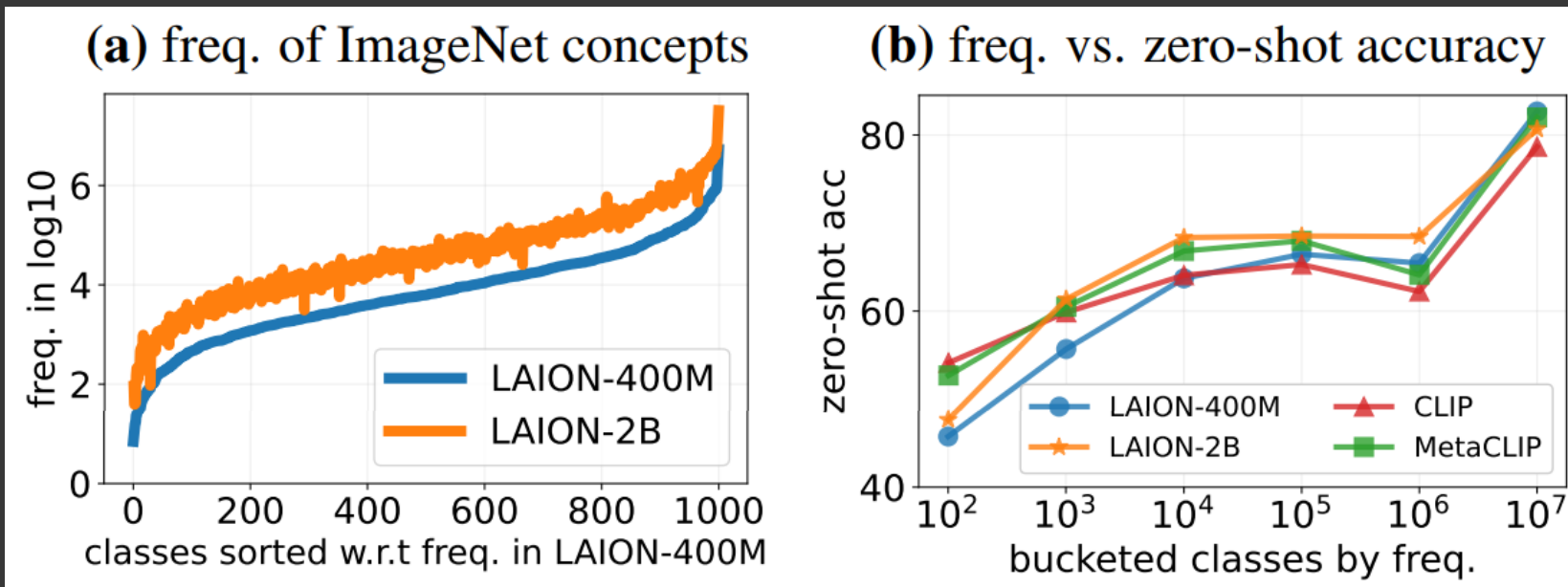
Lexical variation, e.g., synonyms

Linguistic ambiguity



Measure concept frequency

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

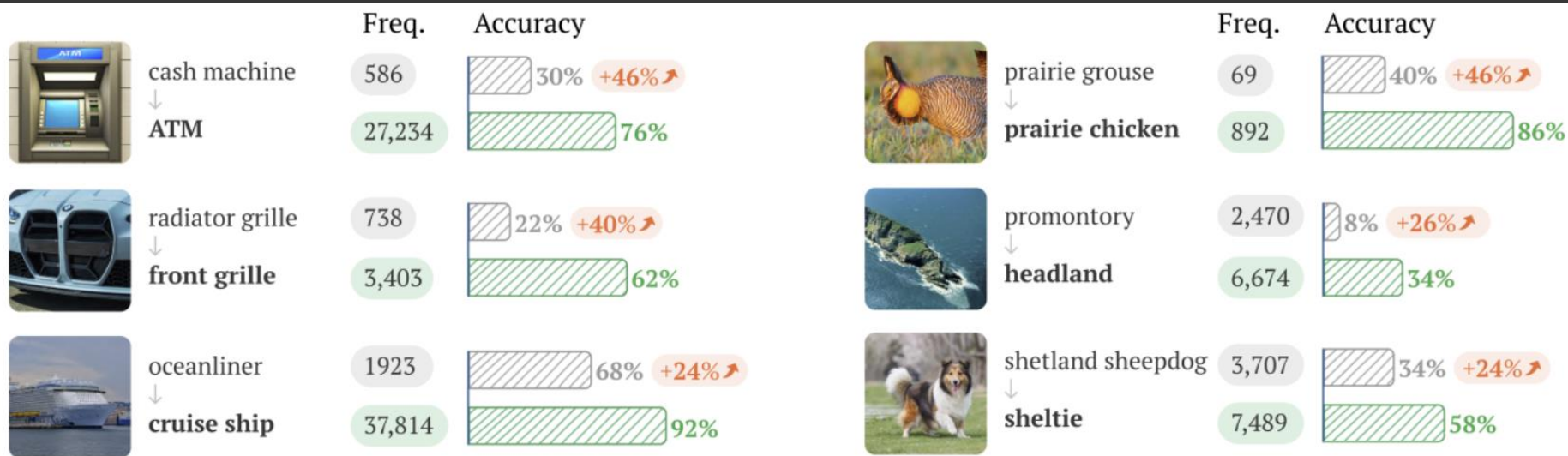


Arch 4A-E Poster #324

Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

Insight 1: prompt VLM using the most frequent synonym

This simple change significantly boosts zero-shot accuracy!

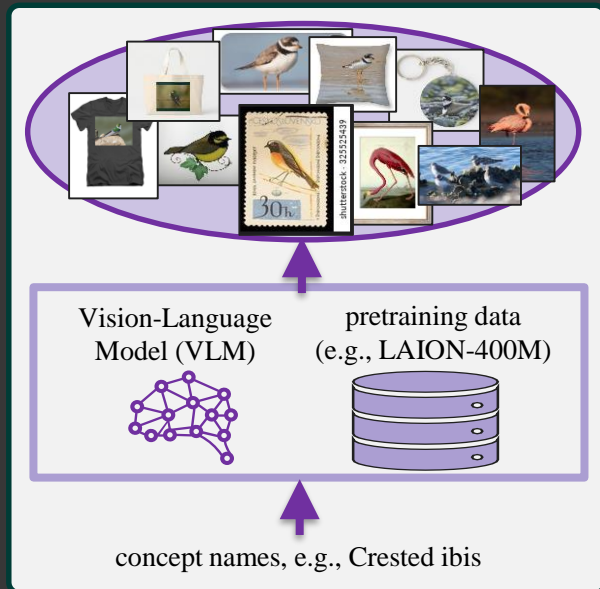


Arch 4A-E Poster #324

Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

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

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

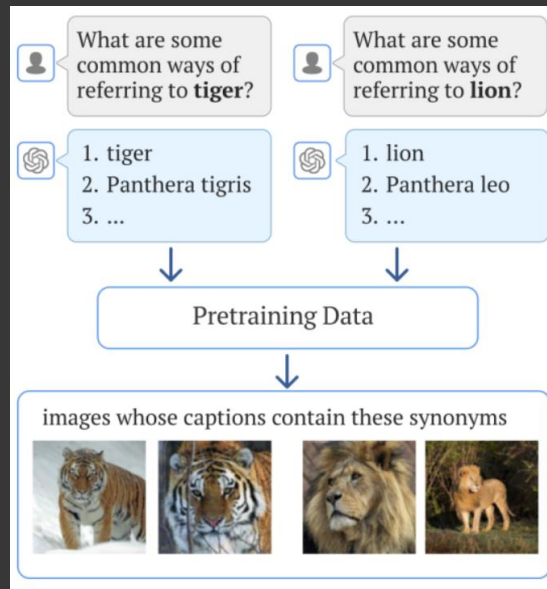
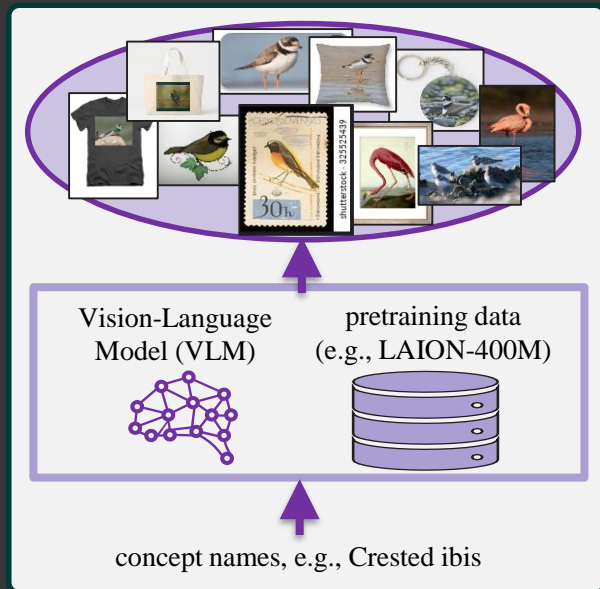


[REACT] H. Liu, K. Son, J. Yang, C. Liu, J. Gao, YJ Lee, C. Li. "Learning customized visual models with retrieval-augmented knowledge", CVPR, 2023

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

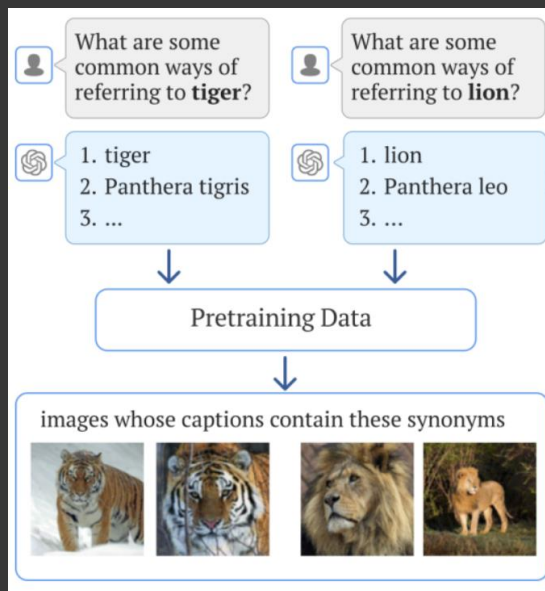
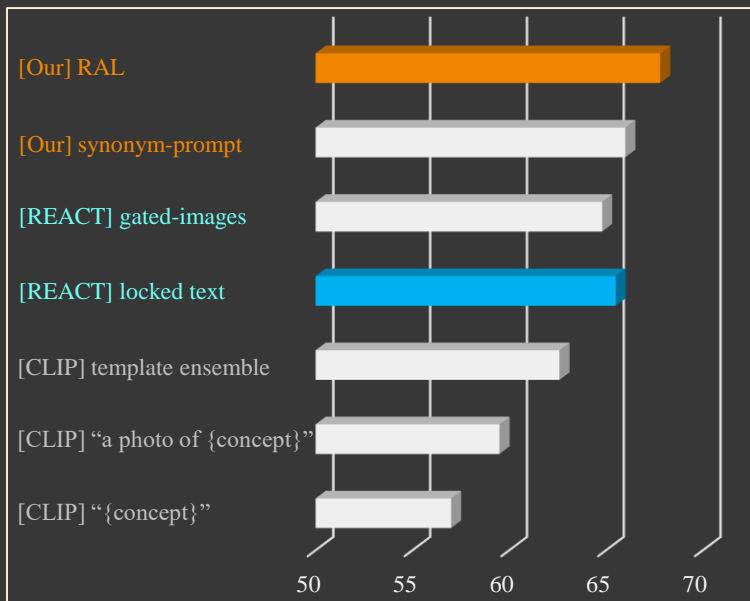
[Our] S. Parashar, Z. Lin, T. Liu, X. Dong, Y. Li, D. Ramanan, J. Caverlee, S. Kong, "The Neglected Tails of Vision-Language Models", CVPR, 2024

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

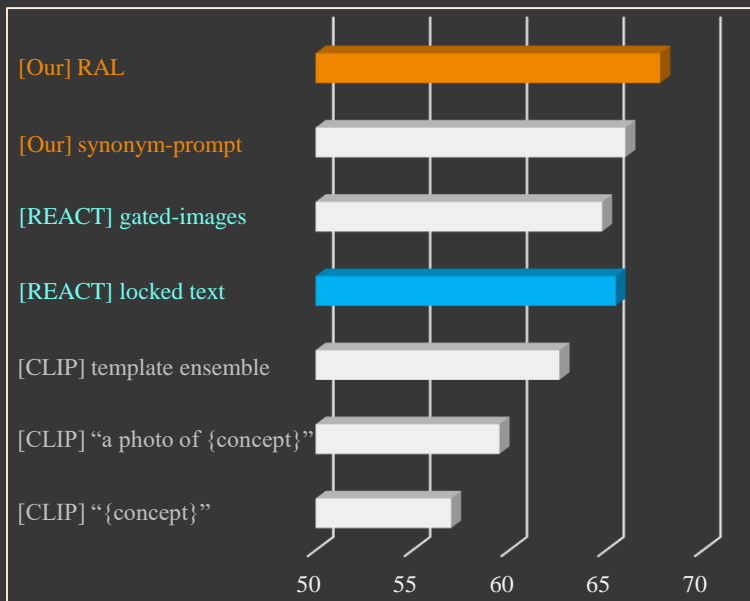
[CLIP] A. Radford, et al. "Learning transferable visual models from natural language supervision." ICML, 2021

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.


















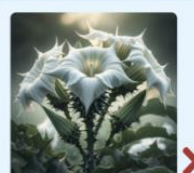










Stage	Resource	REACT	[Our] RAL	Relative Cost
Retrieval	retrieved examples	400M	0.5M	0.1%
	time	200 hrs	6 hrs	3%
	storage	10 TB	25 GB	0.25%
Learning	training images	10M	0.5M	5%
	time	256 hrs	2 mins	0.01%
	# of learned parameters	87M	0.5M	0.6%
	GPU memory	256 GB	2 GB	0.8%

Arch 4A-E Poster #324

Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

Insight 3: use the most frequent synonym in image generation

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

Arch 4A-E Poster #324

Fri 21 Jun 1:30 a.m. CST — 3 a.m. CST

Remarks

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

