The Neglected Tails in Vision-Language Models

Shubham Parashar* Zhiqiu Lin* Tian Liu* Xiangjue Dong Yanan Li Deva Ramanan James Caverlee Shu Kong

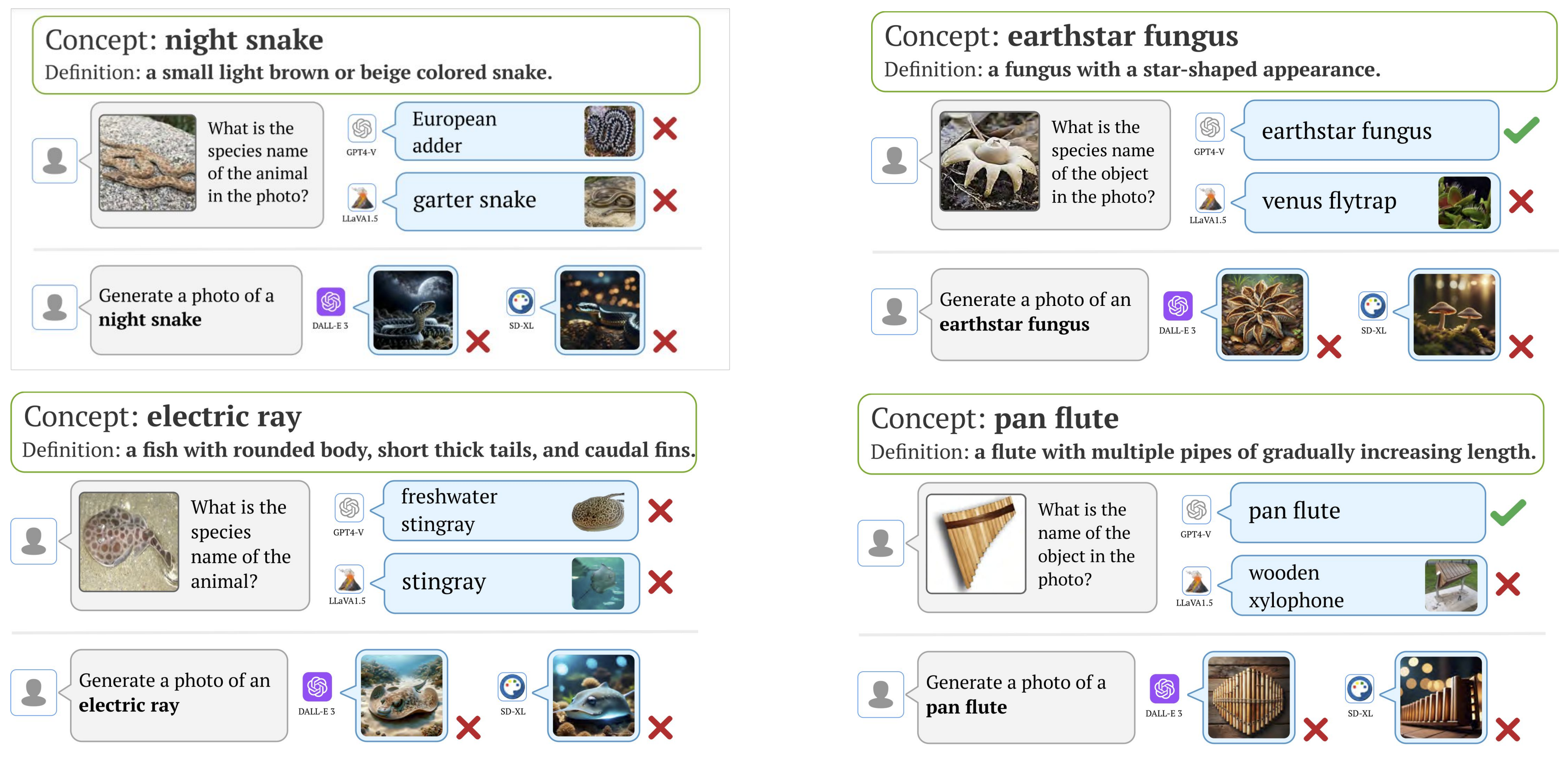
CVPR JUNE 17-21, 2024 SEATTLE, WA

Motivation

Challenges

1. Vision-language models (VLMs) excel in zero-shot recognition but their performance varies across visual concepts. For example, CLIP achieves 60-80% accuracy on ImageNet but drops to <10% for concepts like `night snake`, presumably due to limited presence in VLM's pretraining data.
2. We discover rare concepts in the VLM's pretraining data that downstream applications, such as multimodal chatbots (e.g., GPT-4Vision) and text-to-image models (e.g., Stable Diffusion), fail to recognize or generate.

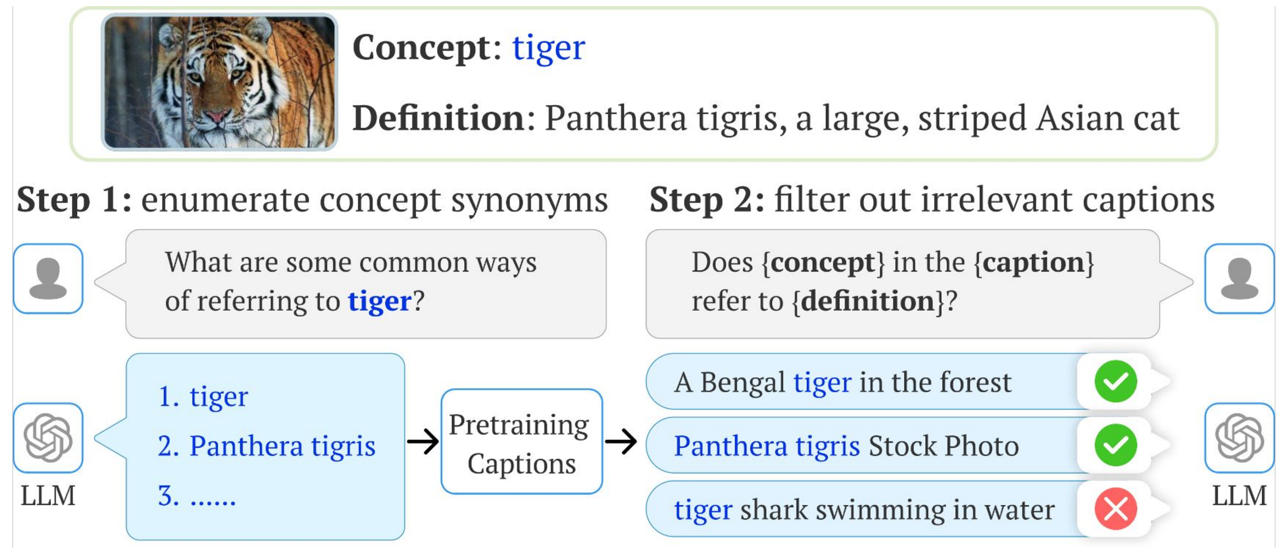
Failures of VLM-based applications on rare concepts



Measuring Concept Frequency in Pretraining Data

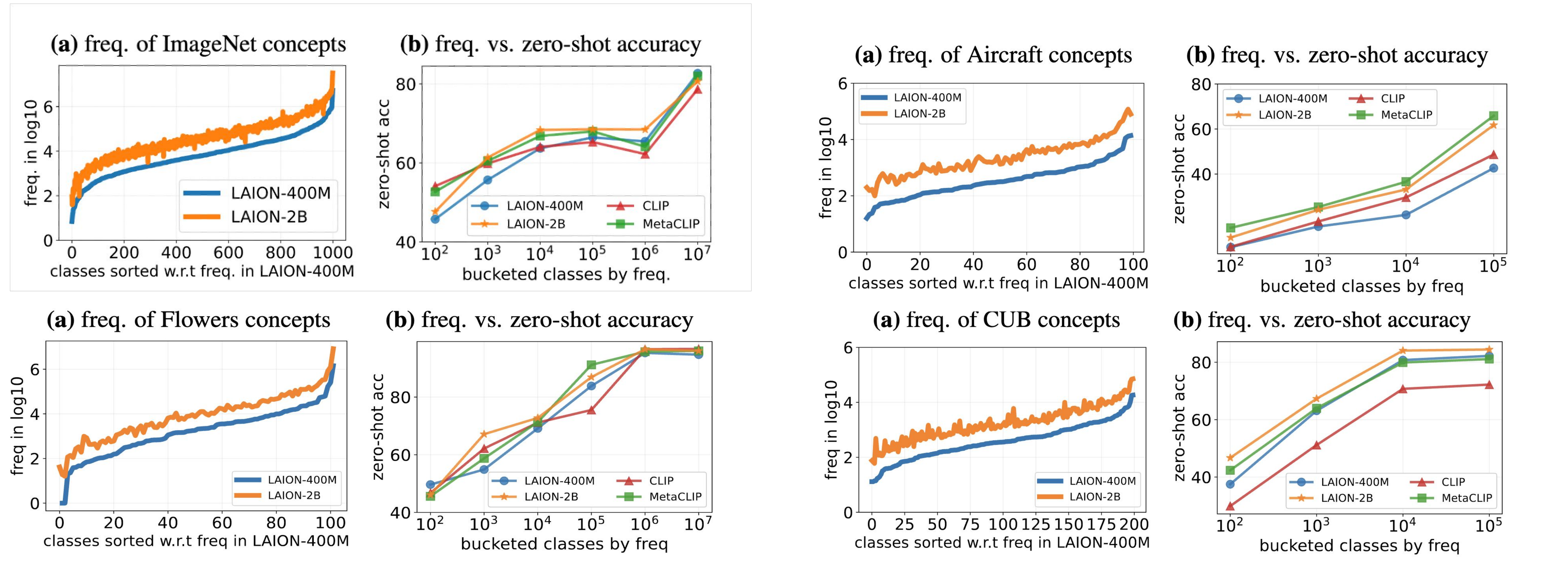
Approach

1. Use an LLM to enumerate all synonyms for a given visual concept.
2. Retrieve all pretraining captions that contain any of these synonyms.
3. Filter out irrelevant captions with an LLM like LLaMA-2.
4. Count the number of relevant captions as concept frequency.



The Long-Tail and Biased Performance

Visual concepts in VLM pretraining data (LAION-400M and LAION-2B) exhibit a long-tailed distribution, strongly correlating with the biased zero-shot accuracy.

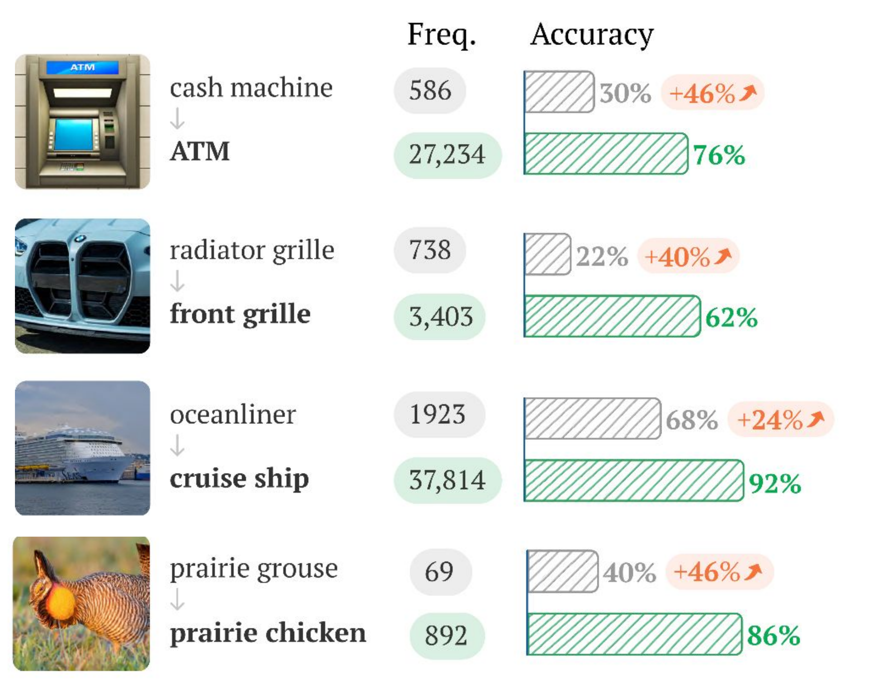


How to Address the Biases?

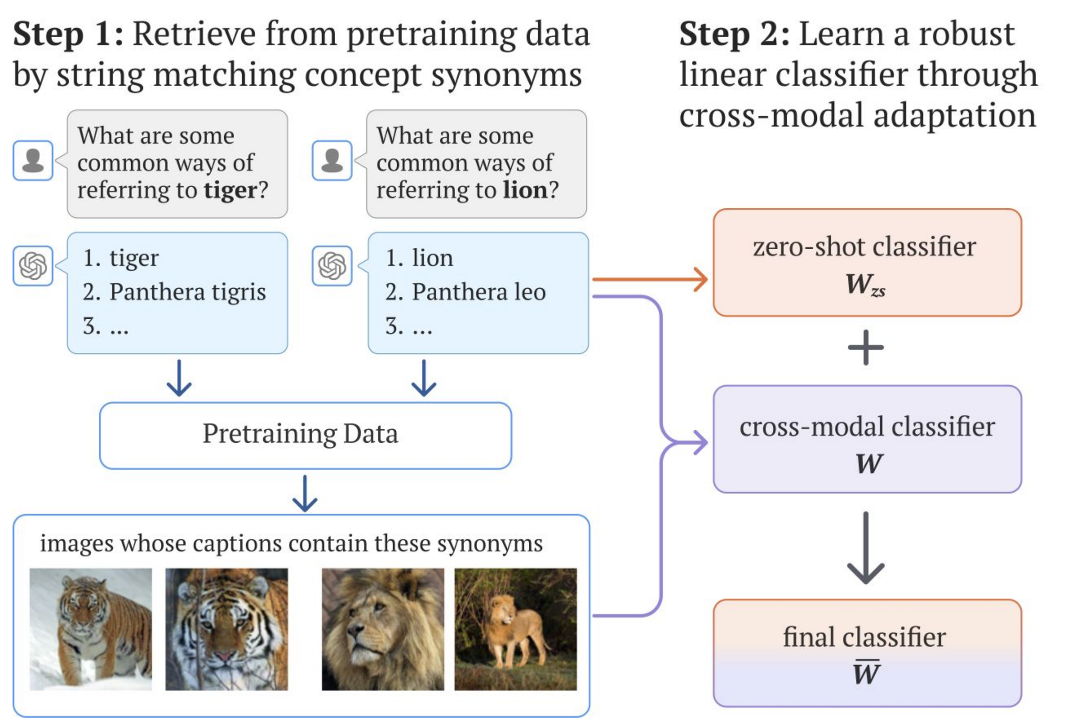
Motivation:

To mitigate the biased zero-shot classification performance of VLMs, we propose **RE**trieval **A**ugmented **L**earning. REAL consists of two novel solutions, a prompt-based approach and a retrieval augmented strategy.

REAL-prompt prompts with the most frequent synonyms found in the pretraining captions.



REAL-linear retrieves relevant pretraining images in the VLM's pretraining dataset, and learns a linear classifier on such data.



Results

State-of-the-art zero-shot recognition performance

	Method	ImageNet	Flowers	Cars	Aircraft	Pets	Food	DTD	EuroSAT	Avg
Zero-Shot Prompting	prompt template									
	"{*concept*}"	60.7	63.8	78.1	12.6	83.3	80.1	48.8	28.6	57.0
	"*a photo of* {*concept*}"	62.5	66.5	77.2	15.8	84.0	80.3	52.8	36.6	59.5
	OpenAI templates	62.9	68.0	79.2	16.7	86.7	80.9	54.5	51.5	62.6
	DCLIP [1]	62.1	—	—	—	84.6	80.1	51.9	36.8	—
	CuPL [2]	**63.7**	65.8	80.0	17.8	87.4	79.5	59.1	—	—
	REAL-Prompt	63.6	**76.6**	**82.7**	**18.0**	**88.8**	**81.0**	**59.9**	**57.5**	**66.0**
Retrieval Augmented	REACT (10K) [3]									
	Locked-Text	65.7	73.1	**88.5**	24.5	89.2	81.8	49.8	51.1	65.5
	Gated-Image	64.2	72.3	88.1	24.8	89.5	**83.0**	51.4	45.4	64.8
	REAL-Linear (500)	**65.9**	**78.8**	84.4	**29.6**	**89.5**	81.4	**61.5**	**51.5**	**67.8**

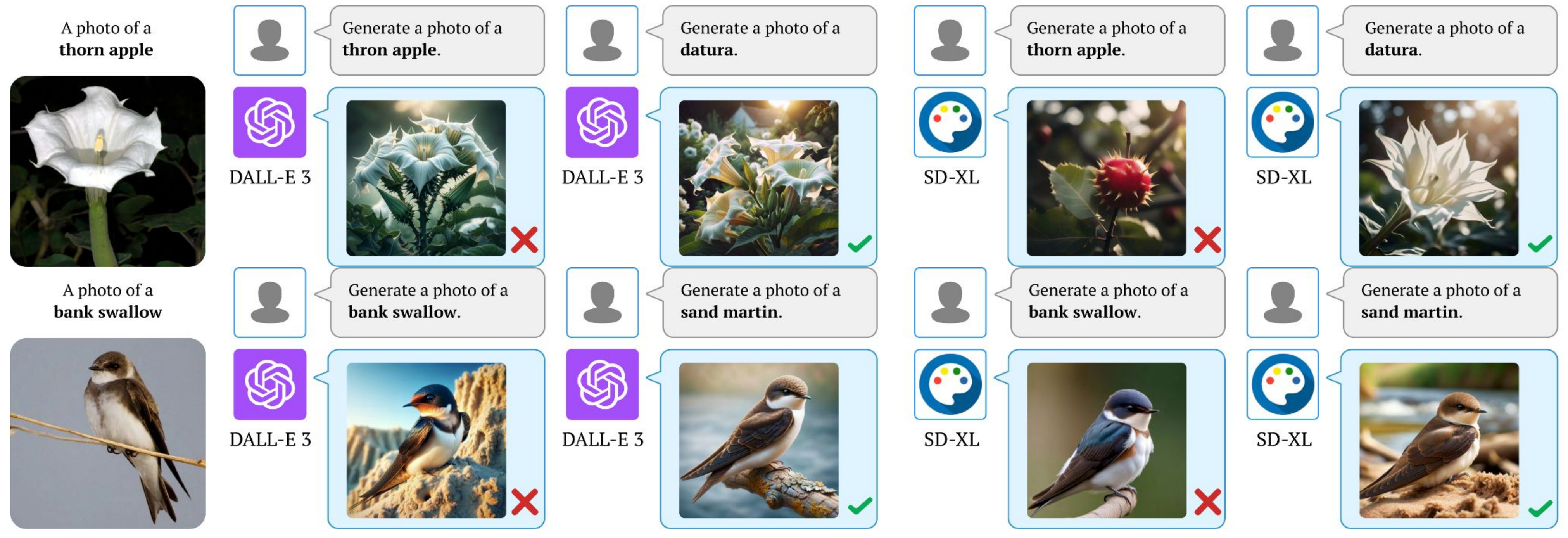
Exceptional efficiency

REAL-Linear is significantly more efficient than REACT [3], the previous state-of-the-art method of zero-shot recognition (which is based on retrieval augmented learning.

Stage	Resource	REACT [3]	REAL (500)	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%

Improving Text to Image Generation

We show that prompting SD-XL and DALLE-3 with the most frequent synonym found by REAL-Prompt leads to more accurate generations.



References

[1] Menon, Sachit, and Carl Vondrick (2023). "Visual classification via description from large language models." In: ICLR.
[2] Pratt, Sarah, et al (2023). "What does a platypus look like? generating customized prompts for zero-shot image classification." In: CVPR.
[3] Liu, Haotian, et al. (2023). "Learning customized visual models with retrieval-augmented knowledge." In: CVPR.