Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification

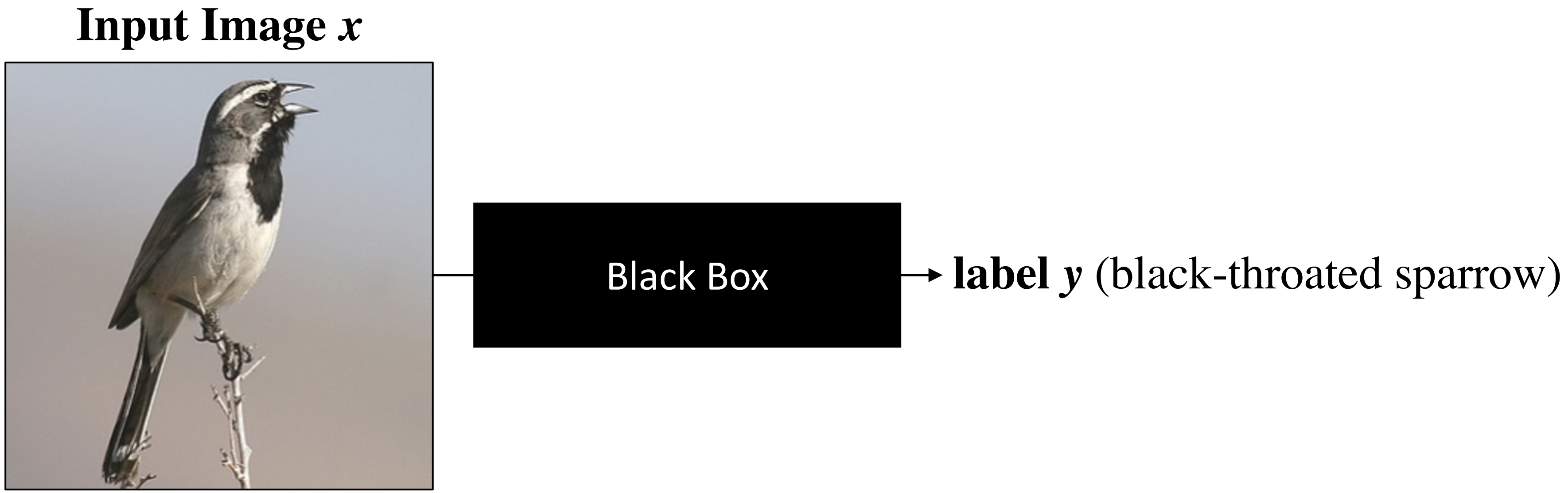
Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin,
Chris Callison-Burch, Mark Yatskar

University of Pennsylvania

Penn Engineering
University of Pennsylvania

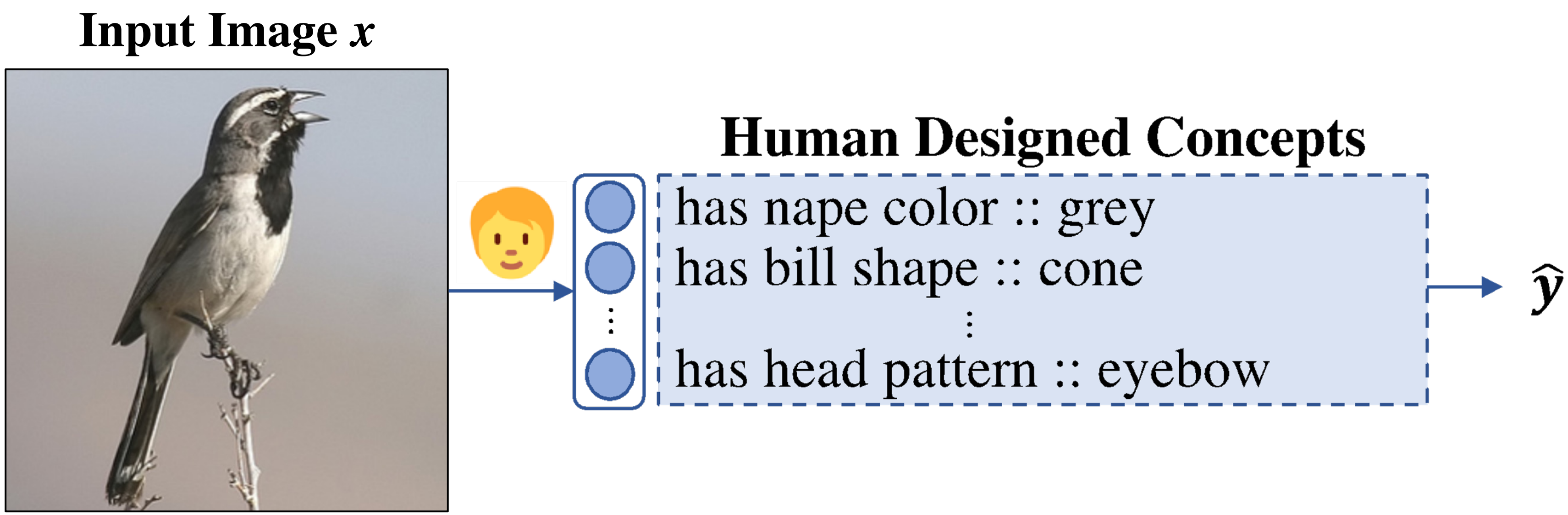
JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

End-to-end Neural Models



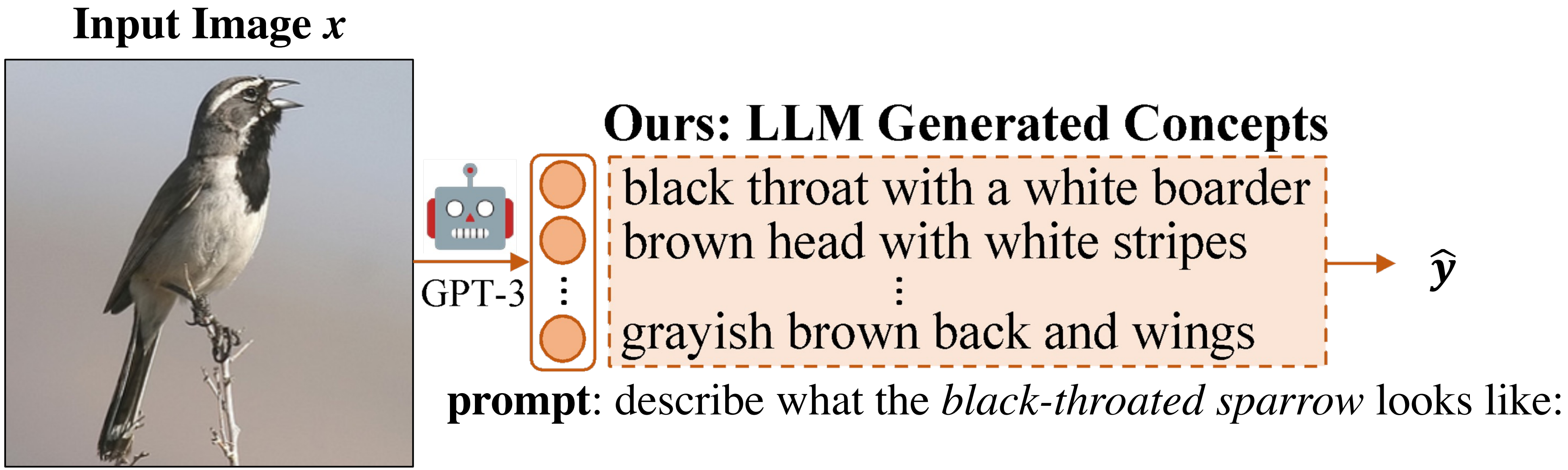
Concept Bottleneck Models (CBM)

Koh et al., *Proceedings of the 37th International Conference on Machine Learning,* 2020

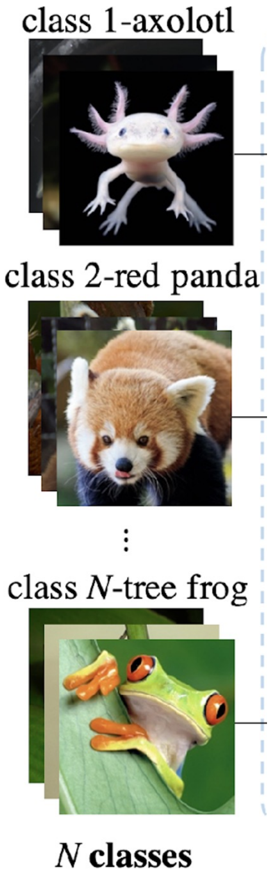


- Challenges:
  - Require heavy human annotation.
  - Underperform end-to-end models.

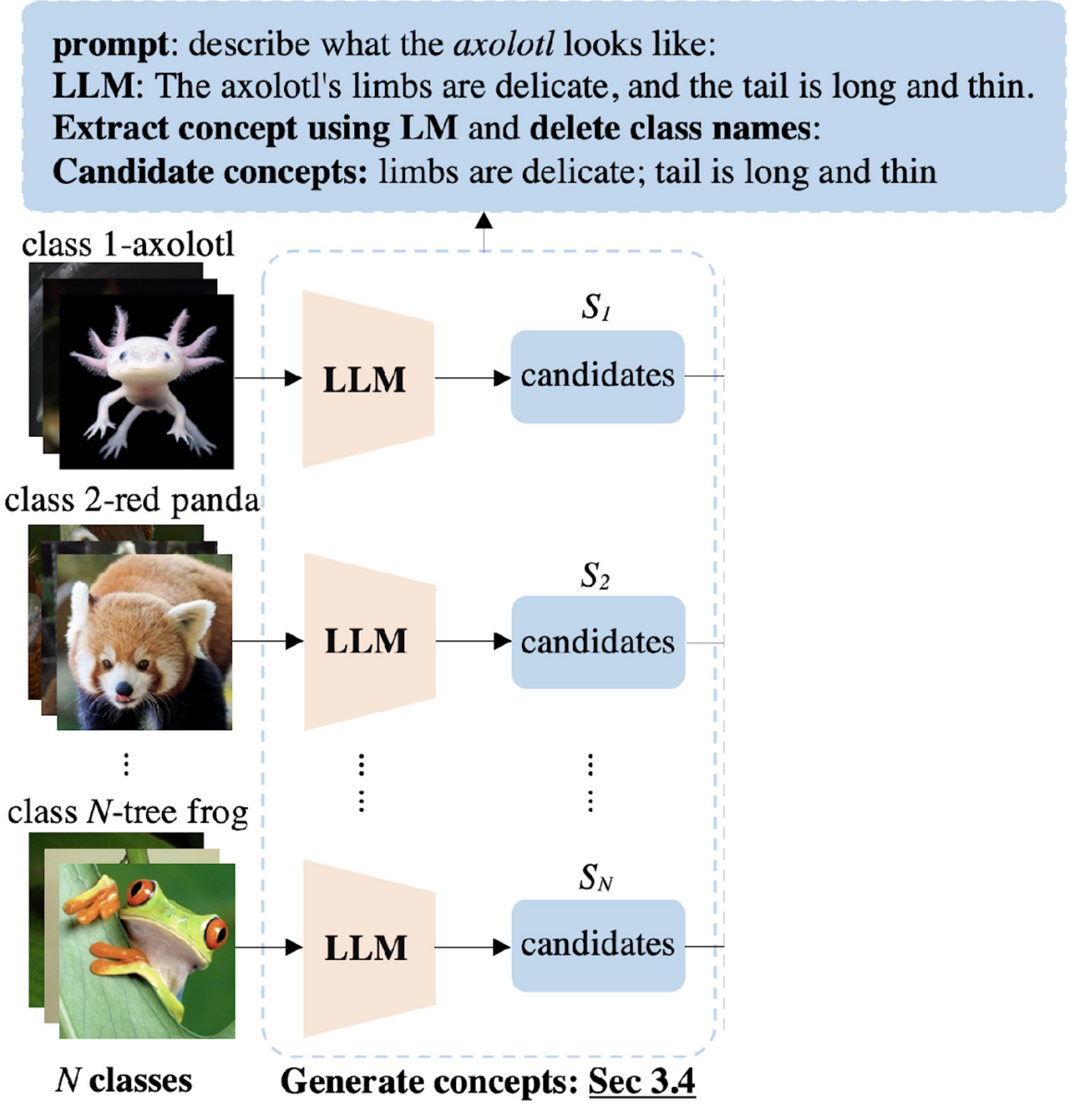
Language Model Guided Concept Bottlenecks



Prompt LLM to generate concepts

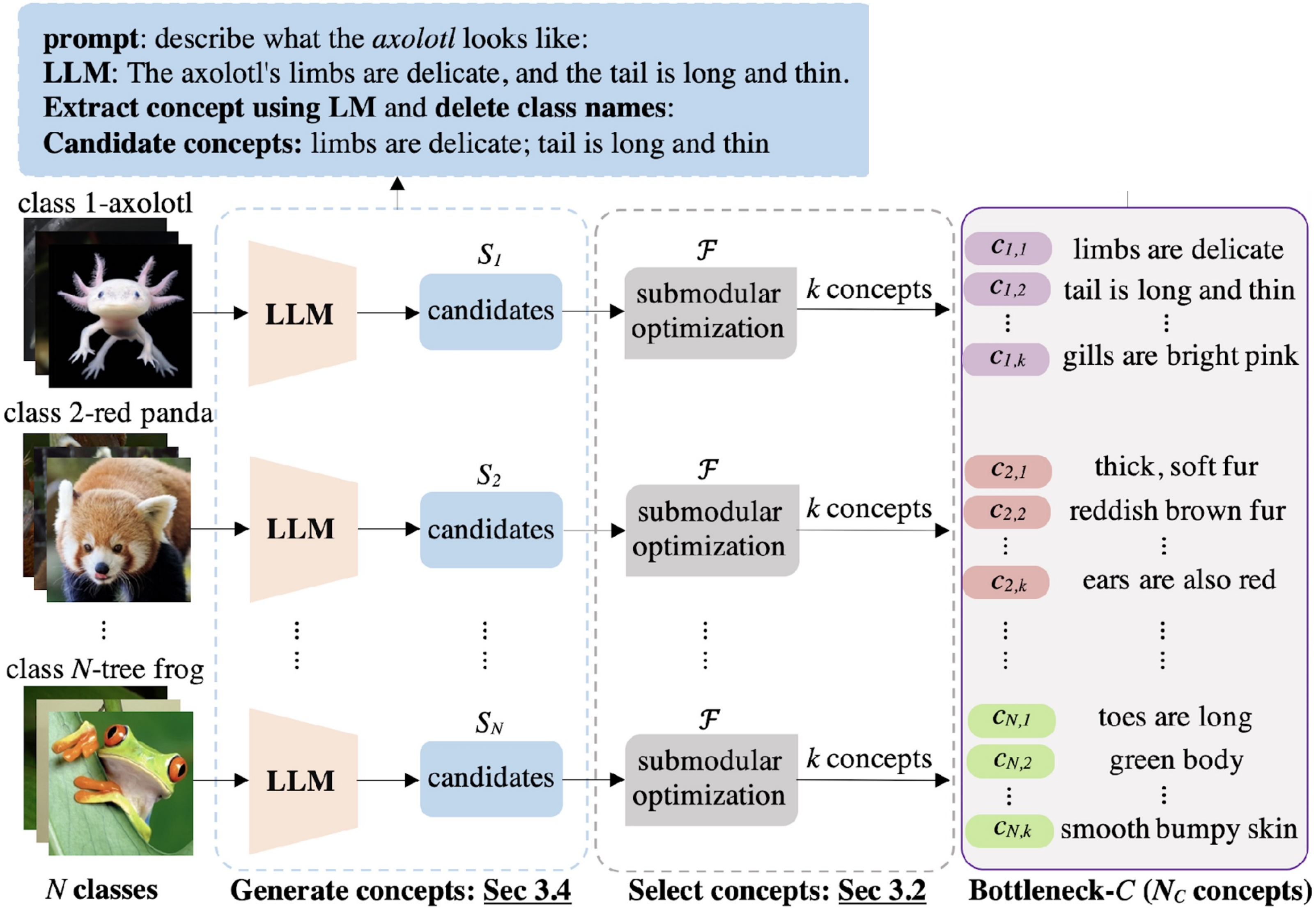


Submodular Concept Selection

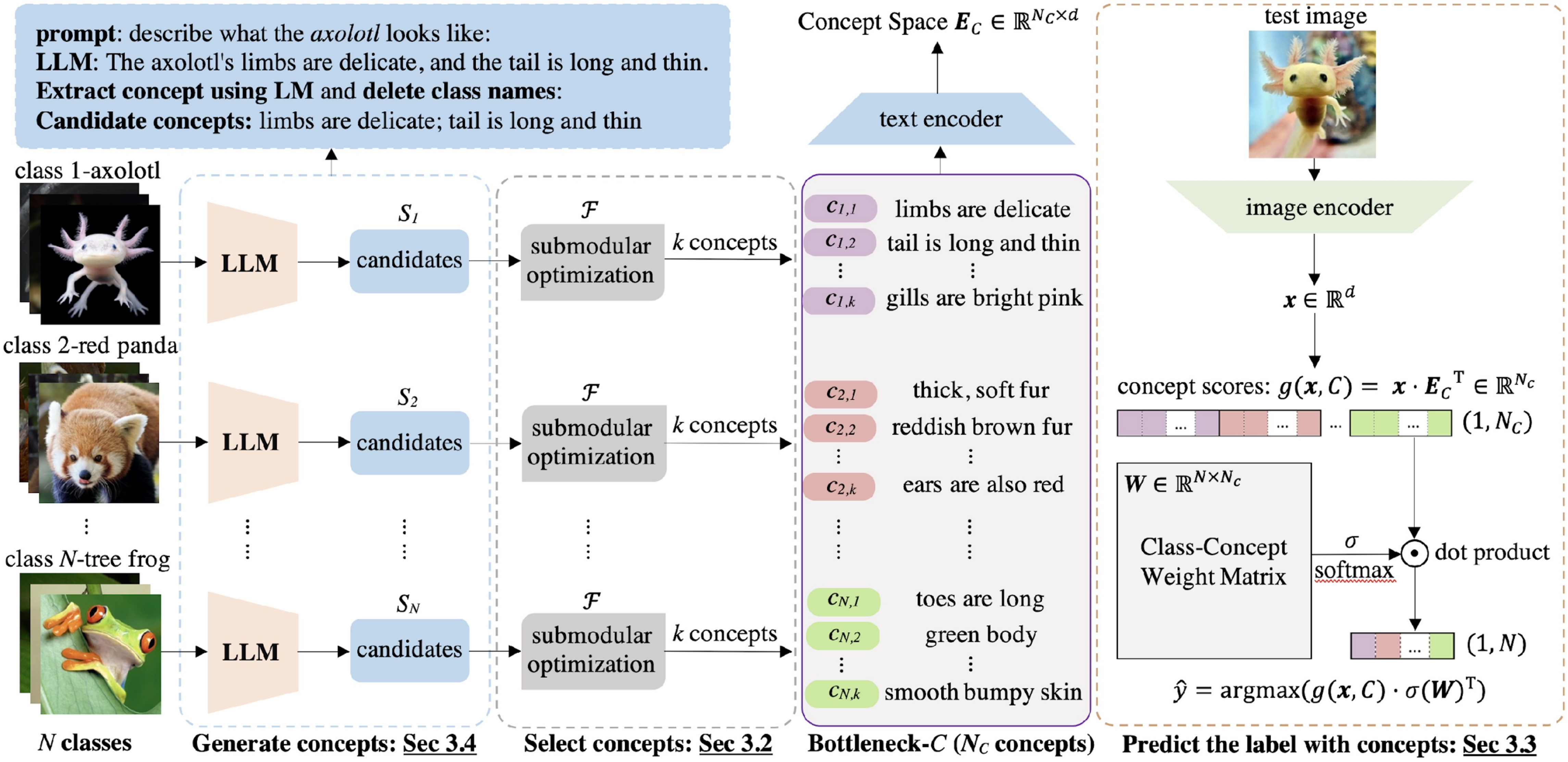


- Ensure the concepts selected for the bottleneck are **discriminative** and **diverse**.
  - *General concepts*: ~~This is an animal.~~
  - *Repetitive concepts*:
    Gills are bright pink./ ~~Pink gills.~~

Compute Concept Scores



Predict the Target

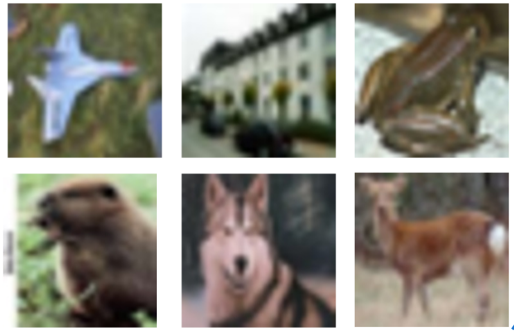


Datasets

Common Objects

ImageNet1K

CIFAR-10/CIFAR-100



Fine-grained Objects

Flower-102

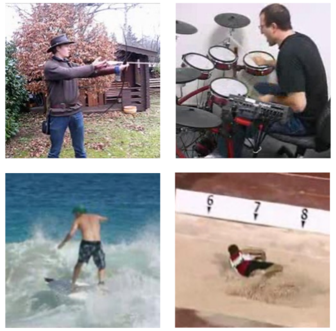
Food-101

Aircraft

CUB

Action

UCF-101

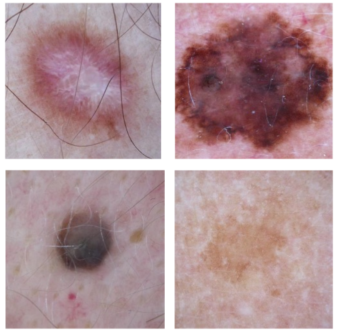


Textures

DTD

Skin Tumors

HAM10000



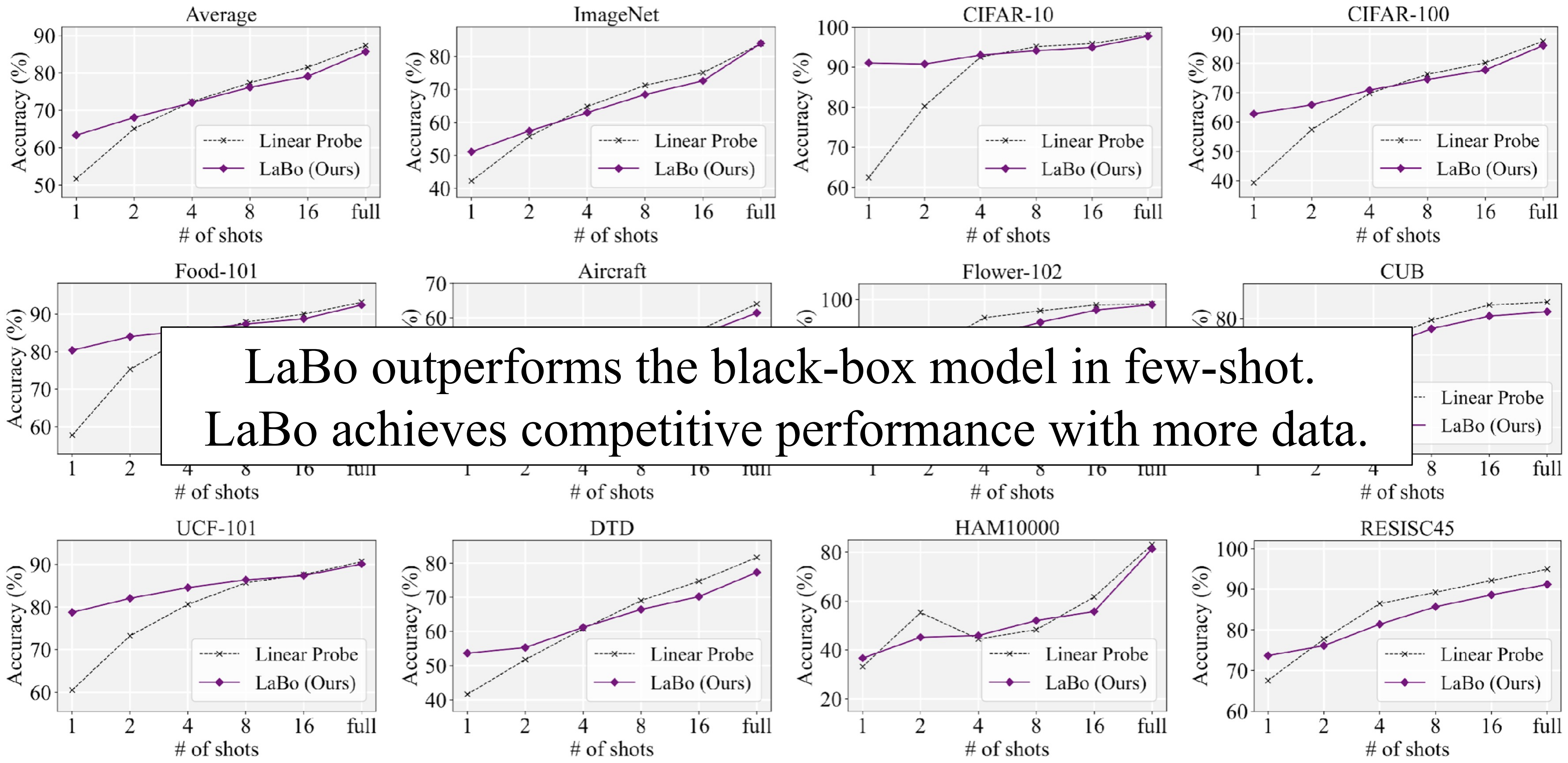
Satellite

RESISC45

Experimental Setup

- **Baselines**:
  - Linear Probe: logistic regression on the image features.
  - PCBM: Post-hoc CBM (Yuksekgonul et al., 2022)
    - Ensemble CBM prediction with end-to-end prediction.
  - ComDL: Compositional Derivation Learning (Yun et al., 2022)
    - Human designed concepts.
    - Linear layer over CLIP similarity scores.
- **Few-shot/Fully-supervised.**
- **Metric**: accuracy.

Comparison with Blackbox Model



LaBo outperforms the black-box model in few-shot.
LaBo achieves competitive performance with more data.

Figure 3. Test accuracy (%) comparison between LaBo and Linear Probe on 11 datasets. The x-axis represents the number of labeled images.

Comparison with Previous CBMs

Method	w/ end-to-end	CIFAR-10	CIFAR-100
PCBM [66]	✗	84.5	56.0
LaBo (Ours)	✗	**87.9**	**69.1**
PCBM-h [66]	✓	87.6	69.9
Linear Probe	✓	88.8	70.1

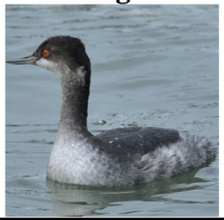
LaBo doesn’t rely on black box predictor.
LaBo doesn’t require human annotations.

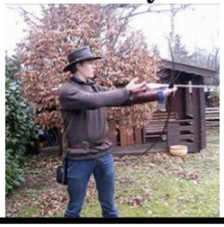
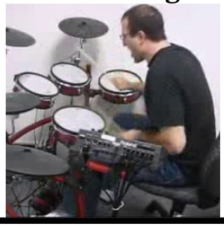
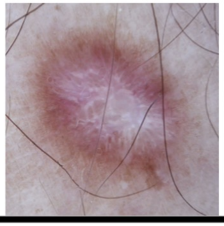
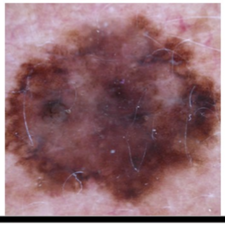
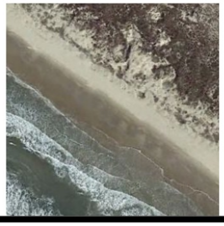
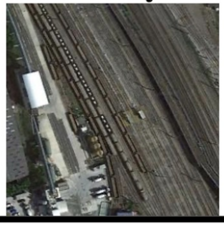
residual predictor from image features to targets.

Method	w/ manual concepts	1	5	Full
CompDL [67]	✓	13.6	33.2	52.6
LaBo (Ours)	✗	**35.1**	**55.7**	**71.8**
Linear Probe	-	28.4	55.4	75.5

Table 3. LaBo and CompDL evaluated on CUB for 1/5/full shots.

Qualitative Results

	Class Name	Top-3 Concepts	Class Name	Top-3 Concepts
ImageNet	**badger** 	1.short legs and long body make it an ex-cellent digger 2. black-and-white striped fur 3. coat is very shaggy	**ant** 	1. black and red stinger 2. small, black insect with six legs 3. long, slender anten-nae that it uses to smell and touch
Food101	**ramen** 	1. garnished with green onions, nori, and other toppings 2. most grocery stores 3. various toppings	**hummus** 	1. chickpeas, tahini, olive oil, garlic, lemon juice 2. made from cooked, mashed chickpeas 3. roasted red peppers
CUB	**eared grebe** 	1. black and white plumage that is strik-ing in the sunlight 2. black body with a long, slender neck 3. red and black bill	**horned lark** 	1. black line running through yellow face 2. head is black with a white horn on each side 3. black horn on each side of their head

UCF-101	**archery** 	1. grip bow tightly in their left hand 2. focused and con-centrated on their task 3. keep bow and ar-rows in safe and dry place when not in use	**drumming** 	1. blur as they fly over the drums 2. sitting on a stool in front of a drum set 3. position the drum-stick so it is resting on your index finger
HAM100000	**dermatofibroma** 	1. generally not painful 2. red, brown, or pur-ple in color 3. thin white halo around them	**melanoma** 	1. dark brown or black in color 2. large and dark 3. flesh-colored, brown, or black
RESISC45	**beach** 	1. waves crashing onto the shore 2. few rocks poking out 3. waves are gentle	**railway** 	1. connected by steel rails 2. tramline that is 3 feet wide and runs along the length of the court 3. faint, twinkling line

Human Evaluation

- **Metrics:**
  - **Factuality**: how accurately the **concepts** describe their designated **class**.
  - **Groundability**: how consistent the model grounds the **concepts** to **images**.

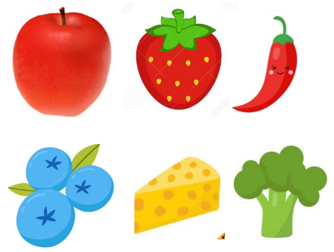
factuality

concept: red → class: apple ✓

concept: red → class: banana ✗

groundability

concept: red



✓

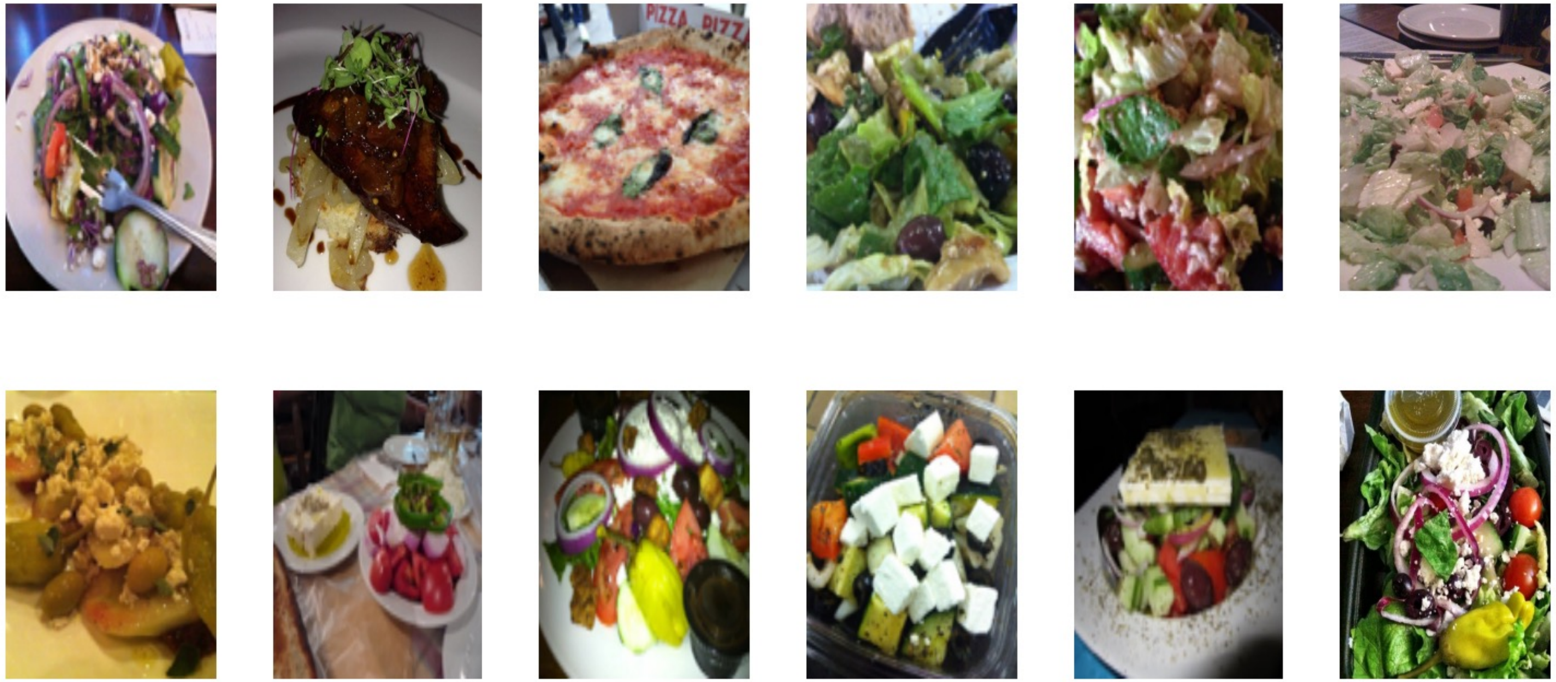
✗

Human Evaluation

$$Factuality(c) = \frac{\text{number of images selected}}{k \text{ ground truth images of the class}}$$

$$Groundability(c) = \frac{\text{number of images selected}}{\text{top-}k \text{ aligned images of the concept}}$$

feta cheese and kalamata olives



If you think that this concept is not good for singling out relevant images, select one or more of the following reasons (if any).

☐ Non-sensical or ungramatical. ☐ Unknown vocabulary ☐ Non visual phrase.

- **Invalid Concepts**
  - Non-sensical
  - Unknown vocabulary
  - Non-visual

Human Evaluation

- Compare with concepts from human-written text:
  - WordNet definition.
  - Wikipedia sentences (Kil and Chao, 2021).

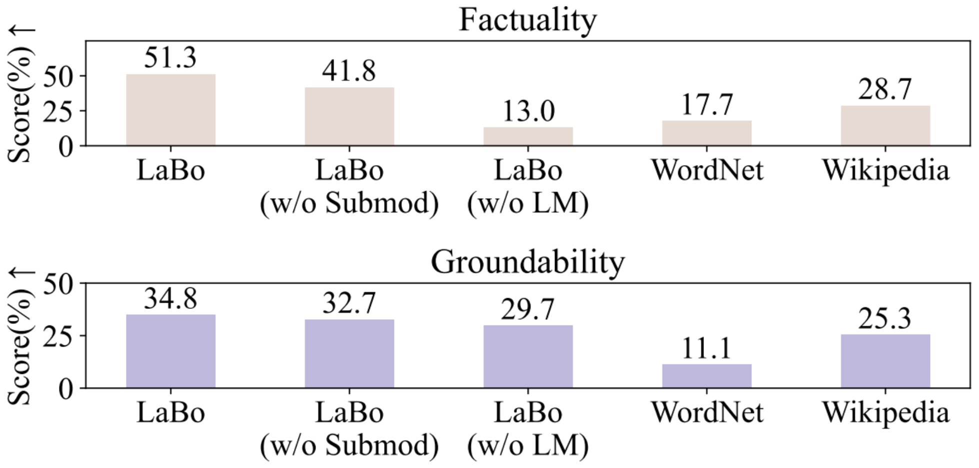


Figure 5. Human evaluation on *Factuality* and *Groundability* for different bottlenecks on ImageNet. "w/o Submod" denotes without submodular function, i.e., random concept selection. "w/o LM" denotes no language prior weight initialization.

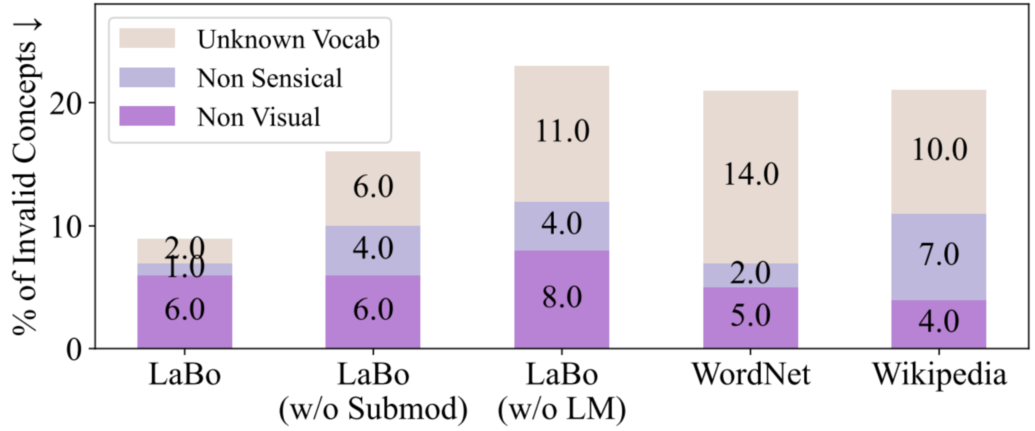


Figure 6. Percentage of invalid concepts identified by humans for different bottlenecks on ImageNet. **Lower** percentage is better.

Conclusion

- We demonstrate that the accuracy and interpretability of vision systems may be less at odds than previously believed.
- Leveraging LLMs was crucial, as they encode important visual knowledge.
- In the future, our approach can easily be enriched with new factors that capture different priors on bottleneck construction.

Thank you!