

LCA-on-the-Line:

Benchmarking Out of Distribution Generalization with Class Taxonomies

Jia Shi¹, Gautam Gare¹, Jinjin Tian¹, Siqi Chai¹, Zhiqiu Lin¹, Arun Vasudevan¹, Di Feng²³, Francesco Ferroni²⁴, Shu Kong⁵⁶

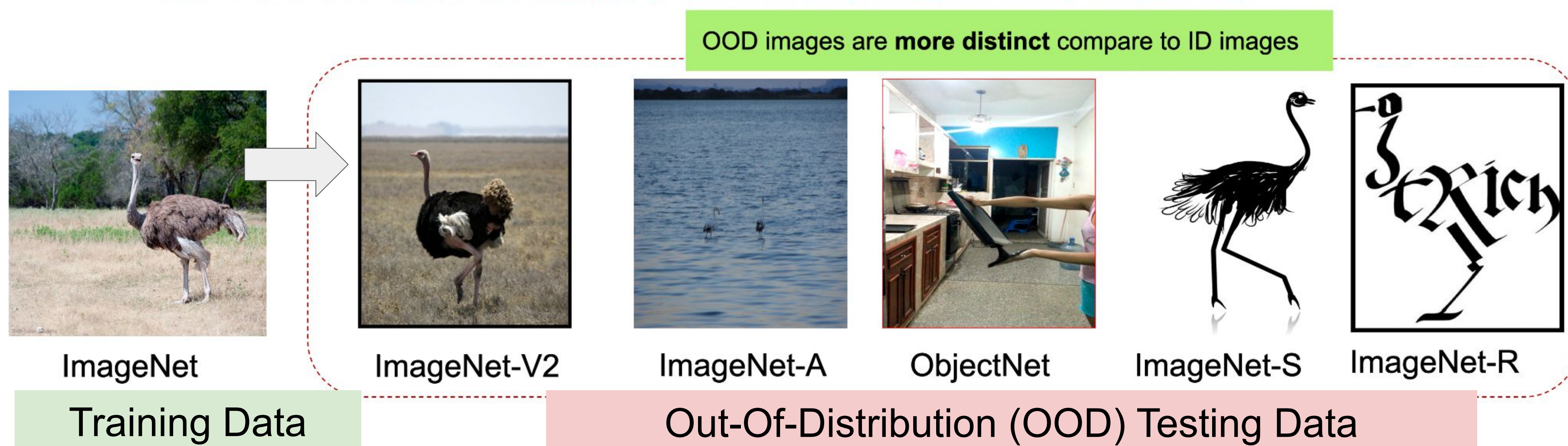
Robotics Institute, Carnegie Mellon University¹, Argo AI², Apple³, Nvidia⁴, Texas A&M University⁵, University of Macau⁶



Introduction

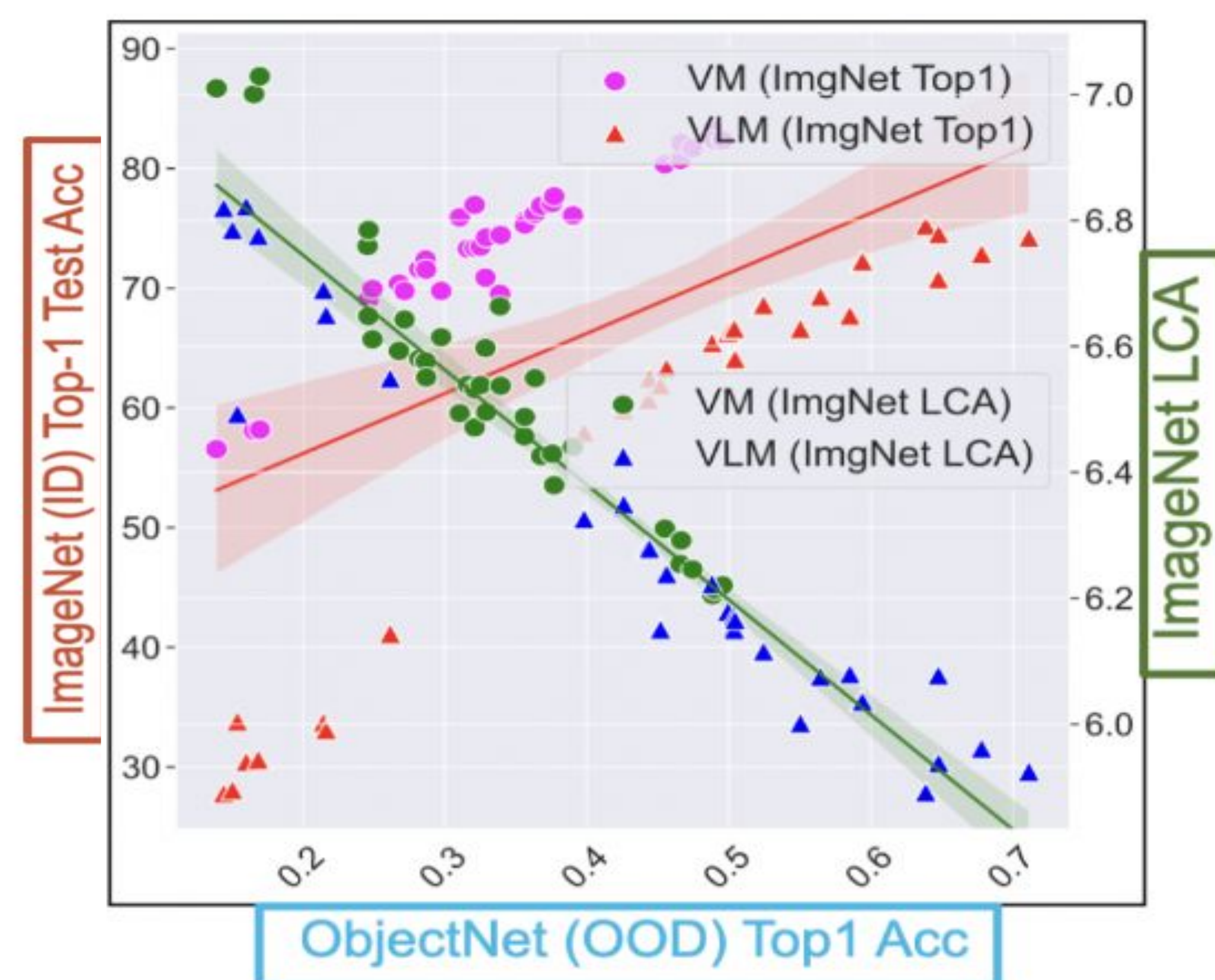
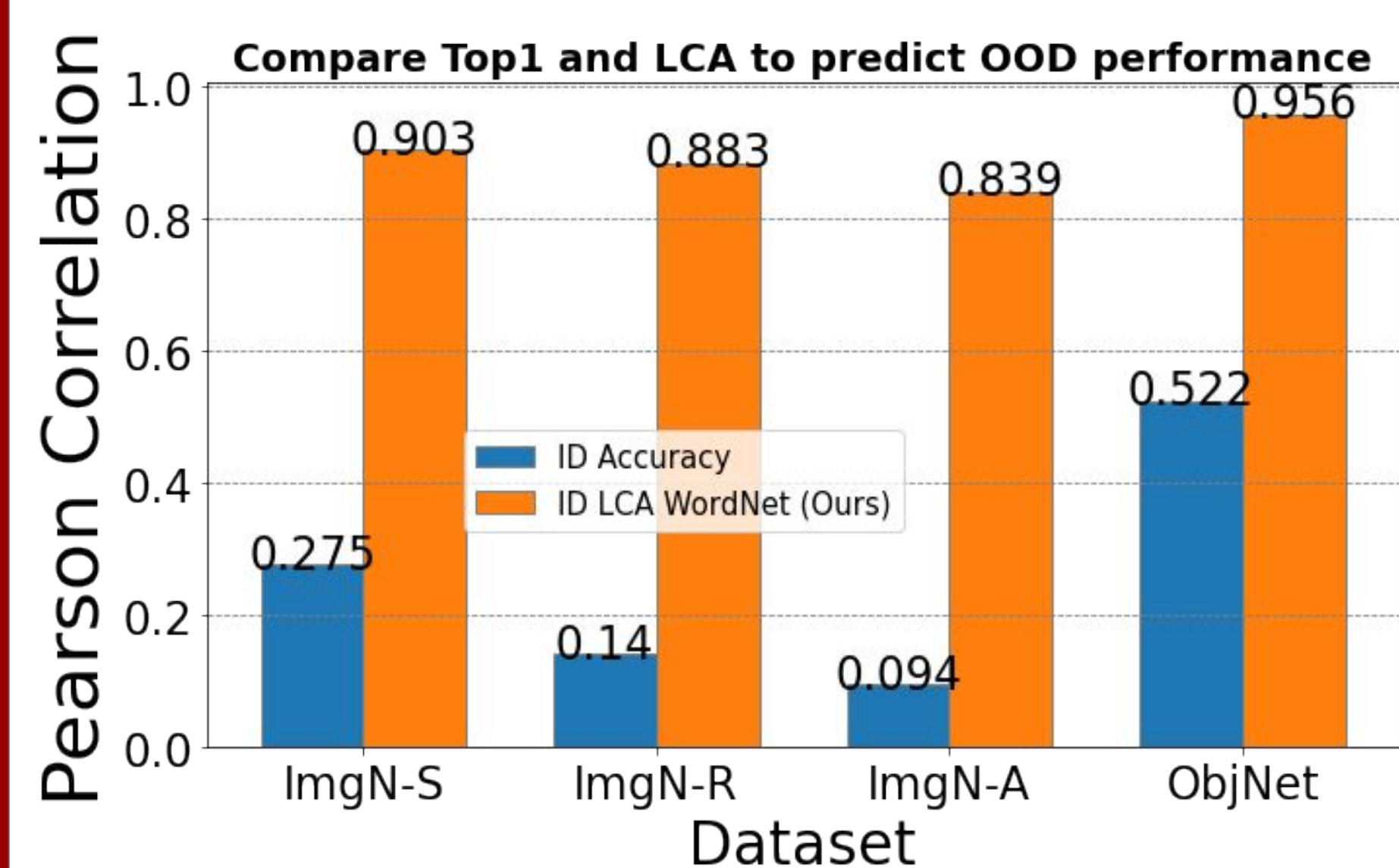
- LCA distance, a metric measuring prediction performance w.r.t an ontology/hierarchy on ID (in-distribution) data, can robustly predict the model's OOD (out of distribution) performance.
- It unifies Vision Models (VMs) and Vision-Language Models (VLMs) across different modalities and training data sources in terms of measuring model generalization, outperforming "accuracy- on-the-line".

LCA-on-the-Line evaluates on severe visual shift datasets



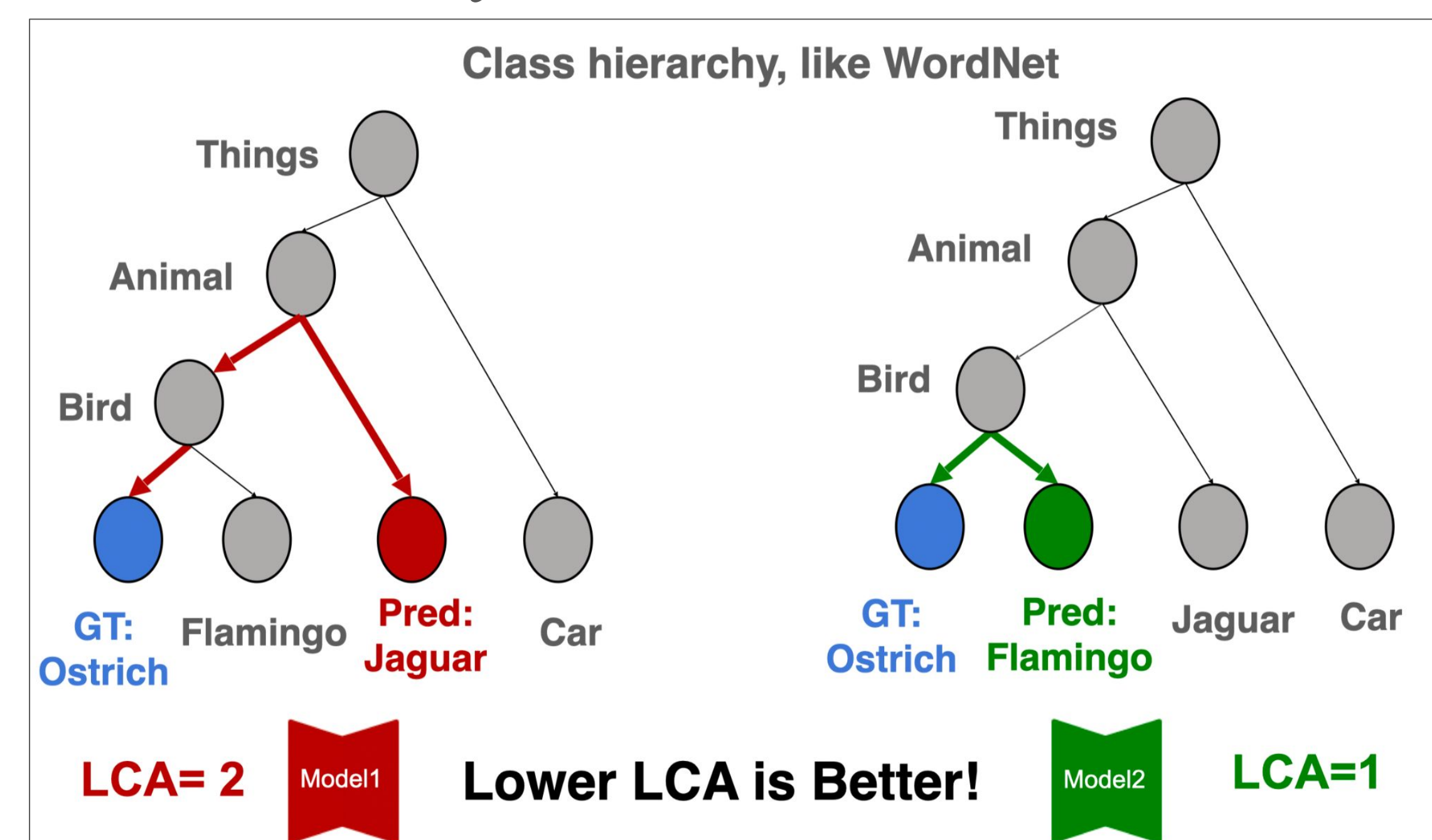
LCA is an indicator to OOD accuracy

- LCA distance achieves strong linear correlation consistently on multiple ImageNet-OOD datasets.
- For VMs + VLMs, ID accuracy is *not on the line*, while ID LCA is *on the line*.



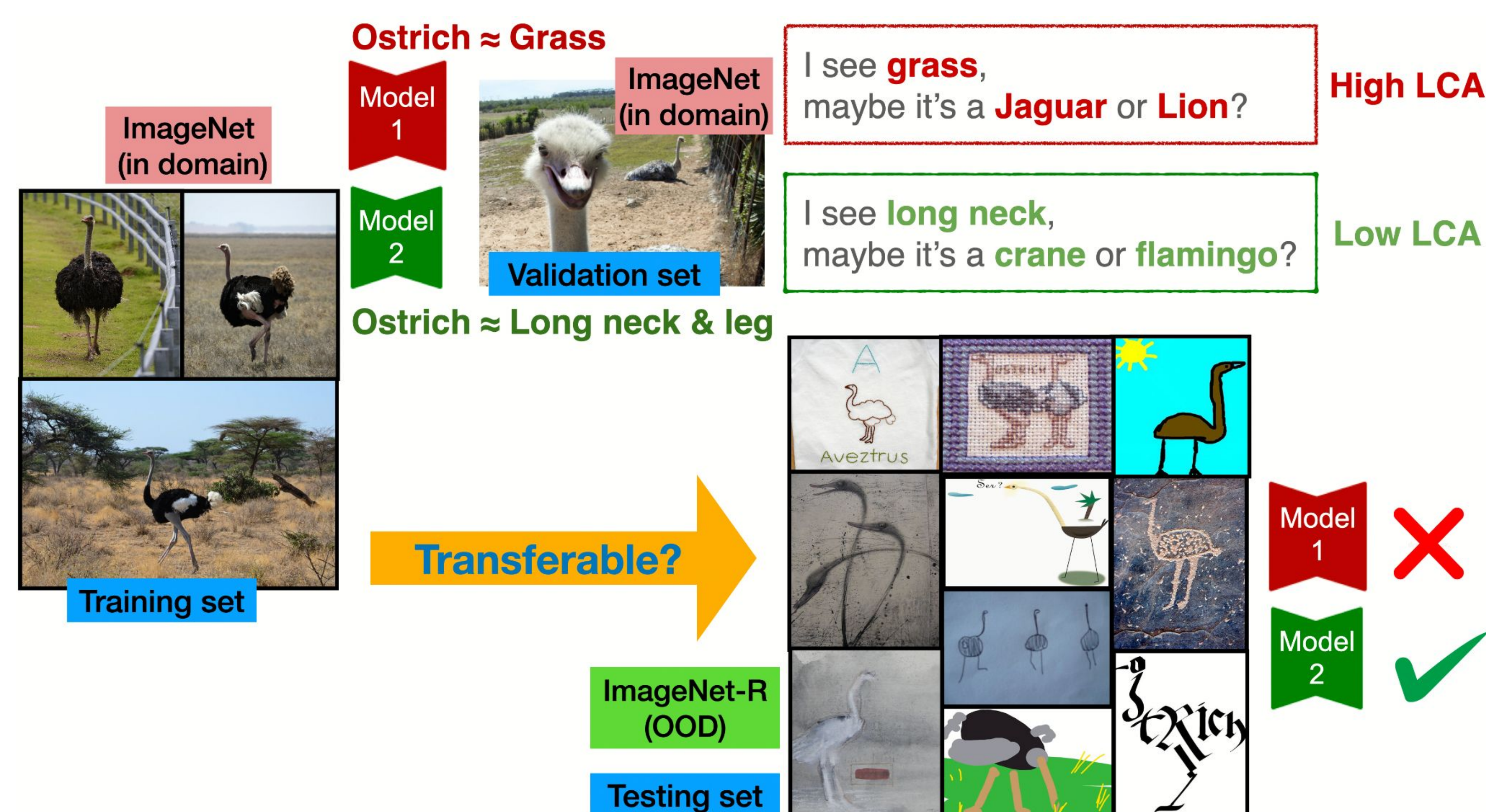
How to calculate LCA distance?

- LCA distance measures class-pairwise distance within a given **ontology/hierarchy**.
- Here, we measure the distance between ground-truth and prediction in a hierarchy, like **WordNet**.



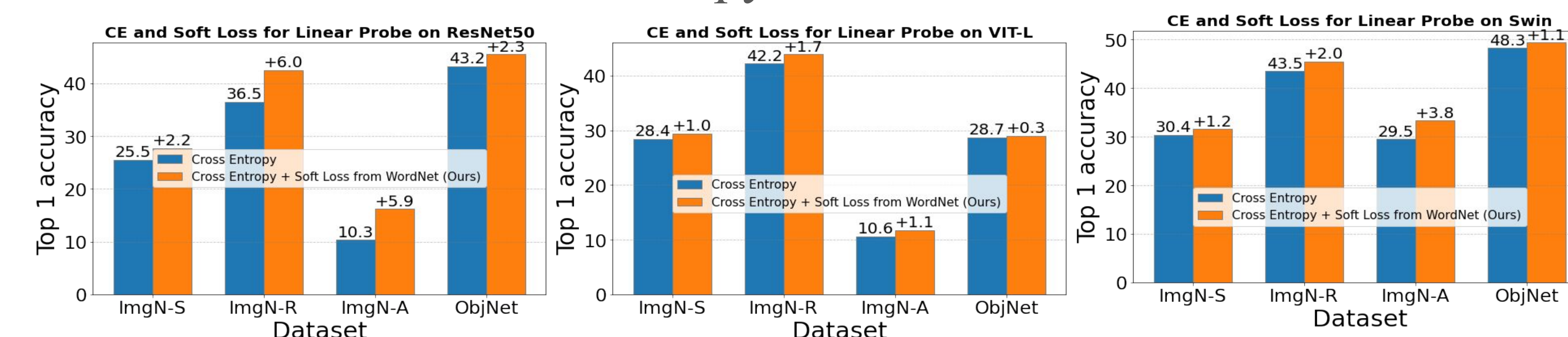
Why does LCA distance work?

- LCA distance reflects model's learned predictive feature, measuring its transferability to OOD.
- Low LCA distance reflects small spurious correlation.



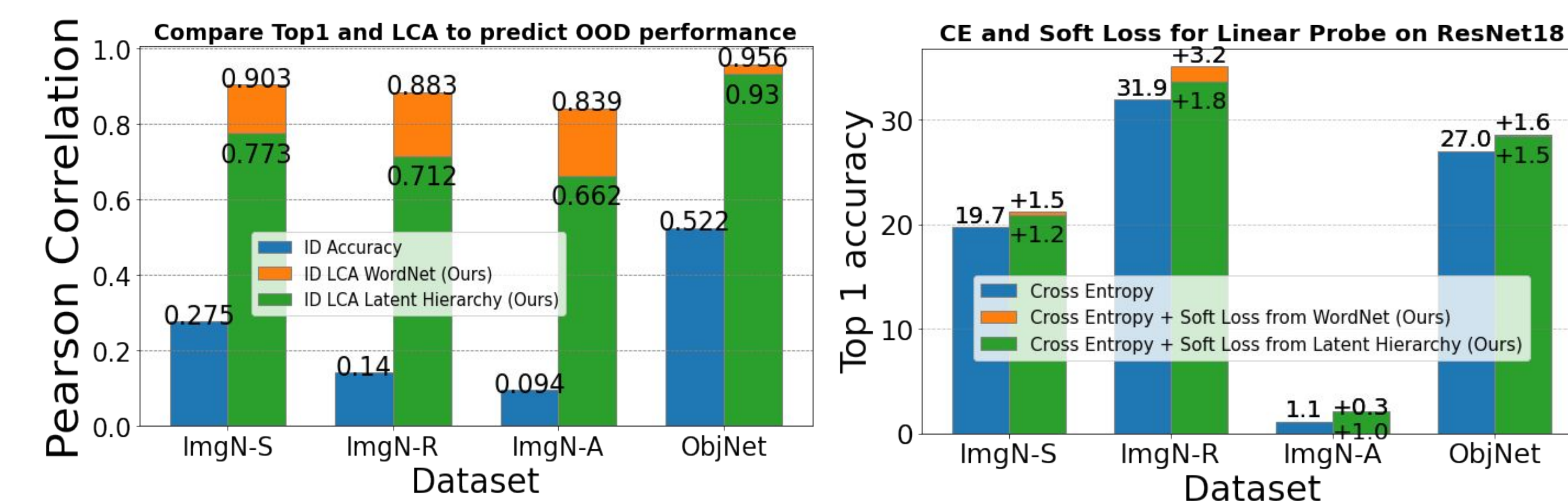
LCA as soft labels improves OOD accuracy

- Hierarchy allows exploiting class-pairwise distances in model training!
- We train with cross-entropy + **soft labels** loss.



Constructing a latent hierarchy on any dataset

- We can construct a hierarchy by clustering per-class mean features using a foundation model (like CLIP).
- Using a latent hierarchy performs as well as WordNet.



New insights to VLM generalization

- Using soft labels from latent hierarchies generated by VLMs yields better OOD results than VMs.
- That said, VLMs have a better human-aligned feature distribution, i.e., their generated labels better align with human-world ontology (WordNet).

