LCA-on-the-Line: Benchmarking Out of Distribution Generalization with Class Taxonomies

Jia Shi[1*†] **Gautam Gare**[1] **Jinjin Tian**[1] **Siqi Chai**[1]
Zhiqiu Lin[1] **Arun Vasudevan**[1] **Di Feng**[2]
Francesco Ferroni [3] **Shu Kong** [4,5] **Deva Ramanan**[1]
[1]Carnegie Mellon University [2]Apple [3]Nvidia
[4]Texas A&M University [5]University of Macau

Abstract

We introduce 'Least Common Ancestor (LCA)-on-the-line' as a method for predicting models' Out-of-Distribution (OOD) performance using in-distribution measurements, without the need for OOD data. We revisit the LCA distance, a concept from the pre-deep-learning era, which calculates the hierarchical distance between labels and predictions in a predefined class hierarchy tree, such as WordNet. Our evaluation of 75 models across five significantly shifted ImageNet-OOD datasets demonstrates the robustness of LCA-on-the-line. It reveals a strong linear correlation between in-domain ImageNet LCA distance and OOD Top-1 accuracy across various datasets, including ImageNet-S/R/A/ObjectNet. Compared to previous methods such as Accuracy-on-the-line [47] and Agreement-on-the-line [2], LCA-on-the-line shows superior generalization across a wide range of models. This includes models trained with different supervision types, such as class labels for vision models (VMs) and textual captions for vision-language models (VLMs). Our method offers a compelling alternative perspective on why vision-language models tend to generalize better to OOD data compared to vision models, even those with similar or lower in-domain (ID) performance. In addition to presenting an OOD performance indicator, we also demonstrate that aligning model predictions more closely with the class hierarchy and integrating a training loss objective with soft-labels can enhance model OOD performance.

1 Introduction

Generalizing models trained on in-distribution (ID) data to out-of-distribution (OOD) conditions is a notoriously challenging task. Distribution shifts can undermine the identical independent distribution (IID) assumption between training and testing data, thereby affecting robust performance. Recent works in OOD detection have targeted shifts in distribution by identifying anomalies [64, 54, 37, 40]. Additionally, numerous OOD datasets have been proposed to study the effects of different interventions, such as temporal shifts [26, 43, 38], artificial noise [21, 1, 33], and natural distribution shifts [23, 21, 3, 52]. Notably, the challenge of maintaining model robustness becomes significantly more difficult with severe visual shifts in the image domain.

Estimating OOD Generalization: In the sphere of model generalization, numerous attempts have been made to predict a model's performance on OOD datasets based on in-domain measurements, following the concept of *effective robustness* [69] (Fig 1). These approaches, referred to as 'XX-on-the-line'[47, 2], suggest that a model's OOD performance is correlated to in-domain accuracy [47, 52, 46, 55] or models consensus on in-domain accuracy [29, 2].

[*]This work was done while Di Feng & Francesco Ferroni was at Argo AI GmbH;
[†]Project page: `https://github.com/ElvishElvis/LCA-on-the-line`

Workshop on Distribution Shifts, 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

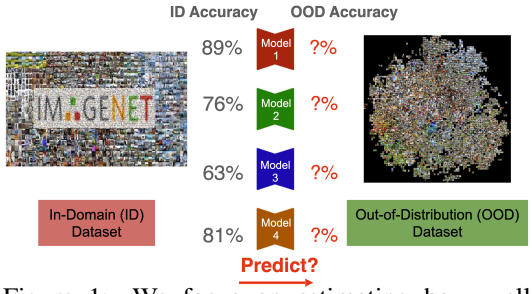

Figure 1: We focus on estimating how well models generalize to unseen, out-of-distribution (OOD) datasets. Specifically, we aim to predict a model's OOD performance from its in-domain performance.

Several methods in prior attempts rely on domain generalization strategies that necessitate prior knowledge of the target domain or require an estimation of OOD domain information [8, 34]. These can lead to computationally intensive processes, particularly when involving multiple models or inferences [2, 13].

Furthermore, many studies evaluate generalization on OOD datasets with limited visual shifts or only involve artificial noise, such as ImageNet-v2 or ImageNet-C [52, 1]. Such datasets fail to fully reflect a model's generalization capability when confronted with severe distribution shifts[23, 21, 3], as there is often limited transfer of robustness from synthetic to natural distribution shifts [69].

Moreover, most prior research has focused solely on estimating generalization among vision models (VMs) supervised on class labels trained on ImageNet [69, 48]. However, the rise of large-scale language models trained on datasets like LAION, particularly given their impressive performance in robust OOD generalization, underscores the necessity to evaluate and compare models across different families under a unified evaluation framework.

Unlike Vision Models (VMs), Vision-Language Models (VLMs) leverage more diverse training data, contrastive base loss, and language supervision. There have been attempts to measure VLM generalization [19, 15, 58, 30], specifically suggesting that training data diversity is an indicator of model generalization. However, collecting or training on such extensive data can be non-trivial [58]. Prior attempts still lack a unified, simple measurement for both VMs and VLMs to explain model generalization and convert it into actionable improvements.

Our experiment observed that prior art, like accuracy-on-the-line[47], fails to explain the increment in effective robustness[69] in VLMs compared to VMs. Recently, [61] observed the same problem and proposed evaluating OOD accuracy using multiple ID test sets, but their method still requires multiple run evaluations.

To address the issues of (1) lack of unified metrics on VLMs; (2) less robust to large domain shift; (3) computation expensive, we propose to adopt the Least Common Ancestor (LCA) score, to measure model generalization. The LCA distance is the taxonomic distance between labels and predictions, given a predefined class hierarchy, such as WordNet. Through a series of empirical experiments involving 75 models of different modalities (36 VMs and 39 VLMs), we show, for the first time to our knowledge, that the in-domain LCA metric **strongly correlates** with multiple ImageNet-OOD datasets under severe visual shifts (ImageNet-Rendition [23], Sketch [21], Adversarial [23], and ObjectNet [3]). This finding may help explain the surprising result that zero-shot vision-language models with poor top-1 accuracy generalize better to novel datasets compared to state-of-the-art vision models, which spurs us to further investigate and discuss the potential of the LCA benchmark for improving model generalization. **Please refer to section 7 for our motivation and hypothesis of adopting LCA, and Fig 2 illustrate settings comparison to prior work**.

2 LCA Distance Measure Mistake Severity

We propose using the in-domain Lowest Common Ancestor (LCA) distance, or taxonomy loss, as a predictor for model generalization. Here, we will formally define how taxonomy loss can be measured using in-domain data. Taxonomy loss measures the class ranking difference between a model's prediction based on class likelihood, and a predefined class order encoded by class taxonomy. Lower taxonomy loss is expected when a model assigns higher likelihood to classes that are semantically closer to the ground truth class, in other words, 'making better mistakes'[5]. Following previous research [5, 11], we utilize WordNet[45], a large-scale lexical database inspired by psycholinguistic theories of human lexical memory [44], to encode class taxonomy. An example of LCA distance is shown in Fig 3.

Given two classes, y (the ground truth class) and y' (the prediction class), we define the **LCA distance** according to [5] as $D_{LCA}(y', y) := f(y) - f(N_{LCA}(y, y'))$, where $f(y) \geq f(N_{LCA}(y, y'))$ and $N_{LCA}(y', y)$ denotes the lowest common ancestor class node for classes y and y' within the hierarchy, and $f(\cdot)$ represents a function of a node, such as the tree depth or entropy. We use the information content as described in [70]. For each sample X_i in the given dataset $\mathcal{M} := X_1, \ldots, X_n$: $D_{LCA}(model, \mathcal{M}) := \frac{1}{n}\sum_{i=1}^{n} D_{LCA}(\widehat{y}_i, y_i) \iff y_i \neq \widehat{y}_i$, where $\widehat{y}_i$ is the predicted class for sample X_i using the model, y_i is the ground truth class for sample X_i, and $y_i \neq \widehat{y}_i$. Intuitively, a model with a lower LCA distance demonstrates a greater semantic understanding of class ontology in WordNet.

3 Experiment

In this section, we present an experiment benchmarking the relationship between Lowest Common Ancestor (LCA) and generalization.

Setup This paper leverages 75 pretrained models sourced from open repositories on GitHub for empirical analysis. Our selection comprises 36 Vision Models (VMs) pretrained on ImageNet and supervised from class labels, alongside 39 Vision-Language Models (VLMs) that incorporate language as part of the supervision. A comprehensive list of model details, ensuring reproducibility, will be provided in Section 6. We use *ImageNet*[11] as the source in-distribution (ID) dataset, while *ImageNet-v2*[52], *ImageNet-Sketch*[21], *ImageNet-Rendition*[23], *ImageNet-Adversarial*[23], and *ObjectNets*[3] are employed as out-of-distribution datasets, exemplifying severe natural distribution shifts. The ImageNet hierarchy, as depicted in [5], is utilized.

For our correlation experiment, we use R^2 *(Coefficient of Determination)* and *PEA (Pearson correlation coefficient)* to measure the strength and direction of linear relationships between two variables. Additionally, we employ *KEN (Kendall rank correlation coefficient)* and *SPE (Spearman rank-order correlation coefficient)* to assess the correspondence of the rankings of two variables.

The importance of these measurements lies in their different focuses. Linearity measures, such as R^2 and PEA, are primarily concerned with the fit of a linear model to data points, allowing us to quantify the predictability of changes in one variable based on the other. Ranking measures, like KEN and SPE, provide insights into how the rankings of variables relate to each other, which is crucial in downstream applications such as image retrievals and search engine optimization, where understanding and predicting the ordering of data points is often more important than predicting their exact values. For prediction experiments, we utilize MAE (Mean Absolute Error) to quantify the absolute difference between predictions and ground truth.

Although *ImageNet-v2* is predominantly deemed an OOD dataset in most prior literature [59, 47, 2], our experiments suggest that *ImageNet-v2* aligns more closely with ImageNet than with other OOD datasets; we delve into these details in Appendix 9.

	Element		**ImageNetv2**		**ImageNet-S**		**ImageNet-R**		**ImageNet-A**		**ObjectNet**	
	ID	OOD	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA
ALL	Top1	Top1	**0.962**	**0.980**	0.075	0.275	0.020	0.140	0.009	0.094	0.273	0.522
	LCA	Top1	0.339	0.582	**0.838**	**0.915**	**0.779**	**0.883**	**0.869**	**0.932**	**0.915**	**0.956**
	Top1	Top5	**0.889**	**0.943**	0.052	0.229	0.004	0.060	0.013	0.115	0.262	0.512
	LCA	Top5	0.445	0.667	**0.883**	**0.940**	**0.738**	**0.859**	**0.909**	**0.953**	**0.924**	**0.961**

Table 1: **Correlation measurement ($R^2 PEA$) of ID LCA/Top1 with OOD Top1/Top5** across 75 models spanning modalities (36 VMs and 39 VLMs) as shown in Figure 4. We demonstrate that LCA has a strong correlation with OOD performance on all datasets (except ImageNet-v2). We take the absolute value of all correlations for simplicity.

3.1 LCA-on-the-Line: In-Domain Taxonomy Distance (LCA) as an Out of Distribution (OOD) Performance Benchmark

Accuracy-on-the-line [47] corroborated that a model's in-distribution (ID) accuracy and its out-of-distribution (OOD) accuracy are largely considered to be strongly correlated. This potent correlation forms a significant baseline for comparison in our research. Unlike the framework presented in [47],

which only compares models within the same modality, our work bridges the gap by contrasting models of different modalities, involving both Vision Models (VM) and Vision-Language Models (VLM). In addition to the Top1 OOD accuracy, we also incorporate Top5 OOD accuracy, yielding a more comprehensive evaluation of model generalization.

As displayed in Table 1, the ImageNet in-domain accuracy [47] forms a robust predictor for most OOD datasets, when the comparison is limited to models with similar setups (VMs or VLMs). However, this predictor fails to provide a unified explanation of generalization across models from both families. As highlighted in Figure 4 (indicated in red), when adhering to 'accuracy on the line' [47], all four OOD datasets plotted showcase two separate linear trends, representing models that belong to each family. This observation aligns with [9], where it was found that VLM models, despite exhibiting significantly lower ID accuracy, could attain higher OOD performance than their state-of-the-art VM counterparts.

As shown in Figure 5, our method, adopting in-domain LCA distance, could unify models from both families. As demonstrated in Table 1 and Figure 4 (colored in green), the severity of in-domain mistakes serves as a more effective indicator of model performance compared to in-domain accuracy. It consistently exhibits a strong linear correlation with all OOD benchmark accuracies for natural distribution shifts (both R^2 and the Pearson correlation coefficient approach 0.9, while [47] drop to 0 in ImageNet-A). Notably, our experiments showed that [47] is a more reliable indicator solely for ImageNet-v2, given its visual similarity to ImageNet. We will further discuss this in Appendix 9. In Section12, we will also include measurements from the KEN and SPE, which similarly demonstrate robust scores in preserving the relative ordering of model OOD performance.

3.2 Predicting OOD Performance with In-Domain LCA

We further highlight the effectiveness of the LCA-on-the-Line' approach by estimating model OOD performance using a linear function derived from in-domain LCA distance. For comparison, we included four competitive baselines: Average Confidence (AC), which leverages OOD logits after temperature scaling; two methods from Agreement-on-the-Line (Aline-D and Aline-S), utilizing consensus of pairs of models on OOD benchmarks; and Accuracy on the Line' (ID Top1), employing in-domain accuracy of established measurement models to fit a linear function. Instead of performing a probit transform as done in [2] and [47], we implemented min-max scaling because LCA does not fall within the [0,1] range.

		ImageNetv2	ImageNet-S	ImageNet-R	ImageNet-A	ObjectNet
ALL	ID Top1 [47]	**0.040**	0.230	0.277	0.192	0.178
	AC [22]	<u>0.043</u>	<u>0.124</u>	<u>0.113</u>	0.324	<u>0.127</u>
	Aline-D [2]	0.121	0.270	0.167	0.409	0.265
	Aline-S [2]	0.072	0.143	0.201	<u>0.165</u>	0.131
	(Ours) ID LCA	0.162	**0.078**	**0.107**	**0.061**	**0.048**

Table 2: **Error Prediction of OOD Datasets** across 75 models of diverse settings with **MAE loss ↓**. Top1 in **bold** and Top2 in <u>underline</u>. Despite ImageNet's in-domain accuracy remaining a significant indicator of ImageNet-v2 accuracy, the in-domain LCA outperforms it as a robust error predictor across four severe distributed OOD datasets, particularly ImageNet-A, which stumps other methods.

As illustrated in Table 2, in-domain LCA distance proves to be a significantly more robust OOD error predictor than other baselines across four OOD benchmarks with varying distribution shifts. This robustness is especially evident for ImageNet-A, an adversarial dataset derived from ResNet50's misclassifications on ImageNet. Consequently, models pre-trained on ImageNet tend to underperform on this dataset, especially those with lower accuracy than ResNet50. This leads to decreased robustness for in-domain indicators like in-domain accuracy [47], methods calibrated from in-domain validation sets [22], and OOD agreement of models from different families [2]. In contrast, LCA, relying solely on the relative ranking of class predictions from a single model, is less sensitive to these issues and thus delivers more consistent performance. This further underscores the efficacy of LCA as a powerful predictor in challenging OOD scenarios.

References

[1] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

[2] Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. (2022). Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, **35**, 19274–19289.

[3] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. (2019). Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, **32**.

[4] Barz, B. and Denzler, J. (2019). Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 638–647. IEEE.

[5] Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., and Lord, N. A. (2020). Making better mistakes: Leveraging class hierarchies with deep networks. In *CVPR*.

[6] Beyer, L., Hénaff, O. J., Kolesnikov, A., Zhai, X., and Oord, A. v. d. (2020). Are we done with imagenet? *arXiv preprint arXiv:2006.07159*.

[7] Bühlmann, P. (2020). Invariance, causality and robustness.

[8] Chen, M., Goel, K., Sohoni, N. S., Poms, F., Fatahalian, K., and Ré, C. (2021). Mandoline: Model evaluation under distribution shift. In *International Conference on Machine Learning*, pages 1617–1629. PMLR.

[9] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2022). Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*.

[10] Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., and Jitsev, J. (2023). Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

[11] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009a). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009b). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

[13] Deng, W., Gould, S., and Zheng, L. (2022). On the strong correlation between model invariance and generalization. *arXiv preprint arXiv:2207.07065*.

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.* (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[15] Fang, A., Ilharco, G., Wortsman, M., Wan, Y., Shankar, V., Dave, A., and Schmidt, L. (2022). Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR.

[16] Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., and Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, **26**.

[17] Gare, G. R., Fox, T., Lowery, P., Zamora, K., Tran, H. V., Hutchins, L., Montgomery, D., Krishnan, A., Ramanan, D. K., Rodriguez, R. L., *et al.* (2022). Learning generic lung ultrasound biomarkers for decoupling feature extraction from downstream tasks. *arXiv preprint arXiv:2206.08398*.

[18] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.

[19] HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, **34**, 5000–5011.

[20] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

[21] Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

[22] Hendrycks, D. and Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

[23] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., *et al.* (2021). The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

[24] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[25] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., *et al.* (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.

[26] Hu, H., Sener, O., Sha, F., and Koltun, V. (2022). Drinking from a firehose: Continual learning with web-scale natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[27] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[28] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

[29] Jiang, Y., Nagarajan, V., Baek, C., and Kolter, J. Z. (2021). Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*.

[30] Kaur, J. N., Kiciman, E., and Sharma, A. (2022). Modeling the data-generating process is necessary for out-of-distribution generalization. *arXiv preprint arXiv:2206.07837*.

[31] Krizhevsky, A., Nair, V., and Hinton, G. (????). Cifar-10 (canadian institute for advanced research).

[32] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, **60**(6), 84–90.

[33] Larochelle, H., Erhan, D., and Bengio, Y. (2008). Zero-data learning of new tasks. In *AAAI*, volume 1, page 3.

[34] Li, C., Zhang, B., Shi, J., and Cheng, G. (2022a). Multi-level domain adaptation for lane detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4380–4389.

[35] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, **34**, 9694–9705.

[36] Li, J., Li, D., Xiong, C., and Hoi, S. (2022b). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

[37] Liang, S., Li, Y., and Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

[38] Lin, Z., Shi, J., Pathak, D., and Ramanan, D. (2021). The clear benchmark: Continual learning on real-world imagery. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[39] Lin, Z., Pathak, D., Wang, Y.-X., Ramanan, D., and Kong, S. (2022). Continual learning with evolving class ontologies. *Advances in Neural Information Processing Systems*, **35**, 7671–7684.

[40] Liu, W., Wang, X., Owens, J., and Li, Y. (2020). Energy-based out-of-distribution detection. *Advances in neural information processing systems*, **33**, 21464–21475.

[41] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

[42] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986.

[43] Lomonaco, V. and Maltoni, D. (2017). Core50: a new dataset and benchmark for continuous object recognition. In *Conference on Robot Learning*, pages 17–26. PMLR.

[44] Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

[45] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, **3**(4), 235–244.

[46] Miller, J., Krauth, K., Recht, B., and Schmidt, L. (2020). The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, pages 6905–6916. PMLR.

[47] Miller, J. P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. (2021). Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR.

[48] Mustafa, B., Riquelme, C., Puigcerver, J., Pinto, A. S., Keysers, D., and Houlsby, N. (2020). Deep ensembles for low-data transfer learning. *arXiv preprint arXiv:2010.06866*.

[49] Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012.

[50] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.* (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[51] Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436.

[52] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

[53] Redmon, J. and Farhadi, A. (2017). Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271.

[54] Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., and Lakshminarayanan, B. (2021). A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*.

[55] Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems*, **32**.

[56] Santurkar, S., Tsipras, D., and Madry, A. (2020). Breeds: Benchmarks for subpopulation shift. *arXiv preprint arXiv:2008.04859*.

[57] Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, **109**(5), 612–634.

[58] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., *et al.* (2022). Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

[59] Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. (2020). Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR.

[60] Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. (2022). Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, **23**, 1–55.

[61] Shi, Z., Carlini, N., Balashankar, A., Schmidt, L., Hsieh, C.-J., Beutel, A., and Qin, Y. (2023). Effective robustness against natural distribution shifts for models with different training data. *arXiv preprint arXiv:2302.01381*.

[62] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*.

[63] Subramanian, J., Annadani, Y., Sheth, I., Ke, N. R., Deleu, T., Bauer, S., Nowrouzezahrai, D., and Kahou, S. E. (2022). Learning latent structural causal models. *arXiv preprint arXiv:2210.13583*.

[64] Sun, Y., Guo, C., and Li, Y. (2021). React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, **34**, 144–157.

[65] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

[66] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

[67] Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

[68] Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2820–2828.

[69] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. (2020). Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, **33**, 18583–18599.

[70] Valmadre, J. (2022). Hierarchical classification at multiple operating points. *arXiv preprint arXiv:2210.10929*.

[71] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778.

[72] Wan, A., Dunlap, L., Ho, D., Yin, J., Lee, S., Jin, H., Petryk, S., Bargal, S. A., and Gonzalez, J. E. (2020). Nbdt: neural-backed decision trees. *arXiv preprint arXiv:2004.00221*.

[73] Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., *et al.* (2022). Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971.

[74] Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2021). Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602.

[75] Zagoruyko, S. and Komodakis, N. (2016). Wide residual networks. *arXiv preprint arXiv:1605.07146*.

[76] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, **64**(3), 107–115.

[77] Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856.

Appendix

4 Figure illustration

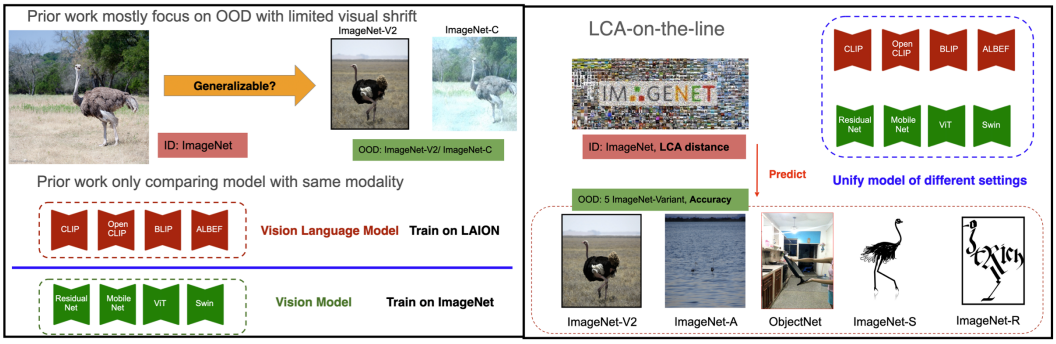

Figure 2: **Illustration of settings comparison to prior work.** Left: prior work settings; Right: our settings. To the best of our knowledge, LCA-on-the-line is the first approach to uniformly measure model robustness across VMs and VLMs, on OOD datasets with significant distribution shifts.

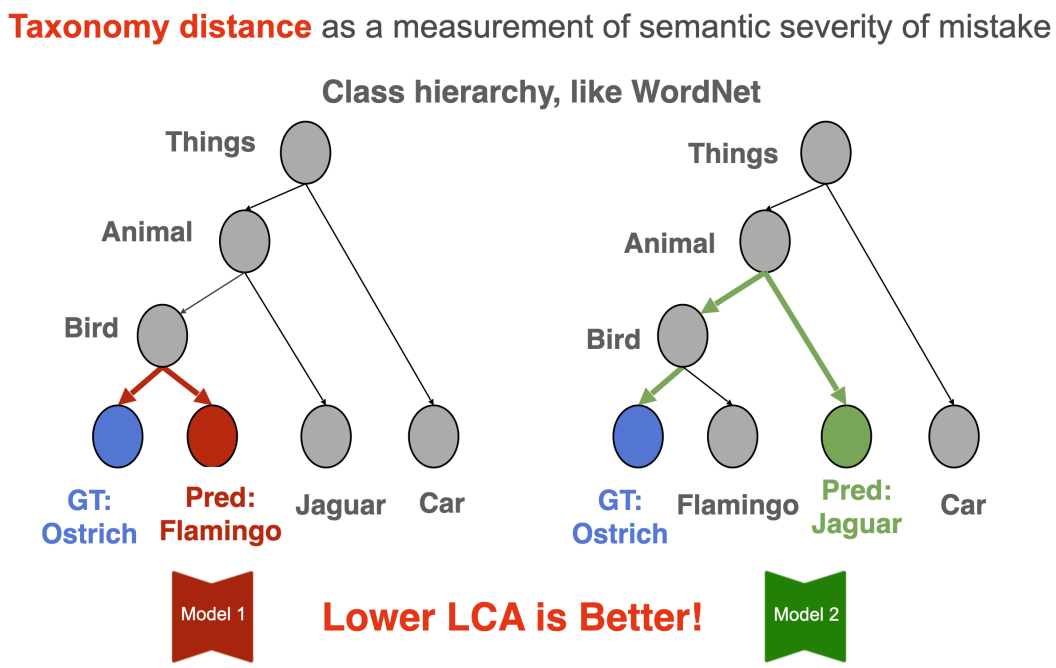

Figure 3: Our method estimates a model's generalization based on its in-domain semantic severity of mistakes. We use the 'Least Common Ancestor' (LCA) distance, the ranking distance between the model's prediction and the ground truth class from a predefined taxonomy hierarchy, like WordNet. The LCA distance is proportional to the shortest path from the prediction to the ground truth class in the hierarchy tree.

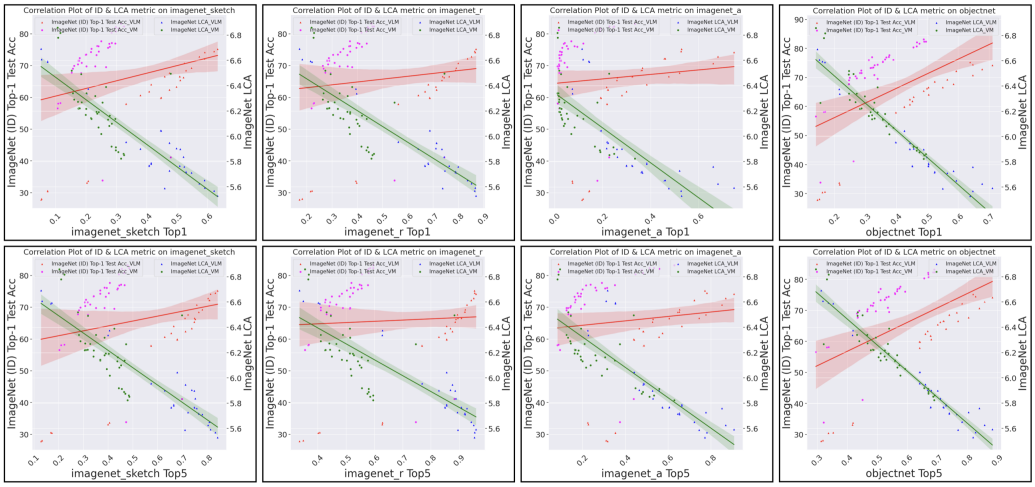

Figure 4: **Correlating OOD Top-1/Top-5 Accuracy (VM+VLM, 75 models) on 4 ImageNet-OOD Datasets.** Following Table 1. The plots clearly demonstrate that the in-domain LCA has a stronger correlation with the model's OOD performance across all OOD datasets over accuracy-on-the-line [47]. Each plot's x-axis represents the OOD dataset metric (with OOD Top-1 in the top row, and OOD Top-5 accuracy in the bottom row); Red represents in-domain classification accuracy (Top-1); Green denotes in-domain taxonomy distance (LCA). As interpreted in Figure 5, accuracy-on-the-line only explains generalization of models with similar settings (VMs or VLMs), but does not unify both model families. For better legibility, please find a PNG of this image in the supplementary material.

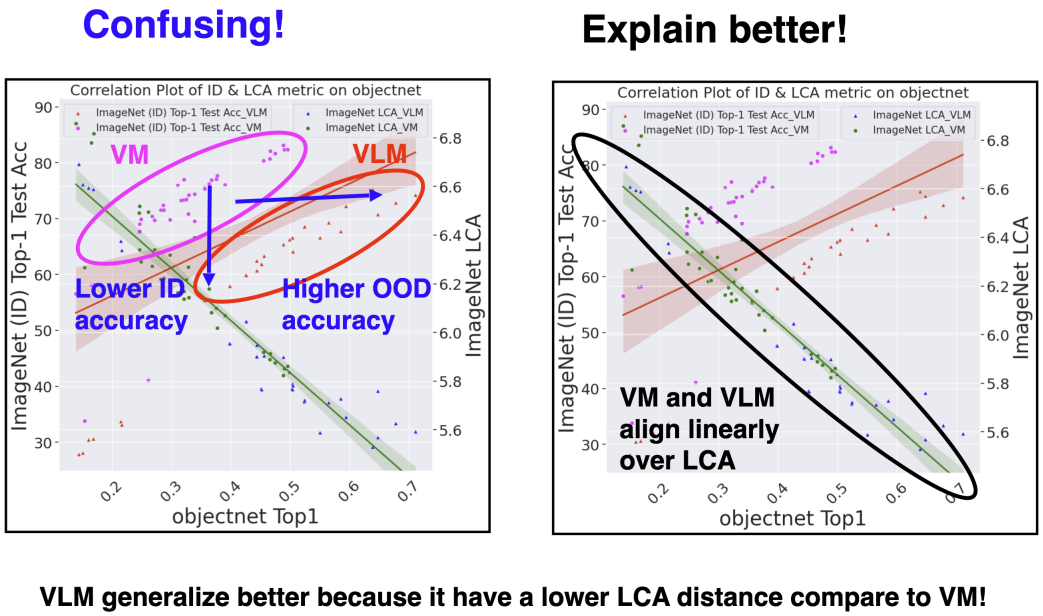

Figure 5: Our method restores the "on-the-line" linear relationship by unifying both VMs and VLMs. Our method provides a compelling alternative to understand why vision-language models with lower in-domain accuracy might generalize better to OOD datasets than vision models.

5 Ethics Statement

This research aims to enhance our understanding of model generalization mechanisms. However, it's crucial to recognize its potential misuse, such as in guiding adversarial attacks that reduce the generalization capabilities of existing models. Although not the intended purpose of our research, the dual potential of our findings in model generalization underscores the need for robust, secure model development and the implementation of ethical guidelines for deploying this knowledge.

6 Model Architectures

We list all models used in ours experiment as follows, including 36 Vision Only Models (VM) and 39 Vision-Language Models (VLM).

Model Category	Architecture	Number of models	Checkpoint Link
VM (Vision-Only-Models)	AlexNet [32]	1	alexnet
	ConvNeXt [42]	1	convnext$_tiny$
	DenseNet [27]	4	densenet121 densenet161 densenet169 densenet201
	EfficientNet [67]	1	efficientnet_b0
	GoogLeNett [65]	1	googlenet
	Inceptionv3 [66]	1	inceptionV3
	MnasNet [68]	4	mnasnet0.5 mnasnet0.75 mnasnet1.0 mnasnet1.3
	Mobilenet-V3 [25]	2	mobilenetv3_small mobilenetv3_large
	Regnet [51]	1	regnet_y_1_6gf
	Wide ResNet [75]	1	wide_resnet101_2
	ResNet [20]	5	resnet18 resnet34 resnet50 resnet101 resnet152
	ShuffleNet [77]	1	shufflenet_v2_x2_0
	SqueezeNet [28]	2	squeezenet1_0 squeezenet1_1
	Swin Transformer [41]	1	swin_b
	VGG [62]	8	vgg11 vgg13 vgg16 vgg19 vgg11_bn vgg13_bn vgg16_bn vgg19_bn
	ViT [14]	2	vit_b_32 vit_l_32
VLM (Vision-Language-Models)	ALBEF [35]	1	albef_feature_extractor
	BLIP [36]	1	blip_feature_extractor_base
	CLIP [50]	7	RN50 RN101 RN50x4 ViT-B-32.pt ViT-B-16.pt ViT-L-14.pt ViT-L-14-336px
	OpenCLIP [10]	30	openCLIP: openCLIP_('RN101', 'openai') openCLIP_('RN101', 'yfcc15m') openCLIP_('RN101-quickgelu', 'openai') openCLIP_('RN101-quickgelu', 'yfcc15m') openCLIP_('RN50', 'cc12m') openCLIP_('RN50', 'openai') openCLIP_('RN50', 'yfcc15m') openCLIP_('RN50-quickgelu', 'cc12m') openCLIP_('RN50-quickgelu', 'openai') openCLIP_('RN50-quickgelu', 'yfcc15m') openCLIP_('RN50x16', 'openai') openCLIP_('RN50x4', 'openai') openCLIP_('RN50x64', 'openai') openCLIP_('ViT-B-16', 'laion2b_s34b_b88k') openCLIP_('ViT-B-16', 'laion400m_e31') openCLIP_('ViT-B-16', 'laion400m_e32') openCLIP_('ViT-B-16-plus-240', 'laion400m_e31') openCLIP_('ViT-B-16-plus-240', 'laion400m_e32') openCLIP_('ViT-B-32', 'laion2b_e16') openCLIP_('ViT-B-32', 'laion2b_s34b_b79k') openCLIP_('ViT-B-32', 'laion400m_e31') openCLIP_('ViT-B-32', 'laion400m_e32') openCLIP_('ViT-B-32', 'openai') openCLIP_('ViT-B-32-quickgelu', 'laion400m_e31') openCLIP_('ViT-B-32-quickgelu', 'laion400m_e32') openCLIP_('ViT-L-14', 'laion2b_s32b_b82k') openCLIP_('ViT-L-14', 'laion400m_e31') openCLIP_('ViT-L-14', 'laion400m_e32') openCLIP_('coca_ViT-B-32', 'laion2b_s13b_b90k') openCLIP_('coca_ViT-L-14', 'laion2b_s13b_b90k')

7 The Suitability of LCA as a Benchmark for Model Generalization

This section explores the hypothesis that links taxonomy loss (LCA) with a model's generalization ability. Furthermore, we discuss how such insights can be put into meaningful, actionable use.

Obstacles to Model Generalization. In deep learning, models form connections between distinguishable image features and class labels. However, these discriminative associations are vulnerable to spurious correlations in training data [76]. An example is erroneously associating the class 'ostriches' with the feature 'grass in the background', as ostriches often appear in grasslands. These correlations may fail when applied to an OOD dataset [76].

Essentials for Model Generalization.

Figure 6 demonstrates a severely shifted OOD dataset, ImageNet-R, where, despite significant distribution shifts, humans can effortlessly identify the correct classes. This is because humans can discern universally transferable distinctions between classes as distinguishable features for classification. Therefore, we posit that a model's generalization capability depends on the transferability of these learned features during training, and only features that align with human understanding of object definitions are universally transferable to any OOD dataset.

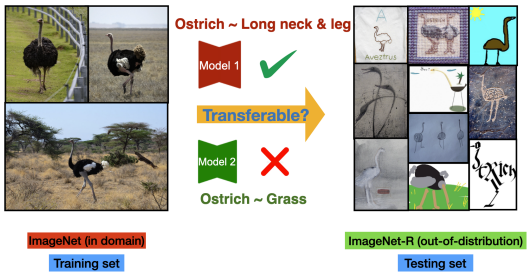

Figure 6: **Capturing transferable features for model generalization**. Despite pronounced distribution shifts, ImageNet-R serves as a valid OOD test set for ImageNet classes, as the images of ostriches, for instance, still maintain shape information [18] like 'long neck', 'big belly', and 'long legs'. We hypothesize that models exhibiting good generalization should capture these transferable features rather than succumb to spurious correlations on features like 'grass'.

But how can we measure what features a model has learned during training? The decision-making process of deep neural networks trained end-to-end has become less interpretable. Attempts to decipher the decision process of models and form decision-tree-like models [72, 17] have been made, but these efforts have not linked this understanding to model generalization.

Alignment to Class Taxonomy as a Representation Measurement.

Ideally, a model that captures more generalizable features tends to 'make better mistakes' by predicting classes that are semantically closer to the ground truth class. As illustrated in Fig 7, a model that learns to associate ostriches with features like 'long legs' and 'long neck', which are more transferable to OOD datasets, will likely predict classes like flamingos or cranes. In contrast, a model influenced by spurious correlations and associating ostriches with grass might predict a semantically distant class, like jaguars or lions, which also often appear on grass.

Our method involves measuring a model's generalization based on its in-domain semantic severity of mistakes. We use the 'Least Common Ancestor' (LCA) distance, the taxonomic distance between the model's prediction and the ground truth class in a predefined taxonomy hierarchy, like WordNet. If a model consistently makes better mistakes on in-domain data, we can reasonably assume that the model has captured more transferable features for class discrimination.

Class Taxonomy and Mistake Severity: Class taxonomy or ontology has been widely utilized in literature to indicate class formation [11, 71] and semantic relationships [16, 4, 72, 53, 39] between classes, offering a hierarchical organization of classes or categories. Following these works, we consider the WordNet class taxonomy [44] as an approximation of natural class taxonomy. The severity of a mistake in many studies is quantified as the shortest path from the prediction node to the least common ancestor (LCA) in a predefined class hierarchy. This metric, known as 'LCA distance' or 'hierarchical error', was used in the early years of the ImageNet [11] challenge. However, it was largely dismissed as it was widely believed to follow the same ordering as Top 1 accuracy [11, 5]. In this work, we revisit this metric and empirically demonstrate that Top 1 accuracy and LCA distance do not always align when VLMs are involved, challenging the common notion. We also appeal for community attention to revisit this metric with its potential usage in measuring a model's feature awareness to indicate generalization.

Causal/Invariant Representation Learning for OOD Generalization. Recently, there has been an increase in OOD generalization research towards formulating training and testing distributions with causal structures [1, 7, 49], where shifts in distribution primarily arise from interventions or confounding factors. Building upon this, methods[57, 60, 63] such as CausalVAE [74] have been proposed, leveraging learned causal representations to capture the causal relationships underlying the data generation process [30], which helps mitigate the distributional shifts caused by interventions.

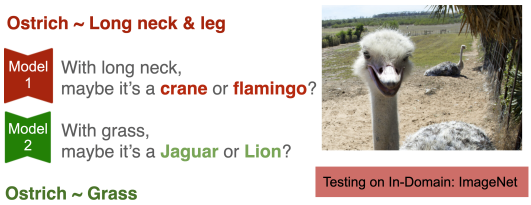

Figure 7: We hypothesize that models capturing more transferable features tend to predict classes that are semantically closer to the ground truth.

While the connection between OOD generalization and causal concepts is not entirely novel, those attempts have focused on the causal structure at the latent or abstract level, lacking both interpretability and transparency. Our method aligns with this growing interest in Causal/Invariant learning, which aims to capture the invariant latent data generation process. One should expect a model prediction that better aligns with the data generation process to be more robust under intervention, thus generalizing better. Although it's less feasible to model the data generation process of natural images (ImageNet), we essentially follow the same intuition and hypothesize that the WordNet class hierarchy serves as an approximation of the invariant relationship between class concepts [56]. WordNet is a widely recognized and effective means of encoding semantic relationships between concepts, making it an appropriate proxy for aligning human semantic knowledge [45]. Unlike previous work, the WordNet hierarchy provides interpretability, adding a level of transparency to our understanding of model generalization.

8 Enhancing Generalization through Class Taxonomy Alignment.

Building upon the earlier discussion, we explore how the devised method can be utilized to enhance a model's generalization capability.

8.1 Improving Generalization by Class Taxonomy Alignment with taxonomy loss

Inferring Class Taxonomy from a Pretrained Model Using K-Means Clustering. In a previous experiment, we adopted the WordNet hierarchy as class taxonomy to calculate LCA distance. While the number of publicly available datasets providing class taxonomy is limited [11, 71], the usefulness of our method is unquestionable. Hence, we propose a method to construct a latent class taxonomy, expanding the potential applications of our work. We show that such a constructed taxonomy could achieve similar performance to the WordNet hierarchy.

Stats among 75 latent hierarchy	Element		ImageNetV2	ImageNet-S	ImageNet-R	ImageNet-A	ObjectNet
	ID	OOD					
Top1	Top1	Top1	**0.980**	0.274	0.141	0.093	0.522
LCA (Mean)	LCA	Top1	0.815	**0.773**	**0.712**	**0.662**	**0.930**
LCA (Min)	LCA	Top1	0.721	0.715	0.646	0.577	0.890
LCA (Max)	LCA	Top1	0.863	0.829	0.780	0.717	0.952
LCA (std)	LCA	Top1	0.028	0.022	0.027	0.025	0.010

Table 3: **Correlation Measurement(R^2) between ID LCA/Top1 and OOD Top1 across 75 Latent Hierarchies Derived from K-means**. Our latent hierarchy construction is robust among 75 different pretrained model adoptions. For each pretrained model, we constructed a 75-class taxonomy hierarchy using the K-means clustering method described previously. We then calculated the LCA for each hierarchy as an in-domain indicator and compared it to OOD accuracy using the same settings as in 1.

The essence of class taxonomy lies in its representation of inter-class distance, encoding class proximity and identifying which classes cluster closely in semantic space. In this spirit, we construct a class taxonomy matrix using K-means clustering. As illustrated in Fig 8, we adopt average class features to cluster data hierarchically at 10 different levels, with an increasing number of clusters to indicate class adjacency. Experiments in Tab 3 show that our method is very robust regardless of which model was used to construct the class hierarchy. Further implementation details in appendix 11.1.

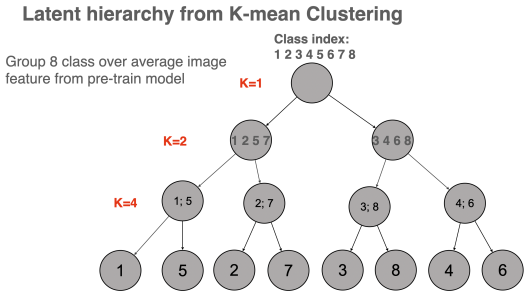

Figure 8: **Visualization of K-mean clustering.** We adopt a pre-trained model and perform K-mean clustering with various numbers of K over encoded image features to construct a latent hierarchy for calculating LCA distance. In Tab 3, we show that robust performance can be achieved among 75 different pretrained models.

Employing Class Taxonomy as Soft Labels. We propose a straightforward approach to demonstrate the potential of LCA as a benchmarking tool for generalization. We encode the normalized pairwise LCA between each class as soft labels and apply linear probing over the pretrained model. Contrary to the rigid probabilistic distribution of single-label classification, we formulate the problem as multi-labeling. Besides, we employ a sigmoid-style [6] BCE loss instead of softmax, relaxing the constraints on inter-class interaction. A more detailed setup will be included in the appendix.

Following the methods above, we constructed class taxonomy matrices for AlexNet [32] and Swin Transformer[41], representing the best and worst performing models on ImageNet in our model pool. Following the intuition of model distillation [24], the hierarchy constructed from the model's pretrained features partially encapsulates the model's interpretation of interclass relationships. As illustrated in Table 4, incorporating more accurate inter-class distances consistently enhances OOD performance across all four OOD benchmarks, albeit with slightly lower Top 1 accuracy.

However, this approach does lead to a slight drop in in-domain accuracy as it less intensively optimizes towards the ground truth class. Inspired by the notion that models are more confident where they excel [73], we apply linear interpolation between linear layers trained from cross-entropy and our proposed loss function. The results suggest that this method strikes a balance, delivering competitive performance on both ID and OOD datasets.

Importantly, we find that models using hierarchies constructed from pretrained models fall short in OOD generalization compared to those utilizing the WordNet hierarchy, even though they exhibit slightly improved ID performance. This indicates that enforcing arbitrary inter-class relationships, derived from in-domain datasets, can negatively affect OOD performance.

	ImageNet	ImageNetv2	ImageNet-S	ImageNet-R	ImageNet-A	ObjectNet
Baseline	**0.690**	**0.5618**	**0.199**	0.322	0.010	0.267
AlexNet Hier	0.665	0.5402	0.189	0.294	0.017	0.247
Swin-T Hier	0.668	0.5429	0.196	0.312	0.023	0.259
WordNet Hier	0.664	0.5387	**0.199**	**0.329**	**0.024**	**0.272**
(CE + CE) Interp	**0.695**	0.5645	0.196	0.325	0.011	0.273
(AlexNet + CE) Interp	0.694	0.5665	0.200	0.325	0.012	0.274
(Swin-T + CE) Interp	**0.695**	**0.5694**	0.202	0.331	0.012	0.274
(WordNet + CE) Interp	0.694	0.5638	**0.2073**	**0.335**	**0.014**	**0.282**

Table 4: **Interpolating Class Taxonomy to Linear Probing on ResNet18 Feature**. Training with a WordNet hierarchy delivers the most significant improvements across OOD benchmarks despite lower Top 1 accuracy, whereas models using hierarchies inferred from pretrained models yield lesser gains. The top table displays results from models trained using latent hierarchy constructed from the indicated model via K-means. The bottom table presents the results of the aforementioned models when interpolated with layers trained from cross entropy in the weight space [73].

8.2 Improving Generalization by Class Taxonomy Alignment with Prompt Engineering

In this section, we discuss results on enhancing model generalization through Taxonomy Integration in Vision-Language Models.

For vision-language models, integrating taxonomy-specific knowledge during zero-shot evaluation is straightforward. The WordNet hierarchy naturally indicates inter-class distances during data generation. For example, 'dalmatian' and 'husky' are semantically close, both originating from the parent node 'dog'. We detail the results with CLIP-vit32 [50] in Tab 5. To test our hypothesis, we explicitly integrated hierarchical taxonomy relationships into the prompt for zero-shot VLM prediction. The prompt was designed as **'A, which is a type of B, which is a type of C'**, informing

the model to make taxonomy-aligned predictions. Additionally, we included two ablation studies: 1) providing the correct taxonomy path without informing the model of the class name relationships **(Stack Parent)**; and 2) informing the model of the hierarchical 'is-a' relationship but providing an incorrect taxonomy relationship randomly sampled from the tree **(Shuffle Parent)**. Our results demonstrate that informing the model of both the correct taxonomy and their hierarchical relationships significantly improves generalization, as evidenced by improvements in Top-1 accuracy, ELCAD, and test-time Cross-Entropy(CE) across all datasets for all tested models.

Model	ImageNet			ImageNetv2			ImageNet-S			ImageNet-R			ImageNet-A			ObjectNet		
	Top1	Test CE	ELCA	Top1	Test CE	ELCA	Top1	Test CE	ELCA	Top1	Test CE	ELCA	Top1	Test CE	ELCA	Top1	Test CE	ELCA
Baseline	0.589	1.635	9.322	0.517	2.014	9.384	0.379	2.817	9.378	0.667	1.348	8.790	0.294	3.098	9.358	0.394	2.631	8.576
Stack Parent	0.381	3.730	9.389	0.347	3.948	9.395	0.219	5.540	9.561	0.438	3.287	9.258	0.223	4.469	9.364	0.148	5.127	9.076
Shuffle Parent	0.483	2.236	9.679	0.432	2.586	9.696	0.329	3.251	9.718	0.557	1.919	9.281	0.236	3.532	9.586	0.329	3.067	8.785
Taxonomy Parent	**0.626**	**1.457**	**9.102**	**0.553**	**1.824**	**9.165**	**0.419**	**2.544**	**9.319**	**0.685**	**1.279**	**8.658**	**0.319**	**2.839**	**9.171**	**0.431**	**2.433**	**8.515**

Table 5: **Accuracy on OOD dataset by enforcing class taxonomy: Baseline**: <*dalmatian*>; **Stack Parent**: <*dalmatian, dog, animal*>; **Taxonomy Parent**:<*dalmatian, which is type of a dog, which is type of an animal*>; **Shuffle Parent**: <*dalmatian, which is type of an organism, which is type of a seabird*>; We demonstrate that integrating both the correct structural information (informing the hierarchical 'is-a' relationship between class names) and valid taxonomy relationships from WordNet significantly boosts model performance and generalization.

9 Discussion

Reestablishing LCA as a Comprehensive Measure of Model Generalization. While Top 1 ID accuracy [47] demonstrates a clear linear trend with OOD datasets in models with similar training mechanisms, this relationship becomes less distinct in vision-only and VLMs. This finding, echoed in earlier studies[15, 73, 9], suggests a more nuanced understanding of how zero-shot VLMs with lower Top-1 accuracy can outperform competitive vision models in generalizing to unfamiliar datasets. While previous works have emphasized the significant impact of data diversity on generalization[15, 58, 30], our results indicate that the LCA offers a more all-encompassing assessment of model generalization. By considering factors such as training data size, architecture, loss, and others, LCA provides a fuller measure of a model's ability to accurately capture semantic distinctions common across ID and OOD benchmarks. This establishes a comprehensive benchmark that encompasses various generalization factors, addressing the issue of inflated VLM effectiveness on "Effective Robustness[69]." Future research should delve into large-scale analytic studies of generalization factors in conjunction with LCA.

ImageNet-v2 Demonstrates Similar Class Discrimination Features to ImageNet. ImageNet-v2, a recollection of ImageNet, is often used as an OOD dataset for ImageNet-based studies[59, 47, 2]. Our experiments indicate that ImageNet-v2 more closely resembles ImageNet than other OOD datasets. We hypothesize that the minimal external intervention in ImageNet-v2's data generation process results in visual similarities to ImageNet, allowing even spurious relationships encoded from ImageNet to transfer successfully to ImageNet-v2. Consequently, models pretrained on ImageNet (VMs) inflate accuracy on ImageNet-v2, disrupting the alignment with trends observed in VLMs.

Is it Possible for a Semantically-Aware (Low LCA) Model to Have Low Top 1 Accuracy? Our empirical analysis indicates a correlation: models not specifically tuned on class taxonomy, with lower Top 1 accuracy, tend to exhibit higher LCA distances. However, this relationship is correlational rather than causal. It remains feasible to design a model adversarially so it consistently predicts the semantically nearest class to the true class. In such cases, the model would show a low LCA distance while maintaining zero Top 1 accuracy. Therefore, while a correlation exists between Top 1 accuracy and LCA, causality cannot be inferred, and this relationship can be disrupted under deliberate adversarial training.

Does ImageNet LCA (Taxonomy Distance) Reflect ImageNet Top 1 Accuracy? It is often suggested that LCA and Top-1 accuracy exhibit similar trends on the same dataset [11, 5]. Intuitively, a high-performing model better fits the data distribution, leading to fewer severe errors. This pattern generally holds true for models under similar settings (either VM or VLM). However, when considering both VM and VLM models, ImageNet and ImageNet-v2 exhibit only a weak correlation between LCA and Top-1 accuracy, whereas other semantically distinct OOD datasets show a stronger

relationship. This finding challenges the prevailing belief that in-domain Top-1 accuracy and LCA maintain the same ranking[12, 5].

10 Metric

In this section, we outline the metrics adopted for our experiment.

10.1 Correlation Measurement

Correlation measurements quantify the degree of association between two variables. This can be further subdivided into linearity and ranking measurements.

10.1.1 Linearity Measurement

Linearity measurement evaluates the strength and direction of a linear relationship between two continuous variables. We use the R^2 and Pearson correlation coefficients to assess linearity.

R^2 (Coefficient of determination): The R^2, or coefficient of determination, quantifies the proportion of the variance in the dependent variable that can be predicted from the independent variable(s). It ranges from 0 to 1, where 1 indicates perfect predictability. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f(x_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where $f(x_i)$ is the prediction of y_i from the model, $\bar{y}$ is the mean of the actual y values, and n is the number of data points.

PEA (Pearson correlation coefficient): The Pearson correlation coefficient, denoted as r, measures the linear relationship between two datasets. It is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the datasets x and y, respectively, and n is the number of data points.

10.1.2 Ranking measurement

Ranking measurement evaluates the degree of correspondence between the rankings of two variables, even when their relationship is non-linear. The Kendall and Spearman rank correlation coefficients are metrics used for this purpose.

KEN (Kendall rank correlation coefficient): Also known as Kendall's tau (τ), this coefficient measures the ordinal association between two variables. It is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)} \tag{3}$$

where n is the number of data points.

SPE (Spearman rank-order correlation coefficient): The Spearman rank-order correlation coefficient, denoted as ρ, assesses the monotonic relationship between two variables. It is defined as:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{4}$$

where d_i is the difference between the ranks of corresponding data points in the two datasets and n is the number of data points.

10.2 Taxonomy Measurement

Taxonomy measurement is designed to assess the alignment between the model-predicted class ranking and the predefined class taxonomy hierarchy tree. This is also referred to as 'mistake severity' or 'taxonomy distance'.

10.2.1 LCA distance

Following [5, 70], we define LCA distance using a predefined hierarchy tree, as indicated in Fig3. We adopt class distance in a hierarchical tree format to denote inter-class relationships, which is necessary to calculate LCA and ELCA. Given a ground truth node y (node 1 in the plot) and a model prediction node y' (node 3 in the plot), their LCA node $LCA(y, y')$ is node 6 in the plot. We define it as:

$$D_{LCA}(y', y) := f(LCA(y', y)) - f(y), \tag{5}$$

where $f(\cdot)$ represents a function for a node's score, such as the tree depth or information content.

Scores as tree depths: We define a function $P(x)$ to retrieve the depth of node x from tree T. Then, LCA distance is defined as:

$$D_{LCA}(y', y)_P := (P(y) - P(LCA(y', y))) + (P(y') - P(LCA(y', y))), \tag{6}$$

where we also append $P(LCA(y', y)) - P(y')$ to counter tree imbalance.

Scores as information: Defining score as tree depth may be vulnerable to an imbalanced hierarchical tree. Thus, we also define a node's score as information to put more weight on nodes with more descendants. Formally, following [70], we apply a uniform distribution p to all leaf nodes in the tree that indicate a class in the classification task. The probability of each intermediate node in the tree is calculated by recursively summing the scores of its descendants. Then, the information of each node is calculated as $I(node) := -log2(p)$. The LCA distance is then defined as:

$$D^I_{LCA}(y', y) := I(y) - I(LCA(y', y)), \tag{7}$$

In this work, we adopt $D^I_{LCA}(y', y)$ for objectNet, ImageNet-R, and ImageNet-v2, and $D^P_{LCA}(y', y)$ for ImageNet-S, and ImageNet-A to achieve optimal performance. Both metrics can significantly outperform Top1 in-domain accuracy.

10.3 Generalizate LCA to Expected LCA

We can also generalize the LCA distance to settings where the model outputs a distribution over all possible classes for each sample (like using softmax). For a sample X_i whose ground truth class is y_i, and the model outputs $(\widehat{p}_{1,i}, \ldots, \widehat{p}_{K,i})$ over the K classes (e.g., 1000 in ImageNet), we define the **Expected Lowest Common Ancestor Distance** (ELCA): $D_{ELCA}(model, \mathcal{M}) := \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} \widehat{p}_{k,i} \cdot D_{LCA}(k, y_i)$. From a probabilistic perspective, D_{ELCA} is a weighted measure of mistake severity according to the model's confidence in each node in the hierarchy. Intuitively, it combines the LCA distance with a cross-entropy measurement.

Model	ImageNet			ImageNetv2			ImageNet-S			ImageNet-R			ImageNet-A			ObjectNet		
	LCA	ELCA	Top1	LCA	ELCA	Top1	LCA	ELCA	Top1	LCA	ELCA	Top1	LCA	ELCA	Top1	LCA	ELCA	Top1
ResNet18 [20]	6.643	7.505	0.698	6.918	7.912	0.573	8.005	9.283	0.202	8.775	8.853	0.330	8.449	9.622	0.011	8.062	8.636	0.272
ResNet50 [20]	6.539	**7.012**	**0.733**	6.863	**7.532**	**0.610**	7.902	**9.147**	0.235	8.779	**8.668**	0.361	8.424	**9.589**	0.018	8.029	**8.402**	0.316
CLIP_RN50 [50]	6.327	~~9.375~~	0.579	6.538	~~9.442~~	0.511	6.775	~~9.541~~	0.332	7.764	~~9.127~~	0.562	7.861	~~9.526~~	0.218	7.822	~~8.655~~	0.398
CLIP_RN50x4 [?]radford2021learning}	**6.166**	9.473	0.641	**6.383**	9.525	0.573	**6.407**	**9.518**	**0.415**	**7.435**	**8.982**	**0.681**	**7.496**	**9.388**	**0.384**	**7.729**	**8.354**	**0.504**

Table 6: **Model performance corresponds to mistake severity. LCA ↓ / ELCA ↓ /Top1 ↑** indicate measurements on a given dataset. We present two pairs of model comparisons from the VMs and VLMs families with different generalization abilities. Note that ELCA should not be compared across modalities, as it is sensitive to logit temperature.

The proposed ELCA distance provides a more generalized metric for assessing model performance compared to Top 1 accuracy, LCA distance, and cross entropy. Top 1 accuracy only considers the top-ranked class; LCA distance measures the Top n class rankings but treats each class equally [5]; Cross-entropy solely focuses on the model's assigned probability to the ground truth class, and ELCA extends it to all classes. The ELCA distance captures the probabilistic distribution of mistake severity across all candidate classes.

In Table 6, we empirically demonstrate that models with better OOD generalization (OOD Top 1 accuracy) also have lower LCA/ELCA distances.

11 Experiment Setup

11.1 K-mean Clustering for Latent Class Hierarchy Construction

As depicted in Fig 8, we begin with a pretrained model M, in-domain image data X, and labels y for k classes. Initially, we extract the in-domain data features $M(X)$. With known labels, we categorize $M(X)$ by y, resulting in k average class features, denoted as kX. Utilizing these per-class features, we perform a 10-layer hierarchical clustering. For kX, we apply the K-means algorithm, setting the number of cluster centers as 2^i, where i ranges from $1, 2, 3, 4, ..., 9$ since $2^9 < 1000$. This procedure results in 9 cluster outcomes. Subsequently, we calculate the pairwise LCA between the k classes, determining the cluster level at which both classes share the same cluster as their LCA height. By definition, all classes share a base cluster level of 10.

11.2 Loss for Linear Probing Experiment

In our linear probing experiment, we define the loss function as follows. For a class with n classes, we first establish an n $\times$ n LCA distance matrix M, where M[i,k] indicates the pairwise LCA distance $D_{LCA}(i, k)$, with LCA calculated using either WordNet hierarchy or the hierarchy derived from the K-mean algorithm (as introduced in the main paper). Next, we scale M by applying an exponential function, MinMax scaling, and normalize to 1 for each row, i.e., M = $normRow(minmaxScaling(M.exp()))$. For the loss computation, we use Binary Cross Entropy (BCE) and adopt the corresponding row value as a soft label. Specifically, if class-i is the ground truth for a given data instance, we use M[i,:] as the soft label.

11.3 LCA Matrix from Pretrained Model

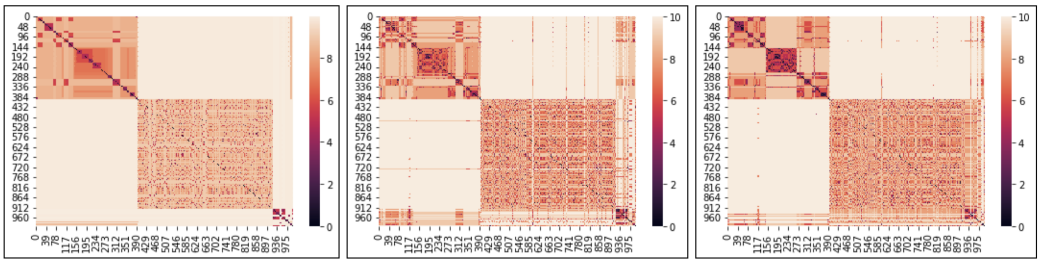

Figure 9: **Comparison between LCA distance matrices**. From left to right: WordNet hierarchy; matrix constructed from AlexNet [32]; and matrix constructed from CLIP ResNet50 [50]. We observe a higher alignment between the CLIP RN50 LCA distance matrix and the WordNet hierarchy as compared to the one from AlexNet.

In Figure 9, we present a comparison of LCA distance matrices, with the diagonal index indicating the shortest distance. Each row signifies the class distance between a specific class and the reference class, arranged in ascending order. Furthermore, we generated 36 LCA distance matrices from pretrained models on ImageNet. The findings illustrated in Figure 10 and Table 7 reveal a moderate correlation between the in-domain LCA of the source model and the generalization capabilities of the linear probe model. They also suggest that a model's generalization ability could be modified by enforcing different inter-class distances, with minimal impact on in-domain accuracy. Our future research will further explore the relationship between inter-class distance in pretrained models and their generalization capabilities.

11.4 Hyperparameters and Computational Resources

In the linear probing experiment, we chose hyperparameters based on the task at hand. The learning rate was set to 0.001, batch size=1024. We used the AdamW optimizer with a weight decay and a cosine learning rate scheduler with a warm-up iteration. The warm-up type was set to 'linear' with a warm-up learning rate of 1e-5. The experiment was run for 50 epochs.

For our computational resources, we utilized a single NVIDIA GeForce GTX 1080 Ti GPU.

	ImageNet		ImageNetv2		ImageNet-S		ImageNet-R		ImageNet-A		ObjectNet	
	PEA	SPE	PEA	SPE	PEA	SPE	PEA	SPE	PEA	SPE	PEA	SPE
LCA ->Hierarchy Linear Prob	0.672	0.462	0.712	0.466	0.719	0.625	0.799	0.733	0.640	0.526	0.622	0.424

Table 7: **Correlation measurement between LCA matrix and In-domain LCA on ResNet18.** Following the algorithm of K-Means Clustering, we construct 36 LCA distance matrices (class hierarchies) from different pretrained models on ImageNet. We then use the LCA distance matrices as soft labels to guide linear probing on ResNet18 features. The table indicates the relationship between the In-domain LCA of the pretrained model and the out-of-distribution (OOD) accuracy on the linear probe model using the corresponding LCA distance matrix. The result is calculated from the average of three random seeds. Visualization is shown in Figure 10.

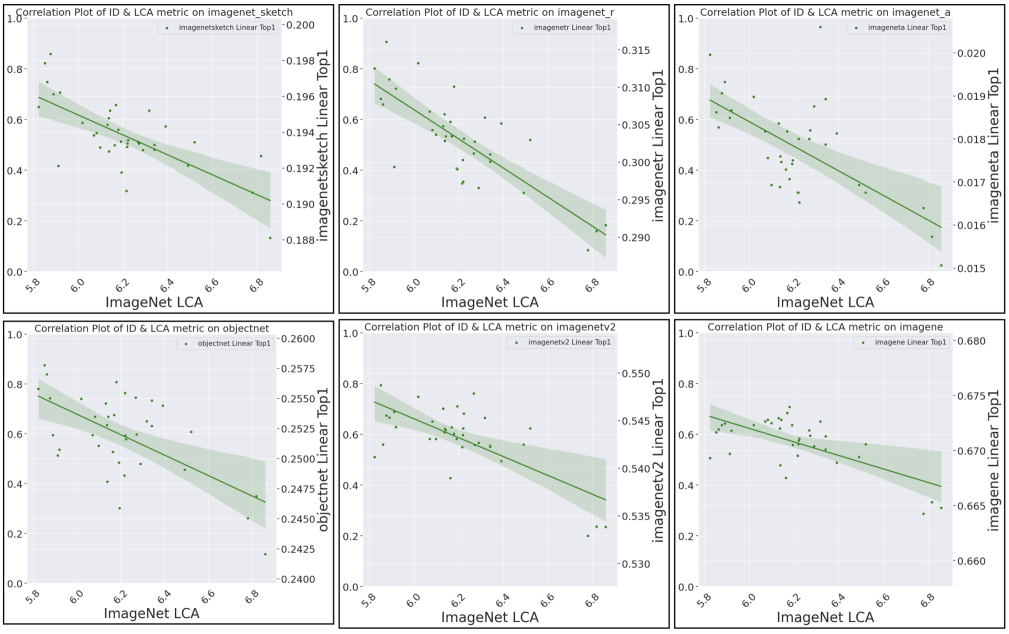

Figure 10: **Coorelation measurement between LCA matrix and In domain LCA on ResNet18.** Visualization on result in Tab 7. Plot shows an intermediate correlation between the two variable. If necessary, please find png of this image in supplementary for better legibility.

12 Supplementary Result

12.1 Comprehensive results from main paper

Extended from Tab 1 and Tab 2 in main paper, we present measurement on only VMs and VLMs in Tab 8 and Tab 9. Equivalently, LCA is also a very good OOD indicator when only involved VM or VLM.

	Element		**ImageNetv2**		**ImageNet-S**		**ImageNet-R**		**ImageNet-A**		**ObjectNet**	
	ID	OOD	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA
ALL	Top1	Top1	**0.962**	**0.980**	0.075	0.275	0.020	0.140	0.009	0.094	0.273	0.522
	LCA	Top1	0.339	0.582	**0.838**	**0.915**	**0.779**	**0.883**	**0.869**	**0.932**	**0.915**	**0.956**
	Top1	Top5	**0.889**	**0.943**	0.052	0.229	0.004	0.060	0.013	0.115	0.262	0.512
	LCA	Top5	0.445	0.667	**0.883**	**0.940**	**0.738**	**0.859**	**0.909**	**0.953**	**0.924**	**0.961**
VLM	Top1	Top1	**0.996**	**0.998**	0.860	0.927	0.851	0.923	0.578	0.761	**0.945**	**0.972**
	LCA	Top1	0.956	0.978	**0.922**	**0.960**	**0.889**	**0.943**	**0.792**	**0.900**	0.936	0.968
	Top1	Top5	**0.988**	**0.994**	0.867	0.931	0.820	0.906	0.740	0.860	**0.970**	**0.985**
	LCA	Top5	0.930	0.964	**0.949**	**0.974**	**0.848**	**0.921**	**0.828**	**0.910**	0.931	0.965
VM	Top1	Top1	**0.996**	**0.998**	0.824	0.908	**0.801**	**0.895**	0.523	0.723	0.900	0.949
	LCA	Top1	0.976	0.988	**0.895**	**0.945**	0.768	0.877	**0.833**	**0.912**	**0.913**	**0.956**
	Top1	Top5	**0.993**	**0.997**	0.829	0.910	**0.821**	**0.906**	0.696	0.834	0.919	0.959
	LCA	Top5	0.970	0.985	**0.925**	**0.962**	0.777	0.882	**0.925**	**0.962**	**0.936**	**0.967**

Table 8: **Correlation measurement of ID LCA/Top1 with OOD Top1/Top5** on 75 models across modality (36 VMs and 39 VLMs) following Fig 4. The 'ALL grouping' demonstrates that LCA has a strong correlation with OOD performance on all datasets (except ImageNet-v2). We take the absolute value of all correlations for simplicity. Equivalently, LCA is also a very good OOD indicator when only involved VM or VLM.

		ImageNetv2	**ImageNet-S**	**ImageNet-R**	**ImageNet-A**	**ObjectNet**
ALL	ID Top1 [47]	**0.040**	0.230	0.277	0.192	0.178
	AC [22]	0.043	0.124	0.113	0.324	0.127
	Aline-D [2]	0.121	0.270	0.167	0.409	0.265
	Aline-S [2]	0.072	0.143	0.201	0.165	0.131
	(Ours) ID LCA	0.162	**0.078**	**0.107**	**0.061**	**0.048**
VLM	ID Top1 [47]	**0.014**	0.077	0.064	0.127	0.052
	AC [22]	0.029	**0.050**	**0.044**	0.217	0.088
	Aline-D [2]	0.151	0.250	0.081	0.296	0.260
	Aline-S [2]	0.070	0.069	0.068	**0.080**	0.153
	(Ours) ID LCA	0.047	0.059	0.062	0.094	**0.043**
VM	ID Top1 [47]	**0.013**	0.099	0.108	0.143	0.068
	AC [22]	0.059	0.204	0.188	0.441	0.168
	Aline-D [2]	0.083	0.427	0.313	0.665	0.364
	Aline-S [2]	0.105	0.182	0.092	0.574	0.216
	(Ours) ID LCA	0.029	**0.079**	**0.113**	**0.080**	**0.056**

Table 9: **Error Prediction of OOD Datasets** across 75 models of diverse settings with **MAE loss ↓**. Top1 in **bold** and Top2 in underline. Despite ImageNet's in-domain accuracy maintain as a significant indicator of ImageNet-v2 accuracy, the in-domain LCA outperforms it as a robust error predictor across four naturally distributed OOD datasets, particularly ImageNet-A, which stumps other methods.

12.2 Does ImageNet LCA (Taxonomy Distance) Reflect ImageNet Top 1 Accuracy?

Here, we present numerical results supporting the discussion in discussion section9. We challenge the common belief that LCA and Top-1 accuracy follow parallel trends within the same dataset[5, 11]. As illustrated in Figures 11 and 10, when including both VM and VLM zero-shot models, ImageNet and ImageNet-v2 show a weak correlation between LCA and Top-1 accuracy, while other semantically distinct OOD datasets exhibit a stronger relationship.

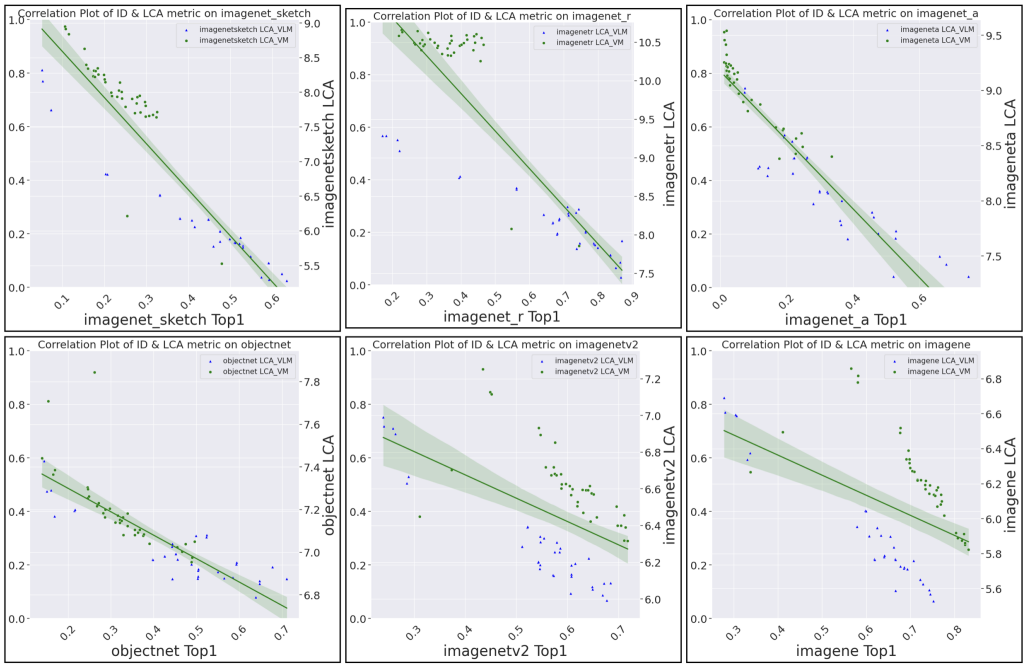

Figure 11: **Predicting LCA (VM+VLM, 75 models) on 6 ImageNet-variant Datasets** As per Tab 10. Each plot's x-axis represents dataset Top-1 accuracy, while the y-axis shows LCA distance. The plots reveal that ImageNet and ImageNet-v2 do not exhibit a strong correlation between LCA and Top-1 accuracy, in contrast to other semantically distinct OOD datasets. This observation challenges the common belief that in-domain Top-1 accuracy and LCA distance maintain the same order[12, 5]. For further details, please refer to the discussion. For better legibility, a png of this image can be found in the supplementary materials.

Model	Group	ImageNet		ImageNetv2		ImageNet-S		ImageNet-R		ImageNet-A		ObjectNet	
Top1->LCA	ALL	*K^2*	*PEA*	*K^2*	*PEA*	*K^2*	*PEA*	*K^2*	*PEA*	*K^2*	*PEA*	*K^2*	*PEA*
		0.237	0.488	0.259	0.509	**0.838**	**0.915**	**0.749**	**0.865**	**0.869**	**0.932**	**0.672**	**0.820**
		KEN	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*
		0.293	0.302	0.298	0.380	**0.828**	**0.937**	**0.600**	**0.795**	**0.813**	**0.948**	**0.727**	**0.901**
	VLM	K^2	PEA	K^2	PEA	K^2	PEA	K^2	PEA	K^2	PEA	K^2	PEA
		0.934	**0.966**	**0.886**	**0.941**	**0.922**	**0.960**	**0.889**	**0.943**	**0.792**	**0.890**	0.570	**0.755**
		KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE
		0.848	**0.955**	**0.684**	**0.853**	**0.867**	**0.959**	**0.686**	**0.861**	**0.689**	**0.879**	0.494	**0.704**
	VM	K^2	PEA	K^2	PEA	K^2	PEA	K^2	PEA	K^2	PEA	K^2	PEA
		0.976	**0.987**	**0.893**	**0.945**	**0.895**	**0.945**	0.095	0.310	**0.833**	**0.913**	**0.913**	**0.956**
		KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE
		0.911	**0.982**	**0.821**	**0.942**	**0.825**	**0.949**	0.149	0.222	**0.782**	**0.917**	**0.838**	**0.957**

Table 10: **Correlation measurement between Top 1 and LCA** on 77 models across modality (37 VM and 40 VLM) on 6 datasets; For instance, Corr(ImageNet Top1 Acc, ImageNet LCA) or Corr(ImageNet-A Top1 Acc, ImageNet-A LCA); Follow Fig 11. We highlight strong correlation indications. We take the absolute value of all correlations for simplicity.

12.3 Ranking Measurement of LCA-on-the-Line

Here we present the numeric result for ranking measures of *KEN (Kendall rank correlation coefficient)* and *SPE (Spearman rank-order correlation coefficient)* in comparision to common use Top1 In domain accuracy in 11. Equalevently, in domain LCA measure present strong result in both preserving linearity and ranking.

	Element		**ImageNetv2**		**ImageNet-S**		**ImageNet-R**		**ImageNet-A**		**ObjectNet**	
	ID	OOD	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE
ALL	Top1	Top1	**0.840**	**0.947**	0.170	0.092	0.146	0.042	0.068	0.037	0.317	0.339
	LCA	Top1	0.421	0.517	**0.828**	**0.937**	**0.761**	**0.911**	**0.813**	**0.948**	**0.867**	**0.967**
	Top1	Top5	**0.672**	**0.818**	0.151	0.059	0.134	0.004	0.108	0.021	0.279	0.297
	LCA	Top5	0.571	0.729	**0.843**	**0.948**	**0.752**	**0.897**	**0.817**	**0.947**	**0.861**	**0.966**
VLM	Top1	Top1	**0.971**	**0.997**	0.840	0.936	**0.864**	**0.943**	0.753	0.915	**0.905**	**0.982**
	LCA	Top1	0.882	0.972	**0.867**	**0.959**	0.762	0.886	**0.800**	**0.942**	0.870	0.972
	Top1	Top5	**0.908**	0.980	0.848	**0.951**	**0.882**	**0.959**	0.753	0.910	**0.842**	**0.964**
	LCA	Top5	0.900	**0.981**	**0.856**	0.950	0.775	0.907	**0.794**	**0.943**	0.829	0.955
VM	Top1	Top1	**0.948**	**0.993**	0.771	0.901	**0.743**	**0.887**	0.735	0.877	0.822	0.927
	LCA	Top1	0.910	0.981	**0.825**	**0.949**	0.705	0.862	**0.782**	**0.920**	**0.838**	**0.957**
	Top1	Top5	**0.939**	**0.992**	0.752	0.894	**0.758**	**0.901**	0.818	0.941	0.815	0.920
	LCA	Top5	0.894	0.977	**0.832**	**0.951**	0.707	0.871	**0.824**	**0.939**	**0.846**	**0.958**

Table 11: **Ranking measurement of ID LCA/Top1 with OOD Top1/Top5** on 75 models across modality(36 VM and 39 VLM); As shown in the 'ALL grouping', LCA shows a much better result in preserve in model relative ranking to model OOD performance on all OOD datasets (with the exception of ImageNet-v2), which indicate the superiority for model selection.

13 Limitations, Conclusions, and Future Directions

While we benchmarked and used LCA based on class hierarchy to measure model generalization, the findings from this work indicate that it is not an effective indicator for datasets visually similar to In-domain data (like ImageNet2). For these datasets, In-domain Top1 remains a strong indicator, which potentially limits the utility of LCA. Also, it's expected that LCA will shows a weaker discrimination between models on datasets with small number of class (like Cifar [31]).

In conclusion, this work reinvigorates LCA distance using class taxonomy like WordNet as a indicator for model OOD generalization. Intuitionally, we discuss such semantic measurement of mistake severity could indicate the transferablility of model's learned feature for classification. Ideally, models that capture more transferable feature should make fewer severe mistakes. Across multiple ImageNet-OOD datasets, we showed that severity of in domain mistakes could served as a unified metric to indicate model generalization among models supervised from either class label(VMs) or captions(VLMs)

This relationship is not reflected when using the widely-accepted in-domain Top 1 accuracy [47]. Furthermore, we demonstrated that aligning model predictions with class taxonomy, whether through prompt engineer or introducing regularization loss, can enhance model generalization. Future direction could focus on provide theoretical justification under LCA-on-the-line, and perform larger scale empirical study regarding this benchmark. This work provides new insights into model generalization using existing resources and encourages further investigation in this direction.