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 3] **Francesco Ferroni** [2 4] **Shu Kong** [5 6]

Abstract

We tackle the challenge of predicting models' Out-of-Distribution (OOD) performance using in-distribution (ID) measurements without requiring OOD data. Existing evaluations with "Effective robustness", which use ID accuracy as an indicator of OOD accuracy, encounter limitations when models are trained with diverse supervision and distributions, such as class labels (*Vision Models, VMs, on ImageNet*) and textual descriptions (*Visual-Language Models, VLMs, on LAION*). VLMs often generalize better to OOD data than VMs despite having similar or lower ID performance. To improve the prediction of models' OOD performance from ID measurements, we introduce the *Lowest Common Ancestor (LCA)-on-the-Line* framework. This approach revisits the established concept of LCA distance, which measures the hierarchical distance between labels and predictions within a predefined class hierarchy, such as WordNet. We assess 75 models using ImageNet as the ID dataset and five significantly shifted OOD variants, uncovering a strong linear correlation between ID LCA distance and OOD top-1 accuracy. Our method provides a compelling alternative for understanding why VLMs tend to generalize better. Additionally, we propose a technique to construct a taxonomic hierarchy on any dataset using K-means clustering, demonstrating that LCA distance is robust to the constructed taxonomic hierarchy. Moreover, we demonstrate that aligning model predictions with class taxonomies, through soft labels or prompt engineering, can enhance model generalization. Open source code in our Project Page.

[1]Carnegie Mellon University [2]Work done at Argo AI GmbH [3]Apple [4]Nvidia [5]Texas A&M University [6]University of Macau. Correspondence to: Jia Shi <jiashi@alumni.cmu.edu>, Shu Kong <skong@um.edu.mo>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

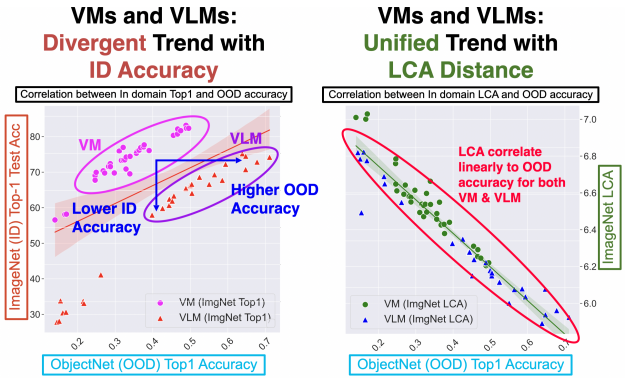

Figure 1. **Correlation between LCA Distance and Out-of-Distribution (OOD) Performance in Vision and Vision-Language Models.** In both panels, the X-axis represents the top-1 accuracy on ObjectNet (OOD test dataset). The Y-axes depict the top-1 accuracy (left-axis) and LCA distance (right-axis) on ImageNet (ID test dataset). The left plot reveals a divergent trend where Vision Models (VMs) show a trade-off between OOD and ID accuracy, while Vision-Language Models (VLMs) tend to maintain higher OOD accuracy regardless of ID performance. The right plot demonstrates a unified, strong positive correlation between LCA distance and OOD accuracy for both VMs and VLMs, showing that LCA distance is a robust metric for evaluating model generalization across different architectures.

1. Introduction

Generalizing models trained on in-distribution (ID) data to out-of-distribution (OOD) conditions is a notoriously difficult task. Distribution shifts undermine the independent and identically distributed (IID) assumption between training and testing data, challenging the model's robustness. Numerous OOD datasets have been proposed to study the effects of different interventions, such as temporal shifts (Hu et al., 2022; Lomonaco & Maltoni, 2017; Lin et al., 2021), artificial noise (Hendrycks & Dietterich, 2019; Arjovsky et al., 2019; Larochelle et al., 2008), and natural distribution shifts (Hendrycks et al., 2021; Hendrycks & Dietterich, 2019; Barbu et al., 2019; Recht et al., 2019). Maintaining model robustness becomes significantly more difficult with severe visual shifts in the image domain. However, many studies evaluate generalization on OOD datasets with lim-

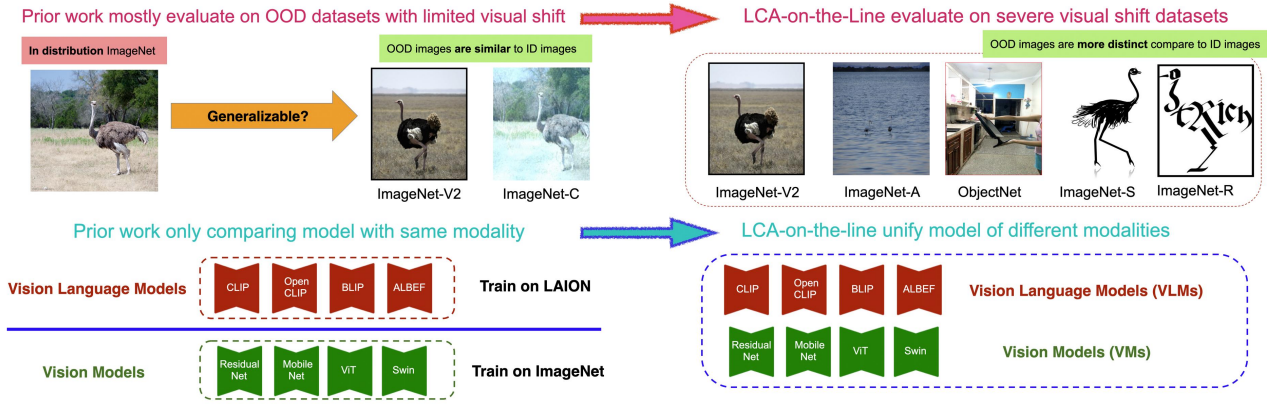

Figure 2. **Comparison of our setting with prior work. Left**: prior work settings such as Accuracy-on-the-line (Miller et al., 2021) and Agreement-on-the-line (Baek et al., 2022). **Right**: our setting. To the best of our knowledge, LCA-on-the-line is the first approach to uniformly measure model robustness across VMs and VLMs on OOD datasets with significant distribution shifts (ImageNet-S/R/A/O).

ited visual shifts or only involve artificial noise, such as ImageNet-v2 or ImageNet-C (Recht et al., 2019; Arjovsky et al., 2019). Such datasets fail to fully reflect a model's generalization capability when confronted with severe distribution shifts (Hendrycks et al., 2021; Hendrycks & Dietterich, 2019; Barbu et al., 2019), as there is often limited transfer of robustness from synthetic to natural distribution shifts (Taori et al., 2020).

In the realm of model generalization, numerous attempts have been made to predict a model's performance on OOD datasets based on in-distribution measurements, following the concept of *effective robustness* (Taori et al., 2020). These approaches, referred to as 'X-on-the-line'(Miller et al., 2021; Baek et al., 2022), suggest that a model's OOD performance is correlated to in-distribution accuracy (Miller et al., 2021; Recht et al., 2019; Miller et al., 2020; Roelofs et al., 2019) or models consensus on in-distribution accuracy (Jiang et al., 2021; Baek et al., 2022).

Moreover, several prior attempts rely on domain generalization strategies that necessitate prior knowledge of the target domain or require an estimation of OOD domain information (Chen et al., 2021; Li et al., 2022a). These can lead to computationally intensive processes, particularly when involving multiple models or inferences (Baek et al., 2022; Deng et al., 2022).

Most prior research has focused solely on estimating generalization among vision models (VMs) supervised on class labels trained on ImageNet (Taori et al., 2020; Mustafa et al., 2020). Emerging large-scale Vision-Language Models (VLMs) trained on datasets like LAION demonstrate exceptional generalization performance on out-of-distribution (OOD) data. However, as shown on the left plot of Fig. 1, existing evaluation (Miller et al., 2021) using ID accuracy fail to explain the effective robustness (Taori et al., 2020) gap between VMs and VLMs. This underscores the necessity to evaluate and compare models across different families under a unified evaluation framework. Recently, (Shi et al., 2023) observed the same problem and proposed evaluating OOD accuracy using multiple ID test sets, but their method requires multiple evaluation runs.

Unlike VMs, VLMs leverage more diverse training data, contrastive loss, and language supervision. There have been attempts to measure VLM generalization (HaoChen et al., 2021; Fang et al., 2022; Schuhmann et al., 2022; Kaur et al., 2022), specifically suggesting that diversity in training data is an indicator of model generalization. However, collecting or training on such extensive data can be non-trivial (Schuhmann et al., 2022).

Prior attempts lack a unified, simple measurement for both VMs and VLMs to explain model generalization and convert it into actionable improvements. To address the issues of (1) lack of unified metrics on VLMs and VMs; (2) need for robustness to large domain shifts; (3) desire for computationally efficient metrics, we propose adopting the Lowest Common Ancestor (LCA) distance 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 (36 VMs and 39 VLMs) (cf. Fig. 2), we show that the in-distribution LCA distance **strongly correlates** with multiple ImageNet-OOD datasets under severe visual shifts (cf. Fig. 1 right plot). *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. This spurs us to further investigate and discuss the potential of the LCA benchmark for improving model generalization.* We also discuss the suitability of LCA as a generalization indicator in section 3.

In summary, we make the following major contributions: (1) We propose the Lowest Common Ancestor (LCA) distance as a new metric for evaluating model generalization. This benchmark utilizes class hierarchies, such as WordNet, which encode relationships between classes. (2) We validate our benchmarking strategy through large-scale experiments, analyzing 75 models across five ImageNet-OOD datasets. Our findings reveal a strong linear correlation between in-distribution LCA and OOD Top-1 performance, thus establishing the 'LCA-on-the-Line' framework. (3) We offer a thorough analysis of the connection between LCA and model generalization, providing new insights to inspire further research in this area. (4) For datasets without a predefined hierarchy, we introduce a method for constructing latent hierarchies using K-means clustering. Our results demonstrate that the LCA distance is robust to variations in underlying taxonomies or hierarchies. (5) We illustrate the potential of this benchmark by demonstrating how model generalization can be enhanced by aligning model predictions with the WordNet hierarchy.

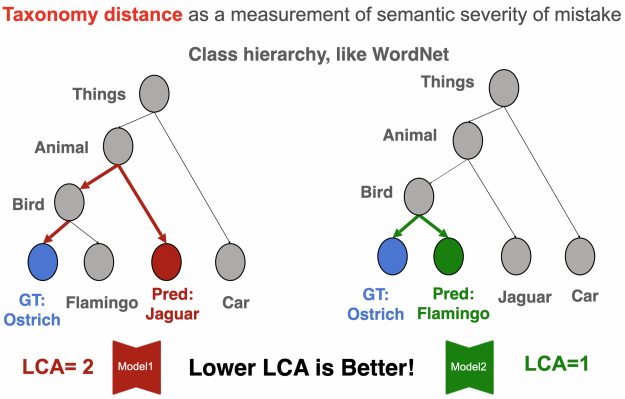

Figure 3. **LCA distance visualization**. Our method estimates a model's generalization based on its in-distribution semantic severity of mistakes. We use the 'Lowest Common Ancestor' (LCA) distance to rank the distance between the model's prediction and the ground-truth class within a predefined taxonomic hierarchy, such as WordNet. The LCA distance is proportional to the shortest path from the prediction to the ground-truth class in the hierarchy.

2. LCA Distance Measure Misprediction Severity

We propose using the in-distribution Lowest Common Ancestor (LCA) distance, also known as taxonomy loss, as a predictor for model generalization. Here, we formally define how taxonomy loss can be measured using in-distribution 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' (Bertinetto et al., 2020). For example, if a cat image is predicted as a dog by model-A and as a car by model-B, model-A would have a lower LCA distance as it makes a better mistake compared to model-B. Following previous research (Bertinetto et al., 2020; Deng et al., 2009b), we use WordNet (Miller et al., 1990), a large-scale lexical database inspired by psycholinguistic theories of human lexical memory (Miller, 1995), to encode class taxonomy. The WordNet taxonomy is well suited for the widely used ImageNet dataset on which it is based on. 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 (Bertinetto et al., 2020) 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 (Valmadre, 2022). 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.

We can also derive the generalized form of LCA distance to settings where the model outputs a distribution over all possible classes for each sample (like using softmax), please refer to appendix D.3 for details.

3. Discussion: The Suitability of LCA as a Benchmark for Model Generalization

This section explores the hypothesis linking LCA distance with a model's generalization ability and discusses how these insights can be meaningfully and actionably applied.

Our primary motivation is to use class hierarchy to capture correlation invariances across training environments, as proposed in the seminal work on 'invariant risk minimization' (Arjovsky et al., 2019). Since the class hierarchy remains consistent across both ID and OOD datasets, it can serve as a surrogate measure of the model's invariant features. Models that generalize well to OOD datasets typically learn universal or non-spurious features from the training dataset that are transferable to OOD datasets (Makar et al.,

2022). Such models are more likely to misclassify an ostrich as another bird rather than a lion. These taxonomy-based mispredictions, quantified using the LCA distance, are shown to be a better indicator of a model's OOD performance in this work.

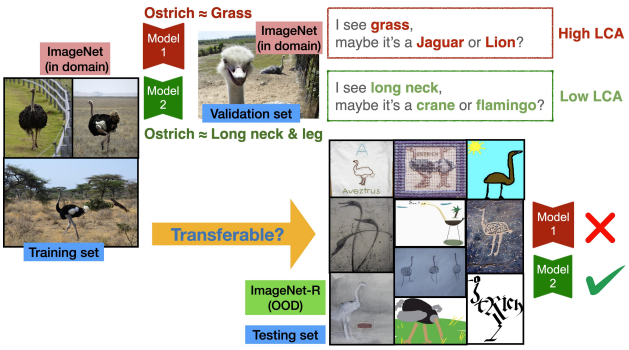

Figure 4. **Capturing transferable features for model generalization.** ImageNet-R maintains shape information (Geirhos et al., 2018) like 'long neck', 'big belly', and 'long legs'. We hypothesize that models with good generalization should capture these transferable features rather than succumbing to spurious correlations such as 'grass', thereby tending to predict classes that are semantically closer to the ground-truth. Such models are expected to have low LCA distances between their predictions and the ground-truth.

Obstacles to Model Generalization. In deep learning, models often learn predictive features from images by creating discriminative associations to class labels. This approach is susceptible to spurious correlations in the training data (Sturm, 2014; Torralba & Efros, 2011; Jabri et al., 2016). For instance, a model might erroneously associate the class 'ostriches' with the feature 'green grass' in the background, as ostriches often appear in grasslands. These correlations may fail when applied to an OOD dataset that only depicts the semantic concept of 'ostriches' (Arjovsky et al., 2019; Zhang et al., 2021).

Essentials for Model Generalization. ImageNet-R is a severely shifted OOD dataset where, despite significant distribution shifts, humans can effortlessly identify the correct classes. This is because humans can discern stable features across environments. A model's generalization capability depends on the transferability of the associations learned during training. Ideally, only features that align with human understanding of object semantics are universally transferable to any constructed OOD dataset. This underscores the importance of identifying transferable features that contribute to robust model generalization.

How can we measure what features a model has learned as predictive during training? The decision-making process of deep neural networks trained end-to-end has become less interpretable. While there have been attempts to decipher this process by forming decision-tree-like models (Wan et al., 2020; Gare et al., 2022) or through learnable activation functions (Liu et al., 2024), these efforts have not linked this understanding to measure model generalization.

Class Taxonomy Alignment as a Representation Measurement. Class taxonomy or ontology has been widely utilized in literature to indicate class formation (Deng et al., 2009b; Van Horn et al., 2018) and semantic relationships between classes (Frome et al., 2013; Barz & Denzler, 2019; Wan et al., 2020; Redmon & Farhadi, 2017; Lin et al., 2022), offering a hierarchical organization of classes or categories.

As WordNet encodes class organization semantically, we hypothesize that transferable features are more likely to be shared among neighboring classes in the hierarchy (e.g., ostrich and crane). In contrast, confounding features are less supported by the hierarchy and tend to appear in less relevant classes that are often more distant in the hierarchy (e.g., lion and ostrich). When a model makes a mistake, its secondary prediction class can reveal the features the model has learned as predictive. Specifically, it reflects that the model perceives the label class and secondary prediction classes to be more similar to each other based on the predictive features it has learned.

Consequently, a model that captures more transferable features tends to 'make better mistakes' by predicting classes that are semantically closer to the ground truth class. As illustrated in Fig 4, 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 model generalization based on the semantic severity of mistakes on in-distribution data. We use the LCA distance, the taxonomic distance between the model's prediction and the ground truth class in a predefined taxonomic hierarchy like WordNet. If a model consistently makes better mistakes on in-distribution data, we can reasonably assume that the model has captured more transferable features for class discrimination.

Class Taxonomy and Mistake Severity. The severity of a mistake in many studies is quantified as the shortest path from the prediction node to the lowest common ancestor (LCA) node in a predefined class hierarchy. This metric, known as 'LCA distance' or 'hierarchical error', was used in the early years of the ImageNet challenge (Deng et al., 2009b). However, it was largely dismissed as it was widely believed to follow the same ordering as Top 1 accuracy (Bertinetto et al., 2020). 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's 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 (Arjovsky et al., 2019; Bühlmann, 2020; Peters et al., 2016), where shifts in distribution primarily arise from interventions or confounding factors. Building upon this, methods (Schölkopf et al., 2021; Shen et al., 2022; Subramanian et al., 2022) such as CausalVAE (Yang et al., 2021) have been proposed, leveraging learned causal representations to capture the causal relationships underlying the data generation process (Kaur et al., 2022), which helps mitigate the distributional shifts caused by interventions.

While the connection between OOD generalization and causal concepts is not entirely novel, previous 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 (Kaur et al., 2022). 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 is 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 invariant correlations between class concepts across environments (Arjovsky et al., 2019; Santurkar et al., 2020), robust to spurious relations in images or shortcuts in learning (Makar et al., 2022). WordNet is a widely recognized and effective means of encoding semantic relationships between concepts, making it an appropriate proxy for aligning human semantic knowledge (Miller et al., 1990). Unlike previous work, WordNet hierarchy provides interpretability, adding a level of transparency to our understanding of model generalization.

LCA Illustration with Simulated Data. To illustrate our hypothesis that LCA can identify features supported by hierarchy, we created a controlled example using a simulated dataset, detailed in Appendix C. In this example, the data generation process is fully controlled. We designed a feature space that includes: 1) transferable causal features supported by hierarchy, 2) non-transferable confounding features not supported by hierarchy, and 3) random noise. Two logistic regression models were trained to mimic models capturing different predictive variables from the training data: one relying on the causal features and the other on the confounding features. The simulation results indicated that the model using causal features supported by hierarchy, which exhibited lower LCA distance, had better out-of-distribution (OOD) accuracy on the in-distribution (ID) test set, despite the model using confounding features achieving better ID accuracy. This example suggests that LCA can effectively identify models that capture relationships aligned with the hierarchical structure. Further details in code snippet.

4. Experiments

We present experiments benchmarking the relationship between Lowest Common Ancestor (LCA) and generalization.

Dataset 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, is provided in Appendix section A. We use *ImageNet* (Deng et al., 2009b) as the source in-distribution (ID) dataset, while *ImageNet-v2* (Recht et al., 2019), *ImageNet-Sketch* (Hendrycks & Dietterich, 2019), *ImageNet-Rendition* (Hendrycks et al., 2021), *ImageNet-Adversarial* (Hendrycks et al., 2021), and *ObjectNets* (Barbu et al., 2019) are employed as out-of-distribution datasets, exemplifying severe natural distribution shifts. The ImageNet hierarchy, as depicted in (Bertinetto et al., 2020), is utilized.

Although *ImageNet-v2* is predominantly deemed an OOD dataset in most prior literature (Shankar et al., 2020; Miller et al., 2021; Baek et al., 2022), our experiments suggest that *ImageNet-v2* aligns more closely with ImageNet than other OOD datasets; we delve into these details in Appendix B.

Note that the terms in-distribution (ID) and out-of-distribution (OOD) are not model-specific in this context. Due to the varying distribution of training data across different models, ImageNet may not necessarily represent ID data for models like CLIP, where the training data distribution is not explicitly known. Instead, ID and OOD are relative concepts. ImageNet is used as a reference anchor dataset, serving as a baseline to evaluate the generalization capabilities of models on OOD datasets. This approach aligns with prior work, allowing us to consistently measure the shift in performance from ID to OOD datasets, despite the differences in the training data distributions of the models.

Metric Setup. 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.

Model	ImgN		ImgN-v2	ImgN-S	ImgN-R	ImgN-A	ObjNet
	LCA ↓	Top1	Top1	Top1	Top1	Top1	Top1
ResNet18	6.643	0.698	0.573	0.202	0.330	0.011	0.272
ResNet50	6.539	**0.733**	**0.610**	0.235	0.361	0.018	0.316
CLIP_RN50	6.327	0.579	0.511	0.332	0.562	0.218	0.398
CLIP_RN50x4	**6.166**	0.641	0.573	**0.415**	**0.681**	**0.384**	**0.504**

Table 1. **Model performance corresponds to mistake severity.** **LCA ↓ /Top1 ↑** indicate measurements on a given dataset. We present model comparisons across VMs and VLMs families. In-distribution LCA distance indicate severely shifted OOD performance (ImageNet-S/R/A/O) better than in-distribution (ImageNet) Top1 accuracy (except for ImageNet-v2). *Full 75 models evaluation in Tab 2.*

Element		ImgN-v2		ImgN-S		ImgN-R		ImgN-A		ObjNet	
ID	OOD	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA
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	**0.816**	**0.903**	**0.779**	**0.883**	**0.704**	**0.839**	**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	**0.811**	**0.901**	**0.738**	**0.859**	**0.799**	**0.894**	**0.924**	**0.961**

Table 2. **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 5. We demonstrate that LCA has a strong correlation with OOD performance on all listed datasets (except ImageNet-v2). We take the absolute value of all correlations for simplicity. *Full table containing results of VMs-only and VLMs-only in Table 10. Measurements from the KEN and SPE show a similar trend as seen in Section F.*

Methods	ImgN-v2	ImgN-S	ImgN-R	ImgN-A	ObjNet
ID Top1 (Miller et al., 2021)	**0.040**	0.230	0.277	0.192	0.178
AC (Hendrycks & Gimpel, 2017)	0.043	0.124	**0.113**	0.324	0.127
Aline-D (Baek et al., 2022)	0.121	0.270	0.167	0.409	0.265
Aline-S (Baek et al., 2022)	0.072	0.143	0.201	0.165	0.131
(Ours) ID LCA	0.162	**0.093**	0.114	**0.103**	**0.048**

Table 3. **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-distribution accuracy remaining a significant indicator of ImageNet-v2 accuracy, the in-distribution LCA outperforms it as a robust error predictor across the four diverse OOD datasets. *Full table containing results of VMs-only and VLMs-only in Table 11.*

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.

4.1. LCA-on-the-Line: In-Distribution Taxonomy Distance (LCA) as an Out-of-Distribution (OOD) Performance Predictor

Accuracy-on-the-line (Miller et al., 2021) 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 (Miller et al., 2021), 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 and 2, the ImageNet in-distribution accuracy (Miller et al., 2021) 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 5 (indicated in red line), when adhering to 'accuracy on the line' (Miller et al., 2021), all four OOD datasets plotted showcase two separate linear trends, representing models that belong to each family. This observation aligns with (Cherti et al., 2022), 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 1, our method, adopting in-distribution LCA distance, could unify models from both families. As demonstrated in Table 2 and Figure 5 (colored in green line), the severity of in-distribution mistakes serves as a more effective indicator of model performance compared to in-distribution 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 exceed 0.7, while (Miller et al., 2021) drop to 0 in ImageNet-A). Notably, our experiments showed that (Miller et al., 2021) is a more reliable indicator solely for ImageNet-v2, given its visual similarity to ImageNet. We will further discuss this in Appendix section B.

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-distribution accuracy might generalize better to OOD datasets than vision models.

4.2. Predicting OOD Performance via ID LCA

We further highlight the effectiveness of the LCA-on-the-Line' by estimating model OOD performance using a linear function derived from in-distribution LCA distance. For comparison, we included four competitive baselines: *Average Confidence (AC)*, which leverages OOD logits after

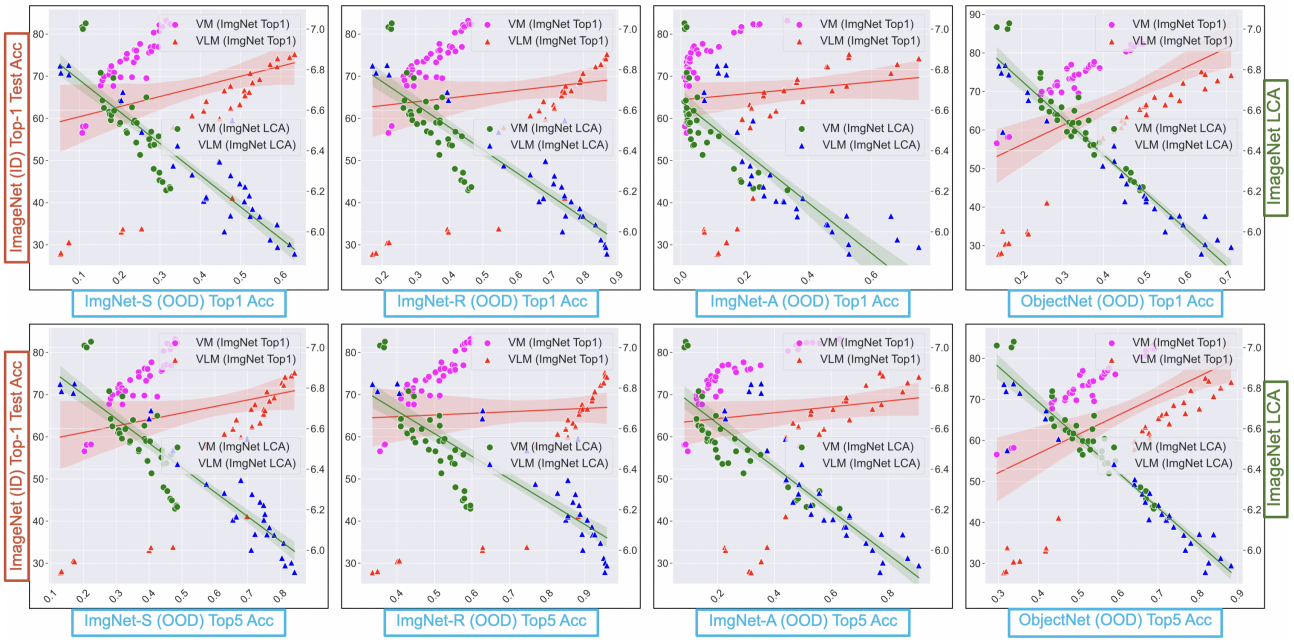

Figure 5. **Correlating OOD Top-1/Top-5 Accuracy (VM+VLM, 75 models) on 4 ImageNet-OOD Datasets.** Visualizing Table 2. The plots clearly demonstrate that the **in-distribution LCA distance** has a stronger correlation with the model's OOD performance across all OOD datasets over **accuracy-on-the-line** (Miller et al., 2021). 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) and y-axis represents ImageNet ID test Top-1 accuracy (left) and LCA (right); **Red line** (Pink dots: VMs and Red dots: VLMs) represents in-distribution classification accuracy (Top-1); **Green line** (Green dots: VMs and Blue dots: VLMs) denotes in-distribution taxonomy distance (LCA). As interpreted in Figure 1, accuracy-on-the-line only explains generalization of models with similar settings (VMs or VLMs), but does not unify both model families.

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-distribution accuracy of established measurement models to fit a linear function. Instead of performing a probit transform as done in (Baek et al., 2022) and (Miller et al., 2021), we implemented min-max scaling because LCA does not fall within the [0,1] range.

As illustrated in Table 3, in-distribution 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-distribution indicators like in-distribution accuracy (Miller et al., 2021), methods calibrated from in-distribution validation sets (Hendrycks & Gimpel, 2017), and OOD agreement of models from different families (Baek et al., 2022). In contrast, LCA, which relies 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.

4.3. Enhancing Generalization via Taxonomy Alignment

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

4.3.1. Inferring Class Taxonomy from a Pretrained Model via 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 (Deng et al., 2009b; Van Horn et al., 2018), the usefulness of our method is unquestionable. Hence, we propose a method to construct a latent class taxonomy given a well-trained model on the task, expanding the potential applications of our work. We show that such a constructed taxonomy could achieve similar correlational performance to the WordNet hierarchy.

The essence of class taxonomy lies in its representation of inter-class distance, encoding class proximity, and identify-

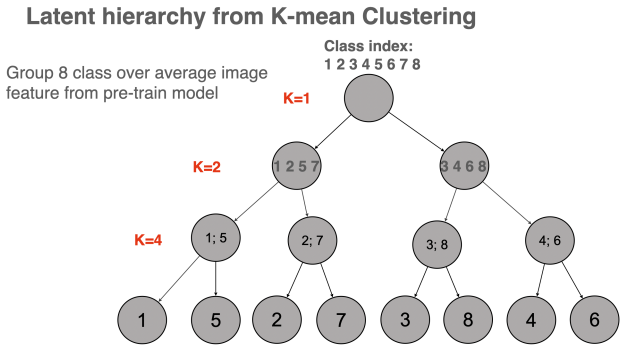

Figure 6. **Hierarchical Structure of Image Feature Clustering Using K-means.** We construct latent hierarchy through K-means clustering on image features extracted from a pre-trained model. K=1 represent the most generalized cluster, then we incrementally increase the granularity by splitting into K=2 and K=4 clusters. Each node in the hierarchy represents a cluster with the number indicating the class indexes assigned to that cluster. Tab 4 show that robust performance can be achieved among 75 latent hierarchy contructed from different pretrained models using clustering.

	Element		ImgN-v2	ImgN-S	ImgN-R	ImgN-A	ObjNet
	ID	OOD					
Top1	Top1	Top1	**0.980**	0.274	0.141	0.093	0.522
LCA (Statistical Measurements calculated from 75 different Latent Hierarchies)							
Mean	LCA	Top1	0.815	**0.773**	**0.712**	**0.662**	**0.930**
Min	LCA	Top1	0.721	0.715	0.646	0.577	0.890
Max	LCA	Top1	0.863	0.829	0.780	0.717	0.952
Std	LCA	Top1	0.028	0.022	0.027	0.025	0.010

Table 4. **Correlation Measurement (PEA) between ID LCA/Top1 and OOD Top1 across 75 Latent Hierarchies Derived from K-means**. Our latent hierarchy construction is robust across 75 different source pretrained models: For each source model, we extracted average class features and applied K-means clustering to construct a latent hierarchy. We then calculated the LCA distance based on each hierarchy, and aggregated the statistical metric of the 75 groups' Pearson correlation coefficient (PEA) to OOD performance (essentially 75 groups of data from Table 2). *We observe that LCA reliably tracks OOD performance even when using different class taxonomies.*

ing which classes cluster closely in feature space. In this spirit, we can construct a class taxonomy matrix using K-means clustering on image features. As illustrated in Fig 6, for the ImageNet dataset, we adopt a well-trained model as the source pretrained model and extract average class features to cluster data hierarchically at different levels (we use n=9 for the 1000-class ImageNet dataset, as $2^9 < 1000$), with an increasing number of clusters to indicate class adjacency. Experiments in Tab 4 show that our method is very robust regardless of which model was used as the source model to construct the class hierarchy. This result demonstrate the potential in practice to use a latent hierarchy constructed by only one well-trained model for evaluating all models on a given task. Further implementation details are provided in appendix E.1.

	ImgN	ImgN-v2	ImgN-S	ImgN-R	ImgN-A	ObjNet
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	**0.207**	**0.335**	**0.014**	**0.282**

Table 5. **Interpolating Class Taxonomy to Linear Probing on ResNet18 Feature**. Training with a WordNet hierarchy delivers the most significant improvements across OOD benchmarks despite slightly lower Top-1 accuracy, whereas models using hierarchies inferred from pretrained models yield lesser gains. The top portion of the table displays results from models trained using latent hierarchies constructed from the indicated model via K-means. The bottom portion presents the results of the aforementioned models when interpolated with layers trained from cross-entropy in the weight space (Wortsman et al., 2022).

4.3.2. Employing Class Taxonomy as Soft Labels

In this preliminary exploration, we investigate the potential for LCA distance to enhance model generalization through improved supervision. We encode the normalized pairwise LCA between each class as soft labels and apply linear probing over a pretrained model. Instead of the rigid probabilistic distribution of single-label classification, we approach the problem as multi-labeling. Additionally, we employ a sigmoid-style (Beyer et al., 2020) BCE loss instead of softmax, which relaxes the constraints on inter-class interactions. A more detailed setup is included in Appendix E.2.

Following these methods, we also constructed latent hierarchies based on AlexNet (Krizhevsky et al., 2017) and Swin Transformer (Liu et al., 2021), representing the best and worst performing models on ImageNet in our model pool. Drawing on the intuition of model distillation (Hinton et al., 2015), the hierarchy constructed from the model's pretrained features partially encapsulates the model's interpretation of interclass relationships.

As shown in Table 5, incorporating more accurate inter-class distances with WordNet enhances OOD performance across all four OOD benchmarks, albeit with slightly lower Top-1 accuracy. However, this approach results in a minor drop in in-distribution accuracy due to less intensive optimization towards the ground truth class. Inspired by the notion that models are more confident where they excel (Wortsman et al., 2022), we applied linear interpolation between linear layers trained from cross-entropy and our proposed loss function. The results indicate that this method balances competitive performance on both ID and OOD datasets.

Notably, 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

Model	ImgN		ImgN-v2		ImgN-S		ImgN-R		ImgN-A		ObjNet	
	Top1 ↑	Test CE ↓	Top1 ↑	Test CE ↓	Top1 ↑	Test CE ↓	Top1 ↑	Test CE ↓	Top1 ↑	Test CE ↓	Top1 ↑	Test CE ↓
Baseline	0.589	9.322	0.517	9.384	0.379	9.378	0.667	8.790	0.294	9.358	0.394	8.576
Stack Parent	0.381	9.389	0.347	9.395	0.219	9.561	0.438	9.258	0.223	9.364	0.148	9.076
Shuffle Parent	0.483	9.679	0.432	9.696	0.329	9.718	0.557	9.281	0.236	9.586	0.329	8.785
Taxonomy Parent	**0.626**	**9.102**	**0.553**	**9.165**	**0.419**	**9.319**	**0.685**	**8.658**	**0.319**	**9.171**	**0.431**	**8.515**

Table 6. **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*>; The Taxonomy Parent method, which includes the full hierarchical relationship, yields the best performance, highlighting the effectiveness of incorporating structured knowledge into model predictions.

suggests that enforcing arbitrary inter-class relationships, derived from in-distribution datasets, can negatively affect OOD performance. We anticipate that future work will involve larger-scale, train-from-scratch validations to fully explore the potential of LCA in improving generalization. Further exploration is detailed in Appendix Section E.3.

4.3.3. Improving Generalization by Class Taxonomy Alignment with Prompt Engineering

In this section, we discuss results on enhancing model generalization through prompt engineering in VLMs.

For vision-language models, integrating taxonomy-specific knowledge during zero-shot evaluation is straightforward. The WordNet hierarchy naturally indicates inter-class distances from class definitions. For example, 'dalmatian' and 'husky' are semantically close, both originating from the parent node 'dog'. We detail the results with CLIP-vit32 (Radford et al., 2021) in Table 6. To test our hypothesis, we explicitly integrated hierarchical taxonomy relationships into the prompt for zero-shot VLM predictions. The prompt was designed as **'A, which is a type of B, which is a type of C'**, guiding the model to make taxonomy-aligned predictions. Additionally, we conducted two ablation studies: 1) **Stack Parent:** providing the correct taxonomy path without informing the model of the class name relationships; and 2) **Shuffle Parent:** informing the model of the hierarchical 'is-a' relationship but providing an incorrect taxonomy relationship randomly sampled from the tree. Our results demonstrate that informing the model of both the correct taxonomy and their hierarchical relationships significantly improves generalization. This improvement is evidenced by enhancements in Top-1 accuracy and test-time Cross-Entropy (CE) across all datasets for all tested models.

5. Limitation

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-distribution data (like ImageNet2, more discussion in Appendix B). For these datasets, in-distribution Top1 remains a strong indicator, which potentially limits the utility of LCA. Also, it's expected that LCA will show a weaker discrimination between models on datasets with small number of classes (like Cifar (Krizhevsky et al.)).

6. Conclusions

This work revitalizes the use of LCA distance, leveraging class taxonomies such as WordNet, to indicate model OOD performance. We assess the severity of model mispredictions in a manner agnostic to model modality or architecture, establishing a comprehensive metric for evaluating model generalization. Our findings, across multiple ImageNet-OOD datasets, highlight the superiority of LCA distance in reflecting the generalization capabilities of models trained with either class labels (VMs) or captions (VLMs), surpassing the traditional reliance on in-distribution Top-1 accuracy (Miller et al., 2021). Additionally, we demonstrate that aligning model predictions with class taxonomies, through soft labels or prompt engineering, can enhance model generalization. To extend the application of LCA distance measurement to any dataset, we introduce a method for creating latent hierarchies using K-means clustering, showcasing the resilience of LCA distance regardless of the applied taxonomy or hierarchy. This work offers new insights into model generalization by leveraging existing resources and encourages further research in this direction.

Future research could focus on providing theoretical justification for the LCA-on-the-Line framework. For instance, exploring causal discovery (Brouillard et al., 2020) methods on the ImageNet dataset to construct a causal graph between classes and underlying variables may offer a more accurate reflection of causal relationships between classes. Additionally, conducting larger-scale empirical studies to further validate this benchmark would be beneficial.

Acknowledgements

Authors thank Prof. Deva Ramanan for insightful discussions, and Hualiang Wang for valuable feedback on the manuscript. The work was partially supported by the CMU Argo Research Center. Prof. Shu Kong is partially supported by the University of Macau (SRG2023-00044-FST).

Impact 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.

References

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

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

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

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

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

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

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Bühlmann, P. Invariance, causality and robustness. 2020.

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

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

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

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009b.

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

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

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

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

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. Learning generic lung ultrasound biomarkers for decoupling feature extraction from downstream tasks. *arXiv preprint arXiv:2206.08398*, 2022.

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

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

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

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

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

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

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

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

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

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

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

Jabri, A., Joulin, A., and Van Der Maaten, L. Revisiting visual question answering baselines. In *European conference on computer vision*, pp. 727–739. Springer, 2016.

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

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

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

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

Larochelle, H., Erhan, D., and Bengio, Y. Zero-data learning of new tasks. In *AAAI*, volume 1, pp. 3, 2008.

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

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

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

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

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

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

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

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

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

Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., and D'Amour, A. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 739–766. PMLR, 2022.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sturm, B. L. A simple method to determine if a music information retrieval system is a "horse". *IEEE Transactions on Multimedia*, 16(6):1636–1644, 2014.

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

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

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

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

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

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

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR*, pp. 1521–1528, 2011.

Valmadre, J. Hierarchical classification at multiple operating points. *arXiv preprint arXiv:2210.10929*, 2022.

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

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

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

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

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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

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

A. Model Architectures

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

Model Category	Architecture	Number of models	Checkpoint Link
VM (Vision-Only-Models)	AlexNet (Krizhevsky et al., 2017)	1	alexnet
	ConvNeXt (Liu et al., 2022)	1	convnext$_{tiny}$
	DenseNet (Huang et al., 2017)	4	densenet121 densenet161 densenet169 densenet201
	EfficientNet (Tan & Le, 2019)	1	efficientnet_b0
	GoogLeNett (Szegedy et al., 2015)	1	googlenet
	Inceptionv3 (Szegedy et al., 2016)	1	inceptionV3
	MnasNet (Tan et al., 2019)	4	mnasnet0.5 mnasnet0.75 mnasnet1.0 mnasnet1.3
	Mobilenet-V3 (Howard et al., 2019)	2	mobilenetv3_small mobilenetv3_large
	Regnet (Radosavovic et al., 2020)	1	regnet_y_1_6gf
	Wide ResNet (Zagoruyko & Komodakis, 2016)	1	wide_resnet101_2
	ResNet (He et al., 2016)	5	resnet18 resnet34 resnet50 resnet101 resnet152
	ShuffleNet (Zhang et al., 2018)	1	shufflenet_v2_x2_0
	SqueezeNet (Iandola et al., 2016)	2	squeezenet1_0 squeezenet1_1
	Swin Transformer (Liu et al., 2021)	1	swin_b
	VGG (Simonyan & Zisserman, 2015)	8	vgg11 vgg13 vgg16 vgg19 vgg11_bn vgg13_bn vgg16_bn vgg19_bn
	ViT (Dosovitskiy et al., 2020)	2	vit_b_32 vit_l_32
VLM (Vision-Language-Models)	ALBEF (Li et al., 2021)	1	albef_feature_extractor
	BLIP (Li et al., 2022b)	1	blip_feature_extractor_base
	CLIP (Radford et al., 2021)	7	RN50 RN101 RN50x4 ViT-B-32.pt ViT-B-16.pt ViT-L-14.pt ViT-L-14-336px
	OpenCLIP (Cherti et al., 2023)	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')

B. Discussion

Reestablishing LCA as a Comprehensive Measure of Model Generalization. While Top 1 ID accuracy (Miller et al., 2021) demonstrates a clear linear trend with OOD datasets in models with similar training mechanisms, this relationship becomes less distinct across VMs and VLMs. This finding, echoed in earlier studies (Fang et al., 2022; Wortsman et al., 2022; Cherti et al., 2022), 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 (Fang et al., 2022; Schuhmann et al., 2022; Kaur et al., 2022), 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 (Taori et al., 2020)". 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 (Shankar et al., 2020; Miller et al., 2021; Baek et al., 2022). 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 collection process results in visual similarities to ImageNet (as ImageNet-v2 is a recollection of ImageNet), allowing even spurious relationships encoded on 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* (Deng et al., 2009b; Bertinetto et al., 2020). 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 separately). 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 (validate in Section F.2). This finding challenges the prevailing belief that in-distribution Top-1 accuracy and LCA maintain the same ranking (Deng et al., 2009a; Bertinetto et al., 2020).

C. LCA illustration with simulated data

To illustrate the hypotheses in Section 3: 1) Transferable features are more likely to be supported by the hierarchy and shared among neighboring classes; 2) Confounding features are less supported by the hierarchy and tend to appear in less relevant classes that are often more distant in the hierarchy; 3) LCA is useful in identifying features supported by the hierarchy, we created a simple example using a simulated dataset.

Consider a feature space $\mathbf{x} := (x_1, x_2, x_3) \in \mathbb{R}^3$ and a latent class $z \in 1, 2, 3, 4$, where class 1 and 2 are similar, and class 3 and 4 are similar. By design, we set the joint distribution of $\mathbf{x}$ and z to follow a mixture of Gaussians, where x_1 = (1,3,15,17), x_2 = (1,17,7,21), x_3 = (0,0,0,0) for each class respectively.

$$
\begin{aligned}
\mathbf{x}|z = 1 &\sim N(\mu_1, \mathbf{I}), \quad \mu_1 = (1, 1, 0) \\
\mathbf{x}|z = 2 &\sim N(\mu_2, \mathbf{I}), \quad \mu_2 = (3, 17, 0) \\
\mathbf{x}|z = 3 &\sim N(\mu_3, \mathbf{I}), \quad \mu_3 = (15, 7, 0) \\
\mathbf{x}|z = 4 &\sim N(\mu_4, \mathbf{I}), \quad \mu_4 = (17, 21, 0)
\end{aligned} \tag{1}
$$

	ID Top1 Error ↓	ID LCA Distance ↓	OOD Top1 Error ↓
g(w. confounding feature)	**0.1423**	2.000	0.7503
f(w. transferable feature)	0.3287	**1.005**	**0.3197**
Diff	+ 0.1864	-0.995	-0.4306

Table 7. **Observation from simulation data with 100 trials.** The average ID test accuracy error (i.e. top 1 error) `ID_Top1_Error` ↓, ID test LCA distance `ID_LCA_Distance` ↓, and OOD test accuracy error `OOD_Top1_Error` ↓ for generalizable "good" prediction model f and non-generalizable "bad" prediction model g over 100 independent trials. Specifically, we design the data generation process as described in (1), and f is "good" as it learns to rely on the transferable causal features supported by hiearachy; while g is "bad" as it instead relies on the non-transferable confounding features not supported by hiearachy. In this example, ID LCA distance is a better indicator of OOD performance than ID Top1 accuracy, and model f display better generalization to OOD dataset despite lower ID Top 1 accuracy.

Given a hierarchy preserving class proximity: `root:(class 1, class 2),(class 3, class 4)`, by design, only feature x_1 supports the class hierarchy, as the distance between classes 1 & 2 and classes 3 & 4 is smaller than those for other pairs. Feature x_2 can distinguish all four classes but is not supported by the class hierarchy. Feature x_3 is random noise with no predictive power for the latent class.

For the in-distribution (ID) data, all three features are observed, while for the out-of-distribution (OOD) data, only x_1 and x_3 are observed. From hypothesis in section 3, x_1 can be considered a transferable causal feature because it is supported by the true class hierarchy and is observable in all datasets. In contrast, x_2 is a non-transferable confounding feature that does not preserve the class hierarchy and is only observable in the ID data. By design (larger μ gap between classes), confounder x_2 display stronger discrimination among four classes than x_1 on ID data.

We trained two logistic regression models on the in-distribution (ID) dataset, mimicking models that captured different features as predictive variables learned from the training data.

- Model f, which trains on the transferable causal feature x_1, and noise feature x_3.
- Model g, which trains on the non-transferable confounding feature x_2, and noise feature x_3.

From simulations (10,000 samples across 100 independent trials), we observed the following results listed in Table 7:

- **Model** g achieved better ID accuracy because it can leverage x_2, which distinguishes all four classes effectively in the ID data.
- **Model** f had better OOD accuracy because x_1 is a transferable feature that is also present in the OOD data, supported by the true class hierarchy that's invariant across ID and OOD data.
- **Model** f showed better (lower) LCA distance on the ID test set, indicating that it captures the class hierarchy better by relying on the transferable causal feature x_1.

This example illustrates the hypothesis presented in Section 3 and provides the expected output in Table 7. The results suggest that LCA can effectively identify models that capture relationships aligned with the hierarchical structure. For further details, please refer to code snippet.

D. Metric

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

D.1. Correlation Measurement

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

D.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:

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

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:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

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.

D.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:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\frac{1}{2}n(n-1)} \tag{4}$$

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:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \tag{5}$$

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.

D.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'.

D.2.1. LCA DISTANCE

Following (Bertinetto et al., 2020; Valmadre, 2022), we define LCA distance using a predefined hierarchy tree, as indicated in Figure 3. We adopt class distance in a hierarchical tree format to denote inter-class relationships, which is necessary to calculate LCA and ELCA (cf. section D.3). Given a ground truth node y (node 1 in the plot), a model prediction node y', and their lowest common ancestor class node $N_{LCA}(y, y')$. We define it as:

$$D_{LCA}(y', y) := f(y) - f(N_{LCA}(y, y')) \tag{6}$$

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:

$$D^P_{LCA}(y', y) := (P(y) - P(N_{LCA}(y', y))) + (P(y') - P(N_{LCA}(y', y))), \tag{7}$$

where we also append $(P(y') - P(N_{LCA}(y', y)))$ 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 (Valmadre, 2022), 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:

$$D^I_{LCA}(y', y) := I(y) - I(N_{LCA}(y', y)), \tag{8}$$

In this work, we adopt $D^I_{LCA}(y', y)$ for experiments.

D.3. ELCA distance

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.

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 (Bertinetto et al., 2020); 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.

For implementation, ELCA is a weighted combination of the LCA distance for each leaf node [1,2,3,4] as in Fig 3, weighted by class probability. Formally, for each prediction node X_i, the probabilistic distribution over all candidate classes can be obtained by applying a softmax function $softmax(x) : \mathbb{R} \to [0, 1]$ to get model outputs probability $(\widehat{p}_{1,i}, \ldots, \widehat{p}_{K,i})$ over the K classes (e.g., 1000 in ImageNet).

In Table 8, we also demonstrate that models with better OOD generalization (OOD Top 1 accuracy) usually also have lower LCA/ELCA distances.

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 (He et al., 2016)	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 (He et al., 2016)	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 (Radford et al., 2021)	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 (Radford et al., 2021)	**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 8. **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.*

E. Experiment Setup

E.1. K-mean Clustering for Latent Class Hierarchy Construction

As depicted in Fig 6, we begin with a pretrained model M, in-distribution image data X, and labels Y for k classes. Initially, we extract the in-distribution 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 average features, we perform a 9-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 $2^9 < 1000$ (ImageNet have 1000 classes). This procedure results in 9 cluster outcomes. Subsequently, we find the LCA node between each pair of the k classes, to determine the cluster level at which both classes exists in the same cluster. We use the height of the common cluster as their pairwise LCA height to be retrieved at training/evaluation. By definition, all classes share a base cluster level of 10.

E.2. Loss for Linear Probing Experiment

This section illustrate loss function used in our linear probing experiment: For a dataset with n classes, we first establish an n × 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.

E.3. Does the Generalization Quality of the Pretrained Source Model Affect the Quality of Soft Labels?

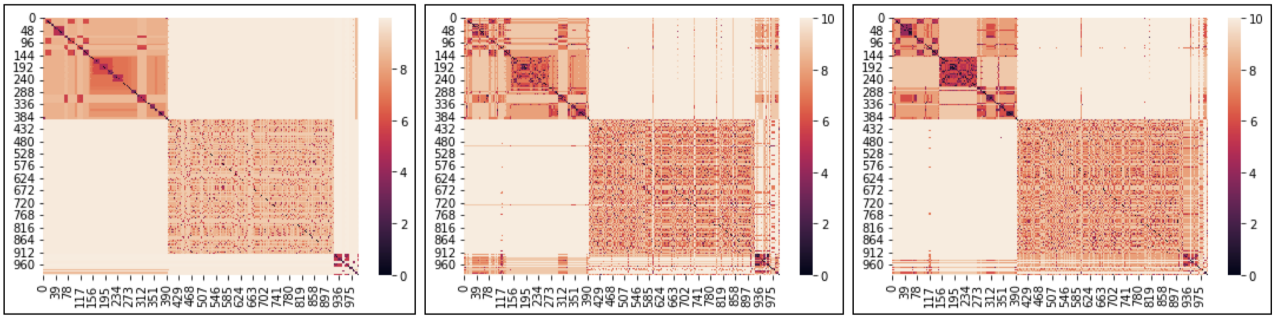

Figure 7. **Visualization of pair-wise LCA distance for ImageNet classes**. From left to right: WordNet hierarchy; matrix constructed from AlexNet (Krizhevsky et al., 2017); and matrix constructed from CLIP ResNet50 (Radford et al., 2021). We observe a higher alignment between the CLIP RN50 LCA distance matrix and the WordNet hierarchy as compared to the one from AlexNet.

This section explores the quality of soft labels as discussed in Section 4.3.2.

In Figure 7, we present a comparison of pair-wise LCA distance matrices visualizations for ImageNet data using three different hierarchies. Each row signifies the LCA distance between a specific class and the reference class, arranged in ascending order, with the diagonal index indicating the shortest distance. The visualizations show that hierarchies derived from different sources will produce different soft labels. Is there a relationship between the pretrained source model we used and the derived soft labels quality?

To test this, we generated 36 LCA distance matrices using 36 latent hierarchies with the proposed K-means clustering. Then, we performed 36 groups of linear probe training using soft labels over ResNet18 as described in Section 4.3.2. We present our findings in Figure 8 and Table 9. The results reveal a moderate-strong correlation between the ID LCA of the pretrained source model on WordNet, and the generalization capabilities of the linear probe model trained from the source-model-derived latent hierarchy. Note that ranking measurement might be noisy due to the small difference in result accuracy. The findings verify that a latent hierarchy derived from a more generalizable model provides higher quality in guiding the linear probe model training to be more generalizable. Our future research will further explore the relationship between inter-class LCA distances in pretrained source models and their generalization capabilities.

	ImageNet	**ImageNetv2**	**ImageNet-S**	**ImageNet-R**	**ImageNet-A**	**ObjectNet**
LCA ->Hierarchy Linear Prob	PEA 0.638	PEA 0.684	PEA 0.715	PEA 0.800	PEA 0.639	PEA 0.598

Table 9. **Correlation Measurement between Source Model Generalization Ability and Soft Labels Quality.** Following the K-Means clustering algorithm, we constructed 36 LCA distance matrices (class hierarchies) from 36 pretrained VMs source models on ImageNet. We then used these LCA distance matrices as soft labels to guide linear probing over ResNet18 features (as described in Section 4.3.2). The table indicates a moderate-strong correlation between the in-distribution LCA of the pretrained source model using WordNet and the out-of-distribution (OOD) accuracy on the linear probe model using the corresponding derived LCA distance matrix. The result is calculated from the average of three random seeds. Visualization is shown in Figure 8.

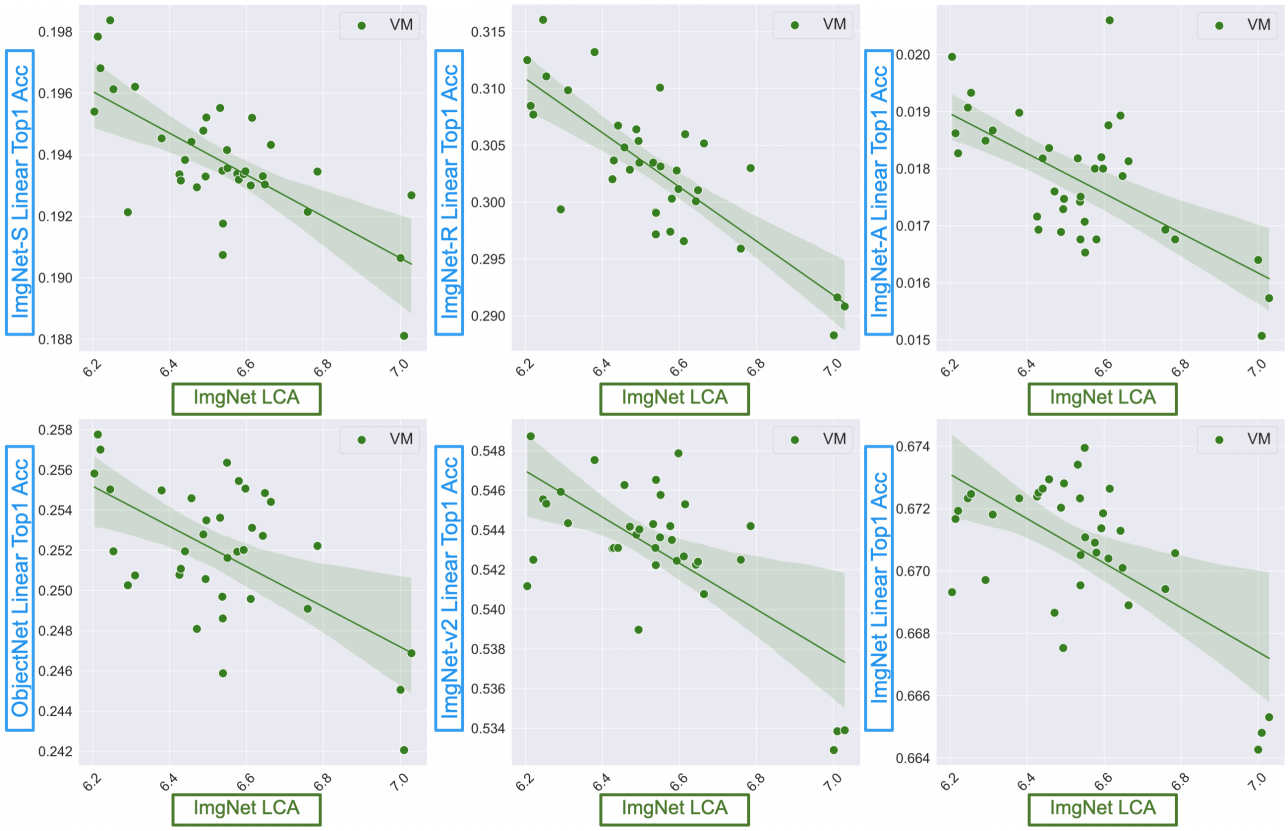

Figure 8. **Correlation Measurement between Source Model Generalization Ability and Soft Labels Quality.** Visualization on result in Tab 9. Plot shows an moderate-strong correlation between the two variable.

E.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.

F. Supplementary Result

F.1. Comprehensive results from main paper

Extended from Tab 2 and Tab 3 in main paper, we present measurement on only-VMs and only-VLMs in Tab 10 and Tab 11. Equivalently, LCA is also a very good OOD indicator when involved only VMs or VLMs.

F.2. Does ImageNet LCA (Taxonomy Distance) Reflect ImageNet Top 1 Accuracy?

Here, we present numerical results supporting the discussion section B. We challenge the common belief that LCA and Top-1 accuracy follow parallel trends within the same dataset. As illustrated in Figures 9 and 12, when including both VM and VLM zero-shot models, ImageNet and ImageNet-v2 show a weak correlation between LCA and Top-1 accuracy on the same dataset, while other semantically distinct OOD datasets exhibit a stronger relationship.

F.3. Ranking Measurement of LCA-on-the-Line

Here we present the numeric result for ranking measures in comparison to common use Top1 in-distribution accuracy in Table 13. Equivalently, in-distribution LCA measure present strong result in both preserving linearity and ranking.

	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	**0.816**	**0.903**	**0.779**	**0.883**	**0.704**	**0.839**	**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	**0.811**	**0.901**	**0.738**	**0.859**	**0.799**	**0.894**	**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	0.850	0.921	**0.867**	**0.931**	**0.691**	**0.832**	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	0.852	0.923	**0.826**	**0.909**	**0.822**	**0.906**	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	0.798	0.893	0.768	0.877	**0.549**	**0.741**	**0.908**	**0.953**
	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	0.797	0.893	0.777	0.882	**0.708**	**0.841**	**0.920**	**0.960**

Table 10. **Correlation measurement of ID LCA/Top1 with OOD Top1/Top5** on 75 models across modality following Fig 5. 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 (Miller et al., 2021)	**0.040**	0.230	0.277	0.192	0.178
	AC (Hendrycks & Gimpel, 2017)	0.043	0.124	**0.113**	0.324	0.127
	Aline-D (Baek et al., 2022)	0.121	0.270	0.167	0.409	0.265
	Aline-S (Baek et al., 2022)	0.072	0.143	0.201	0.165	0.131
	(Ours) ID LCA	0.162	**0.093**	0.114	**0.103**	**0.048**
VLM	ID (Miller et al., 2021)	**0.014**	0.077	0.064	0.127	0.052
	AC (Hendrycks & Gimpel, 2017)	0.029	**0.050**	**0.044**	0.217	0.088
	Aline-D (Baek et al., 2022)	0.151	0.250	0.081	0.296	0.260
	Aline-S (Baek et al., 2022)	0.070	0.069	0.068	**0.080**	0.153
	(Ours) ID LCA	0.047	0.083	0.070	0.105	**0.043**
VM	ID (Miller et al., 2021)	**0.013**	**0.099**	0.108	**0.143**	0.068
	AC (Hendrycks & Gimpel, 2017)	0.059	0.204	0.188	0.441	0.168
	Aline-D (Baek et al., 2022)	0.083	0.427	0.313	0.665	0.364
	Aline-S (Baek et al., 2022)	0.105	0.182	**0.092**	0.574	0.216
	(Ours) ID LCA	0.029	0.102	0.113	0.145	**0.065**

Table 11. **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-distribution accuracy maintain as a significant indicator of ImageNet-v2 accuracy, the in-distribution LCA outperforms it as a robust error predictor across four naturally distributed OOD datasets.

Model	Group	ImageNet		ImageNetv2		ImageNet-S		ImageNet-R		ImageNet-A		ObjectNet	
Top1->LCA	ALL	R^2	*PEA*	R^2	*PEA*	R^2	*PEA*	R^2	*PEA*	R^2	*PEA*	R^2	*PEA*
		0.174	0.417	0.114	0.337	**0.835**	**0.914**	**0.770**	**0.878**	**0.851**	**0.923**	**0.657**	**0.810**
		KEN	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*	*KEN*	*SPE*
		0.280	0.266	0.237	0.294	**0.818**	**0.926**	**0.621**	**0.803**	**0.825**	**0.951**	**0.673**	**0.823**
	VLM	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA
		0.938	**0.969**	**0.891**	**0.944**	**0.945**	**0.972**	**0.878**	**0.937**	**0.725**	**0.851**	0.510	**0.714**
		KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE
		0.880	**0.969**	**0.799**	**0.881**	**0.864**	**0.963**	**0.753**	**0.902**	**0.689**	**0.869**	0.529	**0.720**
	VM	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA	R^2	PEA
		0.973	**0.986**	**0.890**	**0.943**	**0.934**	**0.966**	0.095	0.310	**0.840**	**0.916**	**0.948**	**0.974**
		KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE	KEN	SPE
		0.911	**0.980**	**0.758**	**0.910**	**0.854**	**0.963**	0.149	0.222	**0.839**	**0.952**	**0.854**	**0.960**

Table 12. **Correlation Measurement between Top-1 Accuracy and LCA on the Same Dataset**. This analysis uses 75 models across different modalities (36 VMs and 39 VLMs) on all six ImageNet datasets. While the main paper employs ID LCA to predict OOD performance (*e.g., Corr(ImageNet LCA, ImageNet-A Top-1 Accuracy)*), this setting differs by using LCA to predict Top-1 accuracy on the same dataset (*e.g., Corr(ImageNet-A LCA, ImageNet-A Top-1 Accuracy)*). Following Figure 9, we highlight strong correlation indications. For simplicity, we take the absolute value of all correlations.

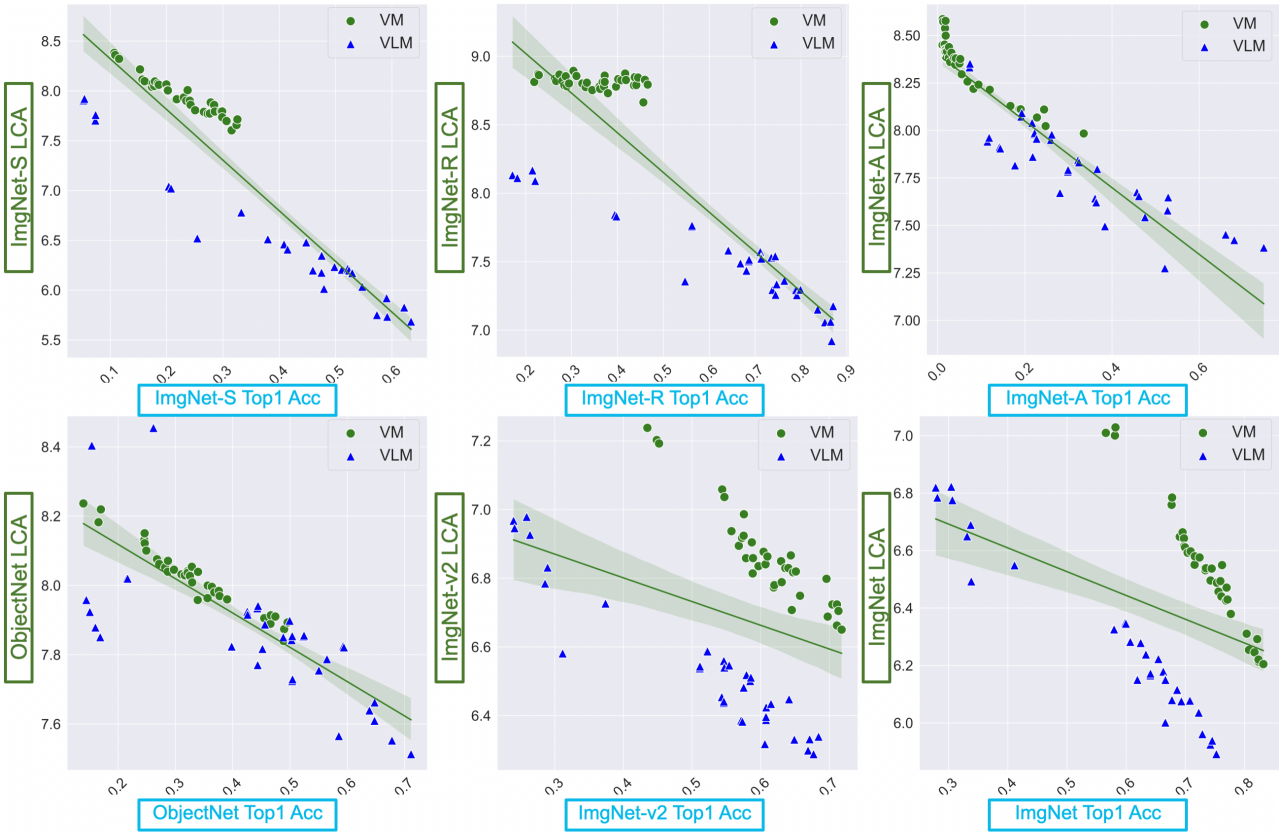

Figure 9. **Predicting LCA (VM+VLM, 75 models) on the same dataset** As per Table 12. Each plot's x-axis represents dataset Top-1 accuracy, while the y-axis shows LCA distance measured on the same datasets. The plots reveal that ImageNet and ImageNet-v2 do not exhibit a strong correlation between LCA and Top-1 accuracy, in contrast to other semantically distinct OOD datasets. This observation challenges the common belief that in-distribution Top-1 accuracy and LCA distance maintain the same order (Deng et al., 2009a; Bertinetto et al., 2020). Details in discussion section.

	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	**0.779**	**0.923**	**0.761**	**0.911**	**0.730**	**0.888**	**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	**0.768**	**0.919**	**0.752**	**0.897**	**0.755**	**0.908**	**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	0.729	0.861	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**	0.746	0.879	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	0.740	0.882	0.705	0.862	0.741	0.851	0.790	0.918
	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	0.733	0.879	0.707	0.871	0.780	0.916	0.783	0.911

Table 13. **Ranking measurement of ID LCA/Top1 with OOD Top1/Top5** on 75 models across modality(36 VMs and 39 VLMs); As shown in the 'ALL grouping', LCA shows a much better result in preserving the model relative ranking to model OOD performance on all OOD datasets (with the exception of ImageNet-v2), which indicates its superiority for model selection.