Aggregation Hides Out-of-Distribution Generalization Failures from Spurious Correlations

Olawale Salaudeen **Haoran Zhang** **Kumail Alhamoud**
Sara Beery **Marzyeh Ghassemi**
Massachusetts Institute of Technology
Correspondence to olawale@mit.edu.

Abstract

Benchmarks for out-of-distribution (OOD) generalization frequently show a strong positive correlation between in-distribution (ID) and OOD accuracy across models, termed "accuracy-on-the-line." This pattern is often taken to imply that spurious correlations—correlations that improve ID but reduce OOD performance—are rare in practice. We find that this positive correlation is often an artifact of aggregating heterogeneous OOD examples. Using a simple gradient-based method, `OODSelect`, we identify semantically coherent OOD subsets where accuracy on the line does not hold. **Across widely used distribution shift benchmarks, the `OODSelect` uncovers subsets, sometimes up to over half of the standard OOD set, where higher ID accuracy predicts lower OOD accuracy.** Our findings indicate that aggregate metrics can obscure important failure modes of OOD robustness. We release code and the identified subsets to facilitate further research.

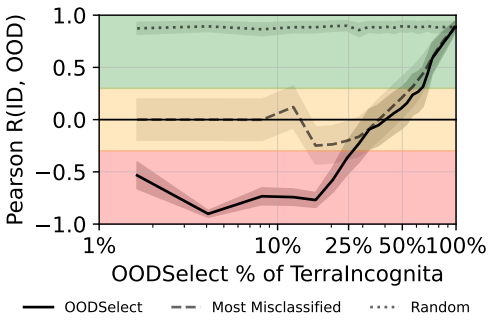
1 Introduction

Benchmarks for out-of-distribution (OOD) generalization have shown a consistent pattern that models performing well on the training distribution also perform well out-of-distribution, a trend known as *accuracy-on-the-line* (AoTL) (Miller et al., 2021; Taori et al., 2020). This pattern has often been interpreted as evidence that spurious correlations—features that improve in-distribution (ID) accuracy but harm OOD performance—are uncommon in practice. We show that this apparent robustness is misleading. When OOD data are disaggregated, large and semantically coherent subsets emerge where higher ID accuracy predicts lower OOD accuracy, a phenomenon we term *accuracy-on-the-inverse-line* (AoTIL). These hidden subsets reveal that aggregation can mask major failures of OOD robustness, suggesting that existing benchmarks may underestimate the prevalence and impact of spurious correlations.

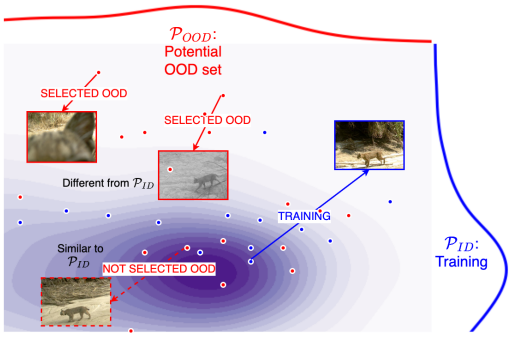
The promise of machine learning lies in generalization, the ability to perform a task on new data with similar effectiveness as on the training data (Blumer et al., 1989; Vapnik, 1999; Shalev-Shwartz and Ben-David, 2014; Zhang et al., 2016; Belkin et al., 2019). Yet models deployed in a dynamic world often encounter data from different distributions (Shimodaira, 2000; Moreno-Torres et al., 2012) and fail. For instance, a medical diagnosis model trained on data from one hospital may perform poorly in another with distinct demographics or equipment (Zech et al., 2018; Yang et al., 2024a), and an animal classifier may misclassify images captured under new conditions (Beery et al., 2018; Xiao et al., 2020). Generalization under such shifts, from in-distribution (ID) training to out-of-distribution (OOD) testing or deployment, defines domain generalization (Zhou et al., 2022; Wang et al., 2022).

These observations motivate a closer examination of what benchmark correlations actually reveal about robustness and when they conceal spurious mechanisms that undermine OOD generalization.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) **AoTIL:** With original OOD data, ID and OOD accuracy across a set of models has a Pearson correlation of 0.89. `OODSelect` finds up to 1000 ($\sim 16\%$) examples from `L46` on the same models with an ID-OOD correlation of -0.77.



(b) `OODSelect` **Strategy:** Excluded examples resemble the training distribution (e.g., centered bobcats in daylight), while included `OODSelect` examples differ (e.g., occluded bobcats, infrared camera capture).

Figure 1: **Aggregation Masking AoTIL.** Consider Terra Incognita, where ID data are drawn from camera traps at locations `L100`, `L38`, `L43`, and OOD data from `L46` (Beery et al., 2018). Aggregation masks the effect of spurious correlations on generalization, such as daylight, even though a substantial number of OOD samples are still systematically misclassified. Note that `OODSelect` examples differ from the most misclassified examples, which always have an ID-OOD accuracy correlation of near zero. Confidence intervals correspond to 95% Fisher z-intervals.

In this work, we establish the existence of large and semantically coherent OOD subsets in state-of-the-art datasets with accuracy on the inverse line. Specifically, **our contributions** are:

- We show that in state-of-the-art domain generalization benchmarks, there exist large, semantically meaningful OOD subsets—sometimes up to over half of the data—with correlations low as -0.9 Pearson R (Figure 2). The familiar *accuracy-on-the-line* trend only emerges once such subsets are aggregated with the rest of the data.
- We show that these subsets are not arbitrary: for example, in Chest X-ray diagnosis tasks, models that improved overall performance performed *worse* on patients with pleural conditions and enlarged cardiomediastinum.
- We propose `OODSelect`, a simple yet effective selection procedure to identify such subsets across datasets, when they exist.
- We provide the identified subsets for state-of-the-art datasets, including those from DomainBed (Gulrajani and Lopez-Paz, 2020) and WILDS (Koh et al., 2021), to facilitate future research (included in the supplementary material).

We provide the code and selected subsets[1] for our proposed OOD selection method and analysis.

2 Background and Related Work

The field of OOD generalization aims to develop models that are robust to spurious correlations (Zhou et al., 2022; Wang et al., 2022). Many of the state-of-the-art methods in domain generalization rely on notions of distributional invariance (Arjovsky et al., 2019; Krueger et al., 2021); often using causal motivations (Peters et al., 2016; Heinze-Deml et al., 2018; Salaudeen and Koyejo, 2024; Salaudeen et al., 2024). Progress in the field of domain generalization has primarily been evaluated by two benchmark suites: DomainBed (Gulrajani and Lopez-Paz, 2020) and WILDS (Koh et al., 2021). However, various studies have suggested that none of the proposed domain generalization methods consistently outperform naive empirical risk minimization on these benchmarks (Gulrajani and Lopez-Paz, 2020; Koh et al., 2021; Yang et al., 2023). Moreover, previous work has suggested that improving ID accuracy tends to improve OOD accuracy, i.e., a strong correlation between ID and OOD accuracy holds, termed *accuracy on the line* (Miller et al., 2021; Taori et al., 2020; Saxena

[1] https://github.com/olawalesalaudeen/OODSELECT

et al., 2024; Sanyal et al., 2024). However, Teney et al. (2023) demonstrate that with a more diverse selection of models, a fraction of real-world datasets do indeed exhibit other correlations besides strong and positive correlations between ID and OOD accuracy. Furthermore, Salaudeen et al. (2025a) provides a theoretical analysis that suggests prioritizing datasets without accuracy on the line; our proposed method provides OOD sets that satisfy such conditions by selecting subsets of existing benchmarks with accuracy on the line.

Existing subset discovery methods—such as `Slice Finder`, `SSD++`, and `DivExplorer`(Polyzotis et al., 2019; Proença et al., 2022; Pastor et al., 2021; Subbaswamy and Saria, 2020)—rely on explicit grouping cues, categorical features, or annotated attributes to define candidate subsets. In contrast, our setting assumes no access to such metadata and requires model-agnosticism, motivating a simple yet effective selection approach. `Influence functions` (Koh and Liang, 2017) may appear suitable at first glance, but they rank *training* points by leave-one-out influence, rather than partitioning the *test/OOD* set. Thus, applying influence functions in this context would still require an additional heuristic to define coherent subsets, while also inheriting known fragilities in modern deep networks (Basu et al., 2020; Epifano et al., 2023; Bae et al., 2022; Grosse et al., 2023; Koh et al., 2019; Hu et al., 2024).

Our proposed method in the next section provides a simple, efficient, yet effective approach.

3 Methodology

First, we define the correlation property that is used to determine AoTL or AoTIL.

Definition 1 (Correlation Property; Miller et al. (2021))**.** *Define $a \in \mathbb{R}$, $\epsilon \geq 0$, and Φ^{-1} as the inverse Gaussian cumulative density function. The* correlation property *is defined as*

$$\left|\Phi^{-1}\left(acc_{P_{ID}}(f)\right) - a \cdot \Phi^{-1}\left(acc_{P_{OOD}}(f)\right)\right| \leq \epsilon \quad \forall f. \tag{1}$$

Definition 1 implies:

$$|\text{Pearson } R(X,Y)| \gtrsim 1 - \frac{\epsilon}{|a| \cdot \sigma_Y}, \tag{2}$$

where σ_Y is the standard deviation of Y. Thus, the correlation property implies that the transformed ID and OOD accuracies lie approximately on a line and are strongly linearly correlated. Moreover, the sign of the Pearson correlation is determined by the sign of a: if $a > 0$, the correlation is positive, and if $a < 0$, the correlation is negative.

Problem Setup. Suppose we have N models f_i and d potential OOD examples. Let $\mathbf{Z} \in \mathbb{R}^{N \times d}$ where $\mathbf{Z}_{ij}$ is 1 if model f_i correctly classifies example j and 0 otherwise. Define $\mathbf{acc_{ID}} \in \mathbb{R}^N$ where $(\mathbf{acc_{ID}})_i$ is the held-out in-distribution accuracy of model f_i. In this work, we are always operating on $0-1$-clipped probit transform of accuracy. Define a sample selection vector $\mathbf{s} \in \{0,1\}^d$ that indicates which examples to select from the candidate OOD set, and denote the selected OOD accuracy for model f_i

$$(\mathbf{acc^s_{OOD}})_i = \frac{\mathbf{Z}_{[i,:]}\mathbf{s}}{\|\mathbf{s}\|_1} \tag{3}$$

and

$$\text{corr}(\mathbf{acc_{ID}}, \mathbf{acc^s_{OOD}}) = \frac{\mathbf{acc_{ID}}^\top \mathbf{acc^s_{OOD}}}{\sqrt{\|\mathbf{acc_{ID}}\|^2 \|\mathbf{acc^s_{OOD}}\|^2}}, \tag{4}$$

where corr is the Pearson correlation between ID and OOD accuracies. Note that the probit-transformed accuracies are mean-centered before computing the correlations.

Objective. We aim to learn a selection vector $\mathbf{s} \in \{0,1\}^d$ (with $S = \|\mathbf{s}\|_1$ fixed) to minimize the correlation between $\mathbf{acc_{ID}}$ and $\mathbf{acc^s_{OOD}}$—ideally with large S.

Importantly, a subset achieving weak or negative correlation may not exist, particularly if there are no spurious correlations with respect to the ID and OOD distributions. Additionally, the change in Pearson R from adding a model or OOD example is bounded by $\mathcal{O}(C/\sqrt{m})$, where m is the number of models or OOD examples and C depends on the accuracy range and the Lipschitzness of Φ^{-1} (theoretical analysis provided in Appendix B Lemma 1-2).

Algorithm 1: `OODSelect`: Selecting OOD subsets without accuracy-on-the-line

Input: $\mathcal{D}_{\text{ID}}^{\text{train}}, \mathcal{D}_{\text{ID}}^{\text{test}}$: in-distribution train/test splits;
$\mathcal{D}_{\text{OOD}}$: out-of-distribution dataset;
$S \in \mathbb{N}_{\leq |\mathcal{D}_{\text{OOD}}|}$: number of OOD samples to select
Output: Subset $\mathcal{D}_{\text{OOD}}^{\mathbf{s}} \subset \mathcal{D}_{\text{OOD}}$ of size S

1: Train N_{models} diverse models on $\mathcal{D}_{\text{ID}}^{\text{train}}$.
2: Let $\mathbf{acc_{ID}} \in \mathbb{R}^{N_{\text{models}}}$ be the vector of probit-transformed accuracies, where $\mathbf{acc_{ID}}_i$ denotes the accuracy of model i on $\mathcal{D}_{\text{ID}}^{\text{test}}$.
3: Construct binary matrix $\mathbf{Z} \in \{0,1\}^{N_{\text{models}} \times |\mathcal{D}_{\text{OOD}}|}$ where:
4:

$$\mathbf{Z}_{ij} = \begin{cases} 1 & \text{if model } i \text{ correctly classifies OOD sample } j \\ 0 & \text{otherwise} \end{cases}.$$

5: Let $\mathbf{acc}_{\mathbf{OOD}}^{\mathbf{s}} \in \mathbb{R}^{N_{\text{models}}}$ denote the per-model average accuracy vector across the OOD examples selected by $\mathbf{s}$; that is,

$$\mathbf{acc}_{\mathbf{OOD}}^{\mathbf{s}} = \frac{1}{\|\mathbf{s}\|_1} \mathbf{Z}_{\mathbf{s}},$$

where $\mathbf{Z}_{\mathbf{s}}$ denotes the columns of $\mathbf{Z}$ indexed by $\mathbf{s}$.
6: Solve the optimization in Equation 5 to find $\mathcal{D}_{\text{OOD}}^{\mathbf{s}}$ from $\mathcal{D}_{\text{OOD}}$.

We consider a constrained objective for selecting S OOD examples:

$$\min_{\mathbf{s} \in \{0,1\}^d} \quad \text{corr}(\mathbf{acc_{ID}}, \mathbf{acc}_{\mathbf{OOD}}^{\mathbf{s}}) \quad \text{subject to} \quad \|\mathbf{s}\|_1 = S.$$

We relax this objective to:

$$\min_{\mathbf{s} \in [0,1]^d} \quad \text{corr}(\mathbf{acc_{ID}}, \mathbf{acc}_{\mathbf{OOD}}^{\mathbf{s}}) + \lambda \cdot \left(S - \|\mathbf{s}\|_1\right)^2, \tag{5}$$

where $\mathbf{s}$ is the output of a sigmoid function in practice.

Soundness of the relaxation and optimization. Our objective is non-convex and non-submodular (Proposition 1), but Lipschitz-continuous (Lemma 3). While global optimization is intractable, the Lipschitz property ensures stable gradients and bounded progress under descent, enabling convergence toward near-binary stationary points that approximate the discrete optima. Non-submodularity also eliminates greedy selection as an optimal strategy. We use the Adam optimizer (Kingma and Ba, 2014) to optimize Equation 5. We use a cosine annealing schedule to adjust the learning rate and λ (Loshchilov and Hutter, 2016). Additional details are available in Appendix A.

On Selected Subsets. Although we are free to choose S examples, a subset that makes the ID-OOD Pearson R negative is not guaranteed to exist. The OOD accuracy of each model is an average over the selected examples. The subset must systematically up- or down-weight groups of examples on which higher-accuracy ID models tend to underperform relative to lower-accuracy ID models. We provide evidence that finding a large sign-flipping subset is evidence of latent structure or spurious shortcuts in the data, not a trivial consequence of sub-sampling. Importantly, we do not select models or alter the ID accuracies; we always correlate the same length-N vectors, only the OOD accuracy values change through the choice of examples.

For brevity, we reserve details of other theoretical analyses for Appendix B, as our results are included for thoroughness, but they are standard (Bertsekas, 1997; Nocedal and Wright, 1999).

Fisher Confidence Intervals. In our estimate of correlations, we compute Fisher z intervals for each correlation estimate, indicating the range of variability expected from estimation; overlapping bars suggest that differences could be arbitrary, while non-overlapping drops signal meaningful differences in correlation.

Table 1: **Dataset Summary.** Each dataset defines a classification task across multiple domains. Full OOD Size refers to the size of the OOD dataset. We select the full dataset to apply `OODSelect` according to splits in DomainBed (Gulrajani and Lopez-Paz, 2020) and WILDS (Koh et al., 2021). WILDSCamelyon-H4/H5 refer to the versions of the dataset where hospitals 4 and 5 are considered OOD distinctly. Importantly, we never train models on either and separate them here because they have distinct properties and results. WILDSCivilComments is a text dataset.

Dataset	Task (# classes)	Domains (#)	Full OOD Size	Largest `OODSelect` Subset w/ AoTIL (-0.3)	# Models
`Chest Xrays`	Finding vs. No Finding (2)	Hospital systems (5)	71433	55000 (75%)	1800
`PACS`	Object classification (7)	Styles (4)	3929	250 (6%)	2804
`VLCS`	Object classification (5)	Visual domains (4)	2656	800 (30%)	4200
`TerraIncognita`	Wildlife classification (10)	Camera traps (4)	6122	1500 (25%)	2980
`WILDSCamelyon-H4`	Tumor vs. Normal (2)	Hospitals (5)	129838	35000 (25%)	944
`WILDSCamelyon-H5`	Tumor vs. Normal (2)	Hospitals (5)	146722	60000 (40%)	944
`WILDSCivilComments`	Toxic vs. Not Toxic (2)	Demographics (8)	52823	25000 (50%)	710

4 Experiments

Procedure. Table 1 summarizes the datasets we study. Given a typical distribution shift benchmark with at least two domains, i.e., $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots\}$, we fix a $\mathcal{D}_{\text{ID}}$, $\mathcal{D}_{\text{OOD}} \subset \mathcal{D}$ pair, which are disjoint sets (concatenated) of domains. This pair denotes an experimental setting. In this work, we focus on the standard $\mathcal{D}_{\text{ID}}$, $\mathcal{D}_{\text{OOD}}$ splits the community uses for each dataset (Gulrajani and Lopez-Paz, 2020; Koh et al., 2021). For each split, we apply our methodology to identify subsets $\mathcal{D}^{\text{s}}_{\text{OOD}}$ with AoTIL—Appendix A Algorithm 1.

Datasets. We consider real-world tasks and distributions such as predicting "Finding"/"No Finding" from **Chest X-rays** where domains ID domains are from `CheXpert` (v1.0-small) (Irvin et al., 2019), `ChestXray8` (Wang et al., 2017), `PadChest` (Bustos et al., 2020), and `VinDr-CXR` (Nguyen et al., 2022). The OOD domain is `MIMIC-CXR-JPG` (Johnson et al., 2019). We also study WILDS (Koh et al., 2021) benchmarks that capture real-world shifts. **WILDS-Camelyon** (Bandi et al., 2018) targets cancer detection from histopathology slides across hospitals. **WILDS-CivilComments** (Borkan et al., 2019; Koh et al., 2021) classifies online comments as toxic or non-toxic across demographic subgroups, with OOD domains defined by shifts in identity attributes such as gender, religion, and race. We also study DomainBed (Gulrajani and Lopez-Paz, 2020) benchmarks reflecting different forms of distribution shift: style, dataset collection, and environment. **PACS** (Li et al., 2017) involves object classification across artistic styles (7 classes across *Photo*, *Art Painting*, `Cartoon`, and `Sketch`), with `Sketch` as OOD. **VLCS** (Fang et al., 2013) spans 5 classes across 4 datasets (`VOC2007` (Everingham et al., 2010), `LabelMe` (Russell et al., 2008), `Caltech101` (Fei-Fei et al., 2004), and `SUN09` (Choi et al., 2010)), capturing collection biases; `LabelMe` is OOD. **Terra Incognita** (Beery et al., 2018) focuses on wildlife recognition across 4 geographic locations, with `L46` as OOD.

Models. We construct a diverse population of models by varying architecture (from VGG to Vision Transformers, listed below), pretraining weights (TorchVision maintainers and contributors, 2016; Deng et al., 2009; He et al., 2019), initialization (from scratch and transfer learning), and hyperparameters. We train up to 4200 models (Figure 7) with various vision architectures, including variants of ResNets (He et al., 2016), DenseNets (Huang et al., 2017), MobileNets (Howard, 2017), ViT (Dosovitskiy et al., 2020), VGG (Simonyan and Zisserman, 2014), and Inception (Szegedy et al., 2015). We do the same for our language experiments, from BERT (Devlin et al., 2019) to GPT-2 (Radford et al., 2019). A full list of models is provided in Appendix A.

Models are split into disjoint train, validation, and test subsets, i.e., the models used for learning the selection, cross-validation, and final testing are non-overlapping. For a given ID/OOD setting, all models are trained on the same ID training data and evaluated on a held-out ID test set and candidate

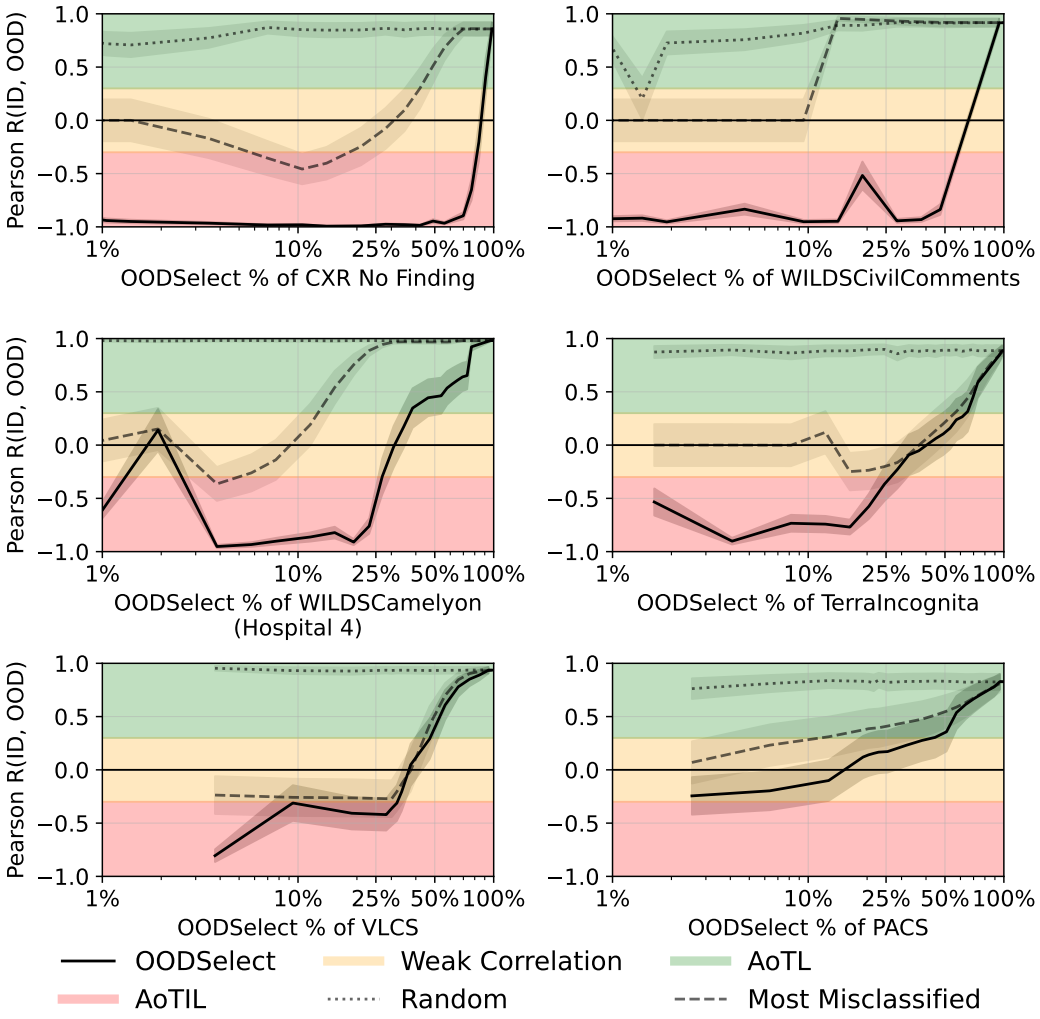
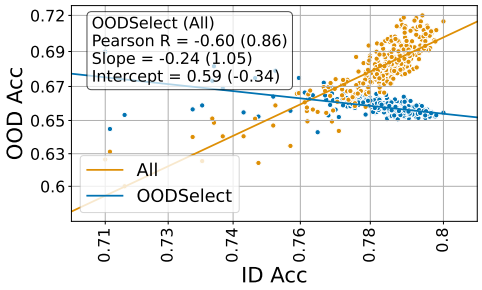


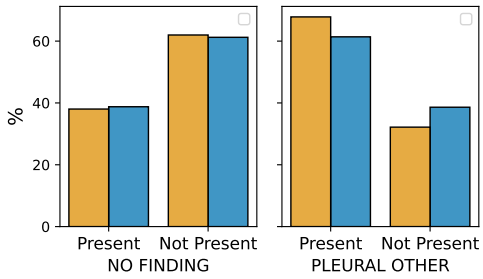
Figure 2: **Comparing AoTL and AoTIL.** Pearson Correlation between ID and OOD accuracy as a function of the number of selected OOD samples. Correlation values above 0.3 indicate AoTL, while below -0.3 is AoTIL—correlations in between are considered weak. We compare a Random Selection of data samples and the Most Misclassified at fixed size intervals from 100 to over 100,000 (normalized to sample size in the figure). Random selections yield strong positive correlation, while misclassified samples have weak correlations; that is, our method does not conflate spurious correlations with general difficulty (e.g., label noise). `OODSelect` identifies subsets where ID and OOD accuracy are negatively correlated—in one case (CXR) for over 70% of the usual OOD dataset. This behavior is dataset-dependent due to differences in distributional properties. Table 4 enumerates detailed correlations.

OOD subsets. The resulting paired ID/OOD accuracies are used to estimate the correlation between ID and OOD performance. Further discussion on implementation is provided in Appendix A.

On the Necessary Quantity of Models. We determine the minimum number of models to sample by thresholding the relative change in ID and OOD accuracy correlation across the full dataset. We select at least a number of models such that adding a new model changes the correlation by less than 1% (Schönbrodt and Perugini, 2013; Bonett and Wright, 2000). Notably, the diversity and quantity of models we consider are orders of magnitude higher than in previous work (Miller et al., 2021); in some cases, tens vs. thousands (ours). This number is also dataset dependent; for instance, 1010 models are needed to satisfy this criterion for the VLCS dataset, while only 610 are needed for WILDS Camelyon. Further details are provided in Appendix A Figure 7.



(a) CXR No Finding accuracy correlations across models has a strong global correlation (0.86) while the OODSelect subset (55000 examples and 77% of the full dataset) has a negative correlation (-0.60).



(b) There is minimal label shift in the full OOD set and the OODSelect subset. No Finding has a prevalence of 39% in the OODSelect samples and 38% in the full dataset. However, the prevalence of "Pleural Other" in the full OOD set and OODSelect subset is 61% and 68%, respectively.

Figure 3: **CXR No Finding.** Figure 3a suggests that poor generalization may arise for a large subset of the OOD population from reliance on spurious correlations. However, aggregation hides this failure mode since the correlation for the full OOD set is strongly positive. This selected subset also has a prevalence shift from the full dataset; statistical significance for the prevalence shift was assessed using bootstrapping with 1000 resamples.

5 Empirical Results and Discussion

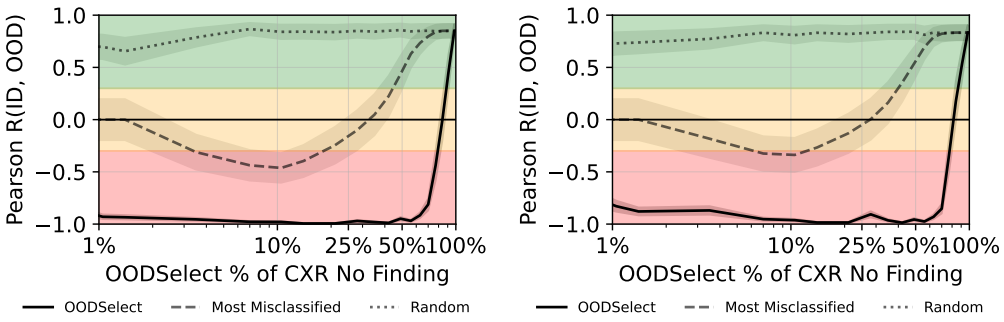
Findings. Overall, we find that many benchmarks contain `OODSelect` subsets of examples that exhibit AoTIL or a weak correlation, though the size of such subsets varies. The same benchmarks exhibit AoTL when all OOD samples are aggregated (Figure 2).

We treat $|R| < 0.3$ as a weak Pearson correlation between ID accuracies and OOD accuracies across models. Using `OODSelect`, our method for selecting the OOD data, we uncover large variance in correlations that are hidden in the full splits. In CXR No Finding, the full OOD set gives a strong positive correlation (Figure 2a), however, `OODSelect` retaining $> 70\%$ of the data has a strong negative correlation (Figure 2). For Terra Incognita, the full OOD set has a strong positive correlation, but a 30% slice from `OODSelect` has a notable negative correlation.

The extent of the existence of such subsets clearly varies across datasets and may not exist in others. For instance, for PACS, a small `OODSelect` size making up 8% of the full dataset has a correlation of -0.33; at 60 % the correlation is already negligible, 0.01, and becomes strongly positive as the size of `OODSelect` grows. The full dataset has a correlation of 0.81.

Focusing specifically on the most misclassified examples, we find that the ID-OOD correlation is near 0 and rarely invert it as the `OODSelect` examples do. This demonstrates that our selection does not conflate spuriousness with general difficulty (e.g., uniform label noise). Random selection consistently preserves a strong positive correlation similar to the full dataset, as expected. Thus `OODSelect` reveals systematic generalization failures due to spurious correlations, which may go undetected under standard evaluation across domains.

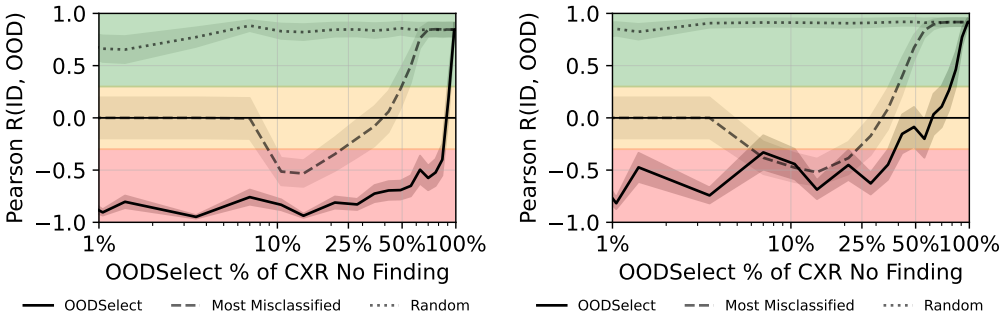
Clearly, to achieve a negative correlation from a positive correlation, we need (i) models that performed well on the full OOD set to perform relatively worse on the `OODSelect` set, and/or (ii) models that performed poorly on the full OOD set to perform relatively better on the `OODSelect` set. For instance, for VLCS, models with relatively low ID accuracy performed better on the `OODSelect` set than the full OOD set. In contrast, the models with relatively high ID accuracy performed better on the full OOD set. For all of our trends, the slope and intercept are determined by these relative performance changes and are dataset dependent. In some datasets, some models still perform near or below chance on the `OODSelect` (Terra Incognita) while in others, all models are above chance (WILDSCamelyon).



(a) CXR No Finding — Pearson R.

(b) CXR No Finding — Spearman Rank.

Figure 4: The correlation directions are not driven by outliers — Spearman rank is robust to outliers while Pearson R is not. Still, the trends are similar (full results in Figure 2 and 8).



(a) Disjoint architecture families — Pearson R.

(b) Disjoint architecture families — Spearman Rank.

Figure 5: Independent Architecture Families. Our findings hold even when train/test models are from disjoint architecture families, e.g., ResNets vs. ViTs.

On the effect of outliers. Some models may be outliers and skew the observed trends. Consequently, we evaluate Spearman rank correlation, which is more robust to outliers than Pearson R. We find that our conclusions remain unchanged (Figure 4). Spearman rank results are provided in Appendix A.

On potential architecture confounds. While we randomly split models into disjoint sets for identifying OOD subsets and computing correlations to simulate i.i.d. sampling from a model population, architectural similarities (e.g., ResNet-50 vs. ResNet-152) could introduce confounding effects. To test this, we perform ablations where model families are disjoint—e.g., ResNets appear only in the training-validation set or only in the test set, but never both—and find that this restriction indeed changes the strength of the correlation, yet does not alter our conclusions. Figure 5 gives an example for CXR No Finding, which has the strongest examples of AOTIL. However, given that architectures have different inductive biases, models may learn different spurious correlations or utilize them differently in decision-making. Sampling from an entirely disjoint population of architectures mitigates the observed strength of spurious correlations learned by model families.

On Vision Language Models (VLM) Trends. We investigate if the same trends hold with vision-language models' zero-shot performance (Shi et al., 2024). We generally find strongly positive correlations between the ID accuracy and the accuracy on `OODSelect` examples. The weakest correlations are: PACS (0.78), VLCS (0.62), TerraIncognita (0.84), WILDSCamelyon (1.00), and CXR (0.94). This should not be interpreted as evidence of VLMs' robustness to spurious correlations. From the point of view of the VLMs in this experiment, both the ID and `OODSelect` examples are OOD, since the VLMs were not explicitly trained on either set. Alternatively, since many of these datasets are publicly available, the VLMs may have been trained on the dataset sets, i.e., all of the examples are in distribution.

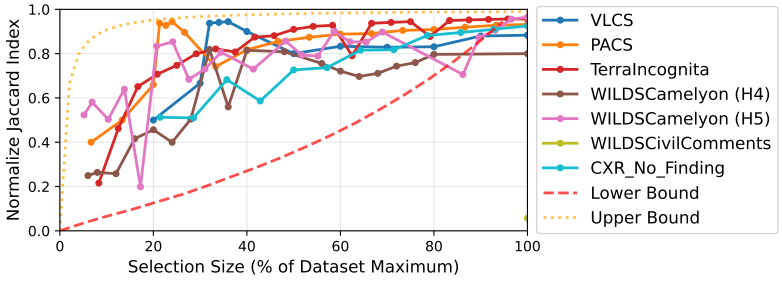


Figure 6: Consistency of selected subsets. Across all datasets and subset sizes, our normalized Jaccard index is greater than would be expected from arbitrary selection (lower bound).

Table 2: Model generated description of selected OOD set vs. ID set.

Dataset	Model Generated Semantic Difference
PACS	extreme wide-angle shots; extreme close-ups; extreme weather conditions.
VLCS	unusual object interactions with urban environments; x-ray or radiographic.
TerraIncognita	frost; motion blur; extreme weather conditions; reflections or glare.

Selection via latent space distance. As a baseline, we implement a selection method that greedily selects the farthest OOD examples from the ID examples in the CLIP embedding space (Radford et al., 2021). Across datasets, this approach often yields positive ID–OOD correlations (e.g., $R = 0.52$ on PACS with $N = 10$), and in some cases even stronger correlations than random selection. However, it consistently fails to capture the weak and negative correlations identified by OODSelect (e.g., $R = -0.92$ on VLCS with $N = 10$). These results show that distance-based selection, while intuitive, overlooks the feature-label correlations that drive OOD errors, and thus cannot uncover the failure modes revealed by our method.

Selection Consistency and Coherence. The identified subsets are also consistent and coherent. We select each S subset independently and do not enforce that smaller subsets are subsets of larger subsets. Still, we find that such consistency holds. We measure consistency with the normalized `Jaccard Index` (Jaccard, 1901) $\in [0, 1]$. For $\mathbf{z} \subset [d]$ examples,

$$\mathcal{J}(\mathbf{z}_i, \mathbf{z}_j) = \frac{|\mathbf{z}_i \cap \mathbf{z}_j|}{|\mathbf{z}_i \cup \mathbf{z}_j|}; \qquad \bar{\mathcal{J}}_{\mathbf{z}} = \frac{1}{T}\sum_{k=1}^{T} \mathcal{J}(\mathbf{z}_k, \mathbf{z}_{k+1}); \qquad \widetilde{\mathcal{J}}_{\mathbf{z}} = \frac{\bar{\mathcal{J}}_{\mathbf{z}} - \bar{\mathcal{J}}_{\min}}{\bar{\mathcal{J}}_{\max} - \bar{\mathcal{J}}_{\min}}, \tag{6}$$

for T selection sizes, where $\widetilde{\mathcal{J}}_{\min}$ is computed with random selections and $\widetilde{\mathcal{J}}_{\max}$ with $\mathbf{z}_i \subset \mathbf{z}_j$ for all i, j, with $i < j$, and the sizes of the sequence of $\mathbf{z}_k$'s are preserved. This normalization is necessary since the subsets are of different sizes. Our selected subsets are indeed consistent (Figure 6).

CXR Semantic Coherence. The CXR dataset, predicting Finding/No Finding[2] in chest X-rays (CXR), is an example where we have demographic and clinical metadata that we can use to study the semantic coherence of our subsets. Figure 3 illustrates how average OOD performance can mask systematic failures in specific subsets. For instance, when selecting a subset with 5000 examples, Figure 3, the ID/OOD accuracy correlation between ID and OOD on the selected subset is strongly negative, while it is strongly positive when we aggregate over the full OOD set.

We then analyze both demographic and clinical attributes. By comparing prevalence rates between the selected subset and the overall OOD pool, we find statistically significant shifts in several attributes, specifically `sex, race, Pleural Other, Support Devices, and Sex-Ethnicity`, determined via bootstrapping with 1000 resamples, Figure 3. However, most datasets have no such metadata. Our normalized Jaccard Index supports consistency and coherence for such datasets.

Potential for model-generated semantic coherence. As a potential future research direction, we investigate the utility of large and vision language models to generate semantic concepts more likely to

[2]"Finding" indicates the presence of a condition from a predefined set and "No Finding" indicates otherwise.

be true for our OODSelect set than the rest of the dataset (Dunlap et al., 2024). We apply the following process. *Step 1:* A VLM generates captions for all images; we use `Qwen2.5-32B-Instruct` (Yang et al., 2024b; Team, 2024). *Step 2:* A large language model (LLM) then proposes candidate natural language descriptions that are more likely to apply to the selected OOD set than others; we use `AIMV2-large-patch14-224-lit` (Fini et al., 2024). *Step 3:* A vision-language model scores and ranks these descriptions based on their distinctiveness to OOD images, identifying interpretable attributes that differentiate the two distributions; we use `CLIP` (Radford et al., 2021).

Table 2 provides example descriptions for natural image datasets. We find that this strategy does not yield consistent and robust results, as the descriptions do not capture feature-label correlations, although some of our findings are promising. Additional details are provided in Appendix C.

Importantly, many potentially spurious predictive features are incomprehensible to humans and may not be expressible in natural language (Szegedy et al., 2013; Goodfellow et al., 2014; Ilyas et al., 2019). As a result, approaches that rely on vision-language models for explanation are likely still insufficient for identifying the subsets we uncover, for instance, through brute-force selection. They also cannot be expected to capture all differences in correlations between ID and OOD examples. This is true for many of the datasets in this work, such as CXR and WILDSCamelyon.

Limitations. Our analysis is computationally intensive, requiring training up to 4200 models per dataset and optimizing a selection objective of up to around 146000 elements. However, this computation is a one-time cost per dataset; we publicly release the resulting selections, covering many state-of-the-art domain generalization benchmarks. Additionally, semantic explanations of the OODSelect set are challenging for datasets such as WILDSCamelyon, whose features are images of tissue cell slides, without extensive metadata. Even when metadata is available, it may not fully represent the signals that capture spurious correlations. Notably, this is also an unstated challenge for the original datasets, where OOD sets are selected based on metadata such as hospital sites, but also contain no information explaining what spurious correlations exist or are expected. Furthermore, it is unclear if we can expect semantic explanations for all spurious correlations since many features models rely on are imperceptible to humans (Szegedy et al., 2013; Goodfellow et al., 2014; Ilyas et al., 2019). For instance, AI systems can predict race from chest X-rays with features that are thus far imperceptible to humans (Gichoya et al., 2022).

Broader Impact. One alternative perspective of our results is that correlations that hold in aggregate are not spurious (Wenzel et al., 2022). We propose that aggregate performance is a narrow view of the effect of spurious correlations. For instance, if spurious statistical associations reflecting historical or structural bias, such as occupation and gender, which can bias the outputs of recommendation systems (Caliskan et al., 2017; Balagopalan et al., 2025), are pervasive in the real world. Then, benchmarks collected *naturally* from real-world distributions whose results are aggregated broadly may preserve such correlations across both training and test environments. As a result, models that rely on such spurious correlations may continue to "perform well OOD," making the correlation appear benign in evaluation. However, this only creates the false impression that spurious correlations are not harmful OOD, even though they degrade performance on affected subsets of the data. Our work surfaces these subsets and advances more robust evaluations of OOD robustness.

6 Conclusion

Spurious correlations do not vanish in the real world; current benchmarks and performance metrics simply hide them through aggregation. By disaggregating OOD data, we revealed large, semantically meaningful subsets where spurious correlations harm performance. The consequential validity (Messick, 1995; Salaudeen et al., 2025b) of distribution shift robustness benchmarks, e.g., robustness to subpopulation shifts (Yang et al., 2023; Sagawa et al., 2019), requires identifying such subsets.

Recommendations. Future work in this area of research should (i) adopt our selection protocol as a robustness check for any new OOD benchmark, (ii) treat identified large OODSelect subsets as first-class evaluation targets, and (iii) design methods that improve *both* average and subset robustness. A discussion on interpreting the results of subset performance can be found in Pfohl et al. (2025). We hope the released code and OODSelect subsets become a stepping stone toward benchmarks and models that confront the adverse effects of spurious correlations.

Acknowledgments

MG acknowledges partial support by the National Science Foundation (NSF) 22-586 Faculty Early Career Development Award (#2339381), a Gordon & Betty Moore Foundation award, a Google Research Scholar award, and the AI2050 Program at Schmidt Sciences. We thank the anonymous reviewers of NeurIPS 2025 for their thoughtful feedback and recommendations. SB acknowledges partial support by the Schmidt Sciences AI2050 Program, NSF Awards No. 2330423 and 2441060, and NSERC Award No. 585136. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, NSERC, or Schmidt Sciences.

References

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1):521–530, 2012.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30(10):2838–2848, 2024a.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

Olawale Salaudeen and Sanmi Koyejo. Causally inspired regularization enables domain general representations. In *International Conference on Artificial Intelligence and Statistics*, pages 3124–3132. PMLR, 2024.

Olawale Elijah Salaudeen, Nicole Chiou, and Sanmi Koyejo. On domain generalization datasets as proxy benchmarks for causal representation learning. In *NeurIPS 2024 Causal Representation Learning Workshop*, 2024.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. *arXiv preprint arXiv:2302.12254*, 2023.

Rahul Saxena, Taeyoun Kim, Aman Mehra, Christina Baek, J Zico Kolter, and Aditi Raghunathan. Predicting the performance of foundation models via agreement-on-the-line. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Amartya Sanyal, Yaxi Hu, Yaodong Yu, Yian Ma, Yixin Wang, and Bernhard Schölkopf. Accuracy on the wrong line: On the pitfalls of noisy data for out-of-distribution generalisation. *arXiv preprint arXiv:2406.19049*, 2024.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71703–71722. Curran Associates, Inc., 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/e304d374c85e385eb217ed4a025b6b63-Paper-Conference.pdf`.

Olawale Elijah Salaudeen, Nicole Chiou, Shiny Weng, and Sanmi Koyejo. Are domain generalization benchmarks with accuracy on the line misspecified? *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856. URL `https://openreview.net/forum?id=fNywRyqPQo`.

Neoklis Polyzotis, Steven Whang, Tim Klas Kraska, and Yeounoh Chung. Slice finder: Automated data slicing for model validation. In *Proceedings of the IEEE Int'Conf. on Data Engineering (ICDE)*, volume 2019, 2019.

Hugo M Proença, Peter Grünwald, Thomas Bäck, and Matthijs van Leeuwen. Robust subgroup discovery: Discovering subgroup lists using mdl. *Data Mining and Knowledge Discovery*, 36(5): 1885–1970, 2022.

Eliana Pastor, Luca de Alfaro, and Elena Baralis. Identifying biased subgroups in ranking and classification. *arXiv preprint arXiv:2108.07450*, 2021.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

Jacob R Epifano, Ravi P Ramachandran, Aaron J Masino, and Ghulam Rasool. Revisiting the fragility of influence functions. *Neural Networks*, 162:581–588, 2023.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B Grosse. If influence functions are the answer, then what is the question? *Advances in Neural Information Processing Systems*, 35:17953–17967, 2022.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Pang Wei W Koh, Kai-Siang Ang, Hubert Teo, and Percy S Liang. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32, 2019.

Yuzheng Hu, Pingbang Hu, Han Zhao, and Jiaqi Ma. Most influential subset selection: Challenges, promises, and beyond. *Advances in Neural Information Processing Systems*, 37:119778–119810, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3): 334–334, 1997.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66: 101797, 2020.

Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):429, 2022.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017.

Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE international conference on computer vision*, pages 1657–1664, 2013.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77:157–173, 2008.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

Myung Jin Choi, Joseph J Lim, Antonio Torralba, and Alan S Willsky. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 129–136. IEEE, 2010.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4918–4927, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Andrew G Howard. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Felix D Schönbrodt and Marco Perugini. At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5):609–612, 2013.

Douglas G Bonett and Thomas A Wright. Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika*, 65:23–28, 2000.

Jia Shi, Gautam Gare, Jinjin Tian, Siqi Chai, Zhiqiu Lin, Arun Vasudevan, Di Feng, Francesco Ferroni, and Shu Kong. Lca-on-the-line: Benchmarking out-of-distribution generalization with class taxonomies. *arXiv preprint arXiv:2407.16067*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579, 1901.

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024b.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Victor Guilherme Turrisi da Costa, Louis Béthune, Zhe Gan, Alexander T Toshev, Marcin Eichner, Moin Nabi, Yinfei Yang, Joshua M. Susskind, and Alaaeldin El-Nouby. Multimodal autoregressive pre-training of large vision encoders, 2024.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4 (6):e406–e414, 2022.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Aparna Balagopalan, Kai Wang, Olawale Salaudeen, Asia Biega, and Marzyeh Ghassemi. What's in a query: Polarity-aware distribution-based fair ranking. In *Proceedings of the ACM on Web Conference 2025*, pages 3716–3730, 2025.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50 (9):741, 1995.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025b.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Stephen R Pfohl, Natalie Harris, Chirag Nagpal, David Madras, Vishwali Mhasawade, Olawale Salaudeen, Awa Dieng, Shannon Sequeira, Santiago Arciniegas, Lillian Sung, et al. Understanding challenges to the interpretation of disaggregated evaluations of algorithmic fairness. *arXiv preprint arXiv:2506.04193*, 2025.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL `https://arxiv.org/abs/2401.04088`.

Appendix Table of Contents

A Empirical Analysis **17**

A.1 Compute 19

A.2 Spearman Rank Results 19

A.3 Table of Results 21

B Theoretical Analysis **27**

B.1 Lemma 1: Bounded effect of New Models on Pearson R 27

B.2 Lemma 2: Bounded effect of New Examples on Pearson R 27

B.3 Lemma 3: Lipschitz Continuity of Selection Objective 28

B.4 Proof of Proposition 1: Non-Submodularity 29

C ID/OOD Explanation **30**

A Empirical Analysis

Vision Models. `AlexNet`, `ResNet-(18/34/50/101/152)`, `DenseNet-(121/161/169/201)`, `MobileNet (V2/V3 Small/V3 Large)`, `EfficientNet-(B0/B1/B3/B7)`, `ConvNeXt-(Tiny/Small/Base/Large)`, `ViT-(B/16, B/32, L/16)`, `Swin Transformer-(Tiny/Small/Base)`, `RegNet-Y (400MF/800MF/1.6GF/3.2GF/8GF)`, `VGG-(11/13/16/19)`, `SqueezeNet (1.0/1.1)`, and `Inception v3`.

Text Models. `BERT-(base/large)`, `SciBERT`, `RoBERTa-(base/large)`, `BioBERT`, `LegalBERT`, `FinBERT`, `ALBERT-(v1/v2, base/large/xlarge/xxlarge)`, `DeBERTa-v2-(xsmall/small/base/large)`, `Longformer-(base/large)`, `DistilBERT-(base/cased, distilled)`, `T5-(small)`, `BART-(base/large/mnli)`, and `GPT-2-(small)`.

See Figure 7 on quantity of models trained.

Dataset. **PACS** involves object classification across artistic styles, with 7 classes ("dog", "elephant", "giraffe", "guitar", "horse", "house", "person") and 4 domains: *Photo*, *Art Painting*, *Cartoon*, and *Sketch*. We consider a setting where *Sketch* is the OOD domain. **VLCS** contains 5 object classes ("bird", "car", "chair", "dog", "person") shared across 4 datasets: *VOC2007* (Everingham et al., 2010), *LabelMe* (Russell et al., 2008), *Caltech101* (Fei-Fei et al., 2004), and *SUN09* (Choi et al., 2010). Each domain reflects a different dataset source with distinct collection biases. We consider a setting where *LabelMe* is the OOD domain. **Terra Incognita** focuses on wildlife recognition from camera trap images, with 10 classes ("bird", "bobcat", "cat", "coyote", "dog", "opossum", "raccoon", "rabbit", "skunk", "squirrel") across 4 geographically distinct domains: *L38*, *L43*, *L46*, and *L100*. The *L46* location is the OOD domain (Gulrajani and Lopez-Paz, 2020).

We also study three WILDS benchmarks that capture distinct real-world distribution shifts. We consider WILDS-Camelyon (Bandi et al., 2018) and WILDS-CivilComments (Borkan et al., 2019). These benchmarks encompass medical imaging, satellite vision, and natural language, providing a diverse evaluation suite for real-world generalization under domain shift. **WILDS-Camelyon** is a histopathology image classification task for detecting cancerous regions in lymph node slides, with domain shifts arising from variations across different hospitals. **WILDS-CivilComments** (Borkan et al., 2019) classifies online comments as toxic or non-toxic across demographic subgroups (*male*, *female*, *LGBTQ*, *Christian*, *Muslim*, *other religions*, *Black*, *White*), with OOD domains reflecting shifts in identity distributions.

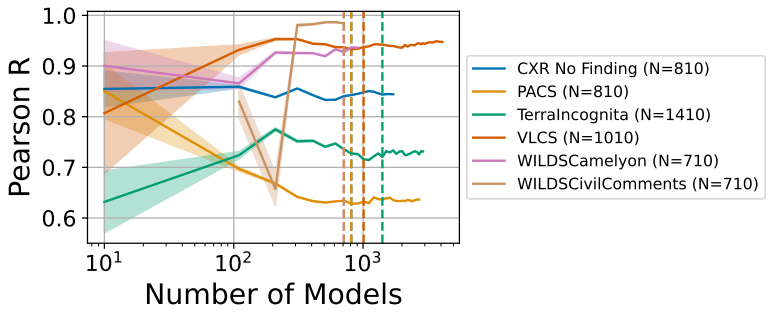


Figure 7: We train over 35 model architectures with varying hyperparameters, pretraining, and data augmentations, yielding an average of [N] trained models per dataset–OOD domain pair. We report the correlation between in-distribution (ID) and out-of-distribution (OOD) accuracy across all models, including standard errors at $\alpha = 0.05$. To ensure stability, we sample enough models such that adding more changes the correlation by less than 1%; vertical dashed lines mark the approximate minimum sample size satisfying this criterion.

CXR ("Finding" vs. "No Finding").[3] This binary classification task predicts whether a chest X-ray shows any abnormal radiological finding. The in-distribution (ID) domains comprise four widely-used datasets—*CheXpert* (v1.0-small) (Irvin et al., 2019), *ChestXray8* (Wang et al., 2017), *PadChest* (Bustos et al., 2020), and *VinDr-CXR* (Nguyen et al., 2022). These sources differ in scanner hardware, patient demographics, annotation guidelines, and prevalence of pathologies. We designate *MIMIC-CXR* (Johnson et al., 2019)—a large, single-institution dataset collected under a distinct clinical workflow—as the out-of-distribution (OOD) domain. This setting captures clinically meaningful shifts (e.g., hospital protocols, imaging devices, disease prevalence) and offers a stringent test of real-world generalization under domain shift.

Train/Val/Test Split. To evaluate generalization, we randomly partition the same set of models into train, validation, and test splits (60/20/20). We optimize our selection objective on the training split and identify the best-performing `OODSelect` configuration using the held-out validation split. Final results are reported on the test split. Importantly, although the selection objective is tuned on one subset of models, the ID and OOD accuracy correlations continue to hold on the held-out test models, demonstrating that the property generalizes across held-out model subsets.

Soundness of the relaxation and optimization. Notably, our objective is non-convex and non-submodular (Proposition 1) yet Lipschitz-continuous (Lemma 3). Consequently, while global optimality is intractable, the Lipschitz property ensures that gradient-based methods with a suitably large exact-penalty parameter admit meaningful descent guarantees; in practice we employ stochastic gradient descent with multiple random restarts, which consistently converges to high-quality feasible solutions. Formal optimization guarantees and proofs are deferred to Appendix B.

Adding the squared regularization term in (5) is an *exact-penalty* reformulation of the original constrained problem. Classical results (Bertsekas, 1997; Nocedal and Wright, 1999)) state that there exists a finite weight $\lambda^\star > 0$ such that, for every $\lambda \geq \lambda^\star$, (i) every global minimiser of the penalised objective satisfies the budget constraint, and (ii) the optimal value coincides with that of the constrained problem.

Moreover, any first-order stationary point that already meets the constraint is *unchanged* by the penalty term, so the relaxation does not create spurious local optima within the feasible region. Hence, gradient-based search on (5) is sufficient: we do not need to solve the penalized problem to global optimality, and any locally optimal feasible solution we find is also locally optimal for the original constrained objective.

We use a cosine annealing schedule to adjust the learning rate and λ (Loshchilov and Hutter, 2016). Additional details are available in Appendix A.

[3]Throughout, we refer to this task as *CXR* for brevity.

Necessary Number of Models. Prior work on accuracy-on-the-line typically trained only tens of models per dataset (Miller et al., 2021). In contrast, we train thousands of models per dataset, spanning architectures from AlexNet to Vision Transformers and incorporating diverse training strategies. To determine how many models are necessary for stable correlation estimates, we incrementally sample models until the Pearson correlation between ID and OOD accuracies changes by less than 1%. Figure 7 shows where this stability threshold is reached for each dataset. Across all datasets, our experiments far exceed this threshold.

On the Size of `OODSelect`. While a detailed analysis of thresholding `OODSelect` is beyond our current scope, we generally recommend choosing the largest `OODSelect` size such that the Pearson correlation is not weak, that is $R \leq -0.3$, following convention (Cohen, 2013). For some datasets, this threshold may yield very small or noisy selections (e.g., PACS), in which case the selected set may not be informative.

A.1 Compute

Table 3: Compute time to reproduce experiments (GPU Hours) — per experiment unit on NVIDIA RTX A6000 GPUs.

Dataset	mean	median	std. dev.	min	max	total
CXR	4	3	3	<1	18	286
PACS	1	1	1	<1	7	109
TerraIncognita	2	2	2	<1	10	292
VLCS	2	2	1	<1	7	183
WILDSCamelyon	4	2	4	1	18	350
WILDSCivilComments	10	9	4	3	18	106

A.2 Spearman Rank Results

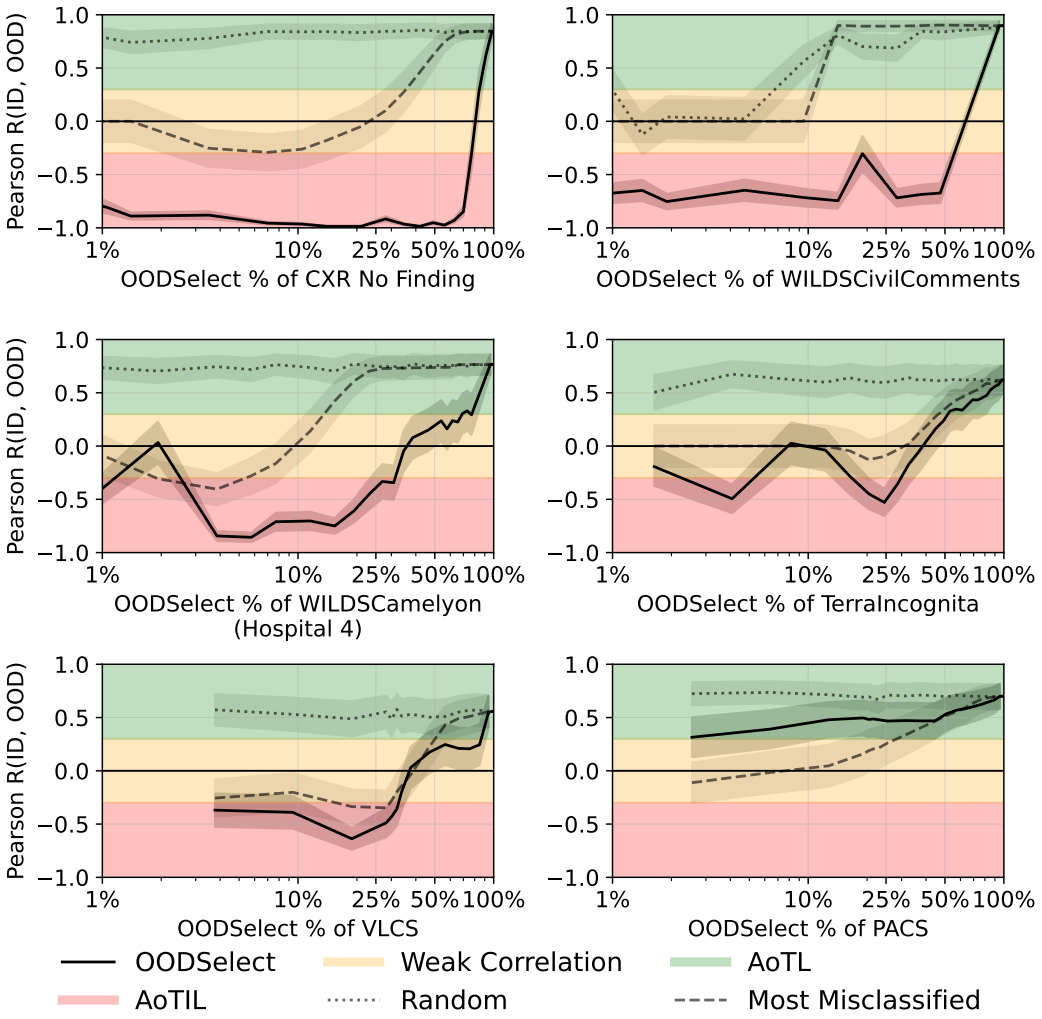


Figure 8: **Comparing AoTL and AoTIL.** Spearman Correlation between ID and OOD accuracy as a function of the number of selected OOD samples.

A.3 Table of Results

Table 4: ID vs. selected OOD accuracy correlations (Pearson R and Spearman ρ) with standard errors over 100 resamplings. For "Random", we randomly select N subsets from candidate OOD examples; for "Hard", we select the N most misclassified examples.

Dataset	**OOD**	**N**	**Pearson R**			**Spearman ρ**		
			Ours	**Random**	**Hard**	**Ours**	**Random**	**Hard**
CXR No Finding	MIMIC-CXR	10	-0.56 (0.12)	0.19 (0.20)	0.00 (0.20)	-0.46 (0.14)	0.07 (0.20)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	20	-0.23 (0.18)	0.20 (0.20)	0.00 (0.20)	-0.33 (0.16)	0.14 (0.20)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	50	-0.59 (0.12)	0.43 (0.18)	0.00 (0.20)	-0.58 (0.12)	0.46 (0.17)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	100	-0.69 (0.09)	0.59 (0.14)	0.00 (0.20)	-0.60 (0.11)	0.52 (0.16)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	250	-0.83 (0.05)	0.70 (0.11)	0.00 (0.20)	-0.72 (0.08)	0.57 (0.15)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	500	-0.84 (0.05)	0.71 (0.11)	0.00 (0.20)	-0.70 (0.09)	0.72 (0.11)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	750	-0.93 (0.02)	0.69 (0.12)	0.00 (0.20)	-0.81 (0.06)	0.75 (0.10)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	1000	-0.93 (0.02)	0.66 (0.13)	0.00 (0.20)	-0.87 (0.04)	0.72 (0.11)	0.00 (0.20)
CXR No Finding	MIMIC-CXR	2500	-0.95 (0.02)	0.78 (0.09)	-0.30 (0.17)	-0.86 (0.04)	0.78 (0.09)	-0.19 (0.18)
CXR No Finding	MIMIC-CXR	5000	-0.98 (0.01)	0.86 (0.06)	-0.40 (0.15)	-0.95 (0.02)	0.84 (0.07)	-0.29 (0.17)
CXR No Finding	MIMIC-CXR	7500	-0.98 (0.01)	0.84 (0.07)	-0.44 (0.15)	-0.96 (0.01)	0.83 (0.07)	-0.30 (0.17)
CXR No Finding	MIMIC-CXR	10000	-0.99 (0.00)	0.84 (0.07)	-0.37 (0.16)	-0.99 (0.00)	0.84 (0.07)	-0.25 (0.18)
CXR No Finding	MIMIC-CXR	15000	-0.99 (0.00)	0.84 (0.07)	-0.22 (0.18)	-0.98 (0.00)	0.83 (0.07)	-0.13 (0.19)
CXR No Finding	MIMIC-CXR	20000	-0.97 (0.01)	0.85 (0.07)	-0.07 (0.19)	-0.91 (0.03)	0.84 (0.07)	0.02 (0.20)
CXR No Finding	MIMIC-CXR	25000	-0.98 (0.01)	0.84 (0.07)	0.08 (0.20)	-0.96 (0.01)	0.84 (0.07)	0.18 (0.20)
CXR No Finding	MIMIC-CXR	30000	-0.99 (0.00)	0.85 (0.07)	0.26 (0.19)	-0.99 (0.00)	0.85 (0.07)	0.37 (0.18)
CXR No Finding	MIMIC-CXR	35000	-0.95 (0.02)	0.85 (0.06)	0.45 (0.17)	-0.95 (0.01)	0.85 (0.07)	0.56 (0.15)
CXR No Finding	MIMIC-CXR	40000	-0.97 (0.01)	0.84 (0.07)	0.64 (0.13)	-0.98 (0.01)	0.83 (0.07)	0.71 (0.11)
CXR No Finding	MIMIC-CXR	45000	-0.91 (0.03)	0.85 (0.07)	0.74 (0.10)	-0.93 (0.02)	0.84 (0.07)	0.80 (0.08)
CXR No Finding	MIMIC-CXR	50000	-0.83 (0.05)	0.85 (0.07)	0.80 (0.08)	-0.87 (0.04)	0.84 (0.07)	0.83 (0.07)
CXR No Finding	MIMIC-CXR	55000	-0.46 (0.14)	0.84 (0.07)	0.84 (0.07)	-0.32 (0.17)	0.84 (0.07)	0.84 (0.07)
CXR No Finding	MIMIC-CXR	60000	-0.02 (0.20)	0.84 (0.07)	0.85 (0.07)	0.18 (0.20)	0.84 (0.07)	0.84 (0.07)
CXR No Finding	MIMIC-CXR	65000	0.46 (0.17)	0.85 (0.07)	0.85 (0.07)	0.56 (0.15)	0.84 (0.07)	0.84 (0.07)
CXR No Finding	MIMIC-CXR	70000	0.85 (0.07)	0.85 (0.07)	0.85 (0.07)	0.84 (0.07)	0.84 (0.07)	0.84 (0.07)
CXR No Finding	MIMIC-CXR	71433	0.85	0.85	0.85	0.84	0.84	0.84
WILDSCivilComments	4	10	-0.33 (0.16)	0.39 (0.18)	0.00 (0.20)	-0.50 (0.13)	0.16 (0.20)	0.00 (0.20)
WILDSCivilComments	4	20	-0.89 (0.03)	0.85 (0.06)	0.00 (0.20)	-0.26 (0.17)	0.13 (0.20)	0.00 (0.20)
WILDSCivilComments	4	50	-0.89 (0.03)	0.39 (0.18)	0.00 (0.20)	-0.55 (0.12)	0.23 (0.19)	0.00 (0.20)
WILDSCivilComments	4	100	-0.56 (0.12)	0.74 (0.10)	0.00 (0.20)	-0.53 (0.13)	0.19 (0.20)	0.00 (0.20)

Continued on next page

Dataset	OOD	N	Pearson R			Spearman ρ		
			Ours	Random	Hard	Ours	Random	Hard
WILDSCivilComments	4	250	-0.79 (0.06)	0.94 (0.03)	0.00 (0.20)	-0.71 (0.09)	0.33 (0.19)	0.00 (0.20)
WILDSCivilComments	4	500	-0.70 (0.09)	0.93 (0.03)	0.00 (0.20)	-0.75 (0.08)	0.46 (0.17)	0.00 (0.20)
WILDSCivilComments	4	750	-0.98 (0.01)	0.80 (0.08)	0.00 (0.20)	-0.65 (0.10)	0.02 (0.20)	0.00 (0.20)
WILDSCivilComments	4	1000	-0.93 (0.02)	0.82 (0.08)	0.00 (0.20)	-0.82 (0.06)	0.08 (0.20)	0.00 (0.20)
WILDSCivilComments	4	2500	-0.78 (0.07)	0.93 (0.03)	0.00 (0.20)	-0.74 (0.08)	-0.01 (0.20)	0.00 (0.20)
WILDSCivilComments	4	5000	-0.98 (0.01)	0.95 (0.02)	0.00 (0.20)	-0.76 (0.07)	0.47 (0.17)	0.00 (0.20)
WILDSCivilComments	4	7500	-0.98 (0.01)	0.97 (0.01)	0.95 (0.03)	-0.80 (0.06)	0.79 (0.09)	0.88 (0.05)
WILDSCivilComments	4	10000	-0.88 (0.04)	0.97 (0.01)	0.99 (0.01)	-0.41 (0.15)	0.70 (0.12)	0.90 (0.05)
WILDSCivilComments	4	15000	-0.97 (0.01)	0.97 (0.02)	0.98 (0.01)	-0.78 (0.07)	0.70 (0.12)	0.89 (0.05)
WILDSCivilComments	4	20000	-0.97 (0.01)	0.98 (0.01)	0.98 (0.01)	-0.76 (0.07)	0.83 (0.07)	0.90 (0.05)
WILDSCivilComments	4	25000	-0.96 (0.01)	0.98 (0.01)	0.98 (0.01)	-0.73 (0.08)	0.84 (0.07)	0.91 (0.04)
WILDSCivilComments	4	50000	0.98 (0.01)	0.98 (0.01)	0.98 (0.01)	0.90 (0.05)	0.90 (0.05)	0.90 (0.05)
WILDSCivilComments	4	52823	0.98	0.98	0.98	0.90	0.90	0.90
WILDSCamelyon	Hospital 4	10	-0.94 (0.02)	0.90 (0.05)	0.00 (0.20)	-0.23 (0.18)	0.33 (0.19)	0.00 (0.20)
WILDSCamelyon	Hospital 4	20	-0.91 (0.03)	0.74 (0.11)	0.00 (0.20)	-0.17 (0.18)	0.25 (0.19)	0.00 (0.20)
WILDSCamelyon	Hospital 4	50	-0.93 (0.02)	0.49 (0.16)	0.00 (0.20)	-0.32 (0.17)	0.50 (0.16)	0.00 (0.20)
WILDSCamelyon	Hospital 4	100	-0.92 (0.03)	0.94 (0.03)	0.00 (0.20)	-0.34 (0.16)	0.52 (0.16)	0.00 (0.20)
WILDSCamelyon	Hospital 4	250	-0.88 (0.04)	0.96 (0.02)	0.00 (0.20)	-0.42 (0.15)	0.37 (0.18)	0.00 (0.20)
WILDSCamelyon	Hospital 4	500	-0.96 (0.01)	0.97 (0.02)	0.00 (0.20)	-0.53 (0.13)	0.49 (0.17)	0.00 (0.20)
WILDSCamelyon	Hospital 4	750	-0.96 (0.01)	0.98 (0.01)	0.00 (0.20)	-0.61 (0.11)	0.75 (0.10)	0.00 (0.20)
WILDSCamelyon	Hospital 4	1000	-0.91 (0.03)	0.99 (0.01)	0.00 (0.20)	-0.51 (0.13)	0.74 (0.10)	0.00 (0.20)
WILDSCamelyon	Hospital 4	2500	0.07 (0.20)	0.98 (0.01)	0.08 (0.20)	0.00 (0.20)	0.71 (0.11)	-0.51 (0.13)
WILDSCamelyon	Hospital 4	5000	-0.99 (0.00)	0.99 (0.01)	-0.32 (0.17)	-0.93 (0.02)	0.77 (0.09)	-0.31 (0.17)
WILDSCamelyon	Hospital 4	7500	-0.98 (0.01)	0.99 (0.01)	-0.23 (0.18)	-0.95 (0.02)	0.78 (0.09)	-0.24 (0.18)
WILDSCamelyon	Hospital 4	10000	-0.98 (0.01)	0.98 (0.01)	-0.10 (0.19)	-0.88 (0.04)	0.77 (0.10)	-0.09 (0.19)
WILDSCamelyon	Hospital 4	15000	-0.97 (0.01)	0.99 (0.01)	0.23 (0.19)	-0.82 (0.06)	0.78 (0.09)	0.23 (0.19)
WILDSCamelyon	Hospital 4	20000	-0.95 (0.02)	0.99 (0.01)	0.52 (0.16)	-0.80 (0.06)	0.79 (0.09)	0.48 (0.17)
WILDSCamelyon	Hospital 4	25000	-0.88 (0.04)	0.99 (0.01)	0.70 (0.12)	-0.49 (0.14)	0.80 (0.08)	0.64 (0.13)
WILDSCamelyon	Hospital 4	30000	-0.90 (0.03)	0.99 (0.01)	0.83 (0.07)	-0.50 (0.13)	0.80 (0.09)	0.76 (0.10)
WILDSCamelyon	Hospital 4	35000	-0.54 (0.13)	0.99 (0.01)	0.92 (0.04)	-0.28 (0.17)	0.80 (0.08)	0.79 (0.09)
WILDSCamelyon	Hospital 4	40000	-0.08 (0.19)	0.99 (0.01)	0.98 (0.01)	-0.37 (0.16)	0.78 (0.09)	0.79 (0.09)
WILDSCamelyon	Hospital 4	45000	0.08 (0.20)	0.99 (0.01)	0.98 (0.01)	-0.02 (0.20)	0.78 (0.09)	0.79 (0.09)
WILDSCamelyon	Hospital 4	50000	0.27 (0.19)	0.99 (0.01)	0.98 (0.01)	0.12 (0.20)	0.81 (0.08)	0.79 (0.09)
WILDSCamelyon	Hospital 4	60000	0.30 (0.19)	0.99 (0.01)	0.98 (0.01)	0.17 (0.20)	0.78 (0.09)	0.80 (0.08)

Continued on next page

Dataset	OOD	N	Pearson R			Spearman ρ		
			Ours	Random	Hard	Ours	Random	Hard
WILDSCamelyon	Hospital 4	70000	0.36 (0.18)	0.99 (0.01)	0.98 (0.01)	0.24 (0.19)	0.80 (0.09)	0.80 (0.08)
WILDSCamelyon	Hospital 4	75000	0.40 (0.18)	0.99 (0.01)	0.98 (0.01)	0.18 (0.20)	0.79 (0.09)	0.80 (0.08)
WILDSCamelyon	Hospital 4	80000	0.41 (0.18)	0.99 (0.01)	0.98 (0.01)	0.24 (0.19)	0.80 (0.08)	0.80 (0.08)
WILDSCamelyon	Hospital 4	85000	0.45 (0.17)	0.99 (0.01)	0.98 (0.01)	0.23 (0.19)	0.81 (0.08)	0.80 (0.08)
WILDSCamelyon	Hospital 4	90000	0.47 (0.17)	0.99 (0.01)	0.99 (0.01)	0.28 (0.19)	0.80 (0.08)	0.81 (0.08)
WILDSCamelyon	Hospital 4	95000	0.50 (0.16)	0.99 (0.01)	0.99 (0.01)	0.29 (0.19)	0.81 (0.08)	0.81 (0.08)
WILDSCamelyon	Hospital 4	100000	0.91 (0.04)	0.99 (0.01)	0.99 (0.01)	0.26 (0.19)	0.80 (0.08)	0.81 (0.08)
WILDSCamelyon	Hospital 4	125000	0.99 (0.01)	0.99 (0.01)	0.99 (0.01)	0.81 (0.08)	0.81 (0.08)	0.81 (0.08)
WILDSCamelyon	Hospital 4	129838	0.99	0.99	0.99	0.81	0.81	0.81
WILDSCamelyon	Hospital 5	10	-0.63 (0.10)	0.25 (0.19)	0.00 (0.20)	-0.03 (0.19)	0.63 (0.14)	0.00 (0.20)
WILDSCamelyon	Hospital 5	20	-0.83 (0.05)	0.43 (0.18)	0.00 (0.20)	0.13 (0.20)	0.48 (0.17)	0.00 (0.20)
WILDSCamelyon	Hospital 5	50	-0.74 (0.08)	0.73 (0.11)	0.00 (0.20)	0.16 (0.20)	0.39 (0.18)	0.00 (0.20)
WILDSCamelyon	Hospital 5	100	-0.81 (0.06)	0.71 (0.11)	-0.18 (0.18)	0.21 (0.20)	0.36 (0.18)	-0.26 (0.17)
WILDSCamelyon	Hospital 5	250	-0.91 (0.03)	0.75 (0.10)	-0.18 (0.18)	-0.53 (0.13)	0.48 (0.17)	-0.46 (0.14)
WILDSCamelyon	Hospital 5	500	-0.87 (0.04)	0.82 (0.08)	-0.51 (0.13)	0.33 (0.19)	0.36 (0.18)	-0.63 (0.11)
WILDSCamelyon	Hospital 5	750	-0.81 (0.06)	0.80 (0.08)	-0.45 (0.14)	0.36 (0.18)	0.47 (0.17)	-0.64 (0.10)
WILDSCamelyon	Hospital 5	1000	-0.83 (0.05)	0.84 (0.07)	-0.38 (0.16)	0.35 (0.18)	0.57 (0.15)	-0.59 (0.11)
WILDSCamelyon	Hospital 5	2500	-0.61 (0.11)	0.80 (0.08)	-0.48 (0.14)	0.26 (0.19)	0.46 (0.17)	-0.74 (0.08)
WILDSCamelyon	Hospital 5	5000	-0.43 (0.15)	0.80 (0.08)	-0.44 (0.15)	0.40 (0.18)	0.49 (0.16)	-0.63 (0.11)
WILDSCamelyon	Hospital 5	7500	-0.44 (0.15)	0.81 (0.08)	-0.38 (0.16)	-0.11 (0.19)	0.47 (0.17)	-0.51 (0.13)
WILDSCamelyon	Hospital 5	10000	-0.68 (0.09)	0.81 (0.08)	-0.32 (0.17)	-0.39 (0.15)	0.48 (0.17)	-0.41 (0.15)
WILDSCamelyon	Hospital 5	15000	-0.58 (0.12)	0.80 (0.08)	-0.15 (0.19)	-0.71 (0.09)	0.47 (0.17)	-0.22 (0.18)
WILDSCamelyon	Hospital 5	20000	-0.84 (0.05)	0.81 (0.08)	0.01 (0.20)	0.02 (0.20)	0.48 (0.17)	0.01 (0.20)
WILDSCamelyon	Hospital 5	25000	-0.81 (0.06)	0.80 (0.08)	0.14 (0.20)	0.09 (0.20)	0.47 (0.17)	0.17 (0.20)
WILDSCamelyon	Hospital 5	30000	-0.77 (0.07)	0.81 (0.08)	0.25 (0.19)	0.15 (0.20)	0.48 (0.17)	0.28 (0.19)
WILDSCamelyon	Hospital 5	35000	-0.68 (0.09)	0.81 (0.08)	0.32 (0.19)	0.23 (0.19)	0.48 (0.17)	0.35 (0.18)
WILDSCamelyon	Hospital 5	40000	-0.71 (0.08)	0.80 (0.08)	0.38 (0.18)	0.23 (0.19)	0.47 (0.17)	0.40 (0.18)
WILDSCamelyon	Hospital 5	45000	-0.20 (0.18)	0.80 (0.08)	0.43 (0.17)	0.18 (0.20)	0.47 (0.17)	0.41 (0.18)
WILDSCamelyon	Hospital 5	50000	-0.66 (0.10)	0.80 (0.08)	0.48 (0.17)	0.27 (0.19)	0.47 (0.17)	0.42 (0.18)
WILDSCamelyon	Hospital 5	60000	-0.62 (0.11)	0.80 (0.08)	0.54 (0.16)	0.28 (0.19)	0.47 (0.17)	0.43 (0.17)
WILDSCamelyon	Hospital 5	70000	0.13 (0.20)	0.81 (0.08)	0.59 (0.15)	0.28 (0.19)	0.48 (0.17)	0.46 (0.17)
WILDSCamelyon	Hospital 5	75000	0.29 (0.19)	0.80 (0.08)	0.67 (0.12)	0.33 (0.19)	0.48 (0.17)	0.47 (0.17)
WILDSCamelyon	Hospital 5	80000	0.31 (0.19)	0.81 (0.08)	0.73 (0.11)	0.38 (0.18)	0.48 (0.17)	0.47 (0.17)
WILDSCamelyon	Hospital 5	85000	0.40 (0.18)	0.80 (0.08)	0.77 (0.09)	0.35 (0.18)	0.48 (0.17)	0.47 (0.17)

Continued on next page

Dataset	OOD	N	Pearson R			Spearman ρ		
			Ours	Random	Hard	Ours	Random	Hard
WILDSCamelyon	Hospital 5	90000	0.37 (0.18)	0.80 (0.08)	0.79 (0.09)	0.40 (0.18)	0.48 (0.17)	0.47 (0.17)
WILDSCamelyon	Hospital 5	95000	0.40 (0.18)	0.80 (0.08)	0.81 (0.08)	0.43 (0.18)	0.48 (0.17)	0.47 (0.17)
WILDSCamelyon	Hospital 5	100000	0.56 (0.15)	0.80 (0.08)	0.82 (0.08)	0.41 (0.18)	0.48 (0.17)	0.47 (0.17)
WILDSCamelyon	Hospital 5	125000	0.66 (0.13)	0.81 (0.08)	0.82 (0.08)	0.45 (0.17)	0.48 (0.17)	0.48 (0.17)
WILDSCamelyon	Hospital 5	130000	0.70 (0.12)	0.80 (0.08)	0.81 (0.08)	0.42 (0.18)	0.48 (0.17)	0.48 (0.17)
WILDSCamelyon	Hospital 5	135000	0.72 (0.11)	0.81 (0.08)	0.81 (0.08)	0.43 (0.17)	0.48 (0.17)	0.48 (0.17)
WILDSCamelyon	Hospital 5	140000	0.76 (0.10)	0.81 (0.08)	0.81 (0.08)	0.44 (0.17)	0.48 (0.17)	0.48 (0.17)
WILDSCamelyon	Hospital 5	145000	0.81 (0.08)	0.80 (0.08)	0.80 (0.08)	0.48 (0.17)	0.48 (0.17)	0.48 (0.17)
WILDSCamelyon	Hospital 5	146722	0.80	0.80	0.80	0.48	0.48	0.48
TerraIncognita	L46	10	-0.86 (0.04)	0.45 (0.17)	0.00 (0.20)	-0.36 (0.16)	0.45 (0.17)	0.00 (0.20)
TerraIncognita	L46	20	-0.90 (0.03)	0.92 (0.04)	0.00 (0.20)	-0.36 (0.16)	0.49 (0.17)	0.00 (0.20)
TerraIncognita	L46	50	0.40 (0.18)	0.92 (0.04)	0.00 (0.20)	-0.18 (0.18)	0.44 (0.17)	0.00 (0.20)
TerraIncognita	L46	100	-0.58 (0.12)	0.87 (0.06)	0.00 (0.20)	-0.33 (0.16)	0.46 (0.17)	0.00 (0.20)
TerraIncognita	L46	250	-0.91 (0.03)	0.90 (0.04)	0.00 (0.20)	-0.46 (0.14)	0.66 (0.13)	0.00 (0.20)
TerraIncognita	L46	500	-0.77 (0.07)	0.87 (0.06)	0.03 (0.20)	-0.05 (0.19)	0.58 (0.15)	0.02 (0.20)
TerraIncognita	L46	750	-0.74 (0.08)	0.89 (0.05)	0.12 (0.20)	-0.01 (0.20)	0.59 (0.15)	-0.03 (0.20)
TerraIncognita	L46	1000	-0.77 (0.07)	0.89 (0.05)	-0.24 (0.18)	-0.22 (0.18)	0.61 (0.14)	-0.12 (0.19)
TerraIncognita	L46	1250	-0.59 (0.11)	0.90 (0.04)	-0.24 (0.18)	-0.43 (0.15)	0.60 (0.14)	-0.15 (0.19)
TerraIncognita	L46	1500	-0.40 (0.15)	0.90 (0.04)	-0.22 (0.18)	-0.54 (0.13)	0.58 (0.15)	-0.10 (0.19)
TerraIncognita	L46	1750	-0.26 (0.17)	0.87 (0.06)	-0.19 (0.18)	-0.38 (0.16)	0.60 (0.14)	-0.06 (0.19)
TerraIncognita	L46	2000	-0.12 (0.19)	0.89 (0.05)	-0.12 (0.19)	-0.23 (0.18)	0.62 (0.14)	0.02 (0.20)
TerraIncognita	L46	2250	-0.08 (0.19)	0.89 (0.05)	-0.05 (0.19)	-0.12 (0.19)	0.59 (0.14)	0.08 (0.20)
TerraIncognita	L46	2500	-0.02 (0.20)	0.89 (0.05)	0.04 (0.20)	-0.00 (0.20)	0.61 (0.14)	0.17 (0.20)
TerraIncognita	L46	2750	0.04 (0.20)	0.89 (0.05)	0.11 (0.20)	0.10 (0.20)	0.59 (0.14)	0.25 (0.19)
TerraIncognita	L46	3000	0.09 (0.20)	0.90 (0.05)	0.18 (0.20)	0.17 (0.20)	0.60 (0.14)	0.32 (0.19)
TerraIncognita	L46	3250	0.15 (0.20)	0.89 (0.05)	0.25 (0.19)	0.27 (0.19)	0.59 (0.14)	0.36 (0.18)
TerraIncognita	L46	3500	0.25 (0.19)	0.90 (0.05)	0.30 (0.19)	0.25 (0.19)	0.62 (0.14)	0.39 (0.18)
TerraIncognita	L46	3750	0.29 (0.19)	0.89 (0.05)	0.37 (0.18)	0.24 (0.19)	0.60 (0.14)	0.43 (0.18)
TerraIncognita	L46	4000	0.34 (0.19)	0.90 (0.05)	0.43 (0.17)	0.31 (0.19)	0.61 (0.14)	0.45 (0.17)
TerraIncognita	L46	4250	0.45 (0.17)	0.90 (0.04)	0.51 (0.16)	0.36 (0.18)	0.61 (0.14)	0.47 (0.17)
TerraIncognita	L46	4500	0.58 (0.15)	0.89 (0.05)	0.59 (0.14)	0.36 (0.18)	0.61 (0.14)	0.49 (0.16)
TerraIncognita	L46	4750	0.64 (0.13)	0.89 (0.05)	0.66 (0.13)	0.38 (0.18)	0.59 (0.14)	0.52 (0.16)
TerraIncognita	L46	5000	0.70 (0.12)	0.89 (0.05)	0.72 (0.11)	0.41 (0.18)	0.61 (0.14)	0.54 (0.15)
TerraIncognita	L46	5250	0.75 (0.10)	0.89 (0.05)	0.77 (0.09)	0.48 (0.17)	0.61 (0.14)	0.54 (0.16)

Continued on next page

Dataset	OOD	N	Pearson R			Spearman ρ		
			Ours	Random	Hard	Ours	Random	Hard
TerraIncognita	L46	5500	0.80 (0.08)	0.89 (0.05)	0.81 (0.08)	0.52 (0.16)	0.60 (0.14)	0.57 (0.15)
TerraIncognita	L46	5750	0.84 (0.07)	0.89 (0.05)	0.85 (0.07)	0.55 (0.15)	0.60 (0.14)	0.60 (0.14)
TerraIncognita	L46	6000	0.89 (0.05)	0.89 (0.05)	0.88 (0.05)	0.61 (0.14)	0.60 (0.14)	0.60 (0.14)
TerraIncognita	L46	6122	0.89	0.89	0.89	0.60	0.60	0.60
VLCS	LabelMe	10	-0.91 (0.03)	0.80 (0.08)	0.00 (0.20)	-0.28 (0.17)	0.16 (0.20)	0.00 (0.20)
VLCS	LabelMe	20	-0.90 (0.03)	0.90 (0.04)	0.00 (0.20)	-0.30 (0.17)	0.27 (0.19)	0.00 (0.20)
VLCS	LabelMe	50	-0.07 (0.19)	0.93 (0.03)	-0.13 (0.19)	-0.29 (0.17)	0.43 (0.18)	-0.15 (0.19)
VLCS	LabelMe	100	-0.82 (0.05)	0.96 (0.02)	-0.22 (0.18)	-0.39 (0.15)	0.57 (0.15)	-0.21 (0.18)
VLCS	LabelMe	250	-0.27 (0.17)	0.94 (0.03)	-0.25 (0.18)	-0.37 (0.16)	0.57 (0.15)	-0.28 (0.17)
VLCS	LabelMe	500	-0.40 (0.15)	0.94 (0.03)	-0.21 (0.18)	-0.62 (0.11)	0.52 (0.16)	-0.29 (0.17)
VLCS	LabelMe	750	-0.40 (0.15)	0.95 (0.03)	-0.25 (0.18)	-0.50 (0.13)	0.59 (0.15)	-0.33 (0.16)
VLCS	LabelMe	800	-0.33 (0.16)	0.94 (0.03)	-0.23 (0.18)	-0.43 (0.15)	0.51 (0.16)	-0.26 (0.17)
VLCS	LabelMe	850	-0.28 (0.17)	0.95 (0.03)	-0.18 (0.18)	-0.33 (0.16)	0.61 (0.14)	-0.18 (0.18)
VLCS	LabelMe	900	-0.17 (0.18)	0.95 (0.03)	-0.10 (0.19)	-0.17 (0.18)	0.56 (0.15)	-0.10 (0.19)
VLCS	LabelMe	1000	0.09 (0.20)	0.95 (0.03)	0.03 (0.20)	0.08 (0.20)	0.59 (0.15)	0.03 (0.20)
VLCS	LabelMe	1250	0.33 (0.19)	0.94 (0.03)	0.47 (0.17)	0.23 (0.19)	0.55 (0.15)	0.28 (0.19)
VLCS	LabelMe	1500	0.65 (0.13)	0.95 (0.03)	0.73 (0.11)	0.30 (0.19)	0.55 (0.15)	0.49 (0.16)
VLCS	LabelMe	1750	0.81 (0.08)	0.95 (0.03)	0.87 (0.06)	0.26 (0.19)	0.60 (0.14)	0.56 (0.15)
VLCS	LabelMe	2000	0.87 (0.06)	0.95 (0.03)	0.92 (0.04)	0.26 (0.19)	0.62 (0.14)	0.58 (0.15)
VLCS	LabelMe	2250	0.91 (0.04)	0.95 (0.02)	0.94 (0.03)	0.31 (0.19)	0.62 (0.14)	0.59 (0.14)
VLCS	LabelMe	2500	0.95 (0.03)	0.95 (0.02)	0.94 (0.03)	0.60 (0.14)	0.61 (0.14)	0.60 (0.14)
VLCS	LabelMe	2656	0.95	0.95	0.95	0.61	0.61	0.61
PACS	Sketch	10	-0.48 (0.14)	0.37 (0.18)	0.17 (0.20)	-0.05 (0.19)	0.39 (0.18)	-0.06 (0.19)
PACS	Sketch	20	-0.33 (0.16)	0.71 (0.11)	0.17 (0.20)	-0.41 (0.15)	0.34 (0.19)	0.00 (0.20)
PACS	Sketch	50	-0.47 (0.14)	0.84 (0.07)	0.00 (0.20)	0.23 (0.19)	0.47 (0.17)	-0.12 (0.19)
PACS	Sketch	100	-0.33 (0.16)	0.73 (0.11)	0.11 (0.20)	0.29 (0.19)	0.70 (0.12)	-0.08 (0.19)
PACS	Sketch	250	-0.30 (0.17)	0.79 (0.09)	0.19 (0.20)	0.35 (0.19)	0.70 (0.12)	-0.04 (0.19)
PACS	Sketch	500	-0.22 (0.18)	0.83 (0.07)	0.28 (0.19)	0.41 (0.18)	0.69 (0.12)	0.07 (0.20)
PACS	Sketch	750	0.01 (0.20)	0.82 (0.08)	0.33 (0.19)	0.43 (0.17)	0.65 (0.13)	0.18 (0.20)
PACS	Sketch	800	0.05 (0.20)	0.81 (0.08)	0.33 (0.19)	0.42 (0.18)	0.66 (0.13)	0.19 (0.20)
PACS	Sketch	850	0.06 (0.20)	0.80 (0.08)	0.33 (0.19)	0.42 (0.18)	0.64 (0.13)	0.22 (0.20)
PACS	Sketch	900	0.08 (0.20)	0.83 (0.07)	0.34 (0.19)	0.42 (0.18)	0.65 (0.13)	0.24 (0.19)
PACS	Sketch	1000	0.10 (0.20)	0.80 (0.08)	0.35 (0.19)	0.40 (0.18)	0.67 (0.12)	0.27 (0.19)
PACS	Sketch	1250	0.16 (0.20)	0.81 (0.08)	0.38 (0.18)	0.41 (0.18)	0.67 (0.12)	0.33 (0.19)

Continued on next page

Dataset	OOD	N	Pearson R			Spearman ρ		
			Ours	**Random**	**Hard**	**Ours**	**Random**	**Hard**
PACS	Sketch	1500	0.21 (0.20)	0.82 (0.08)	0.41 (0.18)	0.42 (0.18)	0.68 (0.12)	0.39 (0.18)
PACS	Sketch	1750	0.24 (0.19)	0.82 (0.08)	0.46 (0.17)	0.42 (0.18)	0.67 (0.12)	0.43 (0.17)
PACS	Sketch	2000	0.29 (0.19)	0.82 (0.08)	0.49 (0.16)	0.48 (0.17)	0.66 (0.13)	0.48 (0.17)
PACS	Sketch	2250	0.48 (0.17)	0.81 (0.08)	0.54 (0.16)	0.52 (0.16)	0.67 (0.12)	0.51 (0.16)
PACS	Sketch	2500	0.56 (0.15)	0.81 (0.08)	0.58 (0.15)	0.54 (0.16)	0.67 (0.12)	0.56 (0.15)
PACS	Sketch	2750	0.62 (0.14)	0.81 (0.08)	0.63 (0.13)	0.56 (0.15)	0.67 (0.13)	0.61 (0.14)
PACS	Sketch	3000	0.67 (0.12)	0.81 (0.08)	0.68 (0.12)	0.58 (0.15)	0.67 (0.12)	0.64 (0.13)
PACS	Sketch	3250	0.71 (0.11)	0.81 (0.08)	0.72 (0.11)	0.61 (0.14)	0.67 (0.13)	0.66 (0.13)
PACS	Sketch	3500	0.75 (0.10)	0.81 (0.08)	0.76 (0.10)	0.63 (0.14)	0.66 (0.13)	0.66 (0.13)
PACS	Sketch	3750	0.81 (0.08)	0.81 (0.08)	0.79 (0.09)	0.67 (0.13)	0.66 (0.13)	0.67 (0.13)
PACS	Sketch	3929	0.81	0.81	0.81	0.67	0.67	0.67

B Theoretical Analysis

B.1 Lemma 1: Bounded effect of New Models on Pearson R

Lemma 1 (Bounded Effect of a New Model on Pearson R)**.** *Let $(\mathbf{z}_i, \mathbf{w}_i)_{i=1}^N \subseteq [\alpha, 1-\alpha]^2$ with $\alpha \in (0,1)$. Define*

$$\mathbf{x}_i = \Phi^{-1}(\mathbf{z}_i), \qquad \mathbf{y}_i = \Phi^{-1}(\mathbf{w}_i),$$

and let

$$\rho_N = \mathit{corr}\big(\mathbf{x}_{1:N}, \mathbf{y}_{1:N}\big)$$

be the sample Pearson correlation of the first N transformed pairs. Add one more pair $(\mathbf{z}_{N+1}, \mathbf{w}_{N+1})$ with $\mathbf{z}_{N+1} \in [\alpha, 1-\alpha]$ and $\mathbf{w}_{N+1} = \beta\, \mathbf{z}_{N+1}$, and denote the updated correlation by ρ_{N+1}. Then

$$\big|\rho_{N+1} - \rho_N\big| \;\le\; \frac{\kappa(1+|\beta|)\, M_\alpha^2}{N}, \qquad M_\alpha = \max\big\{|\Phi^{-1}(\alpha)|, |\Phi^{-1}(1-\alpha)|\big\},$$

where the constant $\kappa > 0$ depends only on α (via the Lipschitz constant of Φ^{-1} on $[\alpha, 1-\alpha]$).

Proof. Let $\mathbf{x}_i = \Phi^{-1}(\mathbf{z}_i)$ and $\mathbf{y}_i = \Phi^{-1}(\mathbf{w}_i)$ for $i = 1, \ldots, N$. Denote $\bar{\mathbf{x}}_N := \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i$ and similarly $\bar{\mathbf{y}}_N$. Let ρ_N be the Pearson correlation between $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$:

$$\rho_N = \frac{\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}_N)(\mathbf{y}_i - \bar{\mathbf{y}}_N)}{\sqrt{\sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}}_N)^2 \sum_{i=1}^N (\mathbf{y}_i - \bar{\mathbf{y}}_N)^2}}.$$

Let $(\mathbf{z}_{N+1}, \mathbf{w}_{N+1})$ be a new pair with $\mathbf{w}_{N+1} = \beta \mathbf{z}_{N+1}$ and both in $[\alpha, 1-\alpha]$. Define $\mathbf{x}_{N+1} = \Phi^{-1}(\mathbf{z}_{N+1})$ and $\mathbf{y}_{N+1} = \Phi^{-1}(\mathbf{w}_{N+1})$. Since Φ^{-1} is L_α-Lipschitz on $[\alpha, 1-\alpha]$, we have $|\mathbf{x}_{N+1}| \le M_\alpha$ and $|\mathbf{y}_{N+1}| \le L_\alpha|\beta| + M_\alpha \le (1+|\beta|)M_\alpha$. Thus, each $|\mathbf{x}_i|, |\mathbf{y}_i| \le (1+|\beta|)M_\alpha$.

Let ρ_{N+1} denote the Pearson correlation after adding $(\mathbf{x}_{N+1}, \mathbf{y}_{N+1})$. A first-order perturbation of the sample Pearson correlation (cf. derivative bounds on correlation statistics) gives:

$$|\rho_{N+1} - \rho_N| \le \frac{C}{N} \cdot \max_i \{|\mathbf{x}_i|, |\mathbf{y}_i|\}^2,$$

where C is a constant depending on the Lipschitz constant L_α and lower bound on variance (which is lower bounded by $(\alpha(1-\alpha)/L_\alpha^2)$ due to the probit transform). Therefore,

$$|\rho_{N+1} - \rho_N| \le \frac{\kappa(1+|\beta|)M_\alpha^2}{N},$$

where $\kappa > 0$ depends only on α. □

B.2 Lemma 2: Bounded effect of New Examples on Pearson R

Lemma 2 (Bounded Effect of a New Example on Pearson R)**.** *Fix $\alpha \in (0,1)$ and assume the per-model accuracies satisfy $\mathbf{z}_i^{(S)}$, $\mathbf{w}_i \in [\alpha, 1-\alpha]$ for $i = 1, \ldots, d$. Define*

$$\mathbf{x}_i = \Phi^{-1}(\mathbf{z}_i), \qquad \mathbf{y}_i = \Phi^{-1}(\mathbf{w}_i),$$

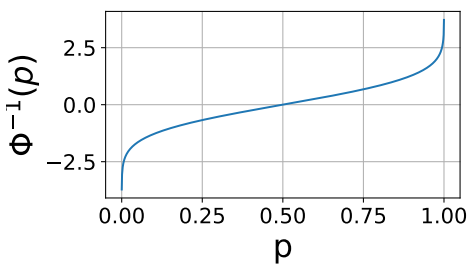
and let

$$\rho_S = \mathit{corr}\big(\widetilde{\mathbf{x}}^{(S)}, \widetilde{\mathbf{y}}\big)$$

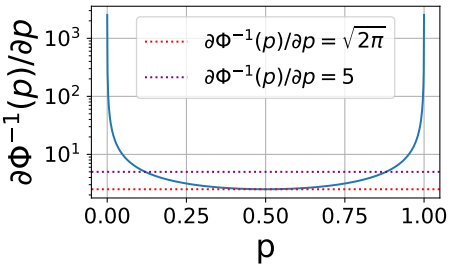
be their sample Pearson correlation. Now add an additional selected example. Note that each average changes by at most $|\mathbf{z}_i^{(S+1)} - \mathbf{z}_i^{(S)}| \le 1/(S+1)$. Write ρ_{S+1} for the resulting correlation between $\mathbf{x}^{(S+1)}$ and $\mathbf{y}$. Then

$$\big|\rho_{S+1} - \rho_S\big| \;\le\; \frac{\kappa\, M_\alpha^2}{S+1}, \qquad M_\alpha = \max\big\{|\Phi^{-1}(\alpha)|, |\Phi^{-1}(1-\alpha)|\big\},$$

where the constant $\kappa > 0$ depends only on α (via the Lipschitz constant of Φ^{-1} on $[\alpha, 1-\alpha]$).



(a) Inverse CDF (Probit) of the Standard Normal



(b) Derivative of the Inverse Normal CDF $= 1/\phi(\Phi^{-1}(p))$ where ϕ and Φ are PDF and CDF, respectively.

Figure 9: The probit transform is indeed Lipschitz.

Proof. Let $\mathbf{x}_i = \Phi^{-1}(\mathbf{z}_i^{(S)})$, $\mathbf{y}_i = \Phi^{-1}(\mathbf{w}_i)$, and let $\rho_S = \text{corr}(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$. Adding a new example changes each $\mathbf{z}_i^{(S)}$ by at most $1/(S+1)$, so $|\mathbf{x}_i^{(S+1)} - \mathbf{x}_i^{(S)}| \le L_\alpha/(S+1)$. Denote $\delta_i = \mathbf{x}_i^{(S+1)} - \mathbf{x}_i^{(S)}$ and $\bar{\delta} = \frac{1}{d}\sum_{i=1}^d \delta_i$. Then

$$|\widetilde{\mathbf{x}}_i^{(S+1)} - \widetilde{\mathbf{x}}_i^{(S)}| = |\delta_i - \bar{\delta}| \le 2L_\alpha/(S+1).$$

Using $\|\widetilde{\mathbf{y}}\|_\infty \le M_\alpha$,

$$\left|\widetilde{\mathbf{x}}^{(S+1)\top}\widetilde{\mathbf{y}} - \widetilde{\mathbf{x}}^{(S)\top}\widetilde{\mathbf{y}}\right| \le \sum_{i=1}^d |\widetilde{\mathbf{x}}_i^{(S+1)} - \widetilde{\mathbf{x}}_i^{(S)}||\widetilde{\mathbf{y}}_i| \le \frac{2dL_\alpha M_\alpha}{S+1}.$$

Also, $\|\widetilde{\mathbf{x}}\|, \|\widetilde{\mathbf{y}}\| \ge \sqrt{dv_\alpha}$ for $v_\alpha := \alpha(1-\alpha)L_\alpha^{-2}$. So

$$|\rho_{S+1} - \rho_S| \le \frac{2dL_\alpha M_\alpha}{(S+1)dv_\alpha} = \frac{2L_\alpha M_\alpha}{(S+1)v_\alpha} = \frac{2L_\alpha^3 M_\alpha}{\alpha(1-\alpha)(S+1)}.$$

Setting $\kappa := \frac{2L_\alpha^3}{\alpha(1-\alpha)}$ gives the result. □

B.3 Lemma 3: Lipschitz Continuity of Selection Objective

Lemma 3 (Lipschitz Continuity of Pearson Correlation w.r.t. $\mathbf{s}$)**.** *Let $\mathbf{X} \in \mathbb{R}^{N\times d}$ be the binary accuracy matrix and $\mathbf{y} \in \mathbb{R}^N$ the held-out training accuracy vector. Define the test-set accuracy for a given selection vector $\mathbf{s} \in [0,1]^d$ (with fixed total mass $\sum_{j=1}^d \mathbf{s}_j = S > 0$) as*

$$\widehat{\mathbf{x}}(\mathbf{s}) = \frac{\mathbf{X}\,\mathbf{s}}{S},$$

and let f be a Lipschitz-continuous probit transformation with Lipschitz constant L_f. Denote

$$\widetilde{\mathbf{x}}(\mathbf{s}) = f\big(\widehat{\mathbf{x}}(\mathbf{s})\big), \quad \widetilde{\mathbf{y}} = f(\mathbf{y}),$$

and the centered versions by

$$\bar{\mathbf{x}}(\mathbf{s}) = \widetilde{\mathbf{x}}(\mathbf{s}) - \frac{1}{N}\mathbf{1}^\top\widetilde{\mathbf{x}}(\mathbf{s}), \quad \bar{\mathbf{y}} = \widetilde{\mathbf{y}} - \frac{1}{N}\mathbf{1}^\top\widetilde{\mathbf{y}}.$$

The Pearson correlation between $\widetilde{\mathbf{x}}(\mathbf{s})$ and $\widetilde{\mathbf{y}}$ is defined as

$$corr\big(\widetilde{\mathbf{x}}(\mathbf{s}), \widetilde{\mathbf{y}}\big) = \frac{\frac{1}{N}\bar{\mathbf{x}}(\mathbf{s})^\top\bar{\mathbf{y}}}{\sqrt{\frac{1}{N^2}\|\bar{\mathbf{x}}(\mathbf{s})\|^2\,\|\bar{\mathbf{y}}\|^2}}.$$

Assume that there exists $\epsilon > 0$ such that $\|\bar{\mathbf{x}}(\mathbf{s})\| \ge \epsilon$ for all admissible $\mathbf{s}$. Then, for any two selection vectors $\mathbf{s}, \mathbf{s}' \in [0,1]^d$ with $\sum_{j=1}^d \mathbf{s}_j = \sum_{j=1}^d \mathbf{s}'_j = S$, there exists a constant $L > 0$ (depending on L_f, $\mathbf{X}$, S, and ϵ) such that

$$\Big|corr\big(\widetilde{\mathbf{x}}(\mathbf{s}), \widetilde{\mathbf{y}}\big) - corr\big(\widetilde{\mathbf{x}}(\mathbf{s}'), \widetilde{\mathbf{y}}\big)\Big| \le L\,\|\mathbf{s} - \mathbf{s}'\|.$$

Proof. Since the test-set accuracy is given by

$$\widehat{\mathbf{x}}(\mathbf{s}) = \frac{\mathbf{X}\,\mathbf{s}}{S},$$

linearity implies that

$$\|\widehat{\mathbf{x}}(\mathbf{s}) - \widehat{\mathbf{x}}(\mathbf{s}')\| \leq \frac{\|\mathbf{X}\|}{S}\,\|\mathbf{s} - \mathbf{s}'\|,$$

where $\|\mathbf{X}\|$ denotes an appropriate operator norm.

Using the Lipschitz continuity of f (with constant L_f), we have for each coordinate,

$$\left|f\big(\widehat{x}_i(\mathbf{s})\big) - f\big(\widehat{x}_i(\mathbf{s}')\big)\right| \leq L_f\,\left|\widehat{x}_i(\mathbf{s}) - \widehat{x}_i(\mathbf{s}')\right|,$$

so in vector form,

$$\|\widetilde{\mathbf{x}}(\mathbf{s}) - \widetilde{\mathbf{x}}(\mathbf{s}')\| \leq L_f\,\|\widehat{\mathbf{x}}(\mathbf{s}) - \widehat{\mathbf{x}}(\mathbf{s}')\| \leq \frac{L_f\,\|\mathbf{X}\|}{S}\,\|\mathbf{s} - \mathbf{s}'\|.$$

Since centering is a linear operation, it follows that

$$\|\bar{\mathbf{x}}(\mathbf{s}) - \bar{\mathbf{x}}(\mathbf{s}')\| \leq \|\widetilde{\mathbf{x}}(\mathbf{s}) - \widetilde{\mathbf{x}}(\mathbf{s}')\| \leq \frac{L_f\,\|\mathbf{X}\|}{S}\,\|\mathbf{s} - \mathbf{s}'\|.$$

Now, note that the Pearson correlation is computed as

$$\operatorname{corr}\big(\widetilde{\mathbf{x}}(\mathbf{s}), \widetilde{\mathbf{y}}\big) = \frac{\bar{\mathbf{x}}(\mathbf{s})^\top \bar{\mathbf{y}}}{\|\bar{\mathbf{x}}(\mathbf{s})\|\,\|\bar{\mathbf{y}}\|}.$$

Since $\bar{\mathbf{y}}$ is independent of $\mathbf{s}$ and by the assumption that $\|\bar{\mathbf{x}}(\mathbf{s})\| \geq \epsilon > 0$, standard arguments (via the mean value theorem and the differentiability of the quotient function on a compact domain) imply that there exists a constant $C > 0$ such that

$$\left|\operatorname{corr}\big(\widetilde{\mathbf{x}}(\mathbf{s}), \widetilde{\mathbf{y}}\big) - \operatorname{corr}\big(\widetilde{\mathbf{x}}(\mathbf{s}'), \widetilde{\mathbf{y}}\big)\right| \leq C\,\|\bar{\mathbf{x}}(\mathbf{s}) - \bar{\mathbf{x}}(\mathbf{s}')\|.$$

Thus, combining the bounds yields

$$\left|\operatorname{corr}\big(\widetilde{\mathbf{x}}(\mathbf{s}), \widetilde{\mathbf{y}}\big) - \operatorname{corr}\big(\widetilde{\mathbf{x}}(\mathbf{s}'), \widetilde{\mathbf{y}}\big)\right| \leq C\,\frac{L_f\,\|\mathbf{X}\|}{S}\,\|\mathbf{s} - \mathbf{s}'\|.$$

Setting $L = C\frac{L_f\,\|\mathbf{X}\|}{S}$ completes the proof. □

B.4 Proof of Proposition 1: Non-Submodularity

Proposition 1 (Non-Submodularity; no diminishing returns (Informal))**.** *Let $\mathbf{s} \in \{0,1\}^d$ be a selection vector over d candidate OOD examples, and write corr($\boldsymbol{acc_{ID}}, acc^{\mathbf{s}}_{OOD}$) for the Pearson correlation in Eq.* (4)*. Define $\mathbf{s}_i \preceq \mathbf{s}_j$ to mean $(\mathbf{s}_i)_t \leq (\mathbf{s}_j)_t$ for all $t \in \{1,\ldots,d\}$. For $k \in \{1,\ldots,d\}$ with $(\mathbf{s})_k = 0$, let $\mathbf{s}^{+k}$ denote the vector obtained by setting the kth coordinate of $\mathbf{s}$ to* 1 *(leaving all others unchanged). Then, in general, there exist $\mathbf{s}_i, \mathbf{s}_j \in \{0,1\}^d$ with $\mathbf{s}_i \preceq \mathbf{s}_j$ and $(\mathbf{s}_i)_k = (\mathbf{s}_j)_k = 0$ such that*

$$corr\big(\boldsymbol{acc_{ID}}, acc^{\mathbf{s}_i^{+k}}_{OOD}\big) - corr\big(\boldsymbol{acc_{ID}}, acc^{\mathbf{s}_i}_{OOD}\big) < corr\big(\boldsymbol{acc_{ID}}, acc^{\mathbf{s}_j^{+k}}_{OOD}\big) - corr\big(\boldsymbol{acc_{ID}}, acc^{\mathbf{s}_j}_{OOD}\big).$$

Let $\mathcal{M}(\mathbf{s}) = \operatorname{corr}(\mathbf{x}_\mathbf{s}, y)$. The exists $\mathbf{s}_i \subseteq \mathbf{s}_j \subseteq \{1,\ldots,d\}$ and $j \neq \mathbf{s}_j$ such that

$$\mathcal{M}(\mathbf{s}_i \cup \{j\}) - \mathcal{M}(\mathbf{s}_i) < \mathcal{M}(\mathbf{s}_j \cup \{j\}) - \mathcal{M}(\mathbf{s}_j).$$

Proof. Let $y \in \mathbb{R}^n$ satisfy $\|y - \bar{y}\mathbf{1}\|_2 > 0$. Choose three candidate columns

$$\mathbf{x}_1 = y, \qquad \mathbf{x}_2 \text{ independent of } y \text{ with } \operatorname{corr}(\mathbf{x}_2, y) = 0, \qquad \mathbf{x}_3 = y.$$

Set the index sets

$$\mathbf{s}_i = \{1\}, \quad \mathbf{s}_j = \{1,2\}, \quad j = 3.$$

Compute the four correlations:

$$\mathcal{M}(\mathbf{s}_i) = \text{corr}(\mathbf{x}_1, y) = 1, \qquad \mathcal{M}(\mathbf{s}_i \cup \{3\}) = \text{corr}\Big(\tfrac{\mathbf{x}_1+\mathbf{x}_3}{2}, y\Big) = 1,$$

$$\mathcal{M}(\mathbf{s}_j) = \text{corr}\Big(\tfrac{\mathbf{x}_1+\mathbf{x}_2}{2}, y\Big) = \tfrac{1}{2}, \qquad \mathcal{M}(\mathbf{s}_j \cup \{3\}) = \text{corr}\Big(\tfrac{\mathbf{x}_1+\mathbf{x}_2+\mathbf{x}_3}{3}, y\Big) = \tfrac{2}{3}.$$

Hence

$$\underbrace{\mathcal{M}(\mathbf{s}_i \cup \{3\}) - \mathcal{M}(\mathbf{s}_i)}_{=0} \;<\; \underbrace{\mathcal{M}(\mathbf{s}_j \cup \{3\}) - \mathcal{M}(\mathbf{s}_j)}_{=\frac{1}{6}},$$

so $\mathcal{M}$ violates the diminishing-returns (submodularity) condition. □

C ID/OOD Explanation

Algorithm 2: Descriptive Differences Between ID and OOD Subsets (Dunlap et al., 2024)

Input: ID image set $\mathcal{D}_{\text{ID}}$, OOD image set $\mathcal{D}_{\text{OOD}}$
Output: Ranked list of difference descriptions highlighting OOD-specific attributes

Step 1: Generate Captions
Use BLIP-2 Li et al. (2023) to generate captions for all images in $\mathcal{D}_{\text{ID}}$ and $\mathcal{D}_{\text{OOD}}$
Step 2: Generate Candidate Difference Descriptions
Use Mixtral Jiang et al. (2024) to propose a set of natural language descriptions more likely to appear in $\mathcal{D}_{\text{OOD}}$ than in $\mathcal{D}_{\text{ID}}$, based on the generated captions
Step 3: Score and Rank Differences
foreach *difference description* d **do**
    Compute average CLIP Radford et al. (2021) similarity between d and images in $\mathcal{D}_{\text{ID}}$: $\text{sim}_{\text{ID}}(d)$
    Compute average CLIP similarity between d and images in $\mathcal{D}_{\text{OOD}}$: $\text{sim}_{\text{OOD}}(d)$
    Compute difference score: $\Delta(d) = \text{sim}_{\text{OOD}}(d) - \text{sim}_{\text{ID}}(d)$
Rank all descriptions by $\Delta(d)$ in descending order
return *Top-k difference descriptions ranked by distinctiveness to $\mathcal{D}_{OOD}$*

We follow Algorithm 2 to generate semantic difference between selected OOD samples and ID samples. We select 200 samples from the ID and selected OOD sets, respectively. As a motivation for future work, we enumerate our observations from this cursory study.

We evaluated multiple vision-language and language models to generate and summarize conceptual differences between In-Distribution (ID) and Out-of-Distribution (OOD) image groups.

Similarity Scoring Models. We tested CLIP, SigLIP, and AimV2 (CVPR 2025) to score image-caption similarity. AimV2 produced the most meaningful rankings based on manual inspection.

Prompting Strategy. Initial difference captions often described *groups* of images (e.g., "images with different views"), which misaligned with how CLIP-like models are trained (single image-caption pairs). To address this:

- We introduced a detailed prompt discouraging group-level descriptions. Larger models responded well, while smaller models were sensitive and inconsistent.
- We then simplified the prompt but edited examples to avoid phrases like "images of...", resulting in captions more compatible with similarity models.

LLM Comparison for Caption Generation. We compared three LLMs: `mistralai/Mistral-7B-Instruct-v0.2`, `Qwen/Qwen2.5-14B-Instruct`, and `Qwen/Qwen2.5-32B-Instruct`. The 32B model produced the most diverse and generalizable difference captions, while smaller models tended to overfit or focus narrowly on individual images.

Summarizing Conceptual Differences. Finally, we prompted LLMs to summarize conceptual shifts based on similarity deltas and difference captions. Larger models generated the most natural and high-level descriptions; `Qwen2.5-32B-Instruct` performed reasonably well.

Overall, while these results show some promise (Table 2) for settings with images with common semantic properties, they are relatively inconsistent. Moreover, for datasets without such common semantic properties, e.g., medical images, these methods may only work with dedicated foundation models appropriate for those image modalities.

Caption Prompt (Terra Incognita)

Caption this image. I know what object it is. Focus on describing the artistic style, texture, and domain-specific details rather than the object itself. Again, do not mention the object class.

Differences Prompt (TerraIncognita)

I am a machine learning researcher trying to figure out properties of an image beyond the image class. Give me a description of this image that is more specific than the image class. For instance, if this is an image of a bird, I don't want to know if the bird is a sparrow or a crow. I want to know if the bird is flying or sitting on a branch, if the camera is a drone or a ground-level camera, if the image is a macro shot or a close-up, or if the lighting is natural or artificial. A broader list of such properties is what I'm looking for. Give me this description for the image without focusing on the animal class.

Come up with `<number of captions>` distinct concepts that are more likely to be true for the Out-of-Distribution Group compared to the In-Distribution Group. Please write a list of captions (separated by bullet points `"*"`). For example:

- `"unusual lighting conditions"`
- `"visual distortions"`
- `"complex backgrounds"`
- `"non-standard object poses"`
- `"uncommon viewing angles"`
- `"partial views of objects"`
- `"objects in unexpected contexts"`
- `"scenes with high visual clutter"`
- `"images with unusual color schemes"`
- `"low-resolution images"`

Do not talk about the caption itself, e.g., `"caption with one word"`, and do not list more than one concept. The hypothesis should be a caption, so phrasing like `"more of ..."`, `"presence of ..."`, or `"images with ..."` is incorrect. Also, do not enumerate possibilities within parentheses. Here are examples of bad outputs and their corrections:

- INCORRECT: `"various nature environments like lakes, forests, and mountains"` CORRECTED: `"nature"`
- INCORRECT: `"images of household object (e.g. bowl, vacuum, lamp)"` CORRECTED: `"household objects"`
- INCORRECT: `"Presence of baby animals"` CORRECTED: `"baby animals"`
- INCORRECT: `"Different types of vehicles including cars, trucks, boats, and RVs"` CORRECTED: `"vehicles"`
- INCORRECT: `"Images involving interaction between humans and animals"` CORRECTED: `"interaction between humans and animals"`
- INCORRECT: `"More realistic images"` CORRECTED: `"realistic images"`
- INCORRECT: `"Insects (cockroach, dragonfly, grasshopper)"` CORRECTED: `"insects"`

Again, I want to figure out what kind of distribution shift there is. List `<number of captions>` properties that hold more often for the images (not captions) in the Out-of-Distribution Group compared to the In-Distribution Group. Answer with a list (separated by bullet points `"*"`).

In-Distribution Group: `<list of in-distribution captions>`

Out-of-Distribution Group: `<list of out-of-distribution captions>`

Your response:

Figure 10: ID/OOD difference prompt for TerraIncognita.

Differences Prompt (PACS)

Caption this image. I know what object it is. Focus on describing contextual and environmental details, such as scene composition, lighting, and background characteristics. Again, do not mention the object class.

Differences Prompt (PACS)

I am a machine learning researcher trying to figure out properties of an image beyond the object class. Give me a description of this image that is more specific than the object class. For instance, if this is an image of a dog, I don't want to know if the dog is a bulldog or a retriever. I want to know if the scene suggests an indoor or outdoor setting, details about the artistic style, or specific texture and lighting. A broader list of such properties is what I'm looking for. Give me this description for the image without mentioning the object class.

Come up with `<number of captions>` distinct concepts that are more likely to be true for the Out-of-Distribution Group compared to the In-Distribution Group. Please write a list of captions (separated by bullet points `"*"`). For example:

- `"unusual lighting conditions"`
- `"visual distortions"`
- `"complex backgrounds"`

- `"non-standard object poses"`
- `"uncommon viewing angles"`
- `"partial views of objects"`
- `"objects in unexpected contexts"`
- `"scenes with high visual clutter"`
- `"images with unusual color schemes"i`
- `"low-resolution images"`

Do not talk about the caption itself, e.g., `"caption with one word"`, and do not list more than one concept. The hypothesis should be a caption, so phrasing like `"more of ..."`, `"presence of ..."`, or `"images with ..."` is incorrect. Also, do not enumerate possibilities within parentheses. Here are examples of bad outputs and their corrections:

- INCORRECT: `"various nature environments like lakes, forests, and mountains"` CORRECTED: `"nature"`
- INCORRECT: `"images of household object (e.g. bowl, vacuum, lamp)"` CORRECTED: `"household objects"`
- INCORRECT: `"Presence of baby animals"` CORRECTED: `"baby animals"`
- INCORRECT: `"Different types of vehicles including cars, trucks, boats, and RVs"` CORRECTED: `"vehicles"`
- INCORRECT: `"Images involving interaction between humans and animals"` CORRECTED: `"interaction between humans and animals"`
- INCORRECT: `"More realistic images"` CORRECTED: `"realistic images"`
- INCORRECT: `"Insects (cockroach, dragonfly, grasshopper)"` CORRECTED: `"insects"`

Again, I want to figure out what kind of distribution shift there is. List `<number of captions>` properties that hold more often for the images (not captions) in the Out-of-Distribution Group compared to the In-Distribution Group. Answer with a list (separated by bullet points `"*"`).

In-Distribution Group: `<list of in-distribution captions>`

Out-of-Distribution Group: `<list of out-of-distribution captions>`

Your response:

Figure 11: ID/OOD difference prompt for PACS.

Caption Prompt (VLCS)

Caption this histopathology image. I know its diagnostic category. Focus on describing tissue morphology, staining patterns, and structural details. Again, do not reveal the diagnostic category.

Differences Prompt (VLCS)

I am a researcher studying domain adaptation. Please describe this image with a focus on properties beyond the object class. For example, if this is an image of a bird, I don't want to know whether it is a sparrow or an eagle. Instead, I want detailed information about the environmental context, scene composition, lighting conditions, and background. Provide such a description without mentioning the object class.

Come up with `<number of captions>` distinct concepts that are more likely to be true for the Out-of-Distribution Group compared to the In-Distribution Group. Please write a list of captions (separated by bullet points `"*"`). For example:

- `"unusual lighting conditions"`
- `"visual distortions"`
- `"complex backgrounds"`
- `"non-standard object poses"`

- `"uncommon viewing angles"`
- `"partial views of objects"`
- `"objects in unexpected contexts"`
- `"scenes with high visual clutter"`
- `"images with unusual color schemes"`
- `"low-resolution images"`

Do not talk about the caption itself, e.g., `"caption with one word"`, and do not list more than one concept. The hypothesis should be a caption, so phrasing like `"more of ..."`, `"presence of ..."`, or `"images with ..."` is incorrect. Also, do not enumerate possibilities within parentheses. Here are examples of bad outputs and their corrections:

- INCORRECT: `"various nature environments like lakes, forests, and mountains"` CORRECTED: `"nature"`

- INCORRECT: `"images of household object (e.g. bowl, vacuum, lamp)"` CORRECTED: `"household objects"`
- INCORRECT: `"Presence of baby animals"` CORRECTED: `"baby animals"`
- INCORRECT: `"Different types of vehicles including cars, trucks, boats, and RVs"` CORRECTED: `"vehicles"`
- INCORRECT: `"Images involving interaction between humans and animals"` CORRECTED: `"interaction between humans and animals"`
- INCORRECT: `"More realistic images"` CORRECTED: `"realistic images"`
- INCORRECT: `"Insects (cockroach, dragonfly, grasshopper)"` CORRECTED: `"insects"`

Again, I want to figure out what kind of distribution shift there is. List `<number of captions>` properties that hold more often for the images (not captions) in the Out-of-Distribution Group compared to the In-Distribution Group. Answer with a list (separated by bullet points `"*"`).

In-Distribution Group: `<list of in-distribution captions>`

Out-of-Distribution Group: `<list of out-of-distribution captions>`

Your response:

Figure 12: ID/OOD difference prompt for VLCS.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] .

   Justification: Each claim made in the abstract and introduction can be directly mapped to a matching empirical result and discussion in the subsequent sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes] .

   Justification: We have a paragraph discussion the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes] .

Justification: We include main results in the main text with supporting lemmas and proofs in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes] .

   Justification: We provide the code for our analysis; all data are open source. Our results can be reproduces from these public datasets and our code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code to generate data, run models, and perform analysis is included in an anonymized repository. We also include our identify subsets in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details on design choices, etc., are included in the main text and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Justification: [Yes]

   Guidelines: We include error bars for our analysis when appropriate and detail the significance tests used.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide this information in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes] .

    Justification: We discuss the implications of our work on subgroups who may be ignored in standard evaluation practices.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all open source or commercial code, models, and data used in this data.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets in this work are pointers to selecting examples from already opensourced data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not use any crowdsourcing or human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were

    Answer: [NA]

    Justification: We use no human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA] .

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.