DETECTING COVARIATE SHIFTS WITH VISION-LANGUAGE FOUNDATION MODELS

Alvin Heng[1] **& Harold Soh**[1,2]
[1]Department of Computer Science, National University of Singapore
[2]Smart Systems Institute, National University of Singapore
{alvinh, harold}@comp.nus.edu.sg

ABSTRACT

Deployed machine learning models often encounter significant challenges in-the-wild due to distribution shifts, where inputs deviate from the training distribution. Covariate shifts, a specific type of distribution shift, have traditionally been addressed with robustness-focused approaches; however, existing models still experience substantial performance degradation under such conditions. In this work, we propose reframing covariate shift detection as an out-of-distribution (OOD) detection problem. We leverage vision-language models (VLMs), in particular CLIP, for detecting covariate shifts using zero-shot detection techniques that require no task-specific training. To facilitate this effort, we introduce ImageNet-CS, a comprehensive benchmark comprising six covariate-shifted datasets derived from ImageNet. Our results demonstrate that VLMs outperform traditional supervised methods in detecting covariate shifts, underscoring their promise for improving the reliability of models deployed in the real world.

1 INTRODUCTION

A significant challenge in deploying machine learning systems in-the-wild is ensuring that models do not make erroneous predictions on out-of-distribution (OOD) inputs (Nguyen et al., 2015). These systems operate on the assumption that inputs during inference are consistent with its training distribution. However, real-world environments are inherently complex and unpredictable, often deviating from this assumption. As a result, it is imperative to equip machine learning models with methods to identify and manage OOD inputs.

There are two primary approaches to dealing with OOD inputs: 1) robustness, which seeks to improve generalization on inputs that are outside the training distribution (Taori et al., 2020), and 2) detection, which seeks to detect these inputs (Yang et al., 2021). Robustness is the common approach for covariate shifts, where the label space of inputs are within the model's predictions. Meanwhile, detection is adopted for semantic shifts, where the label space of inputs are outside the model's output space. This necessitates detection as it is impossible for the model to make a correct prediction.

In this work, we advocate for a different perspective towards covariate shifts. Despite significant progress in improving model robustness, machine learning systems still suffer notable performance degradation when exposed to distributions under covariate shifts. Consequently, detecting such inputs to mitigate erroneous predictions emerges as a logical and natural approach (Guille-Escuret et al., 2023). To facilitate such efforts, we introduce ImageNet-CS, a new benchmark specifically designed to capture the most widely recognized covariate shifts in the ImageNet dataset. Our study focuses on detection methods for covariate shifts, without extending to downstream applications such as Selective Classification (Geifman & El-Yaniv, 2017), which we leave as future work.

Much progress in OOD detection involve traditional supervised learning paradigms (Hendrycks & Gimpel, 2016; Lee et al., 2018) or generative modeling techniques (Serrà et al., 2019; Choi et al., 2018; Liu et al., 2023; Heng et al., 2025). The emergence of foundation models—large, versatile models capable of performing a wide range of tasks in a zero-shot manner—offers a promising avenue for OOD detection (Bommasani et al., 2021). We propose to investigate the use of modern vision-language models, particularly Contrastive Language-Image Pretraining (CLIP) (Radford et al.,

2021), for zero-shot covariate shift detection. Unlike traditional methods that require task-specific training or fine-tuning, we aim to utilize CLIP to identify covariate-shifted samples without additional task-specific adaptation. Our experiments show that CLIP outperforms traditional supervised learning methods without being trained on in-distribution data.

2 PRELIMINARIES

2.1 DISTRIBUTION SHIFTS

We aim to detect when a sample comes from a distribution different from the original distribution of interest. Such shifts fall into two main categories: *covariate shift* and *semantic (label) shift* (Yang et al., 2021). Let the input space be denoted as $\mathcal{X}$ and label space as $\mathcal{C}$. We consider a joint distribution over the input-label space $p(\mathbf{x}, \mathbf{c})$, where $\mathbf{x} \sim \mathcal{X}$ and $\mathbf{c} \sim \mathcal{C}$.

Covariate Shift. Covariate shifts refer to a change in the marginal distribution $p(\mathbf{x})$, affecting the input space $\mathcal{X}$, while the label space $\mathcal{C}$ remains constant. For example, a dataset of cat paintings represents a covariate shift relative to a dataset of real cat photographs.

Semantic Shift. Semantic shifts, in contrast, involve changes in *both* $p(\mathbf{c})$ and $p(\mathbf{x})$, as a difference in labels implies the introduction of new categories, which inherently results in a change in the input space. For instance, a dataset of human faces (e.g., CelebA) is semantically shifted from a dataset of natural objects and animals (e.g., CIFAR10).

2.2 OUT-OF-DISTRIBUTION DETECTION

Building on these types of distribution shifts, we now define the formal objective of OOD detection. Let p_{ID} denote a distribution of interest. Given p_{ID}, the goal of OOD detection is to construct a scoring function $S_\theta(\mathbf{x}) \in \mathbb{R}$ which quantifies how much a given test point $\mathbf{x}$ originates from p_{ID}. We adopt the convention that a higher value of $S_\theta(\mathbf{x})$ implies a sample is more likely to be from p_{ID}. We define a decision function $G_\lambda(\mathbf{x})$, parameterized by a threshold λ, as follows:

$$G_\lambda(\mathbf{x}) = \begin{cases} \text{ID} & \text{if } S_\theta(\mathbf{x}) \geq \lambda \\ \text{OOD} & \text{if } S_\theta(\mathbf{x}) < \lambda. \end{cases} \tag{1}$$

We evaluate OOD methods by computing metrics that integrate over all possible values of λ. The most common statistic is the Area Under the Receiver Operating Characteristic Curve (AUROC), which measures the trade-off between true positive and false positive rates across different thresholds.

3 MOTIVATION

3.1 WHY COVARIATE SHIFT DETECTION?

Under covariate shifts, machine learning models can still make correct predictions as the inputs fall within the model's label space. However, models exhibit performance degradation under such shifts, as these distributions are technically *out-of-distribution relative to the model's training distribution*. Prior research on covariate shifts has predominantly focused on *robustness* (Taori et al., 2020; Shi et al., 2023; Zhou et al., 2022), aiming to develop methods that enhance model performance under such shifts. We argue that this robustness-centered perspective is too narrow and advocate for a broader approach that also emphasizes the *detection* of covariate shifts.

3.2 IMAGENET COVARIATE SHIFT (IMAGENET-CS) BENCHMARK

To facilitate studies into covariate shifts, we introduce the ImageNet Covariate Shift (ImageNet-CS) benchmark, a comprehensive suite of datasets designed to evaluate models under covariate shifts. The benchmark is based on the ILSVRC 2012 dataset (commonly known as ImageNet1K). Due to its widespread use, numerous natural and synthetic variations have been developed to study the effects of covariate shifts. To streamline these efforts, we curated six key datasets into this unified benchmark.

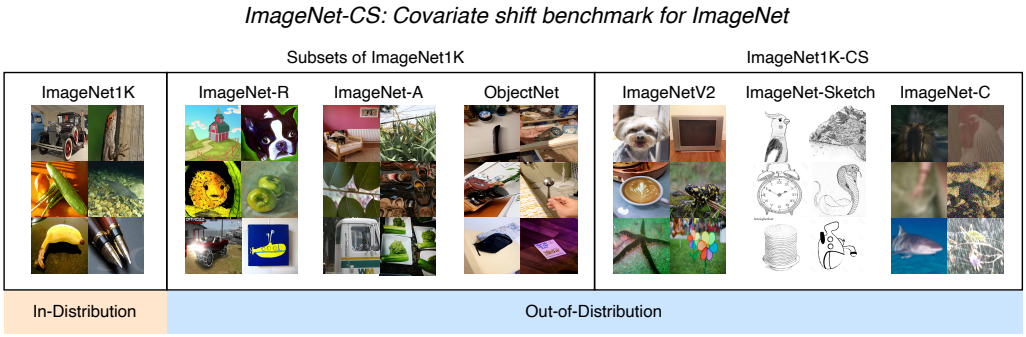

Figure 1: Overview of ImageNet-CS and the various datasets that are considered out-of-distribution. The in-distribution dataset is the ImageNet1K validation set.

ImageNet-CS includes six out-of-distribution (OOD) datasets, each capturing distinct types of covariate shifts: 1) ImageNet-R (Hendrycks et al., 2020), 2) ImageNet-A (Hendrycks et al., 2021), 3) ObjectNet (Barbu et al., 2019), 4) ImageNetV2 (Recht et al., 2019), 5) ImageNet-Sketch (Wang et al., 2019) and 6) ImageNet-C (Hendrycks & Dietterich, 2019). Descriptions of each dataset is provided in appendix Sec. A. ImageNet-R and ImageNet-A contain samples from different 200-class subsets of ImageNet1K, while ObjectNet contains samples from a 113-class subset. When evaluating against these datasets, we use the corresponding subset of ImageNet1K and abbreviate them as ImageNet200 and ImageNet113 respectively.

3.3 Robustness Towards Covariate Shifts Is Insufficient

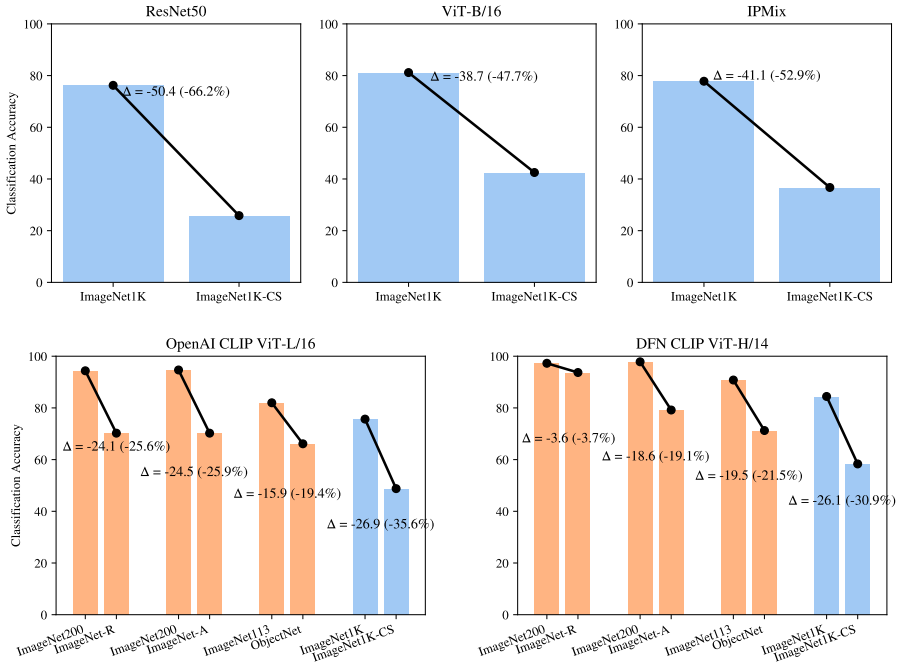

Figure 2: Top-1 classification performance of various architectures on ImageNet-CS datasets compared to original ImageNet. ImageNet1K-CS represents the average performance over ImageNetV2, ImageNet-Sketch and ImageNet-C. The top row consists of supervised models trained on ImageNet1K, while the bottom row are CLIP models of various sizes trained on different datasets: OpenAI's proprietary dataset (Radford et al., 2021) and DFN (Fang et al., 2023).

To justify the necessity for covariate shift detection, we plot the robustness performance of various classifier architectures on ImageNet-CS in Fig. 2. The architectures examined include supervised learning and CLIP models. In addition to ResNet50 and ViT-B/16 trained without dedicated robustness strategies, we also include results from IPMix (Huang et al., 2023) applied to ResNet50, a robust classifier trained with data augmentation techniques. We average the performance of ImageNetV2, ImageNet-Sketch and ImageNet-C and abbreviate it as ImageNet1K-CS as they share the same in-distribution set (ImageNet1K validation set). We evaluate the supervised models only on ImageNet1K-CS as these models are only trained to make predictions on all 1000 classes. In contrast, CLIP models, with their flexibility for zero-shot evaluation, are assessed on both ImageNet1K-CS and datasets with subset classes: ImageNet-R, ImageNet-A, and ObjectNet.

Focusing on ImageNet1K-CS, the most challenging benchmark due to its larger class set, we observe that all models experience significant performance degradation under covariate shifts. Even the most robust model, DFN CLIP, suffers a performance drop of over 30%, while the weakest model, ResNet50, experiences a decline of up to 66%. Although IPMix improves robustness over the standard ResNet50, it still exhibits a substantial performance drop exceeding 50%.In alignment with prior findings (Radford et al., 2021), CLIP models demonstrate greater robustness to covariate shifts than their supervised counterparts despite achieving comparable Top-1 accuracy on ImageNet1K.

Tasks involving covariate-shifted subsets are inherently simpler than ImageNet1K-CS as they require predictions over a reduced class set. This is evident in the overall higher Top-1 accuracies and smaller performance decreases. In the best case, DFN CLIP shows only a 3.7% drop from ImageNet200 to ImageNet-R. However, the overall performance decreases are still notable, averaging around 20%.

These results underscore a key insight: although CLIP models are trained on a vast web-scale corpus that likely includes diverse and covariate-shifted images (e.g., stylized paintings, low-quality images, or corrupted samples), they still experience notable performance degradation under covariate shifts. This underscores the need for better covariate shift detection alongside robustness improvements. As foundation models like CLIP gain prominence, exploring their potential for OOD detection becomes increasingly valuable. Here, we investigate VLMs specifically for covariate shift detection.

4 Related Works

The literature on covariate shifts largely focuses on the robustness of models under such shifts. Several studies have assessed the robustness of traditional supervised learning models. For example, Taori et al. (2020) demonstrated that robustness to synthetic distribution shifts does not translate to natural shifts. Schneider et al. (2020) proposed adjusting batch normalization statistics at test time, while Zhou et al. (2022) identified self-attention in Vision Transformers (ViTs) as critical for robustness. Yang et al. (2023) found that popular semantic shift datasets are contaminated with covariate shifts, showing that modern OOD methods are more sensitive to covariate shifts than semantic shifts. However, unlike our work, these studies do not address covariate shift detection.

In the context of VLMs, Radford et al. (2021) demonstrated that CLIP exhibits greater robustness to covariate shifts compared to supervised learning models. However, Shi et al. (2023) found that CLIP's robustness diminishes when evaluated across multiple ID test sets. Fang et al. (2022) attributed robustness gains in VLMs primarily to diverse training distributions. Crabbé et al. identified the presence of outlier features in CLIP models as an indicator of robustness to ImageNet shifts.

Another relevant line of work is Selective Classification (Geifman & El-Yaniv, 2017), which aims to improve model reliability by allowing classifiers to abstain from making predictions on uncertain samples, linking detection with classification. Prior research has primarily explored selective classification in the context of supervised learning (Geifman & El-Yaniv, 2017; Liang et al., 2017; Galil et al., 2023a) and semantic shifts (Galil et al., 2023b). While Liang et al. (2024) examines selective classification under broad distribution shifts, their focus remains on supervised learning methods and a limited set of covariate shifts. In contrast, our work takes a first step toward a more in-depth study of detection under covariate shifts using VLMs. We see extending this approach to selective classification as a logical and promising future direction.

To our knowledge, only two prior works directly explore covariate shift detection. Hsu et al. (2020) extended ODIN (Liang et al., 2017) and evaluated on the covariate shift dataset DomainNet (Peng et al., 2019). However, DomainNet is less comprehensive than ImageNet-CS proposed in our

work, and their method is tailored to supervised learning models. The most relevant prior work is BROAD (Guille-Escuret et al., 2023), which expands OOD detection to include covariate shifts. BROAD introduces a benchmark that combines semantic and covariate shifts. While BROAD evaluates existing OOD methods for supervised learning, our work diverges by focusing on covariate shifts and detection with VLMs. Moreover, our choice of covariate shift datasets is larger and more widely used than BROAD's counterparts, ensuring broader applicability and relevance.

5 METHODOLOGY

The primary objective of this study is to evaluate the performance of vision-language models (VLMs) in detecting covariate shifts. To accomplish this, we benchmark multiple CLIP models on ImageNet-CS, employing general post-hoc OOD detection techniques that are applicable to the VLM framework. We compare the results against supervised learning baselines, so as to analyze the potential advantages of VLMs over traditional supervised approaches. In this section, we provide an overview of the models and detection methods evaluated in our study.

5.1 MODELS

Supervised Learning Models (SLMs). We investigate two architectures: ResNet50 (He et al., 2016) and vision transformer ViT-B/16 (Dosovitskiy, 2020), both trained on ImageNet1K. The former has 25.6 million parameters and is trained using the default strategy specified in PyTorch[1], while the latter contains 86 million parameters and is trained using a modified version DeIT's (Touvron et al., 2021) training recipe.

Vision-Language Models (VLMs). We investigate three CLIP variants: OpenAI ViT-L/14 (Radford et al., 2021), DataComp ViT-L/14 (Gadre et al., 2024) and DFN ViT-H/14 (Fang et al., 2023). The ViT-L/14 models contain approximately 430 million parameters in total including the tokenizer and vision and text encoders, while the ViT-H/14 model contains approximately 1 billion parameters. OpenAI ViT-L/14 is trained using OpenAI's proprietary dataset comprising 400 million image-text. DataComp ViT-L/14 uses the open-source DataComp-1B dataset (Gadre et al., 2024), which comprises 1.4 billion image-text pairs, while DFN ViT-H/14 is trained on the DFN-5B dataset, which comprises 5 billion images filtered from a pool of 43 billion uncurated image-text pairs using Data Filtering Networks (DFNs) (Fang et al., 2023).

5.2 OOD DETECTION METHODS

To ensure our study is broadly applicable and generalizes to future models and data distributions, we consider post-hoc OOD detection methods that are agnostic to model architecture, does not require OOD-specific training and does not require knowledge of OOD data. We consider methods that are the most well-studied baselines under two broad categories: logit-based and distance-based. Under logit-based methods, we study Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2016), Maximum Concept Matching (Ming et al., 2022) and Energy (EBO) (Liu et al., 2020), while under distance-based methods we study Mahalanobis Distance (MDS) (Lee et al., 2018) and Relative Mahalanobis Distance (RMDS) (Ren et al., 2021). Please see appendix Sec. B for details on the methods and OOD scoring functions. We leave investigation of other methods to future work.

6 EXPERIMENTS

Table 1 compares VLMs and SLMs on ImageNet1K-CS. Since the SLMs are trained on all ImageNet classes, they can be directly evaluated on ImageNet1K-CS. We observe that for SLMs, logit-based methods (MSP/EBO) outperform distance-based methods (MDS/RMDS), whereas the opposite trend holds for VLMs. This finding is consistent with Table 2, where we present results on the full ImageNet-CS dataset. The key difference is that SLM logits are learned from ID data, while VLM logits are crafted at test time using user-defined textual concepts. Notably, in contrast to semantic shift

[1] https://github.com/pytorch/vision/tree/main/references/classification#resnet

Table 1: Average AUROC (↑) on ImageNet1K-CS. Column with MSP/MCM refers to MSP for supervised learning models and MCM for vision-language models.

	MSP/MCM	EBO	MDS	RMDS
Supervised Learning Models				
ResNet50	**0.830**	**0.866**	0.713	0.825
ViT-B/16	0.799	0.770	0.824	0.827
Vision-Language Models				
OpenAI	0.730	0.289	0.849	0.819
DataComp	0.738	0.250	0.807	0.806
DFN	0.725	0.323	**0.878**	**0.834**

Table 2: Average AUROC (↑) on ImageNet-CS.

	MCM	EBO	MDS	RMDS
Vision-Language Models				
OpenAI	0.709	0.336	0.852	0.825
DataComp	**0.731**	0.300	0.802	0.815
DFN	0.730	**0.358**	**0.895**	**0.845**

Table 3: AUROC scores of MCM (logit-based) and MDS (distance-based) with CLIP on datasets in ImageNet-CS.

	ImageNet-R		ImageNet-A		ObjectNet		ImageNetV2		ImageNet-Sketch		ImageNet-C (Avg)	
	MCM	MDS	MCM	MDS	MCM	MDS	MCM	MDS	MCM	MDS	MCM	MDS
OpenAI	0.647	0.924	0.733	0.843	0.621	0.810	0.550	0.584	0.601	0.895	0.808	0.903
DataComp	0.698	0.931	0.809	0.809	0.647	0.634	0.547	0.584	0.616	0.947	0.816	0.828
DFN	0.703	0.960	0.804	0.903	0.712	0.923	0.539	0.655	0.565	0.976	0.812	0.910

detection (Ming et al., 2022), leveraging the language modality appears less beneficial for detecting covariate shifts.

We hypothesize that the challenge lies in describing the in-distribution exhaustively with a finite set of prompts. Even if one carefully crafts an ID prompt such as "a real, high-quality, clear, and clean photo of a ID class", this may fail to capture subtle shifts. Indeed, Table 3 shows that for OOD samples with large perceptual deviations—like those in ImageNet-C—MCM remains reasonably effective, as severe corruptions are relatively easy to detect. However, MCM underperforms on more nuanced shifts like ImageNet-R, which contains diverse renditions (e.g., paintings and sculptures) that still match the nominal ID concepts in certain respects.

An additional observation is the effect of scaling. Among VLMs, DFN (with the largest parameter count) achieves the strongest MDS and RMDS results and also outperforms MSP/EBO in supervised models. Meanwhile, although DataComp and OpenAI VLMs share the same model size, DataComp slightly underperforms OpenAI on MDS/RMDS despite boasting higher ImageNet Top-1 accuracy (79.2% vs. 75.6%). In the supervised setting, ResNet50 surpasses ViT-B/16 under MSP and EBO, despite having fewer parameters and lower Top-1 accuracy. These trends may seem counterintuitive and warrant further investigation.

Finally, note that DFN's EBO scores outperform ResNet50's MCM by about 6% on ImageNet1K-CS (Table 1), although DFN has roughly 25 times more parameters. While this scaling might appear unfavorable, DFN CLIP was not trained on ID data and thus is adaptable to new inlier distributions. Consequently, VLM-based methods become especially attractive when amortized over many OOD detection tasks or when ID training data are scarce or ill-suited for standard supervised approaches.

7 Discussion

In this work, we reframed covariate shift as an out-of-distribution detection problem and introduced a new benchmark, ImageNet-CS, to evaluate performance under widely studied covariate shifts derived from ImageNet. These shifts have traditionally been analyzed in the context of robustness. Our findings reveal that many models exhibit significant performance degradation under such shifts, underscoring the need for effective detection mechanisms. We benchmark supervised learning and foundation models on covariate shift detection and find that foundation models outperform traditional approaches despite never being trained on in-distribution data.

Future work should explore more post-hoc OOD detection methods, as well as larger families of supervised learning and vision-language models. A deeper analysis into the ineffectiveness of language for covariate shift detection, such as at the representation level, could provide valuable insights. Lastly, combining detection with downstream tasks, particularly in the context of selective classification under covariate shifts, presents a promising direction for future research.

8 ACKNOWLEDGEMENTS

This research is supported by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-017).

REFERENCES

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Hyunsun Choi, Eric Jang, and Alexander A Alemi. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*, 2018.

Jonathan Crabbé, Pau Rodriguez, Vaishaal Shankar, Luca Zappella, and Arno Blaas. Interpreting clip: Insights on the robustness to imagenet distribution shifts. *Transactions on Machine Learning Research*.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pp. 6216–6234. PMLR, 2022.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers. *arXiv preprint arXiv:2302.11874*, 2023a.

Ido Galil, Mohammed Dabbah, and Ran El-Yaniv. A framework for benchmarking class-out-of-distribution detection and its application to imagenet. *arXiv preprint arXiv:2302.11893*, 2023b.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Charles Guille-Escuret, Pierre-André Noël, Ioannis Mitliagkas, David Vazquez, and Joao Monteiro. Expecting the unexpected: Towards broad out-of-distribution detection. *arXiv preprint arXiv:2308.11480*, 2023.

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*, pp. 770–778, 2016.

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

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. 2021 ieee. In *CVF International Conference on Computer Vision (ICCV)*, volume 2, 2020.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021.

Alvin Heng, Alexandre H Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. *Advances in Neural Information Processing Systems*, 37, 2025.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10951–10960, 2020.

Zhenglin Huang, Xiaoan Bao, Na Zhang, Qingqi Zhang, Xiao Tu, Biao Wu, and Xi Yang. Ipmix: Label-preserving data augmentation method for training robust classifiers. *Advances in Neural Information Processing Systems*, 36:63660–63673, 2023.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Hengyue Liang, Le Peng, and Ju Sun. Selective classification under distribution shifts. *arXiv preprint arXiv:2405.05160*, 2024.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *International Conference on Machine Learning*, pp. 22528–22538. PMLR, 2023.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1406–1415, 2019.

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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*, 2019.

Zhouxing Shi, Nicholas Carlini, Ananth Balashankar, Ludwig Schmidt, Cho-Jui Hsieh, Alex Beutel, and Yao Qin. Effective robustness against natural distribution shifts for models with different training data. *Advances in Neural Information Processing Systems*, 36:73543–73558, 2023.

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.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

William Yang, Byron Zhang, and Olga Russakovsky. Imagenet-ood: Deciphering modern out-of-distribution detection algorithms. *arXiv preprint arXiv:2310.01755*, 2023.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Animashree Anandkumar, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022.

Supplementary Material for "Detecting Covariate Shifts with Vision-Language Foundation Models"

A IMAGENET-CS DATASETS

Here we provide descriptions of the six OOD datasets that comprise ImageNet-CS.

1. ImageNet-R (Hendrycks et al., 2020): features artistic renditions (e.g., paintings, sketches, and cartoons) of a 200-class subset of ImageNet1K. To ensure consistency, we evaluate on a corresponding 200-class subset of the ImageNet1K validation set, referred to as ImageNet200.
2. ImageNet-A (Hendrycks et al., 2021): contains naturally adversarial examples—images that are frequently misclassified by ResNet50 models, despite not being generated through traditional adversarial attacks. Like ImageNet-R, it includes a 200-class subset of ImageNet1K, though the classes differ from those in ImageNet-R. The validation counterpart, also termed ImageNet200, is determined contextually to match the relevant subset.
3. ObjectNet (Barbu et al., 2019): a dataset characterized by randomized object orientations, backgrounds, and viewpoints. It covers a 113-class subset of ImageNet1K. For evaluation, we utilize a corresponding subset of the ImageNet1K validation set, termed ImageNet113.
4. ImageNetV2 (Recht et al., 2019): introduces three new test sets, sampled independently from ImageNet1K's original data. These sets cover all 1000 classes, with collection methods designed to mirror the original dataset's distribution. In this work, we use the "matched-frequency" subset.
5. ImageNet-Sketch (Wang et al., 2019): contains black-and-white sketch representations for all 1000 ImageNet1K classes.
6. ImageNet-C (Hendrycks & Dietterich, 2019): introduces systematic corruptions across four categories: blur, digital artifacts, noise, and weather effects. Each category comprises several corruption types (e.g., Gaussian noise, motion blur), with five levels of severity. We evaluate on the most challenging corruption level (severity 5) to simulate extreme robustness scenarios. Analysis of lower severity levels is left for future work.

B OOD DETECTION METHODS

Here we provide details of the OOD detection methods that are studied in our work.

Logit-based: MSP, MCM, EBO The scoring function in MSP (Hendrycks & Gimpel, 2016) is $S_\theta(\mathbf{x}) = \max_{\mathbf{c}} p(\mathbf{c}|\mathbf{x})$, where $p(\mathbf{c}|\mathbf{x})$ is the softmax probability for class $\mathbf{c}$ given input $\mathbf{x}$. MSP reflects the model's confidence in its most probable prediction, and is tailored to supervised learning models with defined logits. MCM (Ming et al., 2022) adapts MSP to VLMs by forming "virtual logits" over a predefined set of ID concepts given by $\text{sim}(E_{\mathbf{v}}(\mathbf{x}), E_{\mathbf{w}}(\mathbf{c}))$, where $\text{sim}(\cdot,\cdot)$ refers to the cosine similarity and E_v and E_w refer to the image and text encoders respectively. The scoring function of MCM is given by $S_\theta(\mathbf{c}) = \max_{\mathbf{c}} \text{softmax}\left(\text{sim}(E_{\mathbf{v}}(\mathbf{x}), E_{\mathbf{w}}(\mathbf{c}))/T\right)$, where T is temperature. Energy (Liu et al., 2020) (EBO) defines the score function as $S_\theta(\mathbf{x}) = T \log \sum_i e^{f_i(\mathbf{x})/T}$, where the sum is over all logits $f_i(\mathbf{x})$.

Unlike in semantic shift detection, we cannot form virtual logits using ID concepts based on class information alone (i.e., "a photo of a {ID class}") as OOD samples share the same classes as ID samples. Thus, as ImageNet contains real photographs that are of high-quality, we define the in-distribution concepts as "a real, high-quality, clear, and clean photo of a {ID class}", and investigate if these concepts can distinguish OOD samples that are not real images (e.g., ImageNet-R, ImageNet-Sketch) or are of low-quality (e.g., ImageNet-C).

Distance-based: MDS, RMDS MDS (Lee et al., 2018) computes the Mahalanobis distance $S_\theta(\mathbf{x}) = \max_{\mathbf{c}} -(f(\mathbf{x}) - \boldsymbol{\mu}_{\mathbf{c}})^\top \Sigma^{-1} (f(\mathbf{x}) - \boldsymbol{\mu}_{\mathbf{c}})$, where $f(\mathbf{x})$ is the feature representation of input $\mathbf{x}$ from the model's penultimate layer, $\boldsymbol{\mu}_{\mathbf{c}}$ is the the class-conditional mean of in-distribution features for

class $\mathbf{c}$ and Σ is the shared covariance matrix across all classes. These quantities are calculated from the in-distribution training set. RMDS (Ren et al., 2021) improves MDS by subtracting a correction term representing the Mahalanobis distance to the full training distribution without considering class information.