

A Multifaceted Analysis of Model Generalization in Challenging Environments

Weijie Tu

A thesis submitted for the degree of

Doctor of Philosophy

The Australian National University

School of Computing



Australian
National
University

August 2026

© Copyright by Weijie Tu, 2026

All Rights Reserved

Except where otherwise indicated, this thesis is my own original work.

Weijie Tu
31 August 2026

To family, friends, and the wonderful universe we are intimately connected with

Acknowledgments

First and foremost, I wish to express my sincere gratitude to my supervisors, Prof Tom Gedeon, Dr Sabrina Caldwell, and Dr Zakir Hossain, for their invaluable guidance, encouragement, and mentorship throughout my PhD journey. In particular, I extend my heartfelt thanks to Prof Tom Gedeon, whose trust and encouragement gave me the confidence to pursue a novel and innovative research direction. His support has been a constant source of motivation, and I consider myself extremely fortunate to have had the opportunity to work under his supervision. I would also like to express my sincere appreciation to Dr Weijian Deng and Dr Dylan Campbell for their substantial guidance, insightful discussions, and generous support, which played an important role in shaping this research. Although the journey was demanding, their collective guidance consistently illuminated the path forward.

I would like to express my sincere appreciation to my fellow researchers and colleagues at CSIT-ANU for cultivating a vibrant, collaborative, and intellectually stimulating research environment. I am especially grateful to Jiahao Zhang, Lianling Lai, Dr Yue Yao, Dr Xiaoxiao Sun, Dr Yunzhong Hou, Dr Yuchi Liu, Sasha Pak, and Iko-Ojo Simon for their collegiality, insightful discussions, and continuous support. Throughout this journey, we navigated numerous challenges together, and these shared experiences greatly strengthened both my research skills and resilience. I deeply appreciate your generosity, patience, and thoughtful feedback, all of which have played a significant role in my academic development and personal growth. I would also like to acknowledge my collaborators and mentors—Prof Liang Zheng, Prof Stephen Gould, Dr Yu Yao, Prof Tongliang Liu, and Dr Vijay Kumar—for their valuable insights, constructive guidance, and meaningful contributions that significantly shaped this research.

I am grateful for the generous support of the ANU Postgraduate Research Scholarship, which made my research endeavors possible.

Lastly, I would like to express my heartfelt appreciation to my family and friends for their unwavering love and support throughout this journey. Their encouragement and belief in me provided strength and reassurance during both challenging and rewarding times. My parents, Guangqi Tu and Guiying Luo, have been a constant source of care, understanding, and unconditional support, grounding me at every stage of this process. I am also deeply grateful to my grandparents, Hecai Luo and Longrong Zheng, whose enduring encouragement and quiet strength have continually inspired me. I also wish to thank my friends—Tingyi Gong, Tian Zhu, Ke Xu, Siying Qian, Xuejing Li, Mingtian Cai, Zepeng Zhang, Zhongxuan Jiao, Yafei Jin, Zhuo Huang and Shuai Yang—for their companionship, encouragement, and shared perseverance, which brought motivation, balance, and warmth to this journey.

Abstract

Deep learning models have become central to modern artificial intelligence systems, achieving strong performance across a wide range of visual recognition and reasoning tasks. However, deploying these models reliably in real-world environments remains challenging. In practice, models often encounter distribution shifts, produce unreliable confidence estimates, and must operate in settings where labeled data for validation or model selection are unavailable. These challenges are particularly pronounced for large vision–language and multimodal models, whose complexity and open-ended applications demand a deeper understanding of robustness, uncertainty, and model behavior beyond conventional accuracy-based evaluation. This thesis investigates how trustworthiness can be analyzed and supported in deep models, with a particular focus on vision–language systems operating under realistic conditions.

The thesis first studies the robustness of deep visual and vision–language models across diverse architectures, training distributions, and fine-tuning strategies in **Chapter 3**. Beyond overall accuracy, the analysis examines robustness with respect to specific visual factors, out-of-distribution conditions, and the interaction between vision and language encoders. It further incorporates safety-relevant aspects such as predictive uncertainty, out-of-distribution detection, and sensitivity to 3D perturbations, revealing systematic failure modes that arise under realistic testing scenarios.

Building on these findings, **Chapter 4** further delves into uncertainty estimation and calibration in vision–language models. The results show that strong zero-shot performance does not necessarily imply reliable confidence and that such models are not inherently well-calibrated. Nevertheless, simple post-hoc calibration methods can substantially improve uncertainty estimates, even under distribution shifts, across different label spaces, and with only limited calibration data. These results demonstrate that reliable uncertainty estimation of vision-language models can be achieved in practice without extensive labeled supervision.

The thesis then explores how model performance can be assessed when labeled evaluation data are unavailable. **Chapters 5** and **6** study the problem of ranking models without labels. **Chapter 5** shows that signals derived from softmax prediction probabilities can provide informative indicators of relative model performance on unlabeled data. **Chapter 6** extends this idea to large multimodal models and further investigates the role of uncertainty signals in model ranking. Finally, **Chapter 7** introduces a data-centric perspective through an unsupervised dataset representation that captures semantic structure without labels, supporting reasoning about dataset similarity, training set suitability, and test set difficulty.

Together, this thesis provides a unified perspective on trustworthy machine learning for deep learning models, advancing methods for understanding model generalization in challenging environments.

Contents

Acknowledgments	vii
Abstract	ix
1 Introduction	1
1.1 Model Reliability Analysis	2
1.2 Label-Free Performance Assessment	3
1.3 Thesis Outline	3
1.4 Publications	5
2 Literature Survey	7
2.1 Model Reliability Analysis	7
2.1.1 Robustness	7
2.1.2 Uncertainty Estimation	8
2.1.3 Out-of-Distribution (OOD) Detection	8
2.2 Label-Free Performance Assessment	8
2.2.1 Data-Centric Perspective	8
2.2.2 Model-Centric Perspective	9
3 Robustness Evaluation of Vision-language Models	11
3.1 Motivation	11
3.2 Experimental Setup	14
3.2.1 Models of Interest	14
3.2.2 Test Sets and Metrics	15
3.2.3 Analytical Methodology	16
3.3 Visual Factor-Level Robustness	16
3.3.1 CLIP Models Generally Exhibit Better Factor-Level Robustness Than Other Models	16
3.3.2 Texture Bias <i>v.s.</i> Shape Bias	18
3.4 Out-of-Distribution Detection	21
3.5 Confidence Calibration	22
3.5.1 Zero-Shot CLIPs Are Not Consistently More Calibrated Than Others	23
3.5.2 Temperature Scaling Highlights Well-Calibrated Properties of Zero-Shot CLIP Models	23
3.6 Zero-Shot Retrieval	25
3.7 3D Awareness	25

3.7.1	Correspondence Matching	26
3.7.2	Robustness against 3D corruptions	29
3.8	Visual and Language Encoder Interaction: A Classification Perspective .	30
3.9	Impact of Training and Inference Strategy on Robustness	32
3.9.1	Impact of Test-Time Prompts	32
3.9.2	Effect of Fine-Tuning Procedures	33
3.9.3	Extending to Different Training Paradigms	34
3.9.4	Robustness Evaluation of Dataset Curation	35
3.10	Summary	36
4	Uncertainty Estimation in Vision–Language Models	39
4.1	Motivation	39
4.2	Definition and Notation	41
4.3	Experimental Setup	41
4.4	Factors Affecting Calibration	42
4.4.1	Adaptability to Different Calibration Label Sets	43
4.4.2	Calibration Across Semantic Hierarchy Levels	44
4.4.3	Effect of the Calibration Set Size	46
4.4.4	Effect of Calibration-Target Set Distance	46
4.4.5	Transferability of Calibration Across Prompts	46
4.4.6	Impact of the Calibration Method	48
4.5	Application: Calibration-by-Synthesis	48
4.6	Summary	50
5	Label-Free Model Ranking based on Softmax Prediction Probability	53
5.1	Motivation	53
5.2	Task Formulation	54
5.3	Softmax Probability-based OOD Measures	55
5.3.1	What Makes OOD Measures Interesting?	55
5.3.2	Exploring More Empirical Measures	57
5.3.3	Ours: Softmax Correlation (SoftmaxCorr)	57
5.4	Experiments	58
5.4.1	Experimental Setup	59
5.4.2	Main Observations	61
5.4.3	Sensitivity Analysis on Test Set Size	61
5.4.4	Impacts of Class Distribution Estimator	62
5.4.5	Different Characterizations of Correlation Matrix	63
5.4.6	Effectiveness of SoftmaxCorr for Domain Adaptation	63
5.4.7	SoftmaxCorr Reflects a Model’s Accuracy	64
5.4.8	Evaluation Along Training Trajectory	64
5.5	Discussion and Potential Directions	65
5.6	Summary	65

6	Ranking Multimodal Models Without Labels via Model Uncertainty	67
6.1	Motivation	67
6.2	Task Formulation	70
6.3	Uncertainty for Ranking LMMs	70
6.3.1	What Makes Ranking LMMs Interesting?	70
6.3.2	Assessing Uncertainty by Softmax Probabilities	71
6.3.3	Assessing Uncertainty by Self-Consistency	73
6.3.4	Assessing Uncertainty With Labeled Proxy Data	73
6.4	Experiments	74
6.4.1	Experiment Setup	75
6.4.2	Key Findings	75
6.5	Analysis	77
6.6	Summary and Discussion	79
7	Unsupervised Model Assessment Using Dataset Representation	81
7.1	Motivation	81
7.2	Methodology	83
7.2.1	Bag-of-Words Model Across Communities	83
7.2.2	Bag-of-Prototypes Dataset Representation	84
7.2.3	Measuring Dataset-to-Dataset Similarity	85
7.2.4	Discussion	86
7.3	Comparing Training Suitability of Datasets	87
7.3.1	Experimental Settings	87
7.3.2	Evaluation	88
7.4	Assessing Test Set Difficulty without Labels	90
7.4.1	Experimental Settings	90
7.4.2	Evaluation	92
7.5	Summary	93
8	Conclusion and Future Directions	95
8.1	Contributions and Summary	95
8.1.1	A Unified Framework for Trustworthy Model Assessment	96
8.2	Limitations	97
8.3	Open Problems and Future Directions	98
8.3.1	Future Directions in Model Reliability Analysis	98
8.3.2	Future Directions in Label-Free Model Assessment	99
8.4	Final Remarks	100
A	Appendix: Models and Datasets	101
A.1	Models	101
A.1.1	Zero-shot CLIP	101
A.1.2	Fine-tuned CLIP on ImageNet	103
A.1.3	Other ImageNet Models	104
A.2	Datasets	112

Bibliography

115

List of Figures

1.1	Classical model evaluation <i>vs.</i> model evaluation in the real world. . . .	2
1.2	Overall structure of this thesis.	4
3.1	The models' accuracy on the subset of ImageNet-X annotated with a given visual factor (y-axis) to their overall accuracy on the whole ImageNet-X (x-axis).	17
3.2	Shape bias analysis.	19
3.3	OOD sample identification capability of models <i>vs.</i> ID dataset classification accuracy.	21
3.4	Model calibration performance with respect to classification accuracy .	24
3.5	Image/text zero-shot retrieval <i>vs.</i> classification accuracy.	26
3.6	Correspondence matching performance (Recall $\uparrow$) with respect to their viewpoint change.	27
3.7	Robustness comparison of ViT-based and CNN-based CLIP models under varying 3D-related corruptions.	28
3.8	Influence of test-time prompts on CLIP's visual-factor robustness. . . .	33
3.9	Influence of fine-tuning algorithms on CLIP's robustness, OOD detection, and predictive uncertainty	34
3.10	Visual-factor robustness of different training paradigms from CLIP . . .	35
4.1	Comparing calibration performance of ImageNet-trained models and VLMs.	43
4.2	Adaptability of VLMs to different calibration label sets.	44
4.3	Robustness of VLM calibration to label hierarchy levels.	45
4.4	Data-efficiency of VLM calibration across diverse datasets.	45
4.5	Impact of the distance between calibration and target set distributions on VLM uncertainty estimates.	47
4.6	Efficient calibration with single prompts for VLMs.	47
4.7	Uncertainty estimation performance of VLMs with spline calibration. .	48
4.8	Can VLMs be calibrated without labeled data? A calibration-by-synthesis approach.	49
5.1	Correlation study between MaxPred and accuracy (%) on ImageNet-S and ImageNet-R.	56
5.2	SoftmaxCorr <i>vs.</i> model generalization under ImageNet, CIFAR-10 and WILDS setups.	60

5.3	(a) Impacts of class distribution estimator and (b) SoftmaxCorr <i>v.s.</i> accuracy on ImageNet-C benchmark.	63
5.4	Correlation analysis: SoftmaxCorr and accuracy on CINIC.	64
6.1	Overview of unsupervised ranking for LMMs.	68
6.2	Correlation analysis of model performance across benchmarks.	69
6.3	An example of running one LMM for a VQA task.	72
6.4	Correlation analysis of Fréchet distances and model performance correlation across datasets.	78
7.1	Workflow of BoP representation computation.	83
7.2	Correlation study for training suitability of datasets.	87
7.3	Impact of codebook size on correlation strength for ResNet-101 on four test domains: ClipArt, Painting, Real and Sketch.	89
7.4	Correlation between train-test distance measured by BoP + JS and model accuracy.	90
7.5	Impact of codebook size on correlation strength on CIFAR-10-C.	91

List of Tables

3.1	Summary of evaluation dimensions and key findings.	13
3.2	Shape bias of various fine-tuned CLIP models.	19
3.3	The influence of input resolution on shape bias when fine-tuning CLIP. We also report accuracy on ImageNet-Val(idation) and Stylized ImageNet (SIN).	19
3.4	Comparison of CLIP and LLaVA models on ImageNet-D.	31
3.5	Influence of test-time prompts on CLIP’s classification, OOD detection and calibration.	33
3.6	Performance of various training paradigms on classification, OOD detection, calibration and 3D robustness.	35
3.7	Comparison of CLIP trained with filtered pre-training data on six tasks.	36
4.1	Uncertainty estimation performance on ImageNet-S for calibration-by-synthesis.	50
5.1	Method comparison on ImageNet, CIFAR-10, WILDS and DomainNet.	59
5.2	Comparison between SoftmaxCorr (Ours) and two variants: Certainty (C) and Diversity (D).	62
5.3	Sensitivity analysis of SoftmaxCorr on test set sizes.	62
5.4	Comparison between SoftmaxCorr and Disagreement.	62
6.1	Method comparison across nine multimodal tasks.	74
6.2	Method comparison for ranking different LLaVA-prismatic models on AI2D and TextVQA.	77
6.3	Labeling a subset of target domains to rank models on AI2D, MMMU, TextVQA, and ChartQA.	79
7.1	Comparing BoP with BoW model in natural language processing (NLP) and computer vision (CV).	85
7.2	Compare our methods with other baselines on DomainNet.	88
7.3	Method comparison in predicting classifier accuracy under CIFAR-10 setup.	92
A.1	Statistics of ImageNet-X of each factor.	113

Introduction

Trustworthy machine learning [Mucsányi et al., 2023; Han and Liu, 2025] aims to develop machine learning systems whose behavior remains reliable, robust, and secure when deployed in real-world environments, while providing transparent and well-calibrated predictions that support safe and responsible use. Beyond achieving high predictive performance on benchmark datasets, trustworthy models should produce meaningful estimates of uncertainty and behave predictably when encountering unfamiliar inputs. As deep learning models are increasingly adopted in real-world and risk-sensitive applications, establishing trustworthiness has become a central concern alongside conventional accuracy.

However, achieving such trustworthiness remains challenging in practice. Deep learning models are typically trained and evaluated under controlled assumptions that rarely hold in real-world environments. In particular, machine learning models are commonly evaluated by measuring their performance on a labeled test set that is collected separately from the training data. This evaluation protocol usually relies on the independent and identically distributed (i.i.d.) assumption, where the training and test data are assumed to be independent and drawn from the same underlying distribution. In real-world deployment, however, as illustrated in Fig. 1.1, this assumption often does not hold. Data distributions may differ from the training distribution due to changes in acquisition conditions, domain shifts, or evolving environments. Under such circumstances, evaluating model reliability becomes significantly more difficult, as models may exhibit brittle behavior, produce overconfident predictions, or fail in unexpected ways [Recht et al., 2019; Hendrycks et al., 2021a; Wang et al., 2019; Barbu et al., 2019].

Despite these challenges, model evaluation in practice still relies heavily on classification accuracy. Accuracy summarizes predictive performance on a labeled test set but provides limited insight into broader aspects of model reliability, such as prediction confidence, behavior under distribution shifts, or the ability to recognize unfamiliar inputs. Empirical studies [Miller et al., 2021; Teney et al., 2023; Taori et al., 2020] have shown that models can achieve high accuracy while still producing unreliable predictions when evaluated under different data conditions. Consequently, relying solely on accuracy may obscure important reliability issues and lead to an incomplete understanding of model behavior.

These limitations highlight the need for more comprehensive approaches to eval-

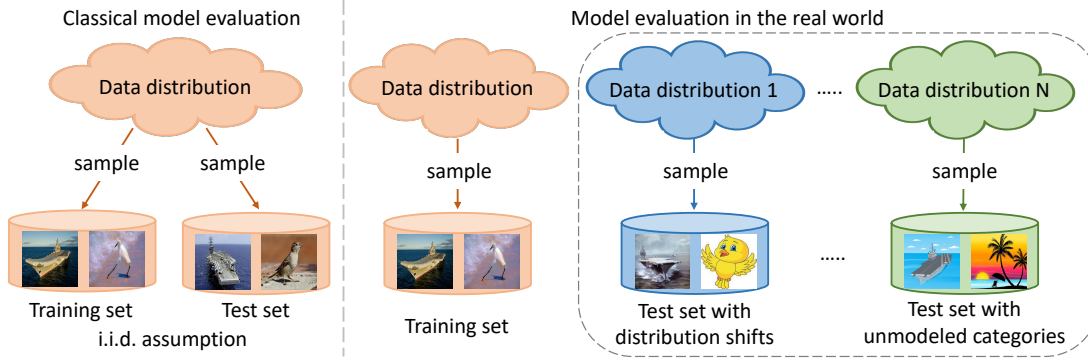


Figure 1.1: **Classical model evaluation vs. model evaluation in the real world.** (Left) Classical model evaluation: training and test sets are independent and identically distributed (i.i.d.). (Right) Unsupervised model evaluation: test sets are drawn from different distributions and may contain samples from unmodeled categories.

uating machine learning models. Beyond aggregate predictive accuracy, model assessment should consider multiple aspects of model behavior, including how models respond to variations in data distributions, semantic changes in inputs, and shifts in task settings. Examining such complementary signals provides deeper insight into model reliability and helps reveal potential failure modes that may not be apparent from overall accuracy alone.

In many real-world deployment scenarios, an additional challenge arises from the limited availability of ground-truth labels. Labels can be costly to obtain, delayed, or entirely absent, making conventional supervised evaluation less practical. Nevertheless, practitioners must still decide which models to deploy and how well they are likely to perform in new environments. This motivates the development of label-free approaches for assessing and comparing models.

In this thesis, we study trustworthy machine learning from perspectives of model reliability analysis and label-free performance assessment. The goal is to better understand the behavior of modern deep and multimodal models and to develop practical approaches for evaluating and selecting models in challenging environments.

1.1 Model Reliability Analysis

Model reliability refers to the degree to which a model’s predictions can be trusted across varying conditions beyond its training environment. A reliable model should not only perform well on in-distribution data but also maintain calibrated confidence estimates, degrade gracefully under distribution shifts, and appropriately handle inputs from unseen categories. As illustrated in Fig. 1.1, real-world deployment exposes models to precisely these challenges, where test data may originate from different distributions or contain categories absent from training.

Studying model reliability therefore requires moving beyond single-metric evaluation. In this thesis, we analyze reliability through multiple lenses: how model

confidence aligns with actual correctness, how performance varies across different distribution shifts, and how models respond to semantically novel inputs. By examining these dimensions jointly, we aim to build a richer understanding of where models succeed and where they remain fragile, ultimately informing better practices for deploying models in safety-critical and open-world settings.

1.2 Label-Free Performance Assessment

In real-world deployment, evaluating models using conventional protocols is difficult because ground-truth labels are often unavailable. While classical model evaluation relies on measuring accuracy on a labeled test set, obtaining such labels in practice can be expensive, time-consuming, or require expert knowledge. For example, annotation in domains such as medical imaging, scientific analysis, or safety-critical applications may require domain specialists and human effort. Moreover, as models are deployed at large scale and continuously encounter new data, it becomes impractical to annotate every dataset that arises during operation.

As a result, classical evaluation based on labeled test sets is often infeasible for assessing model performance in the wild. Nevertheless, practitioners still need to monitor models, compare models, and estimate how well a model is likely to perform in new environments. This motivates the development of label-free performance assessment, which aims to evaluate or compare models using only unlabeled data.

In this thesis, we also investigate methods for assessing and ranking model performance without access to ground-truth labels. By leveraging signals derived from model predictions or dataset characteristics, it becomes possible to infer model behavior and estimate generalization performance even when evaluation data are unlabeled. Such approaches enable more practical evaluation of machine learning systems operating in real-world environments.

1.3 Thesis Outline

This thesis comprises eight chapters, including this introduction. Chapter 2 reviews prior research on trustworthy machine learning for deep learning models. Chapter 3 presents a comprehensive evaluation of the robustness of vision–language models (VLMs), while Chapter 4 extends this analysis by identifying key factors that influence uncertainty estimation performance in VLMs. Building on these findings, Chapters 5 and 6 investigate label-free methods for ranking the generalization of conventional neural networks and modern multimodal models. Finally, Chapter 7 introduces a novel and effective approach to unsupervised model evaluation from a data-centric perspective. The overall structure of this thesis is illustrated in Fig. 1.2. A summary of the remaining chapters is provided below.

Chapter 2: Literature Survey. This chapter reviews related research on trustworthy machine learning for deep learning models and is organized into two parts. The first part surveys studies that analyze model behavior under distribution shifts,

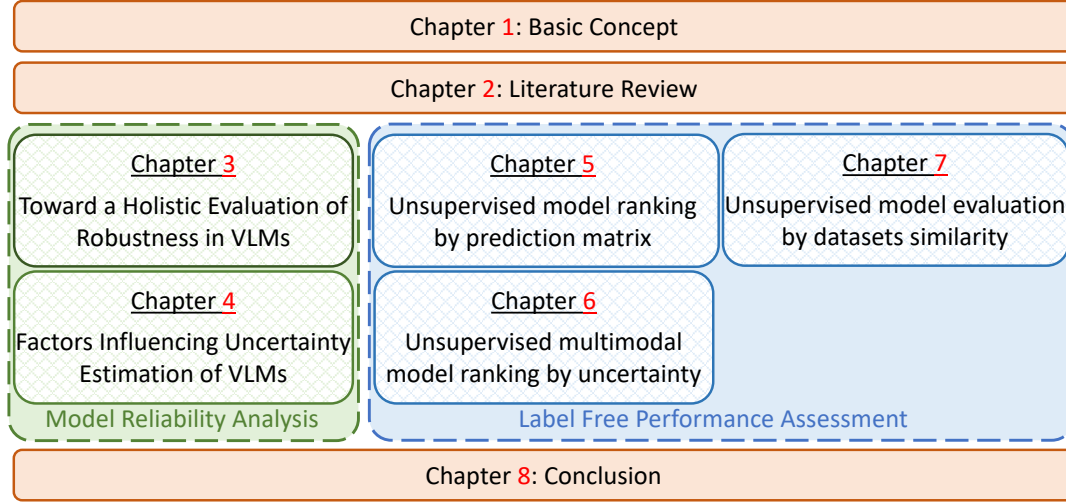


Figure 1.2: **Overall structure of this thesis.** In Chapters 3 and 4, we study the reliability and robustness of vision-language models. In Chapters 5, 6 and 7, we assess and rank model generalization without access to labels.

including work on out-of-distribution generalization, robustness evaluation, and failure mode analysis, as well as research on uncertainty estimation and out-of-distribution detection, which are central to understanding model reliability in realistic deployment settings. The second part discusses recent advances in label-free scenarios, covering emerging directions in model assessment and model selection when ground-truth annotations are unavailable.

Chapter 3: Toward a Holistic Evaluation of Robustness in Vision-language Models. This chapter provides a holistic examination of the robustness of contrastive language-image pre-training (CLIP) models under diverse distribution shifts. It analyzes robustness from multiple perspectives, including sensitivity to specific visual factors, uncertainty estimation and out-of-distribution detection, cross-modal alignment between vision and language, and three-dimensional awareness. The analysis systematically considers the effects of model architecture, training data, fine-tuning strategies, loss design, and test-time prompting, offering a comprehensive view of robustness in vision-language models.

Chapter 4: Uncertainty Estimation in Vision-Language Models. This chapter investigates the uncertainty estimation properties of vision-language models under distribution shifts. It examines how uncertainty estimation performance varies across model architectures, datasets, and training strategies, with particular attention to scenarios in which models are calibrated in one domain, label set, or hierarchy level and evaluated in another. The analysis shows that vision-language models are not inherently well-calibrated, but that simple post-hoc methods such as temperature scaling can consistently improve uncertainty estimation, even under substantial distribution and label shifts and with very limited calibration data.

Chapter 5: Label-Free Model Ranking based on Softmax Prediction Probability. This chapter investigates how uncertainty signals can be leveraged for model

selection in unlabeled environments. It shows that Softmax prediction probabilities on unlabeled test data are informative for estimating relative model performance. Beyond conventional confidence-based measures, the chapter introduces a generalization measure that jointly captures prediction confidence and diversity. Through extensive correlation studies across models and datasets with different types of distribution shift, the chapter demonstrates that the proposed measure enables accurate model ranking without access to ground-truth labels.

Chapter 6: Ranking Multimodal Models Without Labels via Model Uncertainty. This chapter extends label-free model ranking to large multimodal models. It examines uncertainty and probability structures at the token and sequence levels and analyzes how these signals relate to downstream task performance. Building on these observations, the chapter develops and evaluates methods for ranking multimodal models on unlabeled data, demonstrating that uncertainty-aware ranking remains effective in modern foundation model settings.

Chapter 7: Unsupervised Model Assessment Using Dataset Representation. This chapter adopts a data-centric perspective on label-free model assessment, focusing on two tasks: assessing the suitability of a training set for a target domain and evaluating the difficulty of a test set for a given model. To characterize relationships between datasets without relying on labels, the chapter introduces a bag-of-prototypes (BoP) representation that encodes datasets as distributions over semantic prototypes. By quantizing image features using a shared prototype codebook, the BoP representation provides an effective and interpretable summary of dataset semantics, enabling reliable dataset-to-dataset comparison when combined with appropriate divergence measures.

We conclude this thesis in Chapter 8 by summarizing the major contributions and discussing future research directions. Our source codes, models, and datasets have been made publicly available at <https://github.com/Klaus-Tu>.

1.4 Publications

Work described in this thesis has been previously published in conference proceedings and journals as listed below.

1. **Weijie Tu**, Weijian Deng, Tom Gedeon. *Toward a Holistic Evaluation of Robustness in CLIP Models*. Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2025 (Chapter 3)
2. **Weijie Tu**, Weijian Deng, Tom Gedeon. *A Closer Look at the Robustness of Contrastive Language-Image Pre-Training (CLIP)*. Advances in Neural Information Processing Systems (NeurIPS), 2023 (Chapter 3)
3. **Weijie Tu**, Weijian Deng, Dylan Campbell, Stephen Gould, Tom Gedeon. *An Empirical Study Into What Matters for Calibrating Vision-Language Models*. International Conference on Machine Learning (ICML), 2024 (Chapter 4)

4. **Weijie Tu**, Weijian Deng, Liang Zheng, Tom Gedeon. *What Does Softmax Probability Tell Us about Classifiers Ranking Across Diverse Test Conditions?* Transactions on Machine Learning Research (**TMLR**), 2024 (Chapter 5)
5. **Weijie Tu**, Weijian Deng, Dylan Campbell, Yu Yao, Jiyang Zheng, Tom Gedeon, Tongliang Liu. *Ranked from Within: Ranking Large Multimodal Models Without Labels*. International Conference on Machine Learning (**ICML**), 2025 (Chapter 6)
6. **Weijie Tu**, Weijian Deng, Tom Gedeon, Liang Zheng. *A Bag-of-Prototypes Representation for Dataset-Level Applications*. IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**), 2023 (Chapter 7)

Literature Survey

In this chapter, we review existing research on (i) model reliability analysis in the context of trustworthy machine learning and (ii) label-free performance assessment. For model reliability analysis, we survey studies that analyze deep learning models from multiple perspectives, including robustness, uncertainty estimation, and out-of-distribution detection. For label-free performance assessment, we review existing approaches from both data-centric and model-centric perspectives.

2.1 Model Reliability Analysis

2.1.1 Robustness

Machine learning models should generalize from the training distribution to novel testing environments [Djolonga et al., 2021; Koh et al., 2021; Kirsch and Gal, 2022]. One line of work has developed a theoretical framework to investigate model robustness [Ben-David et al., 2006a]. Ben-David et al. [2006a] were the first to propose a generalization bound based on the VC dimension, which quantifies the difference in classifier error between source and target distributions using a divergence measure. Mansour et al. [2009] later expanded this analysis to accommodate more general loss functions, offering improved generalization bounds through Rademacher complexity. To investigate the capability of deep models to various forms of test distributions, a commonly used approach is to introduce artificial transformations onto images, such as style transfer [Geirhos et al., 2019a], corruptions and perturbations [Hendrycks and Dietterich, 2019a; Mintun et al., 2021]. Moreover, many real-world datasets are introduced to assess model robustness under different natural distributional shifts [Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021a; Barbu et al., 2019; Hendrycks et al., 2021b]. For instance, Idrissi et al. [2022] propose ImageNet-X by relabelling the ImageNet validation set to provide detailed labels for naturally occurring factors such as pose, background, and lighting. Zhang et al. [2024a] introduce 3DCC to study the robustness of networks to 3D corruptions.

2.1.2 Uncertainty Estimation

This task aims to calibrate models so that their prediction probabilities align with the empirical frequency of correctness [Nguyen and O'Connor, 2015; Guo et al., 2017]. Much research effort has been devoted to proposing algorithms to improve model calibration performance, such as post-hoc rescaling the prediction probabilities [Guo et al., 2017], ensembling [Lakshminarayanan et al., 2017] and pre-training [Hendrycks et al., 2019]. Another line of research focuses on analyzing the calibration of modern neural networks [Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021; Tu et al., 2023a]. Guo et al. [2017] point out that modern neural networks are poorly calibrated. Ovadia et al. [2019] observe that distribution shifts degrade the performance of calibration methods. Minderer et al. [2021] show that zero-shot CLIP models are well-calibrated given their performance. Tu et al. [2023a] show that zero-shot CLIP models are well-calibrated with temperature scaling. This chapter builds on this prior research by studying a more comprehensive suite of factors that influence the uncertainty estimation performance of vision-language models.

2.1.3 Out-of-Distribution (OOD) Detection

This task targets identifying test data that do not belong to any of classes modeled in the training distribution [Hendrycks and Gimpel, 2016; Yang et al., 2024b; Liu et al., 2020]. A large number of methods are proposed for deep learning models, including generative model-based methods [Kirichenko et al., 2020; Nalisnick et al., 2018; Neal et al., 2018; Oza and Patel, 2019; Ren et al., 2019; Serrà et al., 2019; Xiao et al., 2020] and discriminative model-based methods [Hendrycks and Gimpel, 2016; Liu et al., 2020; Bendale and Boulton, 2016; Hsu et al., 2020; Liang et al., 2017; Lee et al., 2018]. For example, maximum softmax probability [Hendrycks and Gimpel, 2016] is used as the metric to detect OOD samples. Moreover, the above approaches mainly study OOD detection for a task-specific model using only visual information. In contrast, as CLIP models enjoy popularity, zero-shot OOD detection [Ming et al., 2022], is proposed, where the objective becomes filtering out input from tasks of disinterest.

2.2 Label-Free Performance Assessment

2.2.1 Data-Centric Perspective

Unsupervised model evaluation. The goal of unsupervised model evaluation is to estimate the generalization performance of a given model on unlabeled test sets. Since ground-truth labels are unavailable, existing approaches aim to infer model accuracy using signals derived from model predictions, dataset characteristics, or unsupervised objectives. (1) Utilizing model outputs. A widely studied direction is to exploit model outputs on distribution-shifted data to estimate test accuracy. Several works construct relationships between prediction statistics and model performance [Deng et al., 2023; Guillory et al., 2021; Garg et al., 2022a]. Some approaches rely on access to training data to calibrate these relationships [Chen et al., 2021;

Chuang et al., 2020], which limits their applicability in certain deployment settings. More recently, Deng et al. [2023] propose a training-free estimator based on the nuclear norm of the softmax probability matrix, demonstrating that prediction probabilities alone can provide informative signals for estimating model accuracy. (2) Considering distribution discrepancy. Another line of work studies the relationship between model performance and the discrepancy between training and test data distributions [Deng and Zheng, 2021a; Yu et al., 2022c; Tu et al., 2023b; Lu et al., 2023]. These methods typically estimate accuracy by measuring distribution distances between datasets. However, commonly used distribution metrics may not provide stable accuracy estimates under different types of distribution shifts [Guillory et al., 2021; Xie et al., 2023]. In addition, some approaches require access to the training data and can be computationally expensive for large-scale datasets [Deng and Zheng, 2021a]. (3) Constructing unsupervised objectives. A third direction introduces unsupervised evaluation criteria derived from data augmentation, self-supervision, or agreement among multiple classifiers [Jiang et al., 2022; Madani et al., 2004; Platanios et al., 2016, 2017]. While these methods can provide useful signals for performance estimation, they often require specialized model architectures or additional training procedures, which may limit their practicality in real-world applications.

2.2.2 Model-Centric Perspective

Predicting generalization in deep learning. This line of research studies the in-distribution (ID) generalization gap, *i.e.*, the difference between training and test accuracy of deep neural networks. Representative methods develop complexity measures based on model parameters and training data [Jiang et al., 2021, 2020c; Neyshabur et al., 2015; Keskar et al., 2017; Bartlett et al., 2017; Neyshabur et al., 2017b; Liang et al., 2019; Dziugaite et al., 2020; Dinh et al., 2017; Arora et al., 2018a; Smith and Le, 2017; Chuang et al., 2021; Dziugaite and Roy, 2017a]. Examples include the distance between trained weights and their initialization [Nagarajan and Kolter, 2019], the product of weight norms across layers [Neyshabur et al., 2017b], and the sensitivity of model accuracy to perturbations in the training data [Schiff et al., 2021]. However, these approaches typically assume that the training and test data follow the same distribution and rely primarily on information from the training process. As a result, they do not account for the characteristics of the test data and may not provide reliable predictions when models are evaluated under distribution shifts. To address this limitation, this thesis studies model generalization under distribution shifts and develops measures that better reflect model performance on out-of-distribution (OOD) datasets.

Robustness Evaluation of Vision-language Models

In this chapter, we conduct a comprehensive empirical study of the robustness and reliability of contrastive language–image pre-training (CLIP) models under diverse distribution shifts. We analyze CLIP models from multiple perspectives, including robustness to visual factor variations, uncertainty estimation, out-of-distribution detection, cross-modal alignment, and sensitivity to 3D perturbations. To understand the sources of these behaviors, we further examine several factors that may influence model robustness, including model architecture, training distribution, training set size, fine-tuning strategies, contrastive learning objectives, and test-time prompting. Through extensive experiments, we identify several systematic patterns in CLIP models, such as the influence of visual encoder architecture on robustness and the tendency of CLIP models to rely on shape-based cues. We also study the interaction between CLIP vision encoders and language models in modern multimodal systems, showing that models such as LLaVA can improve performance on challenging categories. Overall, this chapter provides insights into the robustness and reliability of CLIP models under distribution shifts.

3.1 Motivation

Leveraging contrastive training to cohesively align images and text within a singular embedding domain, contrastive language–image pretraining (CLIP) [Radford et al., 2021] excels in delivering versatile zero-shot generalizations. This inherent proficiency enables CLIP to handle diverse tasks without the need for task-specific fine-tuning [Radford et al., 2021; Jia et al., 2021]. Remarkably, CLIP models exhibit outstanding zero-shot classification capabilities, even without explicit training on the target dataset. Moreover, they demonstrate commendable robustness against challenging natural distributional shifts [Recht et al., 2019; Wang et al., 2019; Hendrycks et al., 2021a; Barbu et al., 2019; Hendrycks et al., 2021b]. Gaining a deeper understanding of such behaviors in CLIP models is crucial for advancing future image–text foundation models. Contemporary research has delved into multiple facets of CLIP models. This work encompasses areas such as dataset formulation [Nguyen et al., 2022], re-

producibility in scaling laws [Cherti et al., 2023], strategies for fine-tuning [Wortsman et al., 2022], adversarial classification robustness [Zhao et al., 2023] and nuances of the training distribution [Fang et al., 2022; Shi et al., 2023].

Motivated by previous work, we conduct an in-depth analysis of CLIP models, expanding our perspective beyond overall classification robustness. Our analysis includes several key dimensions: (1) robustness to visual factors, where we assess whether CLIP models can maintain performance when encountering variations such as pose, size, color, lighting, and occlusions; (2) out-of-distribution (OOD) detection, evaluating the models’ ability to identify instances with labels not present in the training distribution; (3) predictive uncertainty, examining whether CLIP models provide calibrated predictions that accurately reflect uncertainty under different testing conditions; (4) zero-shot retrieval, assessing the models’ capability to associate novel textual queries with relevant visual content; (5) 3D awareness, evaluating how well CLIP models handle 3D corruptions and maintain multi-view consistency; and (6) interaction between the vision and language encoders, investigating how these components influence classification robustness. Within each of these dimensions, we analyze the impact of several crucial factors on CLIP’s behavior, including variations in training distribution, model architectures, dataset sizes, contrastive loss, fine-tuning, test-time prompts, and dataset curation. This comprehensive analysis provides a thorough assessment of both the strengths and limitations of CLIP models across these critical areas.

To this end, we evaluate 84 zero-shot CLIP models with varying visual encoder architectures, training sources, and dataset sizes, as well as 44 ImageNet fine-tuned CLIP models. To establish a baseline, we compare these models against 127 ImageNet models without language-image pre-training. We examine 10 visual factors variations present in the ImageNet validation set [Idrissi et al., 2022], including object pose, lighting, and background, to assess models’ visual factors-level robustness. As for OOD detection, we employ ImageNet as an in-distribution (ID) set following [Ming et al., 2022] and test on 5 types of OOD scenarios. Then, to investigate predictive uncertainty, we use a set of canonical ImageNet distributions, such as texture, style, and perturbation shifts. We evaluate the effectiveness of data curation methods on the aforementioned datasets. Furthermore, we measure the 3D awareness of CLIP by geometric, semantic correspondence estimation as in [El Banani et al., 2024] and robustness against 3D-related corruptions, such as near focus and motion blur [Kar et al., 2022]. Lastly, to explore the interplay between the visual and text encoders of CLIP, we compare CLIP models with LLaVA [Liu et al., 2024c] in terms of classification performance on the challenging diffusion model-generated ImageNet-D [Zhang et al., 2024a].

We summarize our evaluation dimensions, covering the task, model, and data levels, along with key findings in Table 3.1.

Table 3.1: **Summary of evaluation dimensions and key findings.** The evaluation is conducted at three levels: task-level, model-level, and data-level, each examining different aspects of model behavior and robustness.

Evaluation Dimension	What We Evaluated	Key Findings
I. Task-Level Analysis		
Visual Factor Robustness (Sec. 3.3)	Performance under visual variations (e.g., shape, texture, and size)	CLIP models outperform ImageNet-trained models on six of ten factors. The training distribution significantly affects robustness to visual variations.
OOD Detection (Sec. 3.4)	Novelty detection capability (e.g., NINCO)	Zero-shot CLIP performs competitively with existing methods. The training distribution strongly influences detection accuracy.
Calibration (Sec. 3.5)	Alignment between prediction confidence and correctness	Fine-tuning increases calibration error. LP-FT and WiSE-FT recover calibration with temperature scaling, while FLYP remains overconfident. Data distribution and dataset size both influence calibration quality.
Retrieval (Sec. 3.6)	Image-text matching accuracy	Zero-shot retrieval performance of CLIP aligns with its classification accuracy. Data distribution and augmentation strategies influence retrieval quality.
3D Awareness (Sec. 3.7)	Robustness to 3D corruptions and correspondence matching	CNN-based CLIP models show greater stability than ViT-based CLIP models.
II. Model-Level Analysis		
Vision-Language Interaction (Sec. 3.8)	Comparison between CLIP and LLaVA on difficult category splits	LLaVA outperforms CLIP on ambiguous sets. Stronger language models (Vicuna > Mistral) further improve performance.
Training Paradigm Impact (Sec. 3.9.3)	Comparison of CLIP, BLIP, SigLIP, and ViTamin	No single paradigm consistently dominates. ViTamin is more robust to 3D corruptions, while SigLIP improves robustness but performs poorly on calibration.
Prompt (Sec. 3.9.1)	Sensitivity Impact of prompt set size and quality	LLM-generated prompts improve classification accuracy but do not consistently improve OOD detection or calibration.
Fine-Tuning (Sec. 3.9.2)	Impact Effects of different fine-tuning strategies	LP-FT and WiSE-FT achieve balanced performance. FLYP improves accuracy but harms calibration, while PromptSRC preserves both.
III. Data-Level Analysis		
Dataset Curation Effect (Sec. 3.9.4)	Impact of filtering and dataset diversity	Data curation improves classification, OOD detection, retrieval, and 3D robustness for ViT-CLIP models, but not calibration.

3.2 Experimental Setup

3.2.1 Models of Interest

Contrastive language-image pre-training models: we use **84 zero-shot CLIP models (CLIP)** and **44 ImageNet fine-tuned CLIP models (CLIP-FT)**. They have different visual encoders, including slightly modified ResNet [He et al. \[2016a\]](#), ConvNeXt [Liu et al., 2022](#), ViT [Dosovitskiy et al., 2021a](#) and EVA [Sun et al., 2023](#). There are various training sources, including LAION [Schuhmann et al., 2022](#), WIT [Radford et al., 2021](#) and Conceptual Captions [Sharma et al., 2018](#), and multiple sizes of training datasets from 3 million to 2 billion. In addition, some models are trained on subsets of CommonPool [Gadre et al., 2023](#), ranging from 14 million, 140 million to 1 billion. CommonPool draws its data from the same source as LAION, which is Common Crawl. These models allow us to validate and expand our findings in a medium-to-low accuracy regime. We also assess the performance of very recent CLIP models which are trained on filtered high-quality pre-training datasets using dataset curation techniques [Fang et al., 2024](#); [Xu et al., 2024](#). To compare the performance with LLaVA [Liu et al., 2024c](#), we also include SigLIP [Zhai et al., 2023](#).

For the CLIP-FT models, the vision encoder of CLIP is fine-tuned on ImageNet-1K. We consider different fine-tuning approaches, including directly fine-tuned on ImageNet-1K [Deng et al., 2009](#), first fine-tuned on ImageNet 12K, a subset of ImageNet 22K before fine-tuning on ImageNet-1K, and also fine-tuned by parameter-efficient fine-tuning methods [Zhou et al., 2022b](#); [Zhang et al., 2022b](#). We use the default prompt from [Radford et al. \[2021\]](#) for zero-shot CLIPs unless specified.

Models compared: we use 127 ImageNet models with various architectures, including Convolutional Neural Networks (e.g., ResNet [He et al., 2016a](#) and ConvNeXt [Liu et al., 2022](#)), Vision Transformers (e.g., ViT [Dosovitskiy et al. \[2021a\]](#) and Swin [Liu et al., 2021](#)) and all-MLP architectures [Ding et al., 2022](#); [Tolstikhin et al., 2021](#)) (e.g., MLP-Mixer [Tolstikhin et al., 2021](#)). Following [Taori et al. \[2020\]](#), we divide them into three categories: (i) **Standard Models**. This group consists of models supervised on the ImageNet training set. (ii) **Contrastive learning models**. This category contains 8 models pre-trained by contrastive learning. There are 6 training algorithms investigated, including InsDis [Wu et al., 2018](#), MoCo [He et al., 2020](#), SimCLR [Chen et al., 2020a](#); (iii) **Pre-trained on more data**. This group contains models pre-trained on a significantly larger dataset (e.g., ImageNet-21K) than the ImageNet training set. All the above models, including CLIP, are publicly available on TIMM [Wightman, 2019](#), OpenCLIP [Ilharco et al., 2021](#).

Modern vision language models: This chapter considers LLaVA [Liu et al., 2024c](#), which combines a frozen CLIP vision encoder and a large language model (e.g., Vicuna) for general-purpose visual and language understanding. In our study, we consider six LLaVA models: the visual encoders used are CLIP-L/14-336 and SigLIP, paired with three large language models: Mistral-instruct-V2 [Jiang et al., 2023](#), Llama-Chat [Touvron et al., 2023](#), and Vicuna-V2-7B [Touvron et al., 2023](#), resulting in a total of six LLaVA models. These models are available on HuggingFace, as

provided by [Karamcheti et al. \[2024\]](#).

3.2.2 Test Sets and Metrics

I. Robustness. We first pinpoint model failure patterns by testing on ImageNet-X [\[Idrissi et al., 2022\]](#), which is a relabeling of ImageNet validation by 16 naturally occurring factors. To ensure sufficiently reliable factor-level evaluation, we focus on the 10 factors with an adequate number of test samples and exclude the remaining factors whose sample sizes are too small: *Pose, Background, Pattern, Color, Smaller, Shape, Partial View, Subcategory, Texture* and *Larger*. The metric is accuracy, and high is better. In addition, we include cue-conflict stimuli and Stylized-ImageNet [\[Geirhos et al., 2019b\]](#) to measure the model bias towards the shape or texture.

II. OOD detection. We use a large-scale OOD detection benchmark which is built up on ImageNet: in-distribution (ID) ImageNet *v.s.* {iNaturalist [\[Van Horn et al., 2018\]](#), SUN [\[Xiao et al., 2010\]](#), PLACES [\[Zhou et al., 2017\]](#), TEXTURE [\[Cimpoi et al., 2014\]](#), and ImageNet-O [\[Hendrycks et al., 2021b\]](#) (OOD). Metrics are the area under the receiver operating characteristic curve (AUROC) and the higher is better; false positive rate (FPR@95) when the true positive rate is at 95% and a lower score is better. To evaluate OOD detection across diverse conditions, we employ the NINCO dataset [\[Bitterwolf et al., 2023\]](#), which is ID-contamination-free and comprises OOD classes from various existing OOD datasets. We report mean AUROC and FPR@95.

III. Calibration. We study ID and OOD datasets, where ImageNet validation is ID dataset and OOD datasets are: ImageNet-V2 [\[Recht et al., 2019\]](#), ImageNet-Rendition [\[Hendrycks et al., 2021a\]](#), ImageNet-Adversarial [\[Hendrycks et al., 2021b\]](#), ImageNet-Sketch [\[Wang et al., 2019\]](#), ObjectNet [\[Barbu et al., 2019\]](#) and ImageNet-Vid-Robust [\[Shankar et al., 2021\]](#). Metrics include the expected calibration error (ECE) [\[Naeini et al., 2015\]](#) and the negative log-likelihood (NLL). Lower ECE or NLL values indicate better calibration.

IV. Retrieval. We evaluate zero-shot retrieval performance on Flickr30K [\[Young et al., 2014\]](#) and MSCOCO [\[Chen et al., 2015\]](#) following the evaluation setup and splits from [\[Karpathy and Fei-Fei, 2015\]](#). As in [Radford et al. \[2021\]](#), we compute the cosine similarity between image and text embeddings as the image-text scores. When evaluating image retrieval, we rank the top- K images for each text caption, and vice versa for text retrieval. Recall@ K is the metric with $K = 5$.

V. 3D Awareness. Two tasks are explored for this property: correspondence estimation and robustness to 3D corruptions. We use ScanNet [\[Dai et al., 2017\]](#), NAVI [\[Jampani et al., 2023\]](#) and SPair-71K [\[Min et al., 2019\]](#) as the evaluation datasets for correspondence estimation. The metric is recall. For robustness against 3D corruptions, we use 3DCC [\[Zhang et al., 2024a\]](#), which applies 3D-related corruptions against ImageNet-validation with 5 severity levels. The metric is accuracy.

VI. Comparison to LLaVA. We compare the performance of CLIP and LLaVA on ImageNet-D [\[Zhang et al., 2024a\]](#), which consists of three splits, *Background, Texture*

and *Material*. CLIP is evaluated using standard zero-shot image classification protocol, while LLaVA is assessed by standard visual question answering protocol. They are both required to classify images from four classes and use accuracy as the metric.

3.2.3 Analytical Methodology

Key Factors. To understand the underlying factors that influence the performance of CLIP models, we delve into six primary aspects: 1) training distribution, evaluating the effect of data source; 2) model architecture, looking into the potential effects of different structural choices on model performance; 3) dataset quantity, probing the interplay between the amount of data available for training and the model’s efficiency; 4) contrastive loss, understanding its specific role in training dynamics 5) fine-tuning, 6) test-time prompt, assessing the impact of prompts during the evaluation on model outputs.

We follow the analytical methodology of seminal work [Taori et al., 2020], along with subsequent studies such as [Nguyen et al., 2022; Fang et al., 2022; Miller et al., 2021], to study the influential factor. Within the performance trends observed across all models, any factor causing a deviation from these trends is influential. Notably, in our research, we mainly emphasize and discuss such influential factors within each facet of our investigation. In Table 3.1, we organize our evaluation into task-level, model-level, and data-level dimensions, highlighting the main insights.

3.3 Visual Factor-Level Robustness

Our research builds upon previous findings on the robustness of CLIP models and focuses on the potential failure types of the model. Instead of solely measuring overall accuracy, this section investigates the behavior of CLIP models when faced with varying visual factors such as *Pose*, *Background*, and *Object Scale*.

3.3.1 CLIP Models Generally Exhibit Better Factor-Level Robustness Than Other Models

Factor-level effective robustness. In our study, we introduce the concept of visual factor-level effective robustness based on effective robustness [Taori et al., 2020]. It measures a model’s ability to achieve higher accuracy on the subset annotated by a specific visual factor compared to what is expected based on its overall accuracy on ImageNet-X. Fig. 3.1 displays the accuracy on the subset annotated by a specific visual factor relative to the overall accuracy on ImageNet-X.

(1) CLIP models are generally more robust than other ImageNet models on six out of ten visual factors. Fig. 3.1 highlights several insights into the factor-level robustness of CLIP models. First, we find that CLIP models are more robust than other models on six out of ten visual factors, including *Subcategory*, *Smaller*, *Color*,

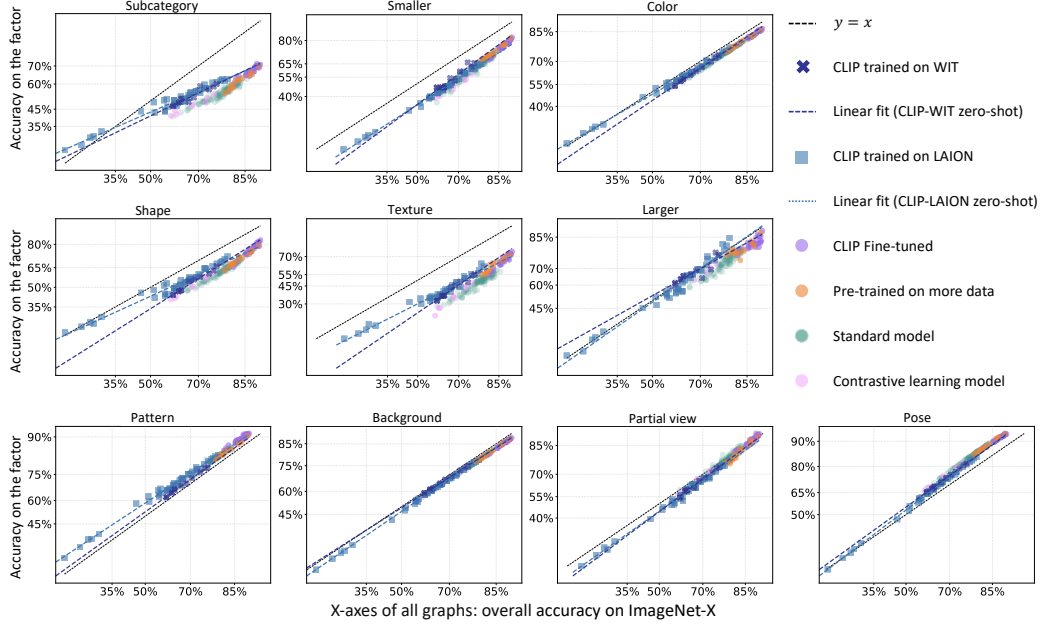


Figure 3.1: **The models’ accuracy on the subset of ImageNet-X annotated with a given visual factor (y-axis) to their overall accuracy on the whole ImageNet-X (x-axis).** Each point represents a model. The x-axis and y-axis are probit transformed following Taori et al. [2020]. The black dashed line represents the ideal robust models whose performance on each visual factor is the same as the overall performance. The blue straight lines are fit with robust linear regression [Huber, 2011]. We include models supervised on ImageNet-1K, pre-trained on more data, contrastive learning models, CLIPs trained on two data distributions, and their fine-tuned counterparts.

Shape, *Texture*, and *Larger*. Specifically, CLIP models exhibit higher factor-level effective robustness than other models on each of these factors. Second, we observe that CLIP models are less robust than other models on *Pose* and *Partial View*. Third, CLIP models show a similar trend to other models on the *Background* factor.

(2) Training distributions lead to different trends in CLIP models. The choice of training distribution impacts the factor-level robustness of CLIP models. Specifically, we find that training on different datasets (*i.e.*, LAION and WIT) forms distinct trends on each visual factor for CLIP, and there is no single training source that always leads to higher factor-level robustness than another. For instance, we observe that CLIP models trained on LAION demonstrate higher robustness on *Shape* factor than those trained on WIT, while this reverses for *Background* and *Pose* factors. The results show a mixed observation on *Large* factor. Furthermore, we further point out that CLIP models trained on different subsets of LAION (LAION-80M, LAION-400M, and LAION-2B) follow the same trend. The above observations highlight the importance of the choice of training source in determining not only the overall accuracy but also the factor-level behaviors of CLIP models. This suggests that factor-level robustness should be considered when choosing the training source.

(3) CLIP fine-tuned models perform slightly better than models pre-trained with more data. We compare CLIP fine-tuned models (CLIP-FT) with other models pre-trained on more data and find that CLIP-FT shows improvement in overall accuracy and robustness on visual factors of *Subcategory*, *Shape*, and *Pattern*. However, no additional robustness gain is observed on other factors. Moreover, CLIP-FT models outperform zero-shot CLIP on variations such as *Pattern* and *Partial View* but perform lower on factors like *Texture* and *Larger*. We speculate that standard fine-tuning introduces spurious correlations [Xiao et al., 2023]. This may lead to a bias for CLIP towards specific visual properties, thereby compromising factor-level robustness on some factors. It would be intriguing to explore fine-tuning techniques to maintain or improve the visual factor-level robustness of CLIP.

(4) Discussion on consistent trends across visual factors. All models exhibit consistent trends across visual factors, despite differences in architecture and training data. Specifically, all models lie below line $y = x$ under *Smaller*, *Shape*, and *Texture* conditions, which involve changes to object geometry, scale, and surface patterns. While such variations do occur in natural datasets, they are neither explicitly annotated nor emphasized, and thus may be underrepresented in the models’ learned feature space. As a result, models tend to rely on statistically dominant but fragile cues, such as canonical shapes, or common textures, rather than learning representations that are robust to these factors. This behavior is consistent with the concept of shortcut learning [Geirhos et al., 2020], where models exploit superficial but predictive patterns that fail under distribution shift. In contrast, performance on *Background* and *Partial View* remains stable, likely due to the abundance of such variations in pretraining data, which encourages models to downweight context and develop object-centric representations. The consistency across models suggests these are not architecture-specific artifacts but shared limitations shaped by training data and objectives.

(5) Pre-training analysis of factor-level coverage for training data selection. Given a candidate training set, we aim to estimate its coverage across key visual factors to anticipate potential robustness gaps before model training. ImageNet-X provides annotations across 16 visual factors, which supports pre-training analysis. Specifically, we compute a prototypical feature for each factor by averaging features from its annotated images using a fixed pretrained backbone (e.g., ResNet-50). For each image in the candidate dataset, we extract its feature and assign it to the nearest factor prototype. By counting how many training data are associated with each factor, we estimate the dataset’s factor-level coverage. This lightweight and scalable analysis enables factor-aware data selection before training models.

3.3.2 Texture Bias v.s. Shape Bias

CLIP exhibits a shape bias. We conducted experiments using the cue-conflict stimuli dataset [Geirhos et al., 2019b] to assess the presence of shape bias in the model’s predictions. Shape bias, in this context, refers to the proportion of correct predictions that are based on the object’s shape rather than texture or other features. Fig. 3.2 visualizes the shape bias exhibited by different models, grouped by training meth-

Backbone	FT methods	Shape bias
ViT-B/32	Zero shot	0.575
	Fine-tune on 1K	0.401
	Contrastive FT	0.561
	CoOp	0.549
	Tip-Adapter	0.579
ViT-B/16	Zero shot	0.473
	Fine-tune on 1K	0.345
	Contrastive FT	0.448
	CoOp	0.472
	Tip-Adapter	0.487

Table 3.2: **Shape bias of various fine-tuned CLIP models.** We include CLIP models fine-tuned using different methods: cross-entropy, contrastive loss [Goyal et al., 2023], and parameter-efficient techniques such as CoOp [Zhou et al., 2022b] and Tip-Adapter [Zhang et al., 2022b].

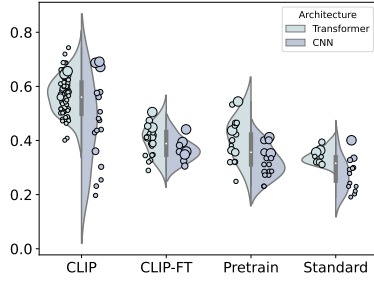


Figure 3.2: **Shape bias analysis** of CLIP, CLIP fine-tuned (CLIP-FT), models pre-trained on more data (Pretrain), and standard models. A larger point denotes a larger model within the group. We observe that CLIP models are more shape-biased.

Source	Backbone	Shape bias	IN-Val	SIN
LAION	ViT/H-14 (336/224)	0.42 / 0.51	0.89 / 0.88	0.28 / 0.32
	ViT/L-14 (336/224)	0.41 / 0.47	0.88 / 0.88	0.27 / 0.31
	ViT/B-16 (384/224)	0.35 / 0.43	0.87 / 0.86	0.23 / 0.25
	ViT/B-32 (384/224)	0.33 / 0.45	0.85 / 0.83	0.21 / 0.22
	ConvNeXt-B (384/224)	0.31 / 0.38	0.87 / 0.86	0.17 / 0.21
WIT	ViT/L-14 (336/224)	0.39 / 0.45	0.88 / 0.88	0.24 / 0.30
	ViT/B-16 (384/224)	0.35 / 0.41	0.87 / 0.86	0.22 / 0.23

Table 3.3: **The influence of input resolution on shape bias when fine-tuning CLIP.** We also report accuracy on ImageNet-Val(uation) and Stylized ImageNet (SIN). The higher value in a model pair is in bold. With the same backbone architecture, the CLIP model fine-tuned with a larger input resolution is more accurate on IN-Val but less shape-biased and less accurate on SIN.

ods (zero-shot, CLIP fine-tuning, additional data pre-training, and standard training) and architecture (transformer versus CNN). Our results show that, among the four training methods, CLIP models exhibit a stronger shape bias compared to the other groups. While previous research has indicated that transformers show a greater shape bias than CNNs [Naseer et al., 2021; Zhang et al., 2022a], we found that CLIP models with CNN-based vision encoders also exhibit a significant shape bias. This suggests that CLIP can align more closely with human visual perception, which is

widely acknowledged to be shape-driven [Geirhos et al., 2019b; Hermann et al., 2020; Li et al., 2024c]. In the following, we provide a more detailed analysis of the shape bias observed in CLIP models and explore the implications of these findings.

(1) Model size does not solely explain the shape bias of CLIP. We further observe that larger CLIP models do not necessarily have higher shape bias than smaller-size ones. For example, two models both trained on LAION-80M, CLIP-ViT/L-14 have 0.54 shape bias, which is 0.09 lower than CLIP-ViT/B-32. This implies that the shape bias of CLIP models cannot be attributed solely to model size. Based on the above, we speculate that the shape bias of CLIP may be attributed to its objective, which involves training the model to associate text and image pairs.

(2) Larger input image resolution during fine-tuning of CLIP results in a stronger bias towards texture. In Table 3.3, we observe that an input resolution during fine-tuning impacts shape bias: increasing input resolution during fine-tuning leads to better accuracy on ImageNet validation but also results in more texture-biased models with lower accuracy on Stylized-ImageNet. Across seven pairs of experiments and two training sources, we observe this pattern consistently. Given that input resolution is a crucial model dimension [Tan and Le, 2019; Bello et al., 2021; Tan and Le, 2021], it would be insightful to study its effects on shape bias beyond classification accuracy when devising scaling strategies.

(3) CLIP models tend to texture bias after fine-tuning. Our study reveals that shape bias in CLIP weakens after fine-tuning on ImageNet. Moreover, the fine-tuned CLIP models exhibit a shape bias comparable to models that are pre-trained on larger datasets. This finding is consistent when using a transformer and CNN as the visual encoder. Moreover, these results illustrate that fine-tuning discards the shape-biased property of zero-shot CLIP, which may affect their overall effective robustness [Geirhos et al., 2019c,b].

(4) Fine-tuning with contrastive loss maintains shape bias. By default, the CLIP-FT models are trained with standard supervised cross-entropy loss. To decouple effects of fine-tuning methods and data source, we use zero-shot CLIP with ViT-B/32 and ViT-B/16, and fine-tune them on ImageNet training set by standard cross-entropy, contrastive loss [Goyal et al., 2023], and parameter-efficient fine-tuning methods (CoOp [Zhou et al., 2022b] and Tip-Adapter [Zhang et al., 2022b]). The shape bias extents are shown in Table 3.2: contrastive fine-tuning on ImageNet maintains the shape bias of CLIP models. This indicates that ImageNet training data might not be the primary cause of shape-bias decrease. We believe that alignment mechanism between visual and textual representations may play a fundamental role in shaping this bias. This is supported by our observation that fine-tuning strategies which preserve image-text embedding association tend to retain or strengthen shape bias.

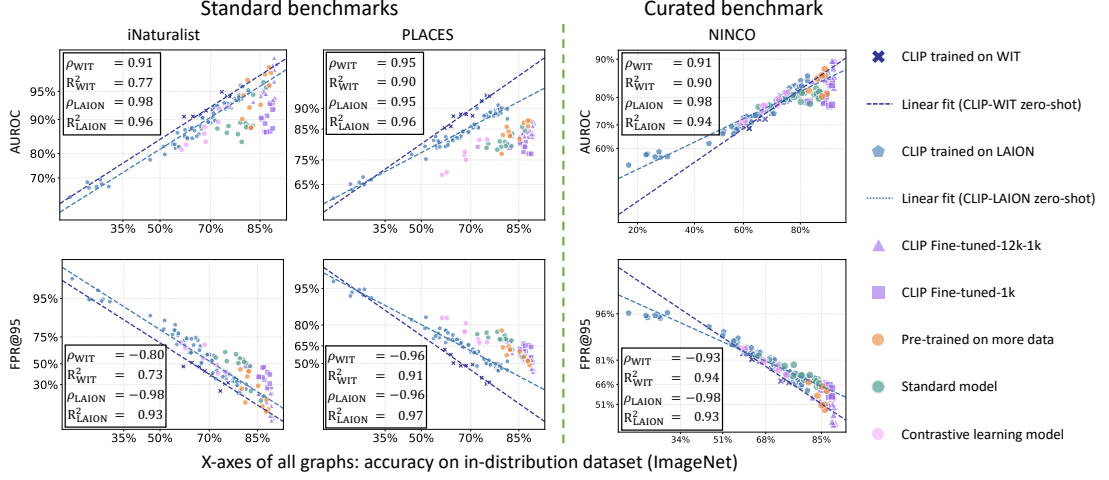


Figure 3.3: **OOD sample identification capability of models vs. ID dataset classification accuracy.** The OOD detection ability is measured by AUROC ($\uparrow$) and FPR@95 ($\downarrow$). Each point represents a model. We plot the results on iNaturalist, PLACES and NINCO. The blue straight lines are fit with robust linear regression [Huber, 2011]. We report spearman’s rank correlation and R^2 to quantify the correlation strength between ID accuracy and OOD detection performance for zero-shot CLIP trained on WIT and LAION. The x-axis and y-axis are probit transformed following Taori et al. [2020]. We observe that training distribution has a greater impact than training dataset quantity on the OOD detection performance of CLIP. Moreover, after additionally fine-tuning on ImageNet-12K, CLIP models are generally better at detecting OOD samples than those only fine-tuned on ImageNet-1K.

3.4 Out-of-Distribution Detection

Zero-shot CLIP allows for a flexible definition of in-distribution (ID) classes without re-training the model. Namely, they can conduct zero-shot OOD detection [Ming et al., 2022]. The current findings suggest that zero-shot CLIP models are competitive with other state-of-the-art models [Ming et al., 2022; Fort et al., 2021]. Based on this finding, we conduct an extensive analysis to determine whether the purported benefits persist across various training sources, subsets, and network architectures. In the experiments, for zero-shot CLIP models, we utilize maximum concept matching [Ming et al., 2022] to detect OOD data. For models trained or fine-tuned on ImageNet-1K, we use maximum softmax score [Hendrycks and Gimpel, 2016].

(1) For CLIP models from the same source, their ID accuracy correlates with OOD detection performance. Our study includes CLIP models trained on WIT or LAION. Given the same training source, our study, conducted across five challenging OOD scenarios, reveals a strong correlation between ID accuracy of zero-shot CLIP models and their OOD detection performance (measured by AUROC and FPR@95). This suggests that zero-shot classification accuracy of CLIP on ID data can serve as a reliable indicator of their OOD detection performance. In contrast, such a trend is not

as strong for both standard models and more data-pre-trained models. Furthermore, CLIP-FT models fine-tuned on ImageNet-1K do not exhibit a clear correlation.

(2) Training source impacts the trend of CLIP. Upon closer examination of the training distribution, we have observed that the correlation trend between ID accuracy and OOD detection performance is largely dependent on the training source. As illustrated in Fig. 3.3, our research shows two distinct trends between CLIP models trained on WIT and those trained on LAION. Moreover, with the same ID accuracy, CLIP models trained on WIT exhibit superior OOD detection performance compared to their counterparts trained on LAION on three OOD scenarios. This further indicates the importance of training sources for CLIP.

(3) Fine-tuning procedure significantly influences the OOD detection ability of CLIP. While fine-tuning generally improves CLIP’s classification performance, this enhancement does not necessarily translate to better OOD detection accuracy. Some fine-tuned CLIP (CLIP-FT) models perform worse in OOD detection compared to their zero-shot counterparts. Our analysis distinguishes between two groups of CLIP-FT models based on their fine-tuning procedures: one group is fine-tuned solely on ImageNet-1K, while the other undergoes additional fine-tuning on ImageNet-12K. We observe that this additional fine-tuning step has a notable impact on OOD detection performance. As shown in Fig. 3.3, despite not yielding significant gains in classification accuracy, CLIP-FT models fine-tuned on ImageNet-12K consistently achieve better OOD detection across all tested scenarios. These findings suggest that fine-tuning dataset plays a critical role in enhancing OOD detection. Future work should further explore alternative fine-tuning strategies that prioritize OOD detection. Additionally, investigating effects of fine-tuning on datasets beyond ImageNet-1K/21K presents an intriguing direction for improving the robustness of CLIPs.

(4) Evaluation on NINCO [Bitterwolf et al., 2023]. To explore the OOD detection across diverse and challenging conditions, we use a new benchmark NINCO. It consists of filtered samples from various existing OOD benchmarks. Fig. 3.3 illustrates the OOD detection performance on NINCO versus ID classification accuracy on the ImageNet validation set. The observations are consistent with those on five standard benchmarks: 1) for CLIP models from the same source, their ID accuracy correlates with OOD detection; 2) training source influences trends of CLIP; 3) additional fine-tuning on ImageNet-12K helps OOD detection ability of CLIP. ImageNet-21K offers broader semantic coverage than ImageNet-1K, which may help bridge the gap between pretraining data (e.g., LAION) and downstream tasks. As an intermediate fine-tuning stage, it could help preserve model generalization, which may explain the improved OOD robustness observed compared to direct fine-tuning on ImageNet-1K.

3.5 Confidence Calibration

To better understand the well-calibrated phenomenon of zero-shot CLIP models reported by Minderer et al. [2021], this section systematically analyzes the calibration

behavior of CLIP models under various training conditions. Specifically, we examine the calibration performance of CLIP models trained on different training distributions, varied training set sizes, and different architectures. Furthermore, we also investigate the calibration performance of CLIP models after fine-tuning.

3.5.1 Zero-Shot CLIPs Are Not Consistently More Calibrated Than Others

(1) Training data distribution and quantity significantly affect CLIP’s calibration. Fig. 3.4 illustrates the calibration of CLIP models concerning classification accuracy under distribution shifts. We find that models trained on different distributions or dataset sizes do not always group consistently. For example, CLIP models trained on WIT and LAION tend to form distinct clusters. Additionally, within subsets of the LAION dataset, models with similar classification accuracy can display varying levels of calibration. While CLIP models are often praised for superior calibration compared to other models [Minderer et al. \[2021\]](#), our analysis shows this is not always the case. Notably, CLIP models trained on the LAION-80M dataset exhibit significantly lower calibration performance compared to standard models. The superior calibration reported by [Minderer et al. \[2021\]](#) is primarily based on CLIP models trained on WIT. However, when we expand the analysis to models trained on the broader LAION dataset and its subsets, we observe more variability.

(2) CLIP fine-tuned models show a trade-off between calibration and classification. As shown in Fig. 3.4, fine-tuning CLIP models consistently results in higher classification accuracy but increased calibration error across all test sets. Furthermore, we did not observe that further fine-tuning CLIP on ImageNet-12K benefits calibration performance, which contrasts with its positive impact on OOD detection. Interestingly, other model groups, including those pre-trained on larger datasets, do not show an obvious trade-off between calibration and classification. Additionally, we observe that few fine-tuned CLIP models achieve better calibration than their zero-shot counterparts, even before applying calibration techniques.

3.5.2 Temperature Scaling Highlights Well-Calibrated Properties of Zero-Shot CLIP Models

Post-hoc calibration methods, such as temperature scaling [\[Guo et al., 2017\]](#), are often employed to correct overconfidence or underconfidence in model predictions. Following the protocol in [\[Gupta et al., 2021\]](#), we split the ImageNet validation set into two halves: one for temperature scaling (ID calibration) and the other for testing. We report results on both in-distribution (ID) and out-of-distribution (OOD) test sets.

(1) Classification accuracy of CLIP models correlates with calibration performance after temperature scaling. In Fig. 3.4, we examine the effects of temperature scaling on both CLIP and non-CLIP models, grouped based on the amount and source of their training data. After applying temperature scaling and evaluating with the negative log-likelihood (NLL) metric, we observe that models with higher classification accuracy generally show better calibration. Importantly, when temperature scaling is

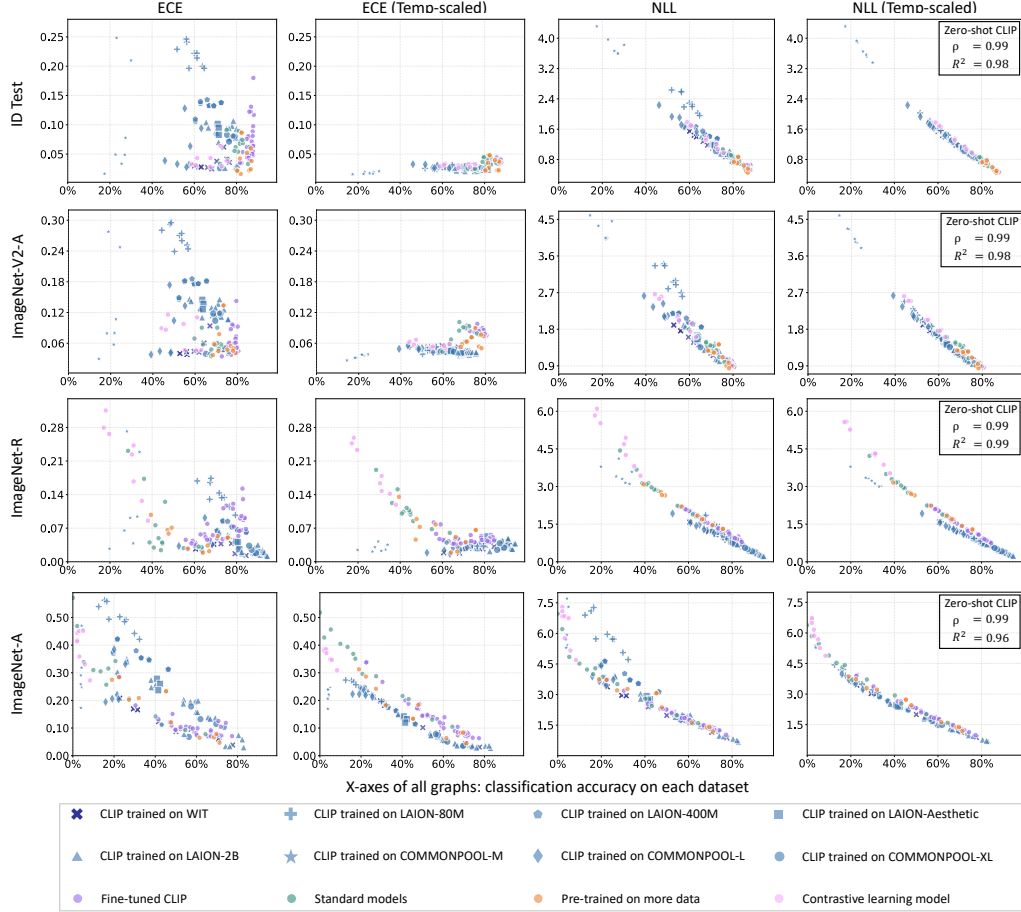


Figure 3.4: **Model calibration performance with respect to classification accuracy.** We report results on in-distribution test set, ImageNet-V2-A, ImageNet-R, and ImageNet-A. Two metrics are considered: ECE ($\downarrow$) and NLL ($\downarrow$), we also include calibration performance after calibration with temperature scaling. Each point represents a model. We use colors to represent model groups. For zero-shot CLIP, we additionally use shapes to indicate training distribution and quantity. CLIP models can have higher ECE than standard models. Also, the training distribution and quantity are the key factors influencing the calibration performance of CLIP models. Moreover, temperature scaling reveals a consistent trend in CLIP models. After using temperature scaling for both CLIP and other models, CLIP models follow a distinct trend from others and show better calibration performance

applied to both CLIP and other models, zero-shot CLIP models consistently outperform other models, including fine-tuned versions, in calibration.

This pattern persists across various testing conditions, including ID and OOD sets, with zero-shot CLIP models demonstrating superior calibration compared to other models. This trend holds across both NLL and ECE metrics.

(2) ID calibration of CLIP models transfers to OOD test sets. While prior studies [Ovadia et al., 2019] report in-distribution (ID) calibration often fails to generalize under distribution shifts, our findings reveal a promising result for CLIP models. After calibrating CLIP models on the ID set, they exhibit improved calibration on OOD test sets. For example, on ImageNet-A, CLIP models exhibit lower calibration error after temperature scaling, a trend not seen in other models. This suggests that CLIP models are relatively easier to calibrate across diverse distributions, indicating their potential for robust and reliable applications in real-world settings.

3.6 Zero-Shot Retrieval

Since CLIP models are trained using contrastive loss to associate text and image pairs, we evaluate their zero-shot retrieval capability on the Flickr30K [Young et al., 2014] and MSCOCO [Chen et al., 2015] datasets in this section.

We have three major observations on the two datasets. **First**, CLIP’s zero-shot retrieval capability correlates with its image classification performance. Fig. 3.5 illustrates image and text zero-shot retrieval (gauged by Recall@5) against their accuracy on ImageNet. We observe that classification ability is predictive of their retrieval capability. **Second**, training distribution deviates from the retrieval performance trend. Specifically, CLIPs trained on WIT slightly deviate from the trend formed by CLIPs trained on LAION, and the training quantity does not affect the trend. **Last**, we observe four specific ConvNeXt-based CLIP models significantly depart from the trend of LAION. We notice that they are trained with a limited random resize crop range (0.9, 1.0). This limited augmentation likely reduces training view diversity, resulting in less variation in object scale and context. In image-text retrieval, where the model must extract consistent global representations that align well with corresponding textual descriptions, this lack of variation can hinder the learning of robust embeddings, ultimately affecting retrieval performance. While this chapter does not consider such training augmentations, it would be interesting to explore their impact on retrieval.

3.7 3D Awareness

CLIP models are trained using contrastive loss to associate text and image pairs in feature space, but this training does not explicitly incorporate 3D understanding, such as recognizing geometric concepts like multi-view consistency and depth. Despite being trained on 2D data, recent studies suggest that models like CLIP can still be effective in 3D-related tasks [El Banani et al., 2024; Zhang et al., 2024b,c]. Building on this insight, this section evaluates the behaviors of CLIP models in 3D-specific

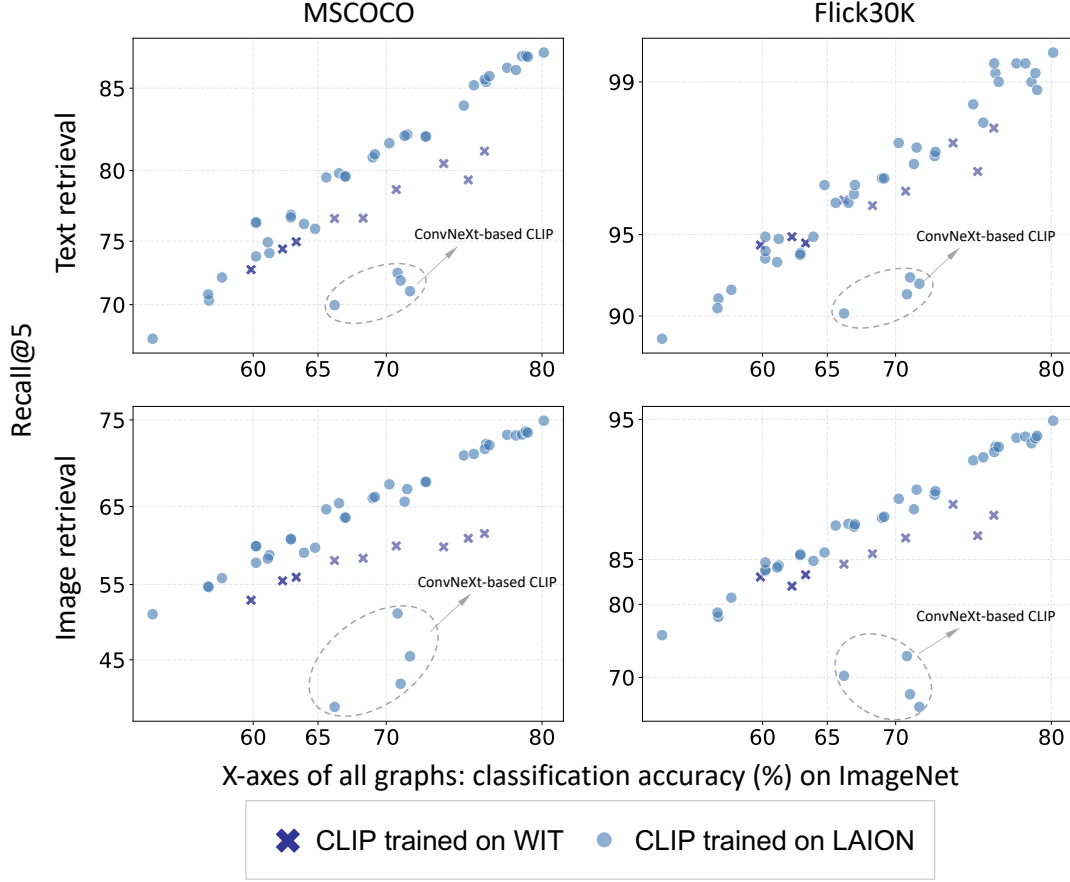


Figure 3.5: **Image/text zero-shot retrieval v.s classification accuracy** on MSCOCO and Flickr30K measured by Recall@5. Classification accuracy is predictive of zero-shot retrieval capability. Moreover, four ConvNeXt-based CLIP models trained with a limited range of random resize crop exhibit much lower retrieval performance.

scenarios, particularly examining their ability to capture 3D geometry and their robustness to 3D distortions.

3.7.1 Correspondence Matching

Geometric Correspondence. Given two views of the same object or scene, the objective of this task is to identify pixels in both views that correspond to the same location in 3D space. We evaluate this using recall on the ScanNet [Dai et al. \[2017\]](#) dataset for object-centric correspondence and NAVI [Jampani et al. \[2023\]](#) for scene-centric correspondence. Correspondence recall measures the percentage of correct correspondences that fall within a defined threshold distance. Following the protocol in [El Banani et al. \[2024\]](#), we categorize performance based on the magnitude of transformation between view pairs.

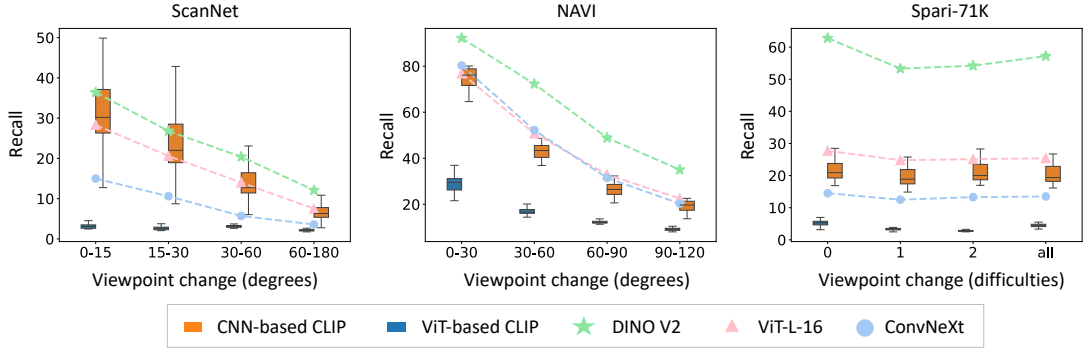


Figure 3.6: **Correspondence matching performance (Recall $\uparrow$) with respect to their viewpoint change.** We report results on geometric correspondence matching (ScanNet, NAVI) and semantic correspondence matching (Spair-71K). CLIP models are grouped by the architecture of the visual encoder into CNN-based and ViT-based. We observe that CNN-based CLIP models consistently outperform ViT-based CLIP models, particularly in scenarios with larger viewpoint variations, and achieve competitive results compared to supervised models like ConvNeXt and ViT-L-16.

Semantic Correspondence. This task generalizes geometric correspondence by requiring matching of semantically similar parts across different instances of the same object class. For example, mapping the left paw of two different dogs. We use the SPair-71K [Min et al. \[2019\]](#) dataset, with performance measured by recall. Similar to geometric correspondence, we group results by the degree of view variation. Fig. 3.6 groups CLIP models based on their visual encoder architectures (CNN-based and ViT-based). For comparison, we also include standard supervised models such as ConvNeXt and ViT-L/16 (DeiT III) [\[Touvron et al., 2022\]](#), which are trained on ImageNet-22K, alongside DINO-V2 [\[Oquab et al., 2024\]](#).

Observations. First, ViT-based CLIP models exhibit weaker performance across three datasets (ScanNet, NAVI, and Spair-71K), falling behind the supervised model (ViT-L-16), which also uses a transformer-based architecture. In contrast, CNN-based CLIPs consistently achieve higher recall scores than their ViT-based counterparts, particularly as viewpoint changes become more extreme. Additionally, CNN-based CLIP models show competitive performance when compared to supervised CNN model ConvNeXt. This suggests the combined effect of the visual encoder architecture and training objective, which plays a crucial role in influencing CLIP’s ability to manage correspondence matching. Second, our study extends the observation of [El Banani et al. \[2024\]](#), showing CNN-based CLIP models not only perform competitively with ViT-L/16 on NAVI but also match DINO-V2 on ScanNet. Note that, DINO-V2 emerges as the top performer across all three datasets. These findings suggest that CNN-based CLIPs generally exhibit stronger correspondence matching than ViT-based CLIPs, especially in scenarios involving significant viewpoint variations.

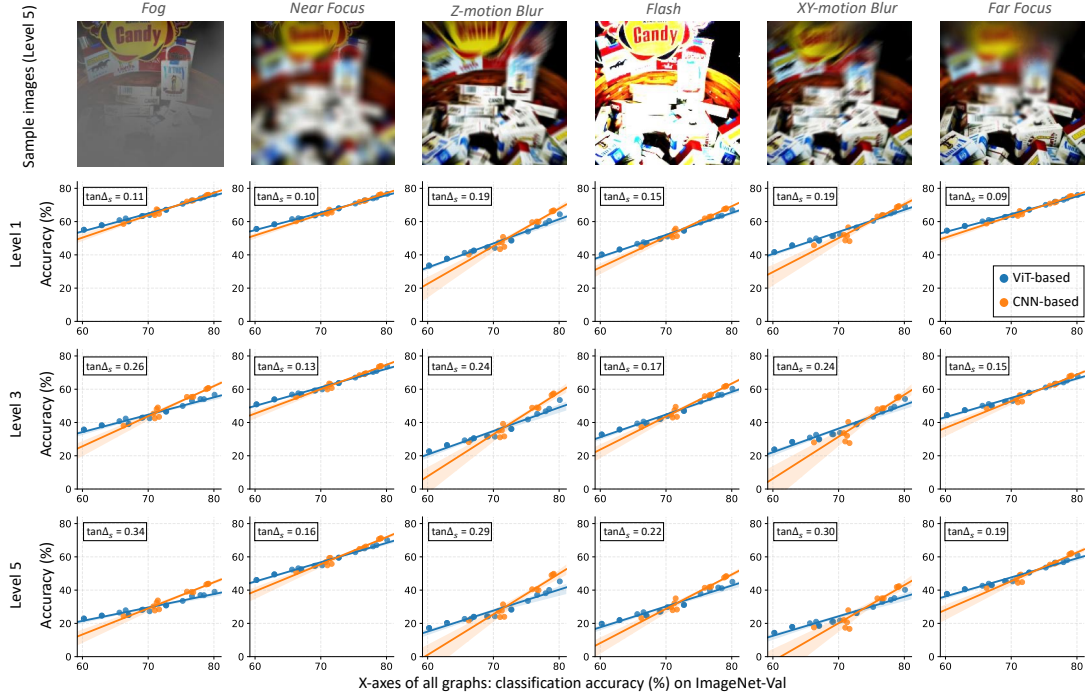


Figure 3.7: Robustness comparison of ViT-based and CNN-based CLIP models under varying 3D-related corruptions. The x-axis represents accuracy on ImageNet-Val, while the y-axis represents accuracy on the corrupted dataset. We show the accuracy of ViT-based and CNN-based CLIP models across six types of 3D-related corruptions: *Fog*, *Near Focus*, *Z-motion Blur*, *Flash*, *XY-motion Blur*, and *Far Focus*, evaluated at three severity levels (Level 1, Level 3, and Level 5). Each column shows that CNN-based models consistently exhibit steeper slopes, indicating greater resilience with less performance degradation as ImageNet-Val accuracy improves. As corruption intensity increases, the gap between the slopes, represented by $\tan(\Delta_s)$, widens, particularly under severe conditions like *Z-motion Blur*. This widening gap highlights the superior robustness of CNN-based models compared to their ViT-based counterparts, especially at higher corruption levels. This reinforces the significant impact of visual encoder architecture on CLIP’s ability to handle 3D-related corruption. Sample images of Level 5 severity for each corruption are provided at the top.

3.7.2 Robustness against 3D corruptions

We further evaluate the ability of CLIP models to handle 3D-related corruptions using the 3D Common Corruptions (3DCC) benchmark [Kar et al., 2022], which applies corruptions based on 3D transformations. Unlike the common corruptions in [Hendrycks and Dietterich, 2019a], these transformations consider the underlying geometry of the scene, producing distortions that are more reflective of real-world conditions. Sample images of corruptions are shown in the last row in Fig. 3.7. For example, the *fog* gets denser further away from the camera. In this study, we analyze six types of 3D-related corruptions, each with five severity levels, and examine only CLIP models pre-trained on LAION to maintain consistency in training dataset distributions. Based on correspondence matching, we categorize the CLIP models into CNN-based and ViT-based groups.

CNN-based CLIP models demonstrate stronger robustness to 3D-related corruptions as corruption intensity increases. Fig. 3.7 shows the performance of ViT-based and CNN-based CLIP models across various 3D-related corruptions (*Fog*, *Near Focus*, *Z-motion Blur*, *Flash*, *XY-motion-blur* and *Far Focus*) at different severity levels (Level 1, Level 3, and Level 5). For each row, the slope of the CNN-based models is consistently steeper than that of the ViT-based models, indicating that CNN-based models experience less degradation in performance as the clean ImageNet validation accuracy increases. This suggests that CNN-based models are more robust in maintaining accuracy under 3D distortions.

Furthermore, as corruption intensity increases (moving from Level 1 to Level 5), the gap between the slopes, represented by $\tan(\Delta_S)$, widens. This increase highlights that the advantage of CNN-based models becomes more pronounced under higher severity of corruptions, particularly for challenging distortions like *Fog* and *Z-motion Blur*. The growing slope difference indicates that CNN-based models are increasingly more capable of handling severe 3D corruptions compared to ViT-based models. These results reinforce the importance of visual encoder architecture in achieving robustness across varying corruption intensities, with CNN-based models consistently outperforming ViT-based models, especially as the corruption severity escalates. When considered alongside the results from the correspondence matching, these findings underscore the pivotal role that visual encoder architecture plays in enhancing robustness to 3D corruptions. Lastly, ViT-based CLIP models struggle with 3D geometric understanding, whereas DINO models perform better. This has implications for downstream multimodal models like LLaVA [Liu et al., 2024c], which typically rely on CLIP backbones. Combining DINO with CNN-based CLIP features could improve spatial reasoning, as suggested by Kar et al. [2024].

3.8 Visual and Language Encoder Interaction: A Classification Perspective

Modern large multimodal models (LMMs), such as LLaVA [Liu et al., 2024c], typically use a frozen pre-trained visual encoder from CLIP as their visual backbone, with instruction fine-tuning applied to the linear projector and the language model components. This raises an important question: how does the interaction between a shared visual encoder and distinct language models affect the classification performance of LLaVA compared to CLIP-like models?

Driven by this, we compare classification accuracy of CLIP and LLaVA to investigate how the interaction between the shared visual encoder and their distinct language models influences overall performance. In this section, “LLaVA” and “CLIP” refer to their training paradigms rather than specific model implementations. We also include SigLIP [Zhai et al., 2023] as another representative of CLIP-like models.

Our evaluation is conducted on three splits of the ImageNet-D dataset [Zhang et al., 2024a]: *Background*, *Texture*, and *Material*. This dataset, generated by a text-to-image diffusion model, poses significant classification challenges. We adopt a VQA-style approach for LLaVA’s classification, providing it with a category list per image and prompting it to select the correct category. The list includes the ground truth (GT) category and three “failure” categories—incorrect categories ranked with the highest confidence by a pretrained category selection model—ensuring a unique category list for each image. We evaluate the role of the category selection model using ResNet-50, CLIP-ViT-L/14-336, and SigLIP-SO-14.

To explore the interaction between the language encoder and CLIP vision encoders, we consider six LLaVA models, combining two types of visual encoders: CLIP-ViT-L/14-336 and SigLIP-SO-14 and three language encoders: Mistral-Instruct-V2 [Jiang et al., 2023], Llama2-Chat [Touvron et al., 2023], and Vicuna-V2-7B [Touvron et al., 2023]. For a fair comparison, CLIP is given the same category list using the default prompt template by Radford et al. [2021] (e.g., “a photo of [category]”). LLaVA’s prompt format is:

What is the main object in this image? Choose from the following list:
 A.[Candidate class 1]
 B.[Candidate class 2]
 C.[Candidate class 3]
 D.[Candidate class 4]
 Please answer the question using the choice from the list.

Observations: We report the results on ImageNet-D in Table 6.2 and summarize the observations as follows.

First, extending the findings of Zhang et al. [2024a], which evaluate CLIP (ViT/14) as a category selection model, we find that the interactions between vision and lan-

No.	Category List	Visual encoder	Type	LLM	Background	Material	Texture
1	ResNet-50	CLIP ViT-L/14-336 (WIT)	CLIP	-	0.90	0.92	0.95
			LLaVA	Mistral-Instruct-V2	0.82	0.81	0.80
			LLaVA	Llama2-Chat	0.91	0.87	0.86
			LLaVA	Vicuna-V2-7B	0.92	0.89	0.91
		SigLIP-SO-14	CLIP	-	0.97	0.96	0.99
			LLaVA	Mistral-Instruct-V2	0.84	0.73	0.79
			LLaVA	Llama2-Chat	0.93	0.91	0.93
			LLaVA	Vicuna-V2-7B	0.92	0.90	0.94
2	CLIP SigLIP-SO-14 (Webli)	ViT-L/14-336 (WIT)	CLIP	-	0.23	0.24	0.21
			LLaVA	Mistral-Instruct-V2	0.41	0.35	0.28
			LLaVA	Llama2-Chat	0.52	0.35	0.34
			LLaVA	Vicuna-V2-7B	0.57	0.48	0.42
		SigLIP-SO-14	CLIP	-	0.65	0.61	0.61
			LLaVA	Mistral-Instruct-V2	0.44	0.33	0.35
			LLaVA	Llama2-Chat	0.60	0.47	0.46
			LLaVA	Vicuna-V2-7B	0.59	0.48	0.43
3	CLIP ViT-L/14-336 (WIT)	ViT-L/14-336 (WIT)	CLIP	-	0.14	0.14	0.13
			LLaVA	Mistral-Instruct-V2	0.37	0.35	0.25
			LLaVA	Llama2-Chat	0.49	0.32	0.30
			LLaVA	Vicuna-V2-7B	0.57	0.45	0.42
		SigLIP-SO-14	CLIP	-	0.69	0.67	0.65
			LLaVA	Mistral-Instruct-V2	0.46	0.34	0.36
			LLaVA	Llama2-Chat	0.62	0.48	0.48
			LLaVA	Vicuna-V2-7B	0.59	0.52	0.44

Table 3.4: **Compared CLIP and LLaVA models on ImageNet-D.** We include two visual backbones: CLIP-L/14-336 and SigLIP-SO-14 and three language models for LLaVA: Mistral-Instruct-V2, Llama2-Chat, and Vicuna-V2-7B.

guage components in selection-based networks vary significantly with task difficulty. When the category list is easy for CLIP, LLaVA models using the same visual encoder do not yield consistent improvements and sometimes exhibit slight performance drops. In contrast, when the category list is challenging for CLIP, LLaVA models using the same visual encoder show substantial gains. For example, in row 1, the most confused categories of ResNet-50 are easy for CLIP, and LLaVA brings no improvement. Similarly, in row 2, when SigLIP-SO-14 performs well, LLaVA shows a performance drop. However, in the same row, when the category list becomes difficult for SigLIP-SO-14, LLaVA improves accuracy by over 20% across three splits. A similar pattern is observed in row 3 for CLIP (ViT-L/14-336) and its LLaVA counterpart. Since LLaVA and CLIP share the same visual encoder, we speculate that the observed gains arise when CLIP’s visual-text alignment is weak—likely due to ambiguous categories or limited pre-training coverage. In such cases, LLaVA’s language model may help disambiguate visual features by leveraging external knowledge acquired during multimodal instruction tuning. Conversely, when CLIP already handles the token comparison effectively, the language model may over-interpret the visual input, occasionally leading to reduced performance. **Second**, the choice of language model within LLaVA significantly impacts classification performance. Vicuna-V2-7B consistently outperforms Mistral-Instruct-V2, while the choice of visual encoder also plays a critical role. LLaVA models built on SigLIP-SO-14 outperform those using ViT-L/14-336, echoing recent findings in the literature.

The above suggests that LLaVA’s performance gains are not solely the result of architectural complexity or additional training data, but rather the interaction between the two. The model’s effectiveness depends on how the language model, visual encoder, and learned projection layer work together in response to varying input complexity. This indicates the importance of designing vision-language models with careful attention to how components are integrated and how they respond under different levels of visual-text alignment difficulty.

3.9 Impact of Training and Inference Strategy on Robustness

3.9.1 Impact of Test-Time Prompts

In previous analyses, we used default prompt set provided by [Radford et al., 2021]. Here, we investigate how varying test-time prompts influence CLIP’s performance in out-of-distribution (OOD) detection, visual factor robustness, and predictive uncertainty. We experiment with five additional prompt sets: (1) a single prompt (“a photo of a {label}”); (2) a set generated by GPT-3 following Pratt et al. [2023]; (3) a prompt set generated by GPT-4 [Achiam et al., 2023] using the chain-of-thought strategy [Wei et al., 2022] (VDT) [Maniparambil et al., 2023]; (4) prompts generated by GPT-4 (MPVR-GPT4) or Mistral (MPVR-Mistral) with a target-task-oriented design [Mirza et al., 2024]; (5) prompts generated by GPT-4 with Retrieval-Augmented Learning (REAL) [Parashar et al., 2024]. These prompts are tested across five CLIP models: RN50, RN50×64, ViT-B/16, ViT-B/32, and ViT-L/14-336, all trained on WIT.

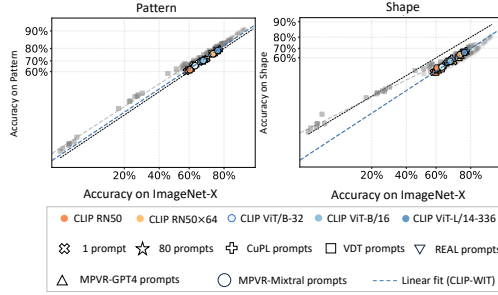


Figure 3.8: Influence of test-time prompts on CLIP’s visual-factor robustness. We evaluate five CLIP models trained on WIT, represented by different colors for architectures and different shapes. The dashed grey line represents robust linear regression Huber [2011] based on the original CLIP-WIT models with 80 prompts. Different prompt sets may influence classification performance but do not significantly impact visual factor robustness because models still lie on the original line.

Backbone	Pre-training dataset	Classification	OOD Detection			Calibration	
			Accuracy (↑)	AUROC (↑)	FPR (↓)	Before-temp ECE (↓)	After-temp ECE (↓)
RN50	1 Prompt	0.41	0.84	0.61	0.08	0.08	
	80 Prompts	0.43	0.83	0.64	0.08	0.08	0.08
	CuPL	0.44	0.83	0.63	0.09	0.09	
	VDT	0.42	0.82	0.65	0.10	0.09	
	MPVR-GPT4	0.43	0.82	0.65	0.09	0.09	
	MPVR-Mistral	0.43	0.82	0.66	0.08	0.09	
	REAL	0.42	0.81	0.69	0.08	0.09	
ViT-B/16	1 Prompt	0.57	0.86	0.55	0.05	0.06	
	80 Prompts	0.59	0.85	0.57	0.05	0.06	
	CuPL	0.60	0.86	0.54	0.06	0.06	
	VDT	0.59	0.86	0.54	0.07	0.06	
	MPVR-GPT4	0.60	0.85	0.58	0.05	0.06	
	MPVR-Mistral	0.60	0.85	0.60	0.05	0.06	
	REAL	0.58	0.83	0.63	0.06	0.06	

Table 3.5: Influence of test-time prompts on CLIP’s classification, OOD detection and calibration. We evaluate CLIP models trained on WIT with ResNet-50 and ViT-B/16 as the visual encoder. We find that prompt sets generated by large language models may improve zero-shot CLIP models’ classification accuracy, but it does not enhance other OOD detection or calibration.

Fig. 4.6 and Table 3.5 summarize the effects of various prompt strategies on CLIP’s classification, robustness, OOD detection, and calibration. Using fewer prompts (*e.g.*, a single prompt) reduces classification accuracy but improves OOD detection and calibration. Factor-level robustness, such as on the Pattern task, remains largely unchanged regardless of prompt type, with models following the original CLIP-WIT trend. Prompt sets generated by large language models—including VDT, MPVR-GPT4, MPVR-Mistral, and REAL—consistently improve classification accuracy, but show no clear advantage in OOD detection or calibration. These results highlight a key challenge: how to design prompt strategies that simultaneously enhance other objectives beyond classification accuracy only.

3.9.2 Effect of Fine-Tuning Procedures

In addition to standard fine-tuning methods (*i.e.*, cross-entropy fine-tuning on ImageNet), we examine seven alternative fine-tuning strategies: contrastive fine-tuning (FLYP) as introduced by [Goyal et al., 2023], two robust fine-tuning methods—WiSE-FT [Wortsman et al., 2022] and LP-FT [Kumar et al., 2022], and four parameter-efficient methods: CoOp [Zhou et al., 2022b], Tip-Adapter [Zhang et al., 2022b], MaPLE [khattak et al., 2023] and PromptSRC [Khattak et al., 2023]. They are applied to fine-tune zero-shot CLIP models pre-trained on WIT.

In Fig. 3.9, we present the performance of fine-tuned CLIP models across visual factor robustness, OOD detection, and calibration. The results reveal mixed effects across different tuning methods. For visual factor robustness, CoOp and PromptSRC

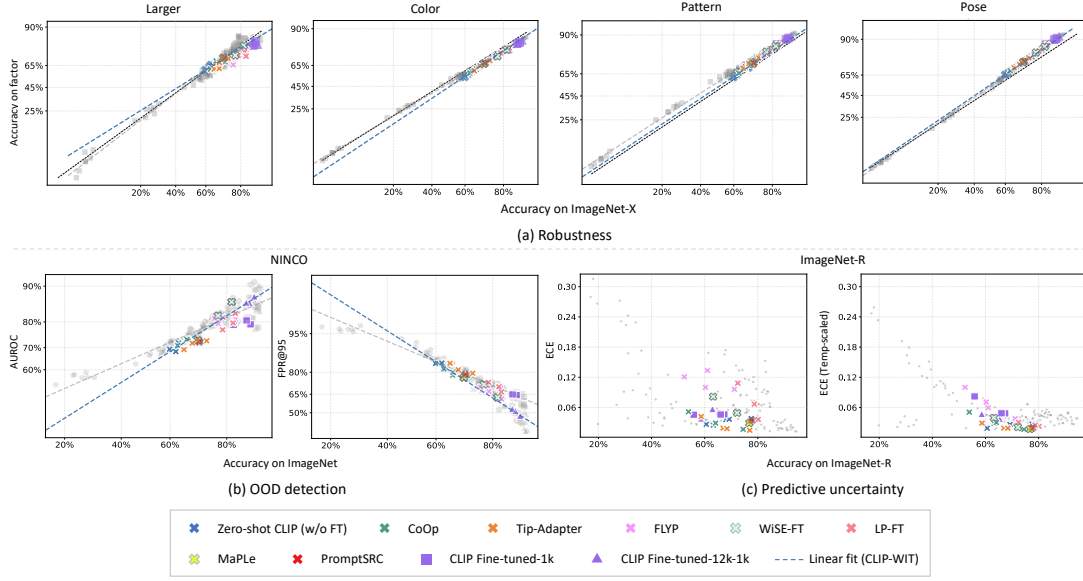


Figure 3.9: **Influence of fine-tuning algorithms on CLIP’s robustness, OOD detection, and predictive uncertainty.** We fine-tune four CLIP models trained on WIT using various algorithms. Different colors represent model architectures, and different shapes denote fine-tuning algorithms. The blue dashed line is fit with robust linear regression [Huber, 2011] for original CLIP-WIT models, while the grey dashed line represents zero-shot CLIP trained on LAION. Results show that contrastive fine-tuning improves overall accuracy but negatively impacts predictive uncertainty.

preserve the properties of zero-shot CLIP, aligning with prior findings that test-time prompts have limited influence on robustness. FLYP and Tip-Adapter improve robustness against the *Pattern* factor but reduce it under *Larger* visual changes. LP-FT maintains robustness across several factors, while WiSE-FT slightly weakens it on *Larger*. On OOD detection, most methods—including FLYP, LP-FT, WiSE-FT, and PromptSRC—enhances both accuracy and detection performance, with LP-FT showing strong generalization as their models lie above the zero-shot CLIP-WIT trend. For calibration, FLYP increases calibration error, while CoOp, Tip-Adapter, and PromptSRC maintain well-calibrated predictions. LP-FT and WiSE-FT increase error slightly before temperature scaling but recover calibration performance afterward, outperforming FLYP in uncertainty estimation.

These findings suggest that while fine-tuning can improve certain aspects of CLIP’s performance, achieving a balance between classification accuracy, OOD detection, and predictive uncertainty remains a challenge, highlighting the need for further research into fine-tuning strategies that can address all of these objectives.

3.9.3 Extending to Different Training Paradigms

We have expanded our analysis beyond CLIP to include four additional vision-language training paradigms: (1) BLIP [Li et al., 2022b], (2) BLIP-2 [Li et al., 2023a],

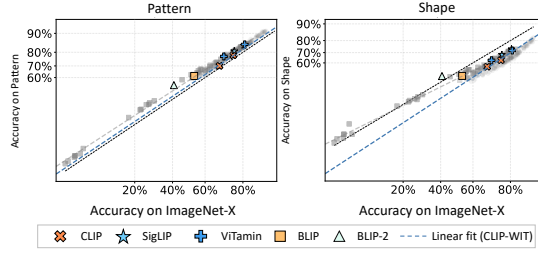


Figure 3.10: **Visual-factor robustness of different training paradigms from CLIP.** We evaluate five vision-language training paradigms: CLIP, BLIP, BLIP-2, SigLIP and ViTamin. The dashed blue line represents robust linear regression [Huber, 2011] based on the original CLIP-WIT models. Different training paradigms effectively impact visual factor robustness.

Training paradigm	Classification Accuracy (↑)	OOD Detection		Calibration		3D Robustness Accuracy (↑)
		AUROC (↑)	FPR (↓)	Before-temp ECE (↓)	After-temp ECE (↓)	
CLIP-ViT-B/16	0.59	0.85	0.57	0.06	0.06	0.34
BLIP-Base	0.51	0.76	0.70	0.07	0.08	0.27
SigLIP-ViT-B/16	0.68	0.89	0.50	0.08	0.05	0.38
ViTamin-Base	0.62	0.84	0.67	0.37	0.23	0.35
CLIP-ViT-L/14	0.72	0.88	0.47	0.06	0.05	0.46
BLIP-2	0.53	0.66	0.89	0.14	0.08	0.37
SigLIP-ViT-L/16	0.75	0.91	0.40	0.07	0.04	0.46
ViTamin-L-256px	0.79	0.89	0.49	0.21	0.17	0.53

Table 3.6: **Performance of various training paradigms on classification, OOD detection, calibration and 3D robustness.** We evaluate five vision-language training paradigms: CLIP, BLIP, BLIP-2, SigLIP and ViTamin. We find that no training paradigm is the most performant on all safety-related objectives.

(3) SigLIP [Zhai et al., 2023], and (4) ViTamin [Chen et al., 2024b]. We summarized the results in Fig. 3.10 and Table 3.6. We observe that different training paradigms yield trade-offs across robustness, OOD detection, calibration, and 3D performance. SigLIP achieves the best overall balance, with strong OOD detection, low calibration error, and competitive 3D robustness. ViTamin-L-256px leads in classification accuracy and 3D robustness but suffers from poor calibration. In addition, BLIP and BLIP-2 consistently underperform across most metrics. These results reinforce our main claim: no single paradigm excels universally, underscoring the need for multi-dimensional evaluation beyond accuracy alone.

3.9.4 Robustness Evaluation of Dataset Curation

High-quality training sets are crucial for developing CLIP models, and as a result, recent research has increasingly emphasized dataset curation (DC) to create datasets [Gadre et al., 2023; Fang et al., 2024; Xu et al., 2024]. In this chapter, we extend the evaluation of DC techniques to robustness-related tasks, including OOD detection, calibration, visual factor-level robustness, and 3D corruption.

To ensure a clear and fair comparison, we control the architecture of the CLIP models and categorize the methods based on their pretraining dataset sources. We consider four DC techniques: 1) CommonPool [Gadre et al., 2023], which uses a trained CLIP model as a filter; 2) MetaCLIP [Xu et al., 2024], which leverages meta-data for curation and balancing of raw web-sourced data; and 3) DFN-2B [Fang et al., 2024], which employs a network trained on high-quality datasets for filtering; 4) Aesthetic [Schuhmann and Beaumont, 2022], which is filtered using perceptual hashing for deduplication and an aesthetic score threshold.

Table 3.7 shows that DC techniques consistently improve performance in classification, OOD detection, visual factor robustness, and 3D robustness—particularly

Table 3.7: **Comparison of CLIP trained with filtered pre-training data on six tasks.** For classification tasks, we report average accuracy on ImageNet validation, ImageNet-V2-A, ImageNet-S, ObjectNet, ImageNet-A, ImageNet-R and ImageNet-Vid. We report averaged AUROC and FPR on NINCO, iNaturalist, DTD, Place, SUN and ImageNet-O. We report ECE before / after calibration. The calibration set is ID-val and test set is the same as OOD generalization. For visual factor robustness, we evaluate *Larger*, *Shape* and *Color*. We use averaged recall@5 to measure text-to-image and image-to-text retrieval on MSCoCo and Flickr30K. For 3D robustness, we use accuracy to metric their performance on six 3D-related corruptions with severity level 5. The best performance for each architecture is in green. We find that data curation is an effective method for enhancing model performance beyond classification.

Backbone	Pre-training dataset	Data Filtering	Classification	OOD Detection			Calibration		Visual factor robustness			Retrieval	3D robustness
				Accuracy (↑)	AUROC (↑)	FPR (↓)	Before-temp ECE (↓)	After-temp ECE (↓)	Larger Accuracy (↑)	Shape Accuracy (↑)	Color Accuracy (↑)		
ViT-B/16	LAION-400M	No	0.61	0.84	0.65	0.13	0.05	0.67	0.56	0.63	0.82	0.32	
	MetaCLIP-400M	Yes	0.67	0.85	0.62	0.09	0.08	0.75	0.61	0.67	0.83	0.35	
	LAION-2B	No	0.64	0.85	0.64	0.13	0.05	0.69	0.60	0.67	0.84	0.34	
	DFN-2B	Yes	0.70	0.88	0.52	0.12	0.07	0.80	0.66	0.73	0.85	0.40	
	CommonPool-L	No	0.43	0.73	0.86	0.06	0.07	0.45	0.46	0.58	0.64	0.19	
	CommonPool-L-CLIP	Yes	0.53	0.77	0.81	0.11	0.07	0.61	0.53	0.56	0.72	0.26	
ViT-L/14	LAION-400M	No	0.68	0.86	0.59	0.17	0.06	0.75	0.64	0.70	0.85	0.38	
	MetaCLIP-400M	Yes	0.76	0.89	0.50	0.09	0.06	0.74	0.67	0.74	0.85	0.45	
	LAION-2B	No	0.72	0.88	0.52	0.11	0.04	0.82	0.66	0.72	0.87	0.42	
	DFN-2B	Yes	0.78	0.91	0.39	0.07	0.04	0.85	0.74	0.79	0.88	0.50	
	CommonPool-XL	No	0.72	0.87	0.54	0.03	0.04	0.72	0.65	0.70	0.80	0.43	
	CommonPool-XL-CLIP	Yes	0.75	0.88	0.54	0.08	0.03	0.74	0.67	0.74	0.84	0.46	
ConvNeXt-Base	LAION-2B	No	0.64	0.85	0.64	0.12	0.05	0.69	0.59	0.67	0.72	0.35	
	LAION-Aesthetic	Yes	0.65	0.85	0.63	0.14	0.05	0.71	0.62	0.67	0.68	0.33	

for transformer-based models. However, their impact on calibration is limited. We also evaluate CNN-based CLIP models trained on LAION-Aesthetic and observe that while dataset filtering improves classification and robustness, it shows limited benefits for retrieval, calibration, and 3D robustness. These observations suggest that the effectiveness of filtering strategies depends on both model architecture and the specific evaluation objective, emphasizing the need for multi-dimensional assessment beyond classification accuracy.

3.10 Summary

Our research contributes to the ongoing discussion regarding the robustness and capabilities of CLIP models, particularly in response to visual factor robustness, OOD detection, the reliability of uncertainty estimation, zero-shot retrieval capabilities, and 3D awareness. To achieve these insights, we perform comprehensive experiments and comparative analyses, systematically evaluating CLIP models against diverse model families. Through an in-depth exploration of critical factors—including training sources, contrastive learning objectives, network architecture, fine-tuning strategies, and test-time prompt variations—our findings provide substantial insights into the unique advantages CLIP models offer.

Discussion on Dataset Overlap. Given that CLIP models are pretrained on large-scale web-crawled datasets such as LAION-5B, potential overlap with evalu-

ation benchmarks is a valid concern. Prior work suggests that such overlap is unlikely to significantly affect our findings. For classification robustness, LAION-5B paper [Gadre et al., 2023] and OpenAI [Radford et al., 2021] report only isolated cases where overlap impacts performance, and do not view it as a major threat to result validity. For OOD detection, Bitterwolf et al. [2023] show that overlapping class semantics between pretraining (e.g., IN-21K) and test sets (e.g., NINCO) does not substantially alter detection performance. For calibration, since our evaluation datasets are shared with classification and focus on relative model comparison, any sample-level overlap is unlikely to influence conclusions. Moreover, our emphasis on relative trends, rather than absolute scores, further mitigates this concern.

This chapter leaves open many interesting and promising directions and we discuss a few. **First**, we offer an analysis of LLaVA and demonstrate that its large language model can assist in classification where CLIP’s text and visual tokens are misaligned. Future work could explore other modern large vision models (LVMs), such as BLIP-3 [Xue et al., 2024] and Otter [Zhao et al., 2023], to deepen this analysis. Further exploration into the interaction between language models and CLIP’s visual encoder could also yield valuable insights. We see our analysis as a starting point. **Second**, our study includes two academic training sources—WIT and LAION—for CLIP. Future work should investigate whether our findings generalize to other training sources, such as datasets generated by Stable Diffusion [Rombach et al., 2022a], to advance our understanding of multi-modal dataset design. **Lastly**, our analysis reveals a critical need for more refined fine-tuning strategies tailored to CLIP models, aimed at improving both classification accuracy and robustness.

Uncertainty Estimation in Vision–Language Models

Chapter 3 shows that CLIP models can be effectively calibrated using simple temperature scaling. In this chapter, we further study the uncertainty calibration properties of vision–language models (VLMs) under distribution shifts. While VLMs have shown strong performance in zero-shot recognition across diverse scenarios, their reliability in risk-sensitive applications requires a better understanding of their uncertainty estimates. We analyze the calibration behavior of VLMs across different model architectures, datasets, and training strategies, focusing on settings where models are calibrated in one domain, label set, or hierarchy level and evaluated in another. Our results show that VLMs are not inherently well-calibrated, but simple post-hoc methods such as temperature scaling can significantly improve calibration, even under distribution shifts and changes in label sets. Moreover, reliable calibration can be achieved using only a small number of examples. These findings provide insights into improving the reliability of VLM predictions in real-world applications.

4.1 Motivation

Vision–language models (VLMs), such as CLIP [Radford et al., 2021] and ALIGN [Jia et al., 2021], have achieved remarkable results for a wide range of tasks, such as zero-shot image recognition [Wortsman et al., 2022], open-vocabulary object detection [Zhou et al., 2022c; Gu et al., 2021], image captioning [Yu et al., 2022a; Mokady et al., 2021] and egocentric perception [Zeng et al., 2022]. The burgeoning field of VLMs has been characterized by rapid exploration along various dimensions [Nguyen et al., 2022; Fang et al., 2022; Wortsman et al., 2022; Cherti et al., 2023; Tu et al., 2023a], such as dataset creation [Nguyen et al., 2022], reproducible scaling laws [Cherti et al., 2023], compositional relationships between objects and attributes [Yuksekgonul et al., 2022], robust fine-tuning approaches [Goyal et al., 2023], and visual factor-level robustness [Tu et al., 2023a].

However, their application in risk-sensitive domains necessitates a more rigorous understanding of their uncertainty estimation capabilities, an area that remains largely under-explored. Model *calibration* is concerned with ensuring that the model’s

predicted output probabilities correspond to its empirical frequency of correctness (*i.e.*, accuracy). For example, a calibrated model that classifies some images as “cow” with a 50% probability will have misclassified roughly half. Galil et al. [2023] and Minderer et al. [2021] report that CLIP models are better calibrated than other models trained on ImageNet. Notwithstanding this observation, Tu et al. [2023a] point out that they are not always well-calibrated and attribute this to the impact of the training data distribution and quantity. Building on this line of research, we study the calibration properties of various VLMs, each characterized by different architectures, datasets, and training strategies.

In this chapter, we investigate which factors affect the calibration of VLMs. Starting from prior research [Tu et al., 2023a] that demonstrates zero-shot CLIP can be well-calibrated with simple temperature scaling under distribution shifts, we extend this analysis to other CLIP variants and exemplar vision–language models. We then examine whether such a property persists when the calibration dataset varies in (1) distribution, (2) label set (*e.g.*, CIFAR-10 *vs.* ImageNet), (3) hierarchy level (*e.g.*, “Spider” *vs.* “Black widow”), (4) the number of images, and (5) feature-space distance of calibration set with respect to the target test set.

To this end, we evaluate 35 vision–language models. They have various image–text pre-training frameworks, such as CLIP [Radford et al., 2021] and BLIP [Li et al., 2022b]. They also have different visual encoder architectures (*e.g.*, ViT [Shankar et al., 2021] and ConvNeXt [Liu et al., 2022]) and training dataset distributions and quantities. We assay the uncertainty estimation of VLMs on three standard image classification benchmarks: ImageNet [Deng et al., 2009], CIFAR-10 [Krizhevsky et al., 2009] and DomainNet [Peng et al., 2019] and 5 types of distribution shift, including reproduction shift [Recht et al., 2019] and sketch shift [Wang et al., 2019]. Moreover, to study the sensitivity of our findings, we analyze the uncertainty estimation performance of VLMs when the quantity and quality of the calibration set are varied and the text prompts are hand-crafted or machine-generated. Our key observations are:

- VLMs do not exhibit better uncertainty estimation performance than other models, but become better calibrated than the other classes of models after calibration with temperature scaling – a single parameter variant of Platt Scaling [Platt et al., 1999];
- VLMs can be calibrated on datasets with different label sets than the target set and can be calibrated at a higher or lower level of the label hierarchy than the level of the target labels;
- VLMs require very few (*i.e.*, less than 100) samples for calibration;
- VLMs do not require sophisticated prompting strategies for calibration, with “a photo of a <class>” being sufficient to achieve good uncertainty estimation;
- Our findings motivate the use of a synthetic calibration set for VLMs in practical settings where labeled calibration data is lacking.

4.2 Definition and Notation

Let $\mathcal{Y} = 1, \dots, K$ and $\mathcal{X} = \mathbb{R}^d$ denote label and input spaces, respectively. A sample $(\mathbf{x}, y)$ from an unknown distribution in $\mathcal{X} \times \mathcal{Y}$ is input to a neural network classifier $\mathbf{f} : \mathcal{X} \rightarrow \Delta^K$. This classifier outputs a probability distribution over K classes for $\mathbf{x}$, where Δ^K is the $K - 1$ dimensional simplex. We assume $\mathbf{f}$ combines two functions: $\mathbf{f} =: \boldsymbol{\alpha} \circ \mathbf{g}$, where $\mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ is a non-probabilistic n -way classifier, and $\boldsymbol{\alpha} : \mathbb{R}^n \rightarrow \Delta_n$ is the softmax operator $\boldsymbol{\alpha}_i(\mathbf{z}) = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^n \exp(\mathbf{z}_j)}$ for $i \in \mathcal{Y}$. The output $\mathbf{g}(\mathbf{x})$ is referred to as the logits of $\mathbf{x}$ relative to $\mathbf{f}$. For any input instance $\mathbf{x}$, $\mathbf{f}$ assigns the predicted label $\hat{y} =: \arg \max_i f_i(\mathbf{x})$ and the corresponding confidence score $\hat{p} =: \max_i f_i(\mathbf{x})$.

Expected Calibration Error (ECE). A model is perfectly calibrated if $\mathbb{P}(\hat{y} = y \mid \hat{p} = p) = p$ for all p in $[0, 1]$, where y is the actual label, $\hat{y}$ the prediction, and $\hat{p}$ the confidence score. To assess model calibration, we typically use the Expected Calibration Error (ECE) [Guo et al., 2017], lower values indicating better calibration. ECE involves dividing samples into M equal bins by confidence scores, then computing the mean absolute difference between each bin’s accuracy and average confidence: $\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{avgConf}(B_m)|$, with n as the total number of samples

Temperature Scaling. Scaling logits from $\mathbf{g}$ with temperature T modifies output probability sharpness. The new prediction confidence is $\hat{p} = \max_i \frac{\exp(\mathbf{g}_i(\mathbf{x})/T)}{\sum_{j=1}^n \exp(\mathbf{g}_j(\mathbf{x})/T)}$. Higher T softens, and lower T sharpens probabilities. As T approaches 0 or infinity, probabilities trend towards a one-hot vector or uniform distribution, respectively. For a trained classifier $\mathbf{f}$, T is optimized using negative log-likelihood (NLL) on a calibration set. Since T does not impact the softmax maximum, $\hat{y}$, the predicted class, remains the same, preserving classification accuracy.

4.3 Experimental Setup

Compared VLMs. We consider 35 vision–language models. These models consist of exemplar VLMs, such as CLIP [Radford et al., 2021], Flava [Singh et al., 2022] and BLIP [Li et al., 2022b]. Particularly, we evaluate zero-shot CLIPs that are trained on different training distributions, such as the WIT [Radford et al., 2021] and LAION [Gadre et al., 2023] datasets, diverse dataset quantities from 3 million to 2 billion, and curated training datasets [Xu et al., 2024]. CLIP variants with different image encoders are also assessed, including ViT [Dosovitskiy et al., 2021a] and ConvNeXt [Liu et al., 2022] encoders, as well as CLIP variants that modify the training objective, such as SigLIP [Zhai et al., 2023], or the training strategy [Li et al., 2023b]. Unless specified, for each model, we use their default prompt sets from Radford et al. [2021].

Compared non-VLMs. We compare VLM calibration performance with models trained on ImageNet to show that VLMs are well-calibrated despite the distribution

shift after temperature scaling. We consider convolutional neural networks, such as ResNet [He et al., 2016a] and ConvNeXt [Liu et al., 2022], and vision transformers, exemplified by ViT [Dosovitskiy et al., 2021a] and Swin [Liu et al., 2021]. These models are trained solely on ImageNet [Deng et al., 2009] or pre-trained on a significantly larger dataset (e.g., ImageNet-21K [Ridnik et al., 2021]). All mentioned models are publicly accessible by OpenCLIP [Ilharco et al., 2021] and TIMM [Wightman, 2019].

Test sets. We evaluate the calibration of VLM on three standard image classification benchmarks: ImageNet [Deng et al., 2009], CIFAR-10 [Krizhevsky et al., 2009] and DomainNet [Peng et al., 2019]. Following the protocol in Gupta et al. [2021], we divide the validation set of ImageNet into two halves: one for the in-distribution (ID) test set, and the other for learning calibration methods. out-of-distribution (OOD) test sets are ImageNet-V2-A [Recht et al., 2019], ImageNet-R(endition) [Hendrycks et al., 2021a], ImageNet-S(ketch) [Wang et al., 2019], and ObjectNet [Barbu et al., 2019]. For CIFAR-10, its validation set is used for model calibration, and CIFAR-10.1, CIFAR-10.2 [Recht et al., 2018a] and CINIC [Darlow et al., 2018] are used for evaluation. The DomainNet benchmark utilizes the ‘Real’ domain for calibration and evaluates on ‘Painting’ and ‘Sketch’ domains. Note that ImageNet-R and ObjectNet use a reduced subset of classes; we follow the literature [Bello et al., 2021] to select subset of logits for these classes before evaluation.

Calibration method. For model calibration, we by default use temperature scaling [Guo et al., 2017] on calibration sets. We also experiment with the spline post-hoc calibration method [Gupta et al., 2021].

Metrics. (1) Calibration metric: we use ECE as the evaluation metric, where a lower score indicates better calibration performance. Throughout the experiments, we estimate ECE using equal-mass binning and 15 bins. (2) Correlation metric: to examine whether the calibrated prediction probabilities for all models exhibit a correlation with their classification accuracy, we use coefficients of determination R^2 [Nagelkerke et al., 1991] to measure the linearity and utilize Spearman’s rank coefficient ρ [Kendall, 1948] to measure monotonicity. R^2 ranges from 0 to 1, where an R^2 of 1 means that regression predictions perfectly correlate with model performance. The rank coefficient ρ spans $[-1, 1]$, where a value closer to 1 (or -1) indicates a better ranking index, while 0 indicates no correlation.

4.4 Factors Affecting Calibration

The cornerstone of safely deploying vision–language models (VLMs) lies in verifying their decision reliability. Specifically, the prediction probabilities provided by VLMs should accurately reflect their performance. With this objective, we have embarked on a comprehensive set of experiments. These experiments are designed to scrutinize the uncertainty estimation of VLMs under various conditions, including changes in

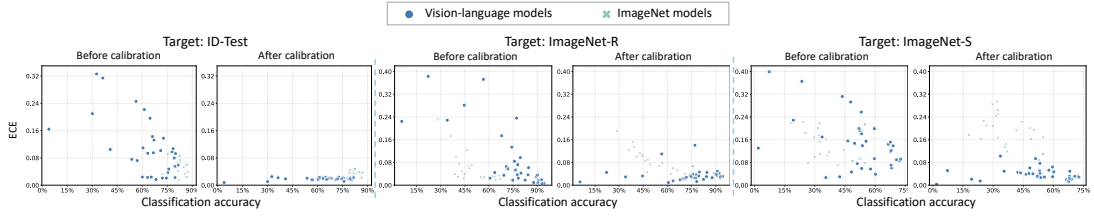


Figure 4.1: **Comparing calibration performance of ImageNet-trained models and VLMs.** We report results on the in-distribution test set (ID-Test) and two out-of-distribution (OOD) test sets: ImageNet-R and ImageNet-S. We plot expected calibration error (ECE) before and after temperature scaling for each model. The blue dots represent VLMs and the green crosses denote ImageNet-trained models. We observe that VLMs are well-calibrated by temperature scaling on both ID and OOD test sets.

(1) distribution, (2) label sets (*e.g.*, CIFAR-10 *vs.* ImageNet), (3) hierarchy levels (*e.g.*, “Spider” *vs.* “Black widow”), (4) the number of images in the dataset, and (5) the feature-space distance between the calibration and target test sets.

VLMs are well-calibrated after temperature scaling across various distributions.

Fig. 4.1 compares the calibration performance of VLMs with models trained on ImageNet. On both ID and OOD test sets (ImageNet-S and ImageNet-R), we observe that before calibration with temperature scaling, VLMs do not necessarily have superior uncertainty estimation performance. For instance, certain ImageNet models demonstrate a lower Expected Calibration Error (ECE) before calibration. However, once temperature scaling is applied, the scenario changes dramatically. VLMs show a marked decrease in their average ECE, dropping to 0.05, whereas ImageNet models see an increase in ECE, rising to 0.15. This suggests that VLMs may benefit more from the calibration process than ImageNet models.

Furthermore, while existing research highlights the challenges in achieving stable calibration results under distribution shifts [Yu et al., 2022b; Zou et al., 2023; Tomani et al., 2023; Ovadia et al., 2019], VLMs manage to maintain consistent and reliable uncertainty estimation after temperature scaling. This is evident in their performance on OOD test sets (ImageNet-S and ImageNet-R), where they exhibit competent uncertainty estimation. This enhanced calibration capability of VLMs, especially after temperature scaling, underscores their potential for more accurate and dependable decision-making in diverse applications.

4.4.1 Adaptability to Different Calibration Label Sets

VLMs can be calibrated on a dataset with a different label set from the target dataset. The zero-shot capability of VLMs facilitates their direct application to a diverse array of downstream classification tasks without the need for explicit training or fine-tuning. The conventional boundary between ID and OOD classes becomes less clear-cut for VLMs, suggesting their potential for cross-label-set calibration. To evaluate the effect of calibration label sets, we conduct experiments calibrating VLMs

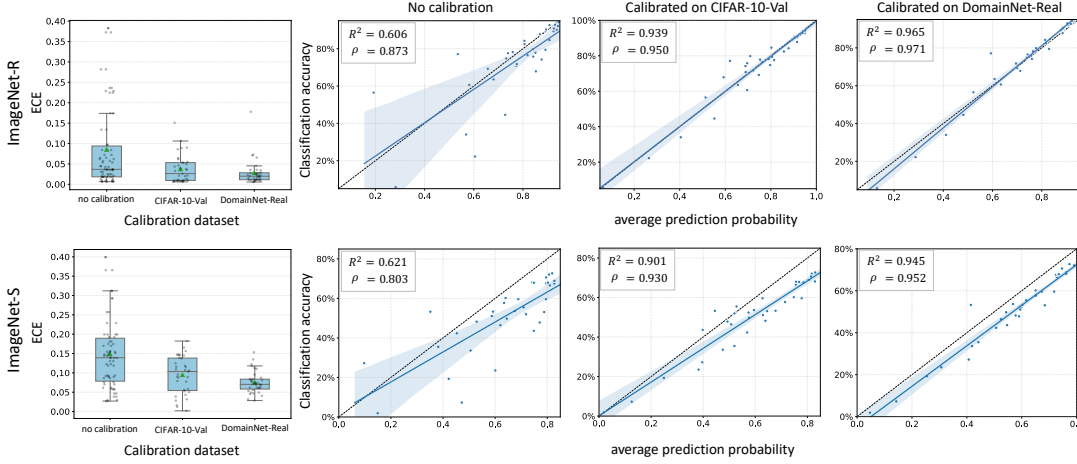


Figure 4.2: **Adaptability of VLMs to different calibration label sets.** **Left: Calibration error reduction.** We observe a significant decrease in expected calibration error for VLMs following cross-label-set calibration, as opposed to when no calibration is applied. **Right: Correlation between VLM prediction probability and classification accuracy.** This graph illustrates classification accuracy of VLMs on ImageNet-R and ImageNet-S against their average prediction probability, before and after calibration with CIFAR-10-Val or DomainNet-Real. Each point represents a model, with the dashed black line indicating perfect calibration ($y = x$). The data showcases a strong linear and rank correlation, even when models are calibrated on label sets different from the target, proving the effectiveness of cross-label-set calibration for VLMs.

on datasets with different label sets. The findings, depicted in Figure 4.2, demonstrate an improvement in the uncertainty estimation of VLMs when calibrated with alternative label sets. For instance, calibrating VLMs on CIFAR-10-Val or DomainNet-Real significantly reduces the ECE on ImageNet-R compared to when no calibration is applied. This trend of reduced ECE is consistently observed on ImageNet-S as well, further validating the effectiveness of the cross-label-set calibration. Furthermore, the calibrated prediction probability strongly correlates with model accuracy, with a linear and rank correlation over 0.90, despite the presence of non-zero ECE. This indicates that, even with label set differences, the calibrated prediction probability is predictive of the rankings of VLMs.

4.4.2 Calibration Across Semantic Hierarchy Levels

VLMs perform zero-shot classification by generating query embeddings for each novel class from their natural language names. Label hierarchy sets have been shown effective in enhancing model accuracy [Novack et al., 2023; Ren et al., 2023]. For example, mapping the predicted sub-class back to its parent to produce the final prediction [Novack et al., 2023]. However, there has been little attention on the influence of a dataset’s label hierarchy for calibration. This section aims to investigate whether VLMs can be effectively calibrated across different levels of a semantic hi-

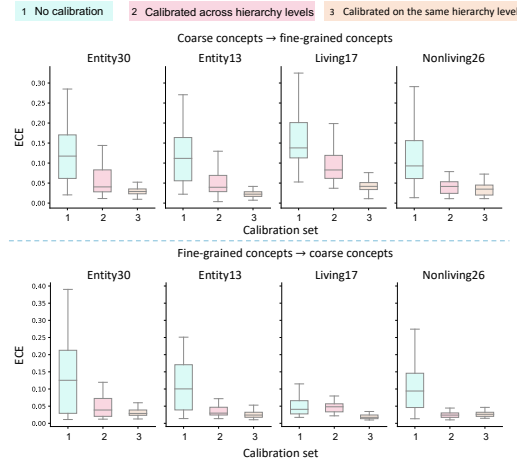


Figure 4.3: **Robustness of VLM calibration to label hierarchy levels.** This figure presents box plots summarizing the calibration errors (ECEs) of VLMs calibrated with label hierarchies differing in granularity from the target dataset (ImageNet-S). The top row shows calibration at a coarser level, and the bottom row at a finer level. Despite not matching the calibration precision of same-level calibration, the minimal differences indicate the robustness of VLM calibration to label granularity.

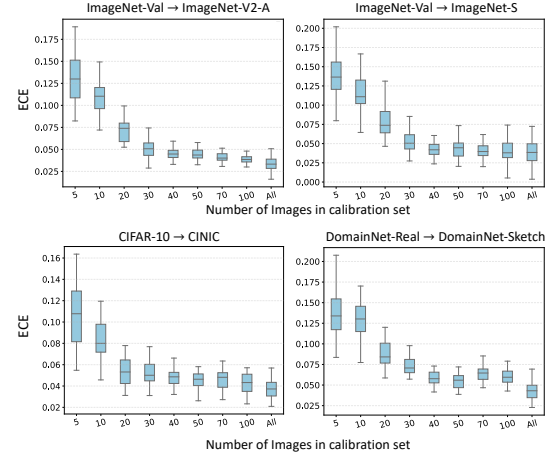


Figure 4.4: **Data-efficiency of VLM calibration across diverse datasets.** This figure displays the ECE of VLMs as a function of the calibration set size across four datasets: ImageNet-V2-A, ImageNet-S, CINIC, and DomainNet. The ECE values, averaged over ten random seeds, plateau after including merely 40–50 images in the calibration set, closely approximating the error obtained using the full set. This trend is observed despite the high number of classes in DomainNet and ImageNet, where many classes may not be represented even in the calibration set.

erarchy. Specifically, we examine whether VLMs can be calibrated using a dataset with coarsely-defined concepts (*e.g.*, “Bag”) while the target dataset comprises fine-grained concepts (*e.g.*, “Backpack”), and vice versa.

To conduct our evaluation, we use four label sets from BREEDS [Santurkar et al., 2020]—Entity13, Entity30, Living17, and Nonliving26—that define a hierarchical mapping between coarse and fine-grained classes. For each set, we adhere to the associated hierarchical mapping to selectively curate and relabel images sourced from the ImageNet validation set. This process yields a *calibration* dataset featuring coarse concepts. Subsequently, we curate the corresponding fine-grained class subset from ImageNet-S, thereby establishing a distinct *target* test dataset with a different distribution. We then repeat this procedure in reverse to calibrate with fine-grained concepts and test on coarse concepts.

VLMs can be calibrated across label hierarchy levels. Fig. 4.3 plots the expected calibration errors of VLMs calibrated at a different label hierarchy level than the target dataset. We see that for all four sets, the ECE decreases considerably when

calibrated at the ‘wrong’ level of the hierarchy. While not as well-calibrated as the models where the calibration and target sets were at the same level of the hierarchy, the difference is not substantial, especially for ‘Nonliving26’. This suggests that calibration is relatively robust to the granularity of the labels in the calibration set.

4.4.3 Effect of the Calibration Set Size

In this section, we investigate whether VLMs demand a large number of images for calibration. We evaluate VLMs on ImageNet-V2-A, ImageNet-S, CINIC and DomainNet. For each target test set, we randomly sample a certain number of images from the corresponding calibration set as described in Section 4.3. The performance for each model is measured by the averaged ECE over ten random seeds.

VLMs can be calibrated with a very small number of images. Fig. 4.4 plots the calibration error of VLMs with respect to the size of the calibration set for four different target datasets. The calibration error plateaus after 40-50 images and is already close to the error after calibration with the entire set. Notably, DomainNet and ImageNet are 345-way and 1000-way classification tasks, and so many classes do not have a single image present in the calibration set. This suggests that calibrating VLMs is very data-efficient: it is a low-dimensional problem.

4.4.4 Effect of Calibration-Target Set Distance

Ovadia et al. [2019] report that uncertainty estimation performance consistently degrades when calibration and test sets have distribution shifts. In this section, we explore whether the same applies to VLMs. Note that, since we have demonstrated that VLMs can be calibrated on a different label set, this experiment compares the calibration performance of VLMs calibrated on a dataset with the same or different label sets. The distribution discrepancy of datasets is computed by Fréchet inception distance (FID) [Heusel et al., 2017]. A zero FID means that the calibration set and target dataset follow a similar distribution.

Fig. 4.5 plots the calibration error of VLMs with respect to the FID between the calibration and target datasets for four target datasets: CIFAR-10.1, ImageNet-V2-A, ImageNet-S, and DomainNet-Painting. We find that FID alone does not entirely explain the calibration error. Instead, it is a combination of the FID and the label set similarity that is more predictive of calibration performance. This observation suggests that we should consider both the distribution shift and label set differences when selecting a calibration set.

4.4.5 Transferability of Calibration Across Prompts

VLMs classify images by comparing image features with class weights computed by text encoders, which take as input the textual prompts describing each class. This suggests that given the same textual class names, a better prompt set may improve

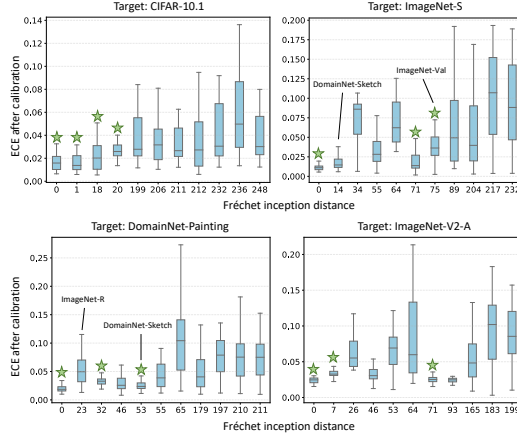


Figure 4.5: **Impact of the distance between calibration and target set distributions on VLM uncertainty estimates.** The distance between the calibration dataset and target dataset is computed by Fréchet inception distance [Heusel et al., 2017]. A green star indicates that the dataset for this column has the same label set as the target dataset. We find that the calibration error has a weak correlation with the FID between the calibration and target datasets, however the label set compatibility also plays a significant role.

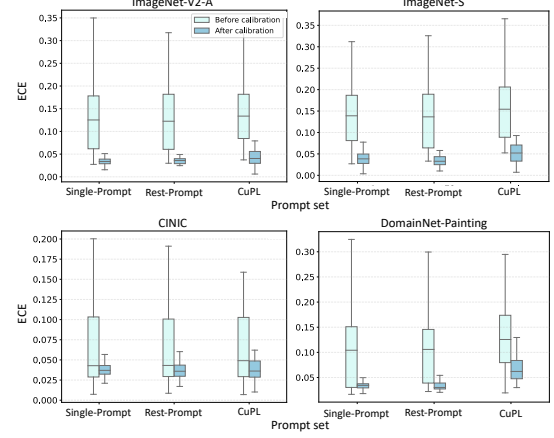


Figure 4.6: **Efficient calibration with single prompts for VLMs.** The figure illustrates the transferability of calibration effectiveness when using a single prompt for VLMs. The temperature scalar estimated using one random prompt, denoted as “Single-Prompt” remains highly effective when applied to other human-designed prompts (“Rest-Prompt”) and machine-generated prompts (“CuPL”). This finding highlights the efficiency of single prompts in achieving robust calibration for VLMs.

a VLM’s discriminative ability, and using specific prompt contexts for the style of images may also enhance model performance [Pratt et al., 2023; Zhou et al., 2022b]. In previous experiments, we use the same set of prompts for classification and calibration. Here, we question whether a temperature scalar determined by one set of prompts can effectively calibrate a different set of prompts used for classification.

To answer this question, we conduct experiments where we randomly selected a single prompt from a larger set for calibration. Surprisingly, we found that calibrating with just a single prompt yielded effective results. The temperature scalar discovered through this single prompt calibration extended its effectiveness to both the rest of human-designed prompts (Rest-Prompt) and machine-generated prompts (CuPL), as depicted in Fig. 4.6. This discovery highlights that VLMs do not necessarily require complex and tailored prompting strategies for calibration. Instead, a straightforward prompt such as “a photo of a <class>” suffices for effective calibration.

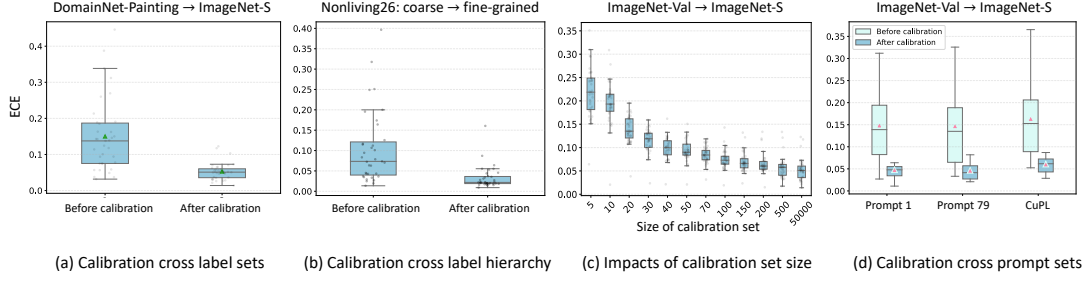


Figure 4.7: **Uncertainty estimation performance of VLMs with spline calibration.** Performance of VLMs using the spline method [Gupta et al., 2021] across four factors: (a) label sets, (b) label hierarchy, (c) calibration set size, and (d) prompt sets. Our results demonstrate that phenomena observed with temperature scaling persist when employing spline calibration. For example, VLMs can be effectively calibrated on datasets with differing label sets or label hierarchies compared to the target test set.

4.4.6 Impact of the Calibration Method

In addition to temperature scaling, we have undertaken an evaluation of another post-hoc calibration method: spline [Gupta et al., 2021]. Spline calibration involves deriving a function through spline-fitting, directly aligning classifier outputs with calibrated probabilities. In this section, we investigate whether the observed uncertainty estimation properties of VLMs hold in the spline setting.

In Fig. 4.7, we present our findings, which indicate that our observations regarding the use of temperature scaling for calibration extend to splines as well. For instance, we can observe that when calibrating VLMs using a dataset with different label sets or hierarchy levels than the target test set, similar trends and outcomes persist. This consistency across different post-hoc calibration methods reinforces the robustness of our observations.

4.5 Application: Calibration-by-Synthesis

In this section, we take some of the calibration robustness findings from Section 4.4 and demonstrate how they might be applied in a realistic scenario. We consider the setting where a VLM-based classifier is to be deployed in a new target domain, but labeled data in that domain is unavailable due to reasons of cost, time, or expertise. Our previous findings give us confidence that such a classifier can still be calibrated even if the calibration set and target domain have a significant distribution shift, if the hierarchy level of the labels differs, and if the calibration set is small.

Our approach is as follows. First, we construct a set of text descriptions for each class by prompting GPT-3 [Brown et al., 2020], a large language model, following the approach of CuPL [Pratt et al., 2023]. Second, we feed these descriptions to Stable Diffusion [Rombach et al., 2022b], a text-to-image model, to synthesize images associated with the descriptions. These are used as a synthetic, automatically-labeled calibration set. We consider two situations: (i) the target task is under-specified, so

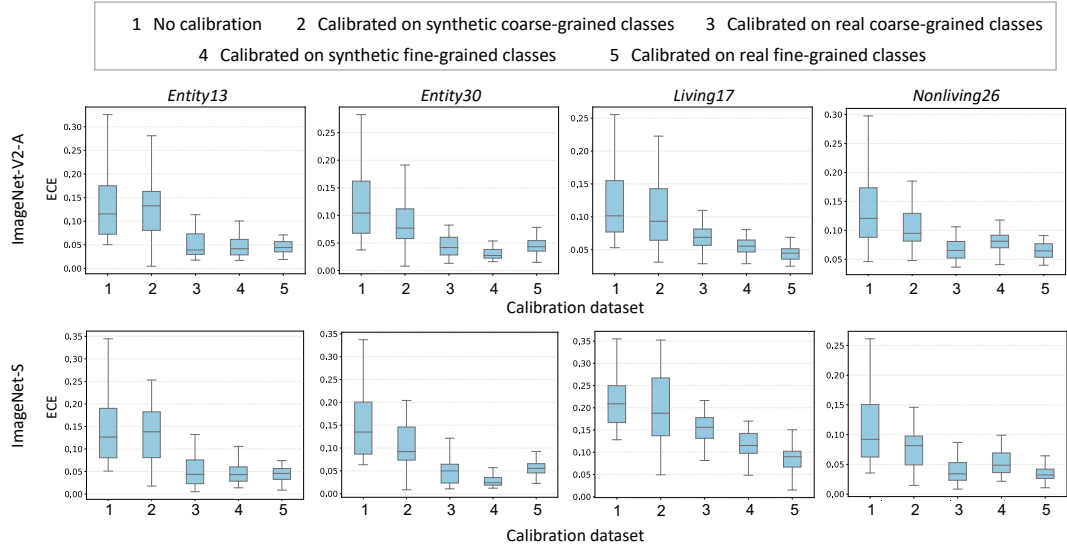


Figure 4.8: Can VLMs be calibrated without labeled data? A calibration-by-synthesis approach. Each column represents a pre-defined coarse-fine hierarchical mapping, and each row corresponds to a test set formed according to this class hierarchy. The uncertainty estimation performance of VLMs is plotted grouped by different calibration datasets: (1) no calibration; (2) a synthetic dataset generated from coarse-grained class labels; (3) a real dataset with coarse-grained labels; (4) a synthetic dataset generated from fine-grained classes; and (5) a real dataset with fine-grained classes. We observe that calibrating on a synthetic dataset with fine-grained classes is a successful strategy in reducing calibration error and is competitive with the performance of using real data for calibration. This indicates that using a small number of synthetic images, generated from and calibrated with detailed and specific class labels, suffices to calibrate VLMs.

we only have coarsely-defined target classes at calibration time; and (ii) the target task is fully-specified, so we have access to the full (fine-grained) target label set at calibration time. For the former, the calibration set has 5 images per coarse class. For the latter, the calibration set has 1 image per fine-grained class. To obtain the coarse-fine label mapping, we use the defined hierarchies (Entity13, Entity30, Living17, and Nonliving26) in the BREEDS benchmark [Santurkar et al., 2020] and form the corresponding fine-grained class subsets of ImageNet-V2-A and ImageNet-S for the two target test datasets.

In Fig. 4.8, we plot the ECE, evaluated on the target dataset with fine-grained class labels, for a set of VLMs grouped by the calibration procedure: (1) no calibration; (2) calibration on a synthetic dataset generated from coarse-grained class labels; (3) calibration on a real labeled dataset (the ImageNet validation set) with coarse-grained class labels; (4) calibration on a synthetic dataset generated from fine-grained class labels; and (5) calibration on a real labeled dataset with fine-grained class labels. Results across four fine-grained label sets show that calibration with synthetic images based on fine-grained classes is remarkably effective, often outperforming calibration

Label set	No calibration			Synthetic calibration		
	MECE	MAE	ρ	MECE	MAE	ρ
Entity13	0.15	15.21	0.77	0.05	5.07	0.93
Entity30	0.16	16.12	0.75	0.04	3.57	0.94
Living17	0.22	21.90	0.74	0.12	11.45	0.94
Nonliving26	0.13	12.73	0.79	0.05	5.29	0.91
Average	0.17	16.49	0.76	0.07	6.35	0.93

Table 4.1: **Uncertainty estimation performance on ImageNet-S for calibration-by-synthesis.** We report the mean expected calibration error (MECE), the mean absolute error (MAE) of the average confidence with respect to model accuracy, and Spearman’s rank correlation (ρ). Calibrating on a synthetic set with fine-grained classes improves the informativeness of prediction probabilities for estimating model performance and ranking.

with real data. For each label set, calibration using fine-grained synthetic classes (Dataset 4) consistently yields lower ECE values compared to uncalibrated models (Dataset 1). Moreover, coarse synthetic label calibration (Dataset 2) is less effective due to compounded domain shift and label granularity discrepancies.

Beyond model calibration, we showcase the tangible benefits of using calibrated prediction probabilities for choosing models and estimating their accuracy, particularly when labeled data is scarce or absent. Table 4.1 assesses how predictive the estimated VLM probabilities are for the model’s performance, before and after calibration with a synthetic dataset with fine-grained classes. Calibration improvements are evident: a reduction in both the mean expected calibration error (MECE) and the mean absolute error (MAE) of average confidence in relation to actual model accuracy, along with a boost in rank correlation with model accuracy. Overall, the above analysis shows how our findings about calibration can be applied to a real scenario where labeled calibration data is not available.

4.6 Summary

In this study, we investigated the factors that affect the uncertainty estimation performance of Vision–Language Models (VLMs). Our experiments, spanning various facets of uncertainty estimation, have revealed insights into the strengths of VLMs. Notably, when coupled with temperature scaling as a calibration method, VLMs surpass other models in their ability to estimate uncertainty accurately. This finding is particularly useful for tasks that require precise uncertainty estimates. Furthermore, we have demonstrated VLMs’ remarkable adaptability—they can be effectively calibrated with datasets that have different label sets or label hierarchy levels.

Additionally, VLMs exhibit efficiency by maintaining calibration quality with a limited number of images and simplified prompts. Real-world applications confirm their potential, showing that VLMs can be calibrated with a small number of syn-

thetic images. Looking ahead, we anticipate exciting avenues for research, including exploring alternative calibration methods and investigating uncertainty estimation properties in domains beyond image classification. These insights, we believe, pave the way for more robust and reliable VLMs, contributing to the broader landscape of algorithmic design and multi-modal understanding.

Label-Free Model Ranking based on Softmax Prediction Probability

After evaluating model reliability from various perspectives, this chapter considers a practical and challenging scenario where test set labels are unavailable. Specifically, we study how to rank classifier performance on unlabeled data under distribution shifts. We first show that conventional uncertainty metrics, such as maximum softmax prediction probability, provide useful signals for predicting model generalization in out-of-distribution (OOD) settings. Building on this observation, we propose a new measure, Softmax Correlation (SoftmaxCorr), which evaluates the similarity between a class–class correlation matrix derived from softmax predictions on unlabeled data and a predefined reference matrix representing ideal class correlations. A higher similarity indicates that the model produces confident and consistent predictions across categories. Extensive experiments on datasets including ImageNet, CIFAR-10, and WILDS demonstrate that SoftmaxCorr effectively predicts relative model performance in both in-distribution and OOD scenarios. We also discuss the limitations of this measure and potential directions for future research.

5.1 Motivation

Machine learning (ML) models typically excel on test sets coming from the same distribution as the training set. However, this assumption seldom holds in real-world deployments, where the test environments often experience distribution shifts caused by factors such as sample bias and non-stationarity. Recognizing this challenge, there is a pressing need to assess ML model performance in unlabeled testing environments where traditional evaluation metrics which require data annotations, such as accuracy and $F1$ score, may become infeasible. Hence, our focus shifts towards the vital yet under-explored task of ranking models under these conditions. Specifically, given a pool of trained models and an unlabeled test set, the objective is to efficiently select the most suitable model for utilization.

Various complexity measures have been proposed to predict the in-distribution (ID) accuracy of models [Jiang et al., 2020b; Neyshabur et al., 2015; Bartlett et al., 2017; Keskar et al., 2017; Nagarajan and Kolter, 2019; Neyshabur et al., 2017a; Chuang

et al., 2021; Jiang et al., 2020c; Smith and Le, 2017; Arora et al., 2018b; Dziugaite and Roy, 2017b; Dinh et al., 2017]. For the OOD test sets, classical domain adaptation theory provides a partial answer by quantifying the distance between the ID and OOD distributions [Ben-David et al., 2006b]. Moreover, the “accuracy-on-the-line” phenomenon [Miller et al., 2021] shows the linear correlation in the probit scale between ID and OOD performance. However, such phenomenon does not always hold on some distributions [Teney et al., 2024]. Building on previous research, our objective is to develop a robust measure capable of ranking models on both ID and OOD datasets *without* testing labels.

Softmax prediction probability has been shown to be useful in analyzing test data in several tasks, such as open-set data detection [Hendrycks and Gimpel, 2016], accuracy estimation [Guillory et al., 2021; Garg et al., 2022a], and misclassified input detection [Deng et al., 2022a]. For example, Hendrycks and Gimpel [2016] and Liang et al. [2018] utilize maximum Softmax prediction probability to identify samples from open-set classes. Driven by these, we develop OOD measures based on Softmax probability. Concretely, given various deep models, we aim to develop probability-based measures that monotonically relate to OOD generalization. To validate the feasibility, we conduct extensive and large-scale correlation studies using various models and different types of dataset shifts. We construct a catalog of empirical prediction probability-based measures and create a wide range of experimental setups. We collect 573 different classification models ranging from standard convolutional neural networks to vision transformers. We cover 11 ID and OOD datasets with various distribution shifts, such as ImageNet-V2 [Recht et al., 2019] with dataset reproduction shift and ImageNet-R [Hendrycks et al., 2021a] with style shift.

Based on experimental results, we first show the demand for measures to rank model performance beyond in-distribution accuracy [Miller et al., 2021]. Then, we observe that empirical uncertainty measures based on prediction probabilities (*e.g.*, maximum softmax probability) are useful in characterizing OOD generalization to some extent. However, we note that these measures solely account for prediction certainty. In response, we introduce SoftmaxCorr, a method designed to harness both prediction certainty and diversity. Confidence pertains to the certainty of individual predictions, whereas dispersity indicates the spread of predictions across all categories. Specifically, for each classifier, we compute a class-class correlation matrix from all prediction vectors in a test set. Then, we calculate its cosine similarity with a predefined reference matrix, designed to represent desirable prediction patterns, to evaluate whether this classifier makes diverse and certain predictions. The broad correlation study shows the efficacy of SoftmaxCorr in ranking models.

5.2 Task Formulation

Task definition. We consider a K -way classification task, and let $\mathcal{Y} = \{1, \dots, K\}$ denote the label space and $\mathcal{X} \in \mathbb{R}^d$ denote the input space. We are given a labeled training set $\mathcal{D}^S$ that contains data i.i.d drawn from a source distribution P_S , and an OOD test

set $\mathcal{D}^T := \{(x_i, y_i)\}_{i=1}^N$ that contains N data i.i.d drawn from another distribution P_T ($P_S \neq P_T$). We train M neural network classifiers $\{\phi_m\}_{m=1}^M$ on $\mathcal{D}^S$. Given a sample (x, y) from $\mathcal{D}^T$, the classifier $\phi_m : \mathcal{X} \rightarrow \Delta^K$ gives Softmax probabilities for x on K classes, where Δ^K denote $K - 1$ dimensional unit simplex. By testing on $\mathcal{D}^T$, ϕ_m yields a prediction matrix $\mathbf{P} \in \mathbb{R}^{N \times K}$, whose each row represents prediction probabilities of a test data. Specifically, the prediction matrix satisfies $\sum_{j=1}^K P_{i,j} = 1 \ \forall i \in 1 \dots N$ and $P_{i,j} \geq 0 \ \forall i \in 1 \dots N, j \in 1 \dots K$, where $P_{i,j}$ indicates the probability that i -th sample is predicted to the j -th class.

The dataset has an evaluation metric (e.g., accuracy) to obtain ground-truth generalization G_m of classifier ϕ_m . The goal is to design a measure to calculate a score S_m for each classifier ϕ_m without access to data annotations. The calculated scores $\{S_m\}_{m=1}^M$ ideally should correlate well with $\{G_m\}_{m=1}^M$, so that we can rank the OOD generalization of models based on the scores.

Evaluation metrics. We use Spearman’s Rank Correlation coefficient ρ [Kendall, 1948] to measure monotonicity between calculated scores and model generalization. In addition, we compute the weighted variant of Kendall’s rank correlation τ_w , which is shown to be a useful measure when selecting the best-ranked item of interest [You et al., 2021]. Both coefficients range from $[-1, 1]$. A value closer to -1 or 1 indicates a strong negative or positive correlation, respectively, and 0 means no correlation. Similar to Miller et al. [2021] and Baek et al. [2022], we apply the same probit scale to both accuracy and SoftmaxCorr in our experiment for a better linear fit.

5.3 Softmax Probability-based OOD Measures

5.3.1 What Makes OOD Measures Interesting?

Beyond Accuracy-on-the-Line (AoL). Miller et al. [2021] report an AoL phenomenon where there exists a strong linear correlation between probit-scaled ID and OOD generalization. This implies that ID accuracy is a good predictor of OOD generalization. However, delving deeper into OOD measures is warranted for three compelling reasons. **First**, Miller et al. [2021] discuss that AoL is not universal. That is, on some datasets, ID, and OOD accuracy do not show a clear positive correlation. This point is further discussed by Wenzel et al. [2022]. Specifically, Wenzel et al. [2022] suggest two patterns preventing this phenomenon: (1) underspecification (e.g., Camelyon17) where the same ID performance leads to different OOD behavior; (2) models do not transfer information from ID to OOD domains (e.g., DomainNet). That is, despite various ID performance, all models perform poorly on OOD datasets. Paradoxically, identifying the failure patterns itself requires labeled OOD data [Teney et al., 2024]. **Second**, it is demanding and sophisticated to design ID test sets [Engstrom et al., 2020], which are expected to be unbiased and representative of distribution to effectively measure model ID accuracy. Further, it is a trade-off to split a full dataset into training, validation and test sets in terms of training and evaluation quality. **Third**, recent advancements in vision-language models, such as CLIP [Radford et al., 2021] and BLIP [Li et al., 2022b], have achieved remarkable zero-shot classification perfor-

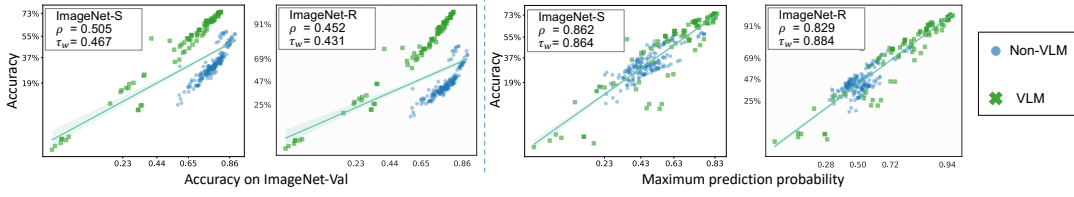


Figure 5.1: **Correlation study between MaxPred and accuracy (%) on ImageNet-S and ImageNet-R.** Every point denotes a classifier. We use 173 ImageNet models and 89 vision-language models introduced in Section 5.4. The straight line is fit with robust linear regression [Huber, 2011] and the shaded region indicates the 95% Clopper-Pearson confidence intervals. We show that MaxPred exhibits a moderate correlation with accuracy, while accuracy on ImageNet-validation shows a relatively low correlation with performance on ImageNet-R. Moreover, Vision-Language Models (VLMs) exhibit varying linear trends in terms of their ID and OOD accuracy compared to standard supervised models.

mance. This means that collecting a dataset for a specific task to train such models is not required. Due to their diverse and large-scale training data, it might not be suitable to use a proxy dataset (e.g., ImageNet validation set) to reflect their performance. As shown in Fig. 5.1, Vision-Language models (VLM) exhibit varying linear trends in terms of their ID and OOD accuracy compared to standard supervised models. Furthermore, even among VLMs themselves, consistency in these linear trends is not always guaranteed. These findings suggest that AoL alone does not suffice to accurately rank CLIP models.

Why Use Softmax Prediction Probability? Neural networks often exhibit a tendency to provide overly confident predictions, as evidenced by various studies [Ovadia et al., 2019; Minderer et al., 2021; Guo et al., 2017; Hein et al., 2019]. Initially, this characteristic might raise doubts about the reliability of using Softmax Prediction as a measure of uncertainty on test data. However, existing research has shed light on its informative nature when analyzing test environments. For instance, Hendrycks and Gimpel [2016] demonstrated that the maximum Softmax prediction probability (MaxPred) for correctly classified samples tends to be higher than that of incorrectly classified or out-of-distribution (OOD) samples. This observation has paved the way for utilizing MaxPred in error/success prediction and ID/OOD detection. Building upon this insight, Guillory et al. [2021] and Garg et al. [2022a] have proposed leveraging MaxPred to estimate the accuracy of trained classifiers on test samples.

Moreover, Softmax probabilities can be efficiently computed without requiring any changes to the network architecture or additional data preprocessing. Additionally, they can be derived solely from unlabeled data. These characteristics endow prediction probability-based measures with substantial practical value as reliable indicators of OOD generalization. Inspired by the above discussion, this chapter aims to validate the feasibility of using model output-based methods (such as softmax probability) for assessing and ranking models.

Proof of concept. The above works imply the prediction probability is likely to be effective in measuring the OOD performance of a pool of models. Given an

OOD test set (ImageNet-R) and various ImageNet models, we conduct a correlation study between MaxPred and classification accuracy. In Figure 5.1, we show that there is a relatively strong correlation between MaxPred and model accuracy ($\rho = 0.829$ and $\tau_w = 0.884$) on ImageNet-R. This indicates that MaxPred is feasible in ranking OOD performance. Based on this observation, we further explore more empirical prediction probability-based measures and develop a more effective measure that exploits more semantics reflected in the prediction probability.

5.3.2 Exploring More Empirical Measures

In addition to accuracy-on-the-line (AoL) and maximum prediction probability (MaxPred), we investigate other empirical measures as follows:

Average Thresholded Confidence with Maximum Confidence (ATC-MC) [Garg et al., 2022a]. This method is used to predict the performance of a specific trained model under distributional shift. Here, we deploy it to rank various models' performance on one particular OOD test set. It firstly identifies a threshold t on $\mathcal{D}^S$ such that the number of samples with confidence score lower than t is equal to the model's error. Then, the ATC on $\mathcal{D}^T$ is given by the number of points whose confidence is less than t . The formula is: $ATC = \mathbb{E}_{x \sim \mathcal{D}^T} [\mathbb{I}[\arg \max_{j \in \mathcal{Y}} \mathbf{P}_{:,j} < t]]$, where $\mathbb{I}[E]$ is the binary indicator of event E .

Softmax Gap (SoftGap). MaxPred only uses the maximal value of Softmax vectors while disregarding values on other entries. Inspired by Baldock et al. [2021], we introduce SoftGap based on MaxPred which further considers the second-largest entry in a prediction vector. Specifically, it calculates the average difference between the largest and second-largest Softmax prediction probabilities over all samples. A high SoftGap indicates a confident prediction, while a low SoftGap suggests confusion between the two classes corresponding to the highest and second-highest probabilities.

5.3.3 Ours: Softmax Correlation (SoftmaxCorr)

Class-class correlation matrix Given the prediction matrix $\mathbf{P} \in \mathbb{R}^{N \times K}$ predicted by ϕ_m , a class correlation matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$ is computed by $\mathbf{C} = \frac{\mathbf{P}^\top \mathbf{P}}{N}$. An entry $C_{i,j}$ indicates a correlation between prediction probabilities of class i and class j over all samples. We define the sum of diagonals of the correlation matrix as intra-class correlation (I_a) and the sum of off-diagonals as inter-class correlation (I_e). The sum of $\mathbf{C}$ is 1, which means $I_a = 1 - I_e$. Moreover, $\text{trace}(\mathbf{P}^\top \mathbf{P}) = \sum_{i=1}^N \sum_{j=1}^K P_{i,j}^2 = \|\mathbf{P}\|^2$, where $\|\cdot\|$ is the Frobenius norm of a matrix. Thus, we can derive $I_a = \frac{\|\mathbf{P}\|^2}{N}$. According to Cui et al. [2020], $\|\mathbf{P}\|$ has strict opposite monotonicity, and the minimum of the entropy $H(\mathbf{P})$ and the maximum of $\|\mathbf{P}\|$ could be achieved at the same value. Therefore, I_a reflects the certainty of predictions.

Motivation: In domain generalization, it has been observed that class-class correlation encodes class confusion, thereby offering the potential for regularization to enhance model generalization ability [Chen et al., 2022; Jin et al., 2020; Li et al., 2021; Cui et al., 2020]. Additionally, existing literature underscores the significance of both

high prediction certainty and prediction diversity in identifying discriminative features [Yang et al., 2021; Wang and Isola, 2020; Asano et al., 2019].

Inspired by these insights, we can propose an OOD metric, which leverages a class-class correlation matrix to take into account the two characteristics:

- Prediction certainty: the model’s ability to produce confident predictions. Given the strict opposite monotonicity between I_a and entropy $H(P)$, prediction certainty can be reflected by a high I_a and consequently a low $I_e = 1 - I_a$ within the class correlation matrix C .
- Prediction diversity: relying solely on prediction confidence may be inadequate. For instance, a model may exhibit high certainty yet be biased towards a single class, making the diagonal entries of C skewed. In such cases, while prediction confidence is high, it does not necessarily indicate high model performance. Driven by this, we further evaluate prediction diversity: each class should be predicted and involved in predictions. To measure this, we expect the diagonals of C to be well-distributed across all classes instead of biased towards few classes.

Combining two characteristics, a desirable class correlation matrix has zeros on the off-diagonal entries, and values on diagonal elements match the distribution of classes. The similarity between the class correlation matrix of an evaluated model and the desirable class correlation matrix reflects the model performance. **Implementation:** To achieve this, we propose **SoftmaxCorr** and define it as the cosine similarity between the class-class correlation matrix C and a reference matrix R : $\cos(C, R) = \frac{\sum_{ij} (C \odot R)_{ij}}{\|C\| \cdot \|R\|}$, where $\odot$ is the element-wise product between matrices. The reference matrix captures the essence of the desirable class correlations: it is a diagonal matrix whose off-diagonal elements are 0 while its diagonal entries mirror the class marginal distribution. To approximate this class distribution, we use the average prediction probability across test data generated by a zero-shot vision-language model (ViT-bigG/14-CLIPA). The range of the SoftmaxCorr is $[0, 1]$. For a fixed reference matrix, the maximum is obtained 1) when a model confidently assigns probability 1 to the predicted class and 2) the proportion of predictions for each class matches the class distribution. The minimal value is achieved when a model is extremely biased towards one specific class with the prediction probability of one, while the estimated distribution for such a class is zero.

5.4 Experiments

In this section, we first describe three setups including ImageNet, CIFAR-10, and WILDS. Then, we analyze the experiment results of prediction probability-based measures on three setups. After that, we study the impacts of class distribution estimator on predictive ability of SoftmaxCorr. Furthermore, we study whether SoftmaxCorr can rank the performance of model checkpoints along the training trajectory. Also, we validate the effectiveness of SoftmaxCorr under the domain adaption setting. Lastly, we study the correlation between SoftmaxCorr and the accuracy of a single model on various OOD test sets.

Setup	Dataset	Validation Required				Validation Free					
		AoL		ATC-MC		MaxPred		SoftGap		SoftmaxCorr	
		ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w
ImageNet	ImageNet-V2	0.954	0.911	0.994	0.961	0.711	0.597	0.796	0.644	0.921	0.758
	ImageNet-A	0.593	0.636	0.830	0.895	0.756	0.813	0.805	0.846	0.964	0.915
	ImageNet-R	0.452	0.431	0.950	0.954	0.829	0.884	0.902	0.911	0.951	0.928
	ImageNet-S	0.505	0.467	0.981	0.959	0.862	0.864	0.887	0.871	0.935	0.909
	ObjectNet	0.619	0.545	0.961	0.821	0.883	0.849	0.908	0.865	0.963	0.895
	ImageNet-Blur	0.637	0.576	0.937	0.905	0.816	0.821	0.844	0.845	0.961	0.907
	Average	0.627	0.594	0.942	0.916	0.810	0.805	0.857	0.830	0.949	0.885
CIFAR-10	CIFAR-10.2	0.983	0.949	0.992	0.966	0.837	0.809	0.854	0.818	0.894	0.836
	CINIC	0.866	0.890	0.952	0.902	0.665	0.663	0.690	0.681	0.821	0.763
	CIFAR-10-Noise	0.641	0.765	0.219	0.461	0.051	0.131	0.139	0.213	0.931	0.917
	Average	0.830	0.868	0.721	0.776	0.518	0.534	0.561	0.571	0.892	0.846
WILDS	Camelyon17-OOD	-0.021	-0.072	-0.111	-0.075	0.192	0.320	0.192	0.320	0.420	0.560
	DomainNet-OOD	0.350	0.219	0.513	0.254	0.403	0.274	0.407	0.258	0.740	0.680
	Average	0.165	0.074	0.201	0.217	0.298	0.597	0.300	0.289	0.580	0.620
Average over three setups		0.598	0.575	0.747	0.751	0.637	0.693	0.675	0.609	0.864	0.824

Table 5.1: **Method comparison on ImageNet, CIFAR-10, WILDS and DomainNet.** We compare SoftmaxCorr with four measures: accuracy-on-the-line (AoL) [Miller et al., 2021], average thresholded confidence with maximum confidence (ATC-MC) [Garg et al., 2022a], MaxPred [Hendrycks and Gimpel, 2016] and SoftGap [Baldock et al., 2021]. Spearman’s rank correlation (ρ) and weighted Kendall’s correlation (τ_w) are metrics. The highest correlation in each row is highlighted in **bold** and the second highest is in **blue**. We show that our method is stable and yields the highest average correlations over three setups.

5.4.1 Experimental Setup

ImageNet setup. We collect 173 models publicly accessible from TIMM [Wightman, 2019]. They are trained or fine-tuned on ImageNet [Deng et al., 2009] and have various architectures, training strategies and training paradigms. We use five OOD datasets for the correlation study. Specifically, OOD datasets are: (1) ImageNet-V2 [Recht et al., 2019]; (2) ObjectNet [Barbu et al., 2019]; (3) ImageNet-S(ketch) [Wang et al., 2019]; (4) ImageNet-Blur severity 5 [Hendrycks and Dietterich, 2019a]; (5) ImageNet-R(endition) [Hendrycks et al., 2021a]; ImageNet-R and ObjectNet contain 200 and 113 ImageNet classes respectively.

CIFAR-10 setup. We collect 65 networks trained with the scheme provided by Wightman [2017] on CIFAR-10 training set [Krizhevsky et al., 2009]. These models have different model architectures. CIFAR-10-Val(idation) is the ID test set. For OOD datasets, we use (1) CIFAR-10.2 [Recht et al., 2018a] (2) CINIC [Darlow et al., 2018] (3) CIFAR-10-Noise with severity 5 [Hendrycks and Dietterich, 2019a]. We use accuracy as the metric of model generalization.

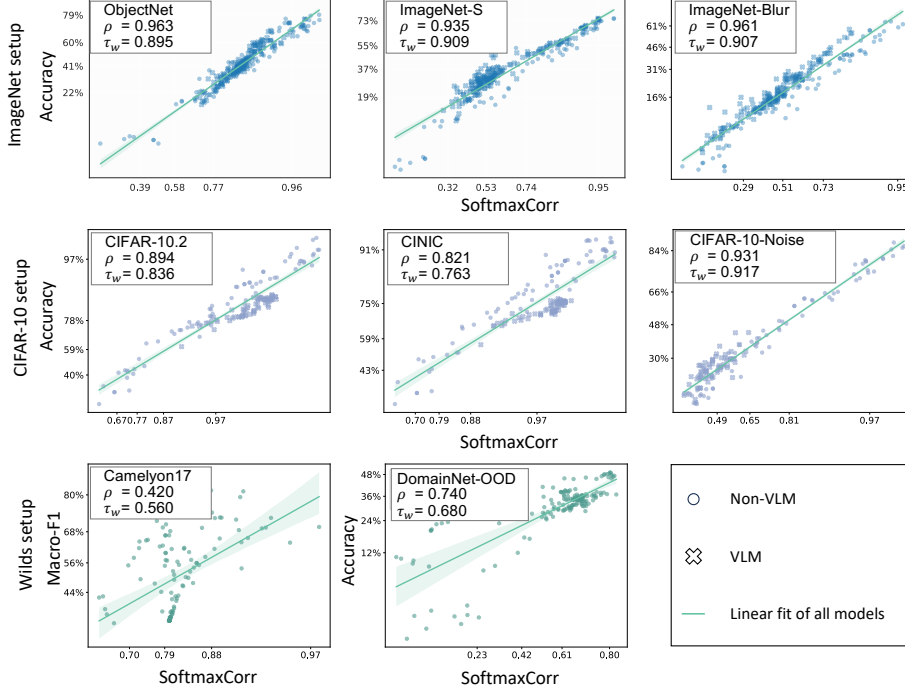


Figure 5.2: **SoftmaxCorr vs. model generalization under ImageNet, CIFAR-10 and WILDS setups.** In each subfigure, each point denotes a model trained for the corresponding task. For ImageNet setup, OOD test sets are ObjectNet and ImageNet-S and ImageNet-Blur. For CIFAR-10 setup, OOD test sets are CIFAR-10.2, CINIC and CIFAR-10-Noise. For WILDS, OOD test sets are Camelyon17-OOD and DomainNet-OOD. The y-axis is top-1 accuracy, top-1 accuracy and macro-F1 for the three setups, respectively. Straight lines are fit with robust linear regression [Huber, 2011]. Axes are probit scaled as described in Section 6.2. We observe that SoftmaxCorr is a reliable and effective metric. Particularly on ImageNet, SoftmaxCorr is predictive of model generalization with strong performance ($\rho > 0.92$).

WILDS setup. We consider a classification task of this setup: Camelyon17 [Bandi et al., 2018]. It is a binary classification dataset where the objective is to classify whether a slide contains a tumor tissue. We use 45 models varying in architectures and random seeds. ID and OOD datasets are the default ID validation set and OOD test set respectively. For each task, we follow the same training scheme provided by Koh et al. [2021] to train or fine-tune models.

Zero-shot vision language models. In addition to models which are trained on the ID training dataset, we also consider 89 zero-shot vision-language models, including CLIP [Radford et al., 2021], SigLIP [Zhai et al., 2023], BLIP [Li et al., 2022b], BLIP-2 [Li et al., 2023a] and Flava [Singh et al., 2022]. We use default prompt set for corresponding models. If the default prompt sets are not provided, “A picture of {class}.” is deployed. Unless specified, we use ViT-bigG/14-CLIPa to estimate class distribution for all setups.

5.4.2 Main Observations

Albeit requiring no access to validation/proxy set, SoftmaxCorr exhibits a strong correlation with model generalization. In Figure 5.2 and Table 6.1, we observe that SoftmaxCorr is indicative of model performance under the three setups. Particularly on the ImageNet setup, SoftmaxCorr has consistently strong correlations with models’ OOD performance. For example, the average Spearman’s Rank Correlation coefficient ρ is 0.949, 0.627, 0.942, 0.810, and 0.857 for our method, AoL, ATC-MC, MaxPred and SoftGap, respectively. We also notice that AoL exhibits very strong correlation with model accuracy on ImageNet-V2 ($\rho = 0.954$), but such a correlation drops noticeably on other OOD datasets. This implies that it is insufficient to rank both supervised models and vision–language models based solely on this phenomenon. On CIFAR-10 and WILDS, while on some test sets it does not present the strongest correlation, we still think that SoftmaxCorr is a preferred measure because it has very competitive average correlation scores. For Camelyon17, we see that all methods are less useful. We speculate this is caused by the under-specification phenomenon [Wenzel et al., 2022] where the model relies on spurious features.

SoftmaxCorr gives stable correlation, while the other four measures have mixed performance on different test sets. On CIFAR-10-Noise, we find that SoftmaxCorr correlates well with model performance ($\rho = 0.931$ and $\tau_w = 0.917$). In contrast, AoL, ATC-MC, MaxPred, SoftGap show a weaker correlation. Although ATC-MC has a higher average weighted Kendall’s correlation than SoftmaxCorr ($\rho = 0.916$ vs. 0.885) on ImageNet setup, it exhibits a weaker correlation on CIFAR-10 and WILDS setups. While in some cases SoftmaxCorr does not present the highest correlation, we emphasize that it overall gives more stable and stronger correlations. Thus, we think SoftmaxCorr is generally indicative of model generalization.

Compared to MaxPred and SoftGap, SoftmaxCorr better utilizes Softmax prediction probabilities. We use SoftGap on top of MaxPred as a simple approach to explicitly consider more entries (the second largest probability) in Softmax predictions. As shown in Table 6.1, SoftmaxCorr and SoftGap both achieve higher correlation results than MaxPred ($\rho = 0.864, 0.675$ and 0.637 , respectively). This indicates that it is helpful to analyze the distribution of softmax outputs. Compared with MaxPred and SoftGap, the class-wise correlation considered in our SoftmaxCorr better reveals the knowledge encoded by Softmax predictions. This is supported by the higher average correlation and more stable performance of SoftmaxCorr.

5.4.3 Sensitivity Analysis on Test Set Size

We study the sensitivity of SoftmaxCorr to test set size. Specifically, we reduce the dataset size by randomly sampling 1%, 5%, 10% and 30% of the original data. We report the averaged Spearman’s correlation of three random runs on six datasets (*i.e.*, ImageNet-V2, ImageNet-R, ObjectNet, CINIC, CIFAR-10-Noise and DomainNet-OOD). As shown in Table 5.3, we observe that when the number of test data is very small (1%), the correlation strength of SoftmaxCorr drops. When the dataset size increases ($\geq 10\%$), SoftmaxCorr exhibits a stable and high correlation with model

Dataset	C	D	Ours
ImageNet-V2-A	0.619	0.551	0.921
ImageNet-R	0.851	0.743	0.951
ObjectNet	0.881	0.812	0.963
CIFAR-10.2	0.881	0.766	0.894
CINIC	0.643	0.543	0.821
DomainNet-OOD	0.412	0.139	0.740

Table 5.2: Comparison between SoftmaxCorr (Ours) and two variants: Certainty (C) and Diversity (D). Rank correlation (ρ) is the metric.

Dataset	1%	5%	10%	30%	100%
ImageNet-V2	0.851	0.791	0.817	0.862	0.921
ImageNet-R	0.913	0.949	0.945	0.951	0.951
ObjectNet	0.906	0.950	0.955	0.963	0.963
CIFAR-10-Noise	0.910	0.933	0.930	0.930	0.931
CINIC	0.596	0.838	0.837	0.832	0.821
DomainNet-OOD	0.758	0.784	0.789	0.789	0.740

Table 5.3: **Sensitivity analysis of SoftmaxCorr on test set sizes.** We test four sampling ratios and report ρ on six datasets. SoftmaxCorr is stable given a reasonable number of samples.

Method	ImageNet-V2-A	ImageNet-R	ObjectNet	CIFAR-10.2	CINIC	DomainNet-OOD
Disagreement	−0.379	0.746	0.021	0.264	0.314	−0.083
SoftmaxCorr	0.921	0.951	0.963	0.894	0.821	0.740

Table 5.4: **Comparison between SoftmaxCorr and Disagreement.** Rank correlation (ρ) is used as the metric.

performance. This suggests that SoftmaxCorr requires a reasonable number of samples to capture model generalization.

5.4.4 Impacts of Class Distribution Estimator

In previous experiments, we solely use a vision-language model as the reference model to estimate class distribution. Another baseline usage of such a reference model is to measure the disagreements between the predictions of evaluated models and the reference model (Disagreement). In Table 5.4, we compare Spearman’s rank correlation of Disagreement and SoftmaxCorr. We observe that SoftmaxCorr consistently gives a stronger correlation than Disagreement. Disagreement explicitly measures the similarity between the evaluated models and the reference model, so the performance of the reference model significantly influences the predictive ability of Disagreement. In contrast, SoftmaxCorr only uses it for estimating the empirical class distribution, which constitutes the diagonal elements of the reference matrix $\mathbf{R}$, with all off-diagonal elements set to zero.

To study the influence caused by the estimator, we further use a less accurate zero-shot vision-language model (ViT-H-14) and an ideal estimator for calculating the marginal class distribution. We evaluate them on ObjectNet, because it has imbalanced class distribution. In Figure 5.3, we observe that with a moderately accurate ViT-H-14, SoftmaxCorr remains predictive, and the reference estimator further enhances SoftmaxCorr. Note that, we only use the average of softmax prediction probability to estimate class weights. Some better-designed algorithms [Lipton et al., 2018; Garg et al., 2020; Sun et al., 2022] may improve SoftmaxCorr’s stability, and we leave it as future work.

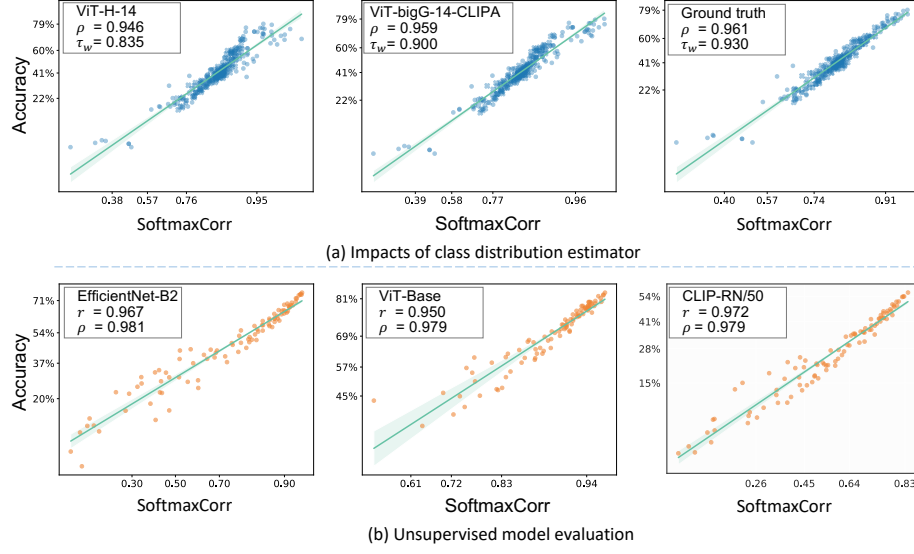


Figure 5.3: **(a) Impacts of class distribution estimator**, we use three estimators: ViT-H-14, ViT-bigG-14-CLIPA and the ground truth. We find SoftmaxCorr is fairly stable. **(b) SoftmaxCorr v.s. accuracy on ImageNet-C benchmark**. In every subfigure, each dot indicates a dataset of ImageNet-C. We see strong correlations between SoftmaxCorr and OOD accuracy across various test sets.

5.4.5 Different Characterizations of Correlation Matrix

To investigate the importance of prediction diversity and certainty for predicting OOD generalization, we compare SoftmaxCorr with two variants: (1) the sum of diagonal entries in the class correlation matrix (Certainty). It measures whether the values in diagonal entries are large, indicating prediction certainty; (2) the Euclidean distance between diagonal elements in the class correlation matrix and estimated class distribution (Diversity). It measures whether models make diverse predictions whose distribution matches the estimated distribution of each class. In Table 5.2, we present Spearman’s rank correlation of three methods on six datasets from three setups. Both variants give weaker correlation strength than SoftmaxCorr. Specifically, SoftmaxCorr is more predictive of OOD generalization than Certainty and Diversity ($\rho = 0.821$ vs. 0.643 vs. 0.543) on CINIC. This indicates that it is important to measure both prediction diversity and certainty for OOD generalization assessment.

5.4.6 Effectiveness of SoftmaxCorr for Domain Adaptation

On ImageNet, CIFAR and WILDS setups, all models are trained by standard empirical risk minimization and do not use the unlabeled OOD samples from training. In some scenarios, domain adaptation (DA) algorithms are employed for learning target-adaptive models with additional unlabeled OOD samples [Kouw and Loog, 2019; Zhou et al., 2022a]. To explore whether SoftmaxCorr is still effective in assessing the generalization of these models, we conduct a correlation study under DomainNet setup [Peng et al., 2019; Sagawa et al., 2021]. The models are trained

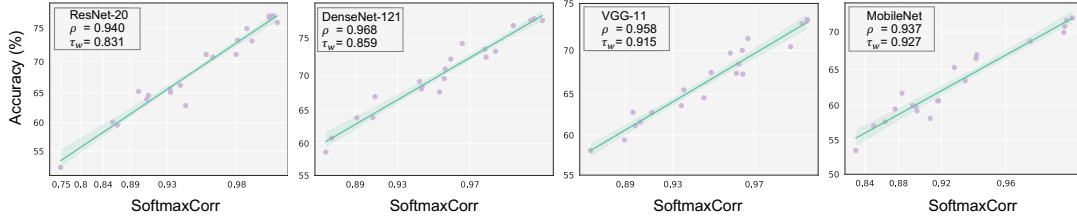


Figure 5.4: **Correlation analysis: SoftmaxCorr and accuracy on CINIC.** Each point represents a checkpoint. We consider CIFAR-10 models: ResNet-20, DenseNet-121, VGG-11 and MobileNet. Axes are probit scaled as in Section 6.2. In each subfigure, every point means a checkpoint of the model along the training process. For four models, we see strong correlations ($\rho > 0.93$). This suggests that SoftmaxCorr is helpful in assessing checkpoints along the training process.

by 9 different DA algorithms (e.g., DeepCORAL [Sun and Saenko, 2016], DANN [Ganin et al., 2016]). In Table 6.1, we observe that SoftmaxCorr performs reasonably on DomainNet-OOD. We also notice that ID accuracy-based methods (AoL and ATC-MC) become less useful. DA algorithms likely focus on improving OOD performance, while ID accuracy may not be enhanced accordingly.

5.4.7 SoftmaxCorr Reflects a Model’s Accuracy

We investigate how a given trained model generalizes to various OOD datasets. In detail, we evaluate a single model on all test sets of ImageNet-C benchmark [Hendrycks and Dietterich, 2019a] and conduct a correlation study between accuracy and SoftmaxCorr. We additionally use Pearson’s correlation (r) to measure the overall linear trend. This coefficient varies in $[-1, 1]$. A value closer to $-1/1$ indicates better negative/positive linearity and 0 means no correlation. We include: EfficientNet-B2 [Tan and Le, 2019], ViT-Base [Dosovitskiy et al., 2021a] and CLIP-RN/50 [Radford et al., 2021]. Figure 5.3 shows a strong linear relationship and high-rank correlation ($r > 0.95$ and $\rho > 0.97$). It indicates that with linear regression, SoftmaxCorr can also help estimate the accuracy of a given model on various test sets.

5.4.8 Evaluation Along Training Trajectory

In previous sections, SoftmaxCorr is used to measure the performance of models varying in different architectures and training strategies. In practice, we are sometimes interested in evaluating the models at different training checkpoints. Hence, we analyze whether SoftmaxCorr is helpful in this case. We collect prediction probabilities on CINIC every 10 epochs along the training process of ResNet-20, DenseNet-121 [Huang et al., 2017], VGG11 [Simonyan and Zisserman, 2014] and MobileNet [Howard et al., 2017] trained on CIFAR-10. In Fig. 5.4, we observe SoftmaxCorr has a high-rank correlation ($\rho > 0.93$) with model performance. This means we can potentially apply SoftmaxCorr to assay model generalization along the training process.

5.5 Discussion and Potential Directions

Discussion on imbalanced test sets. SoftmaxCorr is defined as cosine similarity between correlation matrix C and a reference matrix R_K . The reference matrix is a diagonal matrix and each entry represents the corresponding class distribution. To derive class distribution, a zero-shot vision-language model is used for efficient deployment without training and shows strong robustness towards distribution shifts. Consequently, the predictive ability of SoftmaxCorr correlates to the deployed vision-language model and the method used to estimate the distribution. Hence, it would be beneficial to use advanced label shift estimation techniques [Lipton et al., 2018; Garg et al., 2020; Sun et al., 2022] and employ more performant vision-language models, and we leave it as future work.

Potential OOD measures. This chapter proposes SoftmaxCorr to use class-wise relationships encoded by Softmax prediction probabilities. Here, we discuss other potential ways. **First**, SoftGap computes the difference between the largest and second-largest prediction probabilities. We show that SoftGap exhibits a stronger correlation with performance compared to MaxPred. It would be interesting to improve SoftGap by using more entries (e.g., top five probabilities). **Second**, for a perfectly calibrated model, its MaxPred over a test set corresponds to its accuracy. Yet, calibration methods seldom exhibit desired performance under distribution shift [Ovadia et al., 2019]. That said, it would be promising to study post-hoc calibration methods for OOD datasets, which benefits MaxPred for assessing model generalization. **Last**, this chapter focuses on Softmax prediction probability. We tested our method based on logits but no obvious correlation is exhibited. This may be because the logits of different models vary in significantly different ranges. We also think that studying other model statistics (e.g., weights and feature representations) would be interesting.

5.6 Summary

This chapter studies an important problem of assaying and ranking model generalization under distribution shifts. To this end, we explore the use of Softmax prediction probability for developing OOD measures. We start by identifying the demand for OOD measures beyond accuracy-on-the-line and finding that maximum Softmax prediction probability is to some extent useful to measure the OOD performance. We then propose Softmax Correlation (SoftmaxCorr) which leverages class confusion encoded by the class-class correlation matrix and thus better reflects the overall quality of the classifier predictions. To validate the usefulness of SoftmaxCorr, we compare it with four other empirical measures across 11 datasets under ImageNet, CIFAR-10 and WILDS setups. We observe SoftmaxCorr generally presents a stable and high correlation with model performance on different OOD datasets. This chapter establishes some baseline usage of Softmax prediction probability and a specific improvement, and further investigation is needed.

Ranking Multimodal Models Without Labels via Model Uncertainty

In this chapter, we extend unsupervised model ranking to large multimodal models (LMMs) and investigate whether their performance can be predicted without access to labeled data. As the number of available LMMs continues to grow, selecting suitable models for new tasks becomes increasingly challenging when ground-truth annotations are unavailable. To address this problem, we investigate whether signals derived from model predictions can be used to rank models in an unsupervised manner. We evaluate 47 state-of-the-art LMMs (e.g., LLaVA) across nine visual question answering benchmarks and analyze how uncertainty-based metrics relate to model performance. Our results show that uncertainty signals derived from softmax distributions provide consistent indicators of relative model performance across tasks. These findings suggest that uncertainty-based measures can support effective ranking of LMMs on unlabeled data, offering a practical approach for model selection without requiring manual annotation.

6.1 Motivation

Large multimodal models (LMMs), such as LLaVA [Liu et al., 2024c] and Instruct-BLIP [Dai et al., 2023], have demonstrated remarkable capabilities in handling a wide array of complex multimodal tasks, proving highly adaptable across diverse real-world applications [Liu et al., 2024a; Dai et al., 2023; Wang et al., 2024a; Huang et al., 2024, 2025; Zheng et al., 2025]. From addressing scientific challenges [Lu et al., 2022; Hiippala et al., 2021] and performing optical character recognition [Mishra et al., 2019; Masry et al., 2022; Mathew et al., 2021] to identifying the spatial position of objects [Tong et al., 2024; x.ai, 2024], LMMs are increasingly widespread in practical settings. As LMMs proliferate, the need for rigorous evaluation metrics that accurately capture their capabilities and limitations has become more urgent. Thus, numerous benchmarks have been developed [Lu et al., 2022; Hiippala et al., 2021;

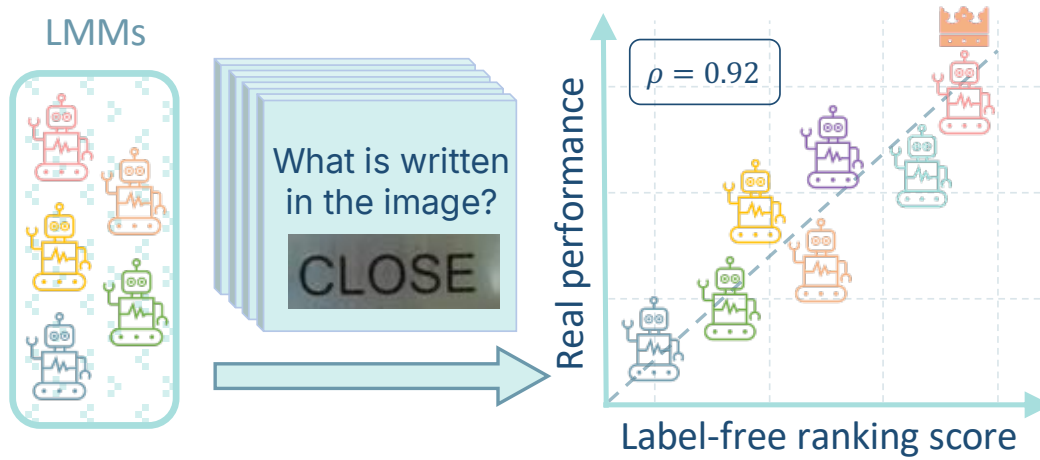


Figure 6.1: **Overview of unsupervised ranking for LMMs.** In scenarios where labeled data is scarce, selecting the best-performing model can be challenging. Our approach introduces a label-free proxy ranking score designed to reflect true performance, achieving a high correlation ($\rho = 0.92$) with actual metrics. This enables unsupervised comparison of LMMs, allowing users to identify the most suitable model without needing labeled data.

Masry et al., 2022; Singh et al., 2019; Mathew et al., 2021; x.ai, 2024; Yue et al., 2024], aiming to provide reliable rankings and guide users in selecting models best suited for specific deployment scenarios.

However, these benchmarks are based on carefully curated datasets that can require substantial resources to develop and label. For many users who may not have access to these resources, assessing model performance can be challenging. Additionally, standard evaluations are often dataset-centric, depending on fixed, human-labeled metrics that may not capture the full range of model capabilities necessary for diverse applications. As LMM tasks continue to diversify and expand, evaluating them effectively has become increasingly complex, with new applications frequently needing additional data curation and specialized capabilities.

Addressing the challenge of efficiently evaluating a set of LMMs is important for their usability and effectiveness in deployment. As illustrated in Fig. 6.1, selecting the most suitable LMM from a range of available options is challenging in the absence of annotations. To address this, this chapter investigates label-free proxy ranking scores that closely align with the true performance ranking, enabling effective model comparison in label-scarce settings. Our first finding is that the naïve approach of using the model’s performance on an existing benchmark to rank its performance in a new target environment could be unstable. We further report that measures of prediction uncertainty are more effective ranking indicators. LMMs generate answers in an open-ended form by producing token sequences, with each token selected from the vocabulary based on the probability. This enables model uncertainty to be assessed using the logits at each token position.

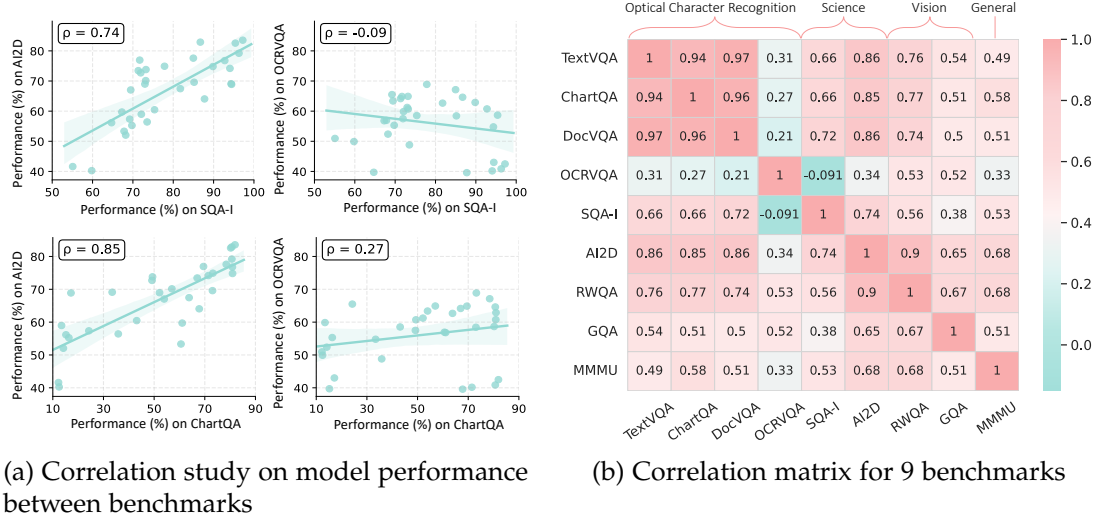


Figure 6.2: Correlation analysis of model performance across benchmarks. (a) Scatter plots illustrating the Spearman’s rank correlation coefficients (ρ) between performance on selected benchmarks, indicating how well performance on one benchmark predicts performance on another. Each point represents a model. The straight lines are fit with robust linear regression [Huber, 2011]. (b) Heatmap of the correlation matrix for performance across nine benchmarks, with color intensity representing the strength of correlation. Higher correlations (closer to 1) appear in red, while weaker correlations approach blue. The varying correlation strength indicates that using performance on one benchmark to rank LMMs in a target deployment environment may be inconsistent or unreliable.

To this end, we evaluate 45 different LMMs that have different training frameworks, *e.g.*, LLaVA-V1.5 [Liu et al., 2024a] and InstructBLIP [Dai et al., 2023], different visual encoders, *e.g.*, CLIP [Radford et al., 2021] and SigLIP [Zhai et al., 2023], and language models, *e.g.*, Vicuna [Team, 2023] and LLaMA [Touvron et al., 2023]. We investigate three categories of model uncertainty approaches: softmax probabilities, self-consistency, and labeled proxy sets. We evaluate the ranking performance on 9 widely-adopted multimodal benchmarks, which span diverse domains, including reasoning scientific questions, recognizing optical characters, and identifying objects’ spatial positions. Our main findings are that

- the performance of models on one dataset may not accurately reflect the ranking of the same models on a different dataset (Section 6.3);
- the effectiveness of ranking methods is influenced by task characteristics (*e.g.*, closed-form or free-form generation), but probability-based variants are typically quite robust and predictive (Section 6.4); and
- when examining correlations in model performance across different dataset pairs, we observe that text prompt similarity better correlates with model performance across datasets than image feature similarity (Section 6.5).

6.2 Task Formulation

Task Definition. Let a multimodal task be represented by a dataset $T = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the input prompt and the corresponding answer for the i -th sample, respectively. We have access to M large multimodal models (LMMs), denoted as $\{f_m\}_{m=1}^M$. Each LMM f_m , with pre-trained weights θ_m , generates a sequence of tokens $\{z_k\}_{k=1}^K$ from the input prompt $\mathbf{x}$ via a decoding process: $z_k = f_m([\mathbf{x}, z_1, z_2, \dots, z_{k-1}] \mid \theta_m)$, where z_k represents the k -th generated token. To assess the performance of each LMM, this task employs an evaluation metric (e.g., accuracy) that determines the ground-truth performance $\{g_m\}_{m=1}^M$ by comparing the generated sequences with the ground-truth answers $\mathbf{y}_i$.

The objective of this chapter is to develop methods for computing a score s_m for each LMM without requiring access to task-specific data annotations. Ideally, these computed scores should closely correlate with ground-truth performance, allowing us to rank and select LMMs based on their performance using only these scores.

Evaluation Metric. We use Spearman’s rank correlation coefficient ρ [Kendall, 1948] to evaluate the monotonic relationship between scores and model performance. Additionally, we calculate Kendall’s weighted rank correlation τ_w [Shieh, 1998], which effectively highlights top-ranked items [You et al., 2021]. Both coefficients range from $[-1, 1]$, with values near -1 or 1 indicating strong negative or positive correlations, and 0 indicating no correlation.

6.3 Uncertainty for Ranking LMMs

This section discusses the unique characteristics of ranking various LMMs compared to conventional ML models. We then introduce three distinct approaches that leverage uncertainty in model predictions for ranking.

6.3.1 What Makes Ranking LMMs Interesting?

Unique Challenges for Ranking LMMs. While LMMs can be considered a subset of machine learning (ML) models, ranking them introduces unique challenges not present in traditional ML models. Below, we discuss the unique characteristics of LMMs and the challenges they bring to the model-ranking process. First, LMMs often have billions of pre-trained parameters, which presents significant challenges for traditional “white-box” analysis for ranking various LMMs. Second, these models are typically trained on large datasets of instruction fine-tuning samples, which may be proprietary data. As a result, risk assessment methods that require access to training data are either unsuitable or should be adjusted. Additionally, although some LMMs may disclose their final model weights, other information, such as training loss and intermediate checkpoints, often remains undisclosed. The lack of access to training details limits the use of techniques in unsupervised accuracy estimation [Deng and Zheng, 2021b; Tu et al., 2023c].

Does Accuracy-on-the-Line (AoL) Suffice? Given these challenges, one may consider leveraging the AoL phenomenon [Miller et al., 2021] as a potential method for ranking LMMs. AoL refers to the strong linear correlation between probit-scaled in-distribution (ID) accuracy and out-of-distribution (OOD) accuracy across various ML models. This suggests that ID accuracy could serve as a reliable predictor of OOD performance, making AoL suitable for selecting LMMs in target testing environments. However, there are several reasons why this is not the case. First, obtaining ID performance data is often impractical, as LMMs are typically trained on large, sometimes proprietary datasets, limiting direct access to ID metrics. Second, while performance on existing benchmarks is more readily available, using this data to rank models for new deployment environments is problematic. Fig. 6.2 illustrates the correlation between proxy benchmark performance and target testing environments, revealing extreme variability in correlation strength. This highlights the unreliability of using benchmark performance as the sole ranking criterion. Third, relying solely on proxy benchmark performance fails to capture the unique statistical characteristics of target testing datasets. AoL tends to rank models identically across different environments, regardless of the actual deployment context.

What can we use for ranking LMMs? Our approach leverages the readily available outputs of LMMs—specifically, token prediction logits and generated tokens—without requiring intricate architectural analysis or complex extraction techniques. Inspired by recent work on uncertainty scores for classifier ranking [Hu et al., 2024; Tu et al., 2024], we introduce a novel adaptation of these techniques tailored to the specific characteristics of LMMs. By analyzing the distribution of prediction logits and the variability in generated outputs, we aim to assess each model’s self-awareness of its limitations. This enables an unsupervised model ranking method, focusing primarily on techniques that utilize these readily accessible LMM outputs.

6.3.2 Assessing Uncertainty by Softmax Probabilities

LMMs generate answers in an open-form manner by producing sequences of tokens. The selection of each token can be viewed as a classification process, where the model selects the token with, *e.g.*, the highest probability over its entire vocabulary. This allows us to assess model uncertainty based on the logits at each token position.

For token-level uncertainty [Huang et al., 2023], we can focus on two specific positions: the first token and the penultimate token (*i.e.*, the token preceding the end-of-sequence token), as illustrated in Fig. 6.3. The logit of the first token reflects the model’s initial response to the input prompt, while the logit of the penultimate token captures the model’s understanding of both the prompt and the generated response. The uncertainty associated with these two positions may provide insight into the model’s overall confidence in answering the question. Beyond token-level uncertainty, sentence-level uncertainty can be calculated by aggregating uncertainty across all tokens in the sequence [Manakul et al., 2023; Huang et al., 2023]. Specifically, sentence-level uncertainty can be quantified using the mean or the maximum

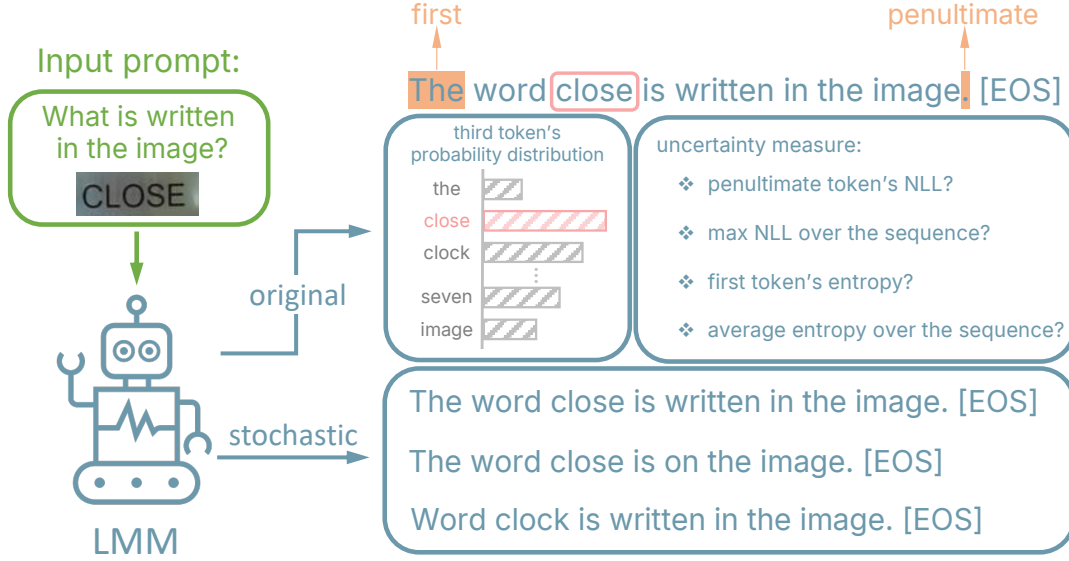


Figure 6.3: **An example of running one LMM for a VQA task.** We also present different token positions, methods to compute token-level uncertainty and the generation of stochastic predictions.

negative log-likelihood (NLL) values across the entire generated sequence:

$$\text{NLL}_{\max} = \max_j (-\log p_{ij}) \quad (6.1)$$

$$\text{NLL}_{\text{avg}} = -\frac{1}{J} \sum_{j=1}^J \log p_{ij} \quad (6.2)$$

where p_{ij} is the likelihood of the word generated by the LMM at the j -th token of the i -th sentence and J is the number of tokens generated in the sentence. Equation 6.1 quantifies a sentence's uncertainty via the least likely token, while Equation 6.2 uses the average per-token log-likelihood, allowing for length-independent comparisons of uncertainty [Malinin and Gales, 2020; Murray and Chiang, 2018]. Moreover, Equation 6.2 relates to perplexity, $\exp \text{Avg}(-\log p)$ [Jelinek et al., 1977; Manning and Schütze, 1999], a standard measure of model quality.

Alternatively, entropy $\mathcal{H}$ can be used instead of the negative log-likelihood to assess uncertainty. In the context of LMMs ranking, we adopt normalized entropy to account for varying vocabulary sizes across models, thus scaling entropy to the interval $[0, 1]$ and making it comparable across different models. The normalized entropy is given by:

$$\mathcal{H}_{ij} = -\frac{1}{\log |\mathcal{W}|} \sum_{w \in \mathcal{W}} p_{ij}(w) \log p_{ij}(w) \quad (6.3)$$

where $p_{ij}(w)$ is the likelihood of the word w being generated at the j -th token of the

i -th sentence, and W is the set of all possible words in the vocabulary.

There are eight variants of output probability-based methods, denoted as $\mathbf{NLL}_F$, $\mathbf{NLL}_P$, $\mathbf{NLL}_{\max}$, $\mathbf{NLL}_{\text{avg}}$, $\mathbf{Ent}_F$, $\mathbf{Ent}_P$, $\mathbf{Ent}_{\max}$, and $\mathbf{Ent}_{\text{avg}}$. “NLL” and “Ent” represent negative log-likelihood and entropy, respectively, while “F” and “P” refer to the first and penultimate tokens.

6.3.3 Assessing Uncertainty by Self-Consistency

Another approach involves examining non-deterministic generations produced by models. The core intuition is that a more accurate model will produce predictions closely aligned with the original answer, while a less accurate model may yield more divergent responses with each inference. In the case of LMMs, the temperature parameter t controls the randomness of predictions: a temperature of zero forces deterministic predictions, where the model selects only the token with the highest probability. When t is larger than 0, the model generates tokens stochastically, selecting tokens based on probabilities above a threshold. As t increases, tokens are sampled from an increasingly uniform distribution [Chen et al., 2023b; Cobbe et al., 2021].

To analyze the consistency in these stochastic predictions, we explore two common methods: BLEU [Papineni et al., 2002] and BERTScore [Zhang et al., 2019]. BLEU is a n-gram-based metric that evaluates the similarity of generated sequences to the reference answer, while BERTScore uses a pre-trained language model to embed answers and measures similarity in embedding space. Then, we use the mean value of similarities to represent the consistency for the input sample, which can be denoted as $\frac{1}{T} \sum_{i=1}^T \text{sim}(P_i, P_{\text{ori}})$, where T is the number of stochastic inferences, $\text{sim}(\cdot)$ is the similarity function (e.g., BLEU), and P_i and P_{ori} are the i -th stochastic prediction and the original answer, respectively. Following the practice outlined in [Chen et al., 2023b; Cobbe et al., 2021; Huang et al., 2023], we collect five stochastic inferences per sample and set $t = 0.7$ to maintain a relatively high degree of stochasticity in LMM generation while keeping compute overhead manageable. We adopt a unigram BLEU and denote the methods using BLEU and BERTScore as **Sample**_{BLEU} and **Sample**_{BERT}, respectively. Furthermore, BERTScore tends to assign high similarity scores between single letters, such as “A” and “B”, which should be considered very different in the contexts of MCVQ. For example, BERTScore gives a high similarity score of 0.998 for “A” and “B”, making it ineffective for distinguishing model performance. To address this, we modified the response to include the full answer text (e.g., “A. North America”), denoted as **Sample**_{BERT}^{*}.

6.3.4 Assessing Uncertainty With Labeled Proxy Data

Beyond the uncertainty score calculated directly on the target dataset, we also explore scores obtained using a proxy labeled dataset. Garg et al. [2022a] proposed average thresholded confidence (ATC), which calculates a threshold δ on a validation set (e.g., CIFAR-10 [Krizhevsky et al., 2009] validation set) and considers an image correctly classified on a new dataset with the same task if its confidence score exceeds the

Method	MCVQ								VQA										Average	
	SQA-I		AI2D		RWQA		MMMU		GQA		ChartQA		OCRQA		TextVQA		DocVQA			
	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w	ρ	τ_w		
AoL	0.52	0.45	0.73	0.61	0.70	0.57	0.54	0.45	0.53	0.32	0.69	0.59	0.30	0.20	0.69	0.57	0.68	0.60	0.60	0.48
NLL _F	0.83	0.78	0.84	0.76	0.56	0.58	0.60	0.59	0.71	0.57	0.63	0.41	0.78	0.64	0.74	0.63	0.84	0.74	0.73	0.63
NLL _p	0.64	0.60	0.61	0.58	0.25	0.37	0.16	0.18	0.58	0.52	0.79	0.67	0.81	0.68	0.71	0.67	0.88	0.78	0.60	0.56
NLL _{max}	0.80	0.76	0.91	0.81	0.63	0.63	0.49	0.44	0.72	0.59	0.94	0.80	0.64	0.63	0.83	0.69	0.92	0.82	0.76	0.69
NLL _{avg}	0.72	0.64	0.88	0.73	0.64	0.56	0.50	0.40	0.67	0.55	0.92	0.75	0.81	0.65	0.81	0.72	0.93	0.82	0.76	0.65
Ent _F	0.59	0.51	0.82	0.59	0.65	0.56	0.64	0.52	0.54	0.20	0.64	0.45	0.69	0.57	0.71	0.56	0.80	0.70	0.68	0.52
Ent _p	0.49	0.39	0.69	0.48	0.43	0.43	0.34	0.24	0.46	0.21	0.82	0.64	0.80	0.64	0.70	0.54	0.86	0.68	0.62	0.47
Ent _{max}	0.58	0.39	0.82	0.59	0.67	0.60	0.66	0.54	0.54	0.21	0.88	0.66	0.53	0.52	0.76	0.60	0.87	0.73	0.70	0.54
Ent _{avg}	0.58	0.33	0.80	0.56	0.62	0.50	0.59	0.41	0.57	0.26	0.91	0.68	0.77	0.62	0.74	0.58	0.88	0.74	0.72	0.52
Sample _{BLEU}	0.65	0.68	0.76	0.59	0.48	0.36	0.37	0.12	0.44	0.41	0.89	0.61	0.47	0.58	0.81	0.60	0.90	0.63	0.64	0.51
Sample _{BERT}	0.46	0.52	0.70	0.55	0.52	0.28	0.45	0.44	0.51	0.47	0.90	0.61	0.75	0.72	0.77	0.59	0.89	0.64	0.66	0.54
Sample _{BERT}	0.67	0.56	0.78	0.62	0.67	0.48	0.59	0.39	0.51	0.47	0.90	0.61	0.75	0.72	0.77	0.59	0.89	0.64	0.73	0.56
ATC	0.73	0.74	0.80	0.72	0.51	0.40	0.41	0.35	0.39	0.20	0.95	0.85	0.28	0.21	0.69	0.64	0.89	0.77	0.63	0.54

Table 6.1: **Method comparison across eight multimodal tasks.** The table presents a comparison of four groups of methods: accuracy-based, output-probability-based, sample-based, and unsupervised model evaluation methods. We evaluate these methods using Spearman’s rank correlation (ρ) and weighted Kendall’s correlation (τ_w). Both coefficients range from -1 to 1 , where values close to -1 or 1 indicate strong negative or positive correlations, respectively, and 0 indicates no correlation. The **AoL** and **ATC** performance is calculated as the average correlation when using the scores computed on the other seven domains and the model performance on the target domain. The highest correlation values for each task are highlighted in **green**, while the second highest values are marked in **blue**. Methods are ranked to **three** decimal places. We use the absolute value of correlation strength in the table for **NLL** and **Ent**. The results indicate that **NLL_{max}** and **NLL_{avg}** are often preferable, as they show greater stability and stronger correlation with model performance.

threshold. With the derived threshold, model performance on the target domain is estimated by the proportion of samples with confidence scores higher than the threshold. For LMMs, we use an existing benchmark as a proxy set to calculate δ . The calculation of ATC is

$$\text{ATC} = \mathbb{E}_{x \in T} [\mathbb{I}[u(f(x)) > \delta]], \quad (6.4)$$

where $\mathbb{I}$ is a binary indicator function, and $u(\cdot)$ represents an uncertainty estimation method, for which we use **NLL_{max}**. ATC provides both a ranking of model performance and an estimate of the expected performance.

6.4 Experiments

In this section, we first introduce the experimental setup including the evaluated datasets and models we considered. Then, we show the results of ranking different large multimodal models (LMMs).

6.4.1 Experiment Setup

Benchmarks. We choose visual question answering, which is a common protocol to evaluate LMMs [Liu et al., 2024a; Dai et al., 2023; Beyer et al., 2024; Lu et al., 2022; Yue et al., 2024]. We evaluate on multiple choice visual question (MCVQ) and visual question answering (VQA) benchmarks. MCQ and VQA are both types of question-answering formats for evaluating LMMs. For the former, the model is provided with several answer options, out of which the correct subset is to be selected. In contrast, the latter is usually open-ended and the models may generate answers freely. We consider 9 widely-adopted MCVQ and VQA benchmarks. They are (1) the subset of ScienceQA [Lu et al., 2022] with images (SQA-I) and AI2D [Hiippala et al., 2021] which assess LMMs’ scientific knowledge; (2) ChartQA [Masry et al., 2022], OCRVQA [Mishra et al., 2019], TextVQA [Singh et al., 2019] and DocVQA [Mathew et al., 2021] to examine their ability to recognize optical character; (3) RealWorldQA (RWQA) [x.ai, 2024] and GQA [Ainslie et al., 2023] which evaluate LMMs’ vision-centric capability; (4) MMMU [Yue et al., 2024] which assays LMMs on multi disciplinary tasks that demand college-level subject knowledge. Note that SQA-I, AI2D, RWQA and MMMU are MCVQ datasets, while the others are VQA.

Models. The goal is to choose the best LMM over different series of LMMs. We include LLaVA-V1.5 [Liu et al., 2024a], ShareGPT4V [Chen et al., 2023a], LLaVA-NeXT [Liu et al., 2024b], InstructBLIP [Dai et al., 2023], LLaVA-NeXT-Interleave [Li et al., 2024b], LLaVA-OneVision [Li et al., 2024a], Eagle [Shi et al., 2024], mPLUG-Owl [Li et al., 2022a], InternVL [Chen et al., 2024c], PaliGemma [Beyer et al., 2024], Mantis [Jiang et al., 2024], DeepSeek-VL2 [Wu et al., 2024] and Qwen2-VL [Wang et al., 2024b]. In total, we collect 32 different models, which all can be accessed on Hugging Face [Wolf et al., 2020].

6.4.2 Key Findings

Accuracy-on-the-Line is unreliable for ranking LMMs in new domains. Table 6.1 summarizes the ranking capability of all methods across eight benchmarks. The AoL performance is calculated as the average correlation strength when using the other seven domains to predict model rankings in the target domain. We observe that AoL does not achieve consistently high correlation with model performance in seven out of the eight benchmarks. Although AoL shows strong results on RealWorldQA, where it performs best, uncertainty-based methods also demonstrate high correlation strength. For instance, $\text{NLL}_{\max}$ exhibits a Spearman’s ρ 0.07 lower but a weighted Kendall’s τ_w 0.06 higher than AoL. These findings suggest that relying on existing benchmarks alone to select models for deployment could be risky and unstable, as they may not well capture the statistical characteristics of the new target domain.

The choice of token matters for output probability-based ranking. We studied four ways of using tokens for estimating model uncertainty. For the two on specific positions, the first and penultimate tokens, their predictive performance depends on the task type (*i.e.*, MCVQ *vs.* VQA). The uncertainty associated with generating the first token is more indicative for MCVQ tasks, while the penultimate token gener-

ally proves more predictive for VQA tasks. This difference is due to the nature of MCVQ tasks, which often prompt models to generate only a single option letter (e.g., “A”), making the first token crucial for evaluating response accuracy. In contrast, VQA tasks require open-form responses consisting of multiple tokens, making the penultimate token—reflecting the model’s understanding of both the question and the answer—more informative.

For the two that consider every token in the generated output, *i.e.*, $\mathbf{NLL}_{\max}$ and $\mathbf{NLL}_{\text{avg}}$, we find that they are more stable and less task-specific compared to variants that consider individual tokens. While $\mathbf{NLL}_{\max}$ and $\mathbf{Ent}_{\max}$ (reflecting the least confident token) is more effective for MCVQ tasks, both methods perform similarly on VQA tasks. The higher ranking correlation of using the first token and the least confident token as ranking indicators for MCVQ suggests that a single token (typically the option letter) holds significant meaning. Considering all potential tokens can sometimes result in an unexpectedly higher confidence level, as LMMs may generate the complete answer alongside the option letter (e.g., “B. Columbia”). The tokens following the option letter often have high confidence levels, leading to an overall increase in estimated uncertainty. This tendency results in models generating entire answers with lower uncertainty scores, yielding higher ranks. These findings underscore the importance of further exploration into the optimal token positions for uncertainty estimation in LMMs. One potential approach is leveraging a language model to identify the position of the option letter and use its softmax probability as the uncertainty measure for the complete response.

The negative log-likelihood (NLL) is more stable and predictive for LMMs ranking than normalized entropy. For seven out of eight benchmarks, NLL consistently shows stronger correlation with model performance than entropy. Both NLL and entropy make use of the softmax probability distribution predicted for each token during generation. However, the NLL-based approaches solely consider the maximum probability in this distribution, while the entropy considers all entries. The lower correlation strength of normalized entropy may therefore stem from the significantly different vocabulary sizes of various LMMs. For instance, LLaVA-V1.5-7B has a vocabulary of 32k tokens, while PaliGemma-3B-mix has a much larger vocabulary of 257k tokens. Despite being normalized by vocabulary size, entropy may still be more susceptible to noise than NLL. Future work could include applying dimension reduction techniques or limiting consideration to the top- k most probable tokens or a pre-defined set of tokens.

Sample-based methods are strong candidates for model ranking without intrinsic access to LMMs. The correlation strength of sample-based methods is influenced by the nature of tasks. They yield higher correlation scores on VQA tasks compared to MCVQ tasks. In VQA, models typically produce more varied responses, making BLEU and BERTScore effective for capturing uncertainty. However, MCVQ tasks constrain models to select from a pre-defined set of answers. While non-zero temperature introduces some randomness, the variation between stochastic inferences is limited. Also, we find $\mathbf{Sample}^*_{\text{BERT}}$ has higher correlation scores than $\mathbf{Sample}_{\text{BERT}}$ on four MCVQ benchmarks and suggests that developing more advanced algorithms to

Method	MCVQ		VQA		Average
	AI2D	MMMU	TextVQA	ChartQA	
AoL	0.42	0.43	0.57	0.64	0.52
NLL _{max}	0.29	0.59	0.97	0.99	0.71
NLL _{avg}	0.26	0.28	0.98	0.98	0.63
Ent _{max}	0.37	0.62	0.97	0.96	0.73
Ent _{avg}	0.11	0.38	0.98	0.98	0.63
Sample _{BLEU}	0.63	0.73	0.76	0.64	0.69
ATC	0.12	0.59	0.85	0.84	0.60

Table 6.2: **Method comparison for ranking different LLaVA-prismatic models on AI2D and TextVQA.** We use Spearman’s rank correlation (ρ) as the metric. Uncertainty-based methods still exhibit moderately high correlation strength, indicating their effectiveness for ranking LLaVA models.

capture uncertainty from multiple stochastic inferences could be beneficial.

Although sample-based methods show weaker overall correlations, they remain competitive without requiring access to model architecture or internal states. This highlights their potential to rank API-based or closed-source LMMs.

ATC can be used for LMMs ranking, but the correlation strength is influenced by the choice of proxy dataset. We compute ATC performance as the average correlation when using the other seven datasets as proxy datasets. Our analysis reveals the choice of proxy dataset is critical, as the scale of uncertainty calculated on different datasets can vary. This variation can lead to an inaccurately estimated threshold for determining instance correctness. A potential improvement involves adopting uncertainty calibration methods, such as temperature scaling [Guo et al., 2017], to calibrate model uncertainty onto a consistent scale.

6.5 Analysis

This section includes three distinct analyses. The first specifically investigates whether the considered methods are effective for ranking models within the same series. We use LLaVA models as a case study because they are widely adopted and representative of LMM architectures. The other two analyses consider models from different series, consistent with the broader evaluation in Section 6.4.

Ranking LMMs from the same series. So far, all experiments have focused on selecting LMMs from different model series. However, in some scenarios, the objective is to choose a training recipe that yields better-performing LMMs within the same model series. These models may be trained with additional fine-tuning steps, varied data augmentations, different training sources, or different language models (e.g., Vicuna and Mistral [Jiang et al., 2023]) and visual encoders, such as CLIP and DINOv2 [Oquab et al., 2024]. For our analysis, we use the LLaVA series due to its widespread adoption. Specifically, we employ 15 different LLaVA prismatic (LLaVA-pri) models [Karamcheti et al., 2024]. We report results on four datasets in Table 6.2:

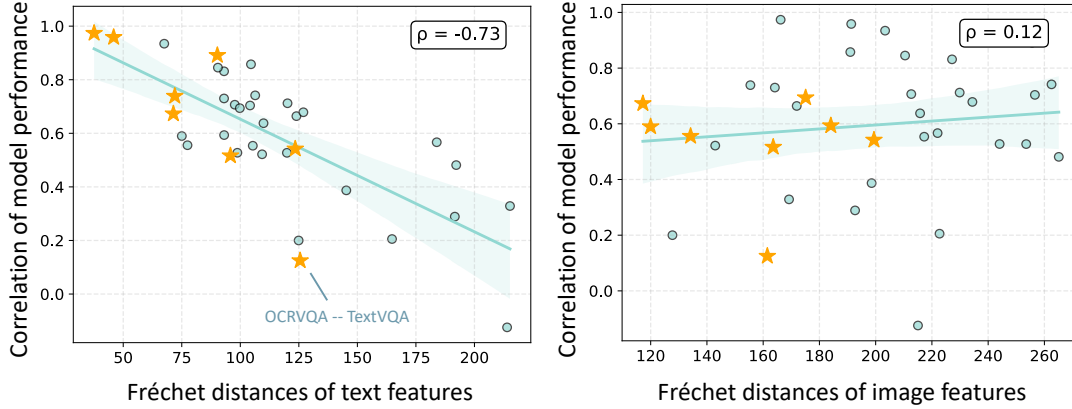


Figure 6.4: **Correlation analysis of Fréchet distances and model performance correlation across datasets.** Orange stars indicate the dataset pairs with the highest similarity for each dataset. Observations reveal that variations in text prompt similarity are more closely aligned with changes in performance correlation than variations in image feature similarity.

AI2D, MMMU, TextVQA, and ChartQA, with AI2D and TextVQA being used to evaluate the performance of different LLaVA variants in the original paper. The **AoL** and **ATC** performance metrics are computed in the same manner as described in Section 6.4, using the average correlation across other three datasets.

Our findings align with those from ranking different series of LMMs. First, using model performance on a single existing dataset may not accurately reflect LMMs rankings on a different domain, even when the models are trained with a similar pipeline and minor variations. Second, we observe that ranking different LLaVA-prior models in MCVQ presents a significant challenge, as the variance in model performance on MCVQ tasks is lower compared to VQA tasks. For instance, the performance gap on AI2D is only 4%, whereas the gap on TextVQA exceeds 20%. This indicates that ranking methods must capture subtle differences between models. We find that **Sample**_{BLEU} remains effective, while NLL and Entropy may not capture these nuances accurately. Additionally, the decrease in correlation between uncertainty estimated by softmax probability suggests that although modifications to the training pipeline may have a slight effect on performance, they can lead to significant variations in the confidence levels for generation. This finding underscores the importance of assessing LMMs more comprehensively, beyond accuracy alone.

Analysis of the weak correlation of AoL. Fig. 6.2(b) illustrates the correlation of model performance across different benchmarks. We observe that while TextVQA, ChartQA, DocVQA, and OCRVQA all aim to assess the capability of LMMs to recognize optical characters, the correlation between them varies significantly. Specifically, model performance on TextVQA, ChartQA, and DocVQA shows strong correlations, whereas performance on OCRVQA consistently exhibits low correlation with the other three datasets. To investigate whether images or text within datasets influence the correlation strength in model performance, we utilize CLIP-L-14-336 [Radford

Method	MCVQ		VQA	
	AI2D	MMMU	TextVQA	ChartQA
#Samples	3088	1050	5000	2500
50 samples	0.89	0.35	0.92	0.98
NLL _{max}	0.92	0.56	0.82	0.93

Table 6.3: **Labeling a subset of target domains to rank models** on AI2D, MMMU, TextVQA, and ChartQA. We report Spearman’s rank correlation. While a small labeled set provides reasonable rankings, it may not fully capture the order. NLL_{max} offers more stable correlations, suggesting its potential for label-free model ranking.

et al., 2021] to extract image and text embeddings. We then use Fréchet distance (FD) [Fréchet, 1957] to measure similarities between datasets based on these features.

Fig. 6.4 presents a correlation study between the FD of dataset pairs and the correlation strength of model performance on those datasets. We observe a strong correlation for FDs computed using text input features, while FDs measured by image features show a weaker correlation. This suggests that text input similarity is likely a more influential factor for model performance correlation than image similarity. Additionally, orange stars are used to label the points representing the lowest FD for each dataset. Moreover, OCRVQA shows a low correlation strength with other datasets (Fig. 6.2). The closest dataset in prompt feature space is TextVQA, with a FD of 124, which is considerably higher than the lowest FDs observed between other dataset pairs, typically around 70 or lower. This difference sheds light on the weak model performance correlation between OCRVQA and other OCR-focused datasets.

Ranking LMMs by a labeled subset of the target domain. Table 6.3 presents the correlation strength between model performance on 50 labeled instances in the target domain and the overall performance across the entire dataset. We observe that a small labeled set can provide a good ranking. However, such an approach may not fully capture the model ranking across the entire dataset, since the sampled data may not be representative [Polo et al., 2024]. In contrast, NLL_{max} gives a more stable correlation, highlighting the potential of uncertainty-based methods for effective model ranking without data annotation.

6.6 Summary and Discussion

This chapter studies whether the performance of large multimodal models in a new target domain can be ranked without the access to target domain labels. Our analysis identified only a weak correlation in model performance across different domains. This motivated the investigation and evaluation of alternative approaches based on uncertainty estimates obtained from model predictions. We evaluate 47 LMMs on closed and open-ended visual question answering tasks, testing various training frameworks, visual encoders, and language models. Our experiments reveal that scores based on the negative log-likelihood of generated tokens serve as

highly effective performance indicators for target domains. We also find that while stochastic sampling can be helpful, it is less effective for multiple-choice tasks, where it requires many repetitions of inference and careful temperature tuning. By establishing a baseline for uncertainty-based LMMs ranking, this study aims to motivate and inspire further research into this important area.

Potential future directions. Beyond uncertainty scores, several promising directions remain for future research. Test-time augmentation and semantic entropy present natural next steps. Additionally, internal states of LMMs [Azaria and Mitchell, 2023; Gottesman and Geva, 2024] offer opportunities for uncertainty estimation or error detection, for example, by training a classifier on these representations. However, in unsupervised LMM ranking, this approach requires training a separate classifier per model, which can be computationally expensive. An alternative is to measure statistical distance between a model’s internal state for a given response and the distribution of internal states from multiple inferences. A larger average distance may signal greater uncertainty and potentially indicate a lower model rank. Exploring how internal states can inform LMM ranking is an open direction. Lastly, the observed asymmetry in impacts of text and image feature dissimilarity on cross-dataset correlations highlights a valuable area for improving multimodal benchmark design.

Unsupervised Model Assessment Using Dataset Representation

This chapter continues the study of label free performance assessment. We take a data-centric view on two dataset-level tasks: assessing training set suitability for a target domain and estimating test set difficulty for a trained model. Both tasks require measuring relationships between datasets without relying on labels. To address this problem, we propose a bag-of-prototypes (BoP) representation that encodes datasets as distributions over semantic prototypes. A prototype codebook is first constructed from a reference dataset, and each dataset is represented as a histogram obtained by quantizing its image features to the prototypes. This representation captures the semantic distribution of a dataset without requiring annotations. Combined with Jensen–Shannon divergence for measuring dataset similarity, BoP enables effective comparison between datasets. Extensive experiments show that this simple representation consistently outperforms existing dataset representations on multiple benchmarks for both tasks.

7.1 Motivation

Datasets are fundamental in machine learning research, forming the basis of model training and testing [Paullada et al., 2021; Deng et al., 2009; Lin et al., 2014; Liang et al., 2022]. While large-scale datasets bring opportunities in algorithm design, there is a lack of appropriate tools to analyze and make the best use of them [Mazumder et al., 2022; Liang et al., 2022; Aroyo et al., 2022]. Therefore, as opposed to traditional algorithm-centric research where improving models is of primary interest, the community has seen a growing interest in understanding and analyzing the data used for developing models [Liang et al., 2022; Mazumder et al., 2022]. Recent examples of such goals include data synthesis [Fu et al., 2022], data sculpting [Eyuboglu et al., 2022; Liang et al., 2022], and data valuation [Mazumder et al., 2022; Aroyo et al., 2022; Ghorbani and Zou, 2019]. These tasks typically focus on individual samples of a dataset. In this chapter, we aim to understand the nature of datasets from a dataset-level perspective.

This chapter considers two dataset-level tasks: suitability in training and diffi-

culty in testing. **First**, training set suitability denotes whether a training set is suitable for training models for a target dataset. In real-world applications, we are often provided with multiple training sets from various data distributions (e.g., universities and hospitals). Due to distribution shift, their trained models have different performance on the target dataset. Then, it is of high practical value to select the most suitable training set for the target dataset. **Second**, test set difficulty means how challenging a test set is for a learned model. In practice, test sets are usually unlabeled and often come from different distributions than that of the training set. Measuring the test set difficulty for a learned model helps us understand the model reliability, thereby ensuring safe model deployment.

The core of the two dataset-level tasks is to measure the relationship between datasets. For example, a training set is more suitable for learning a model if it is more similar to the target dataset. To this end, we propose a vectorization scheme to represent a dataset. Then, the relationship between a pair of datasets can be simply reflected by the distance between their representations. Yet, it is challenging to encode a dataset as a representative vector, because (i) a dataset has a different cardinality (number of images) and (ii) each image has its own semantic content (e.g., category). It is thus critical to find an effective way to aggregate all image features to uncover dataset semantic distributions.

In the literature, some researchers use the first few moments of distributions such as feature mean and co-variance to represent datasets [Sun and Saenko, 2016; Sun et al., 2016; Peng et al., 2019; Deng and Zheng, 2021a; Vanschoren, 2018]. While being computationally friendly, these methods do not offer sufficiently strong descriptive ability of a dataset, such as class distributions, and thus have limited effectiveness in assessing attributes related to semantics. Some methods also learn *task-specific* dataset representations [Peng et al., 2020; Achille et al., 2019]. For example, given a dataset with labels and a task loss function, Task2Vec [Achille et al., 2019] computes an embedding based on estimates of the Fisher information matrix associated with a probe network’s parameters. While these task-specific representations are able to predict task similarities, they are not suitable for characterizing dataset properties of interest. They require training a network on the specific task [Achille et al., 2019] or on multiple datasets [Peng et al., 2020], so they are not effective in assessing the training set suitability. Additionally, they require image labels for the specific task, so they cannot be used to measure the difficulty of unlabeled test sets.

In this chapter, we propose a simple and effective bag-of-prototypes (BoP) dataset representation. Its computation starts with partitioning the image feature space into semantic regions through clustering, where the region centers, or prototypes, form a codebook. Given a new dataset, we quantize its features to their corresponding prototypes and compute an assignment histogram, which, after normalization, gives the BoP representation. The dimensionality of BoP equals the codebook size, which is usually a few hundred and is considered memory-efficient. Meanwhile, the histogram computed on the prototypes is descriptive of dataset semantic distribution.

Apart from being low dimensional and semantically rich, BoP has a few other advantages. **First**, Unlike recent task-specific dataset representations that require

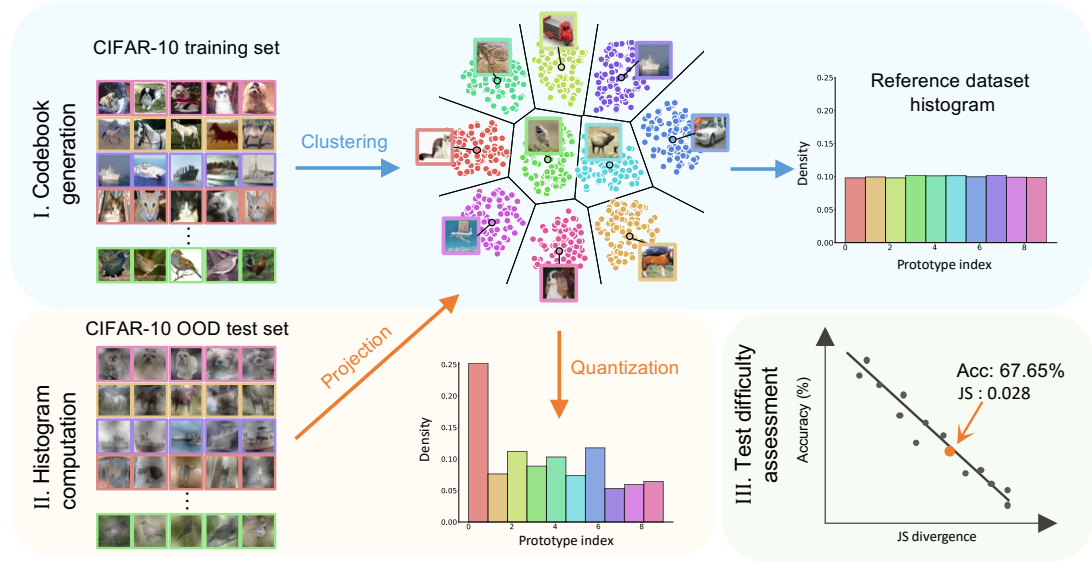


Figure 7.1: **Workflow of BoP representation computation** using CIFAR-10 [Krizhevsky et al., 2009] and one CIFAR-10 out-of-distribution (OOD) test set as an example. **Top:** We group image features of the reference dataset CIFAR-10 into 10 clusters, and the centers are called prototypes. The prototypes constitute the code-book of size 10. **Bottom left:** To encode the OOD test set, we project it onto the code-book by quantizing each image feature to its corresponding prototype. Lastly, we compute the histogram, *i.e.*, BoP representation, of CIFAR-10 OOD test set. **Bottom right:** We regard dataset-to-dataset similarity as Jensen-Shannon divergence between BoP histograms of CIFAR-10 OOD test set and reference dataset. With such similarity, we can measure the test set difficulty for the model trained on reference dataset.

full image annotations and additional learning procedures [Peng et al., 2020; Achille et al., 2019], BoP requires neither annotations nor additional learning. It is relatively efficient and allows for unsupervised assessment of dataset attributes. **Second**, BoP supports dataset-to-dataset similarity measurement through Jensen-Shannon divergence. We show in our experiment that this similarity is superior to commonly used metrics such as Fréchet distance [Fréchet, 1957] and maximum mean discrepancy [Gretton et al., 2006] in two dataset-level tasks.

7.2 Methodology

7.2.1 Bag-of-Words Model Across Communities

In **natural language processing** (NLP) and information retrieval, the Bag-of-Words (BoW) model [McCallum et al., 1998; Joachims, 1998; Lewis and Gale, 1994; Joachims et al., 1997] vectorizes textual data as a word histogram. Specifically, for each word in the dictionary, its occurrences in a document are counted, which fills in the corresponding entry of the BoW feature. This word frequency vector is thus used to represent a document. Numerous improvements of the BoW feature were made

in NLP, such as n-grams [Lewis and Gale, 1994; Joachims et al., 1997] and term frequency-inverse document frequency [Rajaraman and Ullman, 2011].

In the early 2000s, the BoW representation was introduced to the **computer vision** (CV) community to encode hundreds or thousands of local image descriptors [Lowe, 1999; Bay et al., 2006] into a compact vector [Sivic and Zisserman, 2003]. As there is no semantic codebook like in NLP, a visual codebook is constructed by performing clustering (e.g., k-means) on a collection of local image features, where the resulting clustering centers are called “visual words”. Local image descriptors are quantized to their nearest cluster center so that a visual word histogram can be computed. This BoW histogram also have undergone extensive improvements in later years, such as Fisher vector [Perronnin and Dance, 2007; Perronnin et al., 2010], vector of locally aggregated descriptors (VLAD) [Jégou et al., 2010], and the use of principal component analysis and whitening [Jégou and Chum, 2012].

Contribution statement. This chapter contributes a baseline method in adopting the BoW idea study the two basic properties of a dataset. To this end, we propose to represent a dataset using its histogram over a series of prototypes. A comparison between the usage of BoW model in NLP, CV and our dataset-level research is shown in Table 7.1. Specifically, the BoP representation relies on clustering for codebook formation, has a relatively small codebook (depending on the richness of dataset semantics), and has semantically sensible codewords.

7.2.2 Bag-of-Prototypes Dataset Representation

Given a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^N$ where N is the number of images and a feature extractor $\mathbf{F}(\cdot)$ that maps an input image into a d -dimensional feature $f \in \mathbb{R}^d$, we extract a set of image features $\mathcal{F} := \{\mathbf{F}(\mathbf{x}_i)\}_{i=1}^N$. While it is possible to directly use the dataset images (or features) as model input under small N , it becomes prohibitively expensive when N is large. We therefore focus on extracting useful semantic features of $\mathcal{F}$ by encoding its image features into a compact representation. Below we detail the necessary steps for computing the proposed BoP representation (refer Fig. 7.1).

Step I: Codebook generation. Given a reference dataset $\mathcal{D}_r = \{\mathbf{x}_i^r\}_{i=1}^{N_r}$, we extract all of its image features $\mathcal{F}_r := \{\mathbf{F}(\mathbf{x}_i^r)\}_{i=1}^{N_r}$ using a pretrained network, from which a codebook is constructed. Specifically, we adopt standard k-means clustering [MacQueen et al., 1967] to partition the feature space $\mathbb{R}^d$ into K clusters. Each of the K cluster centers is called a “prototype”, because oftentimes each center mainly represents a certain semantic content. See Fig. 7.1 right for exemplar image of each prototype. The prototypes constitute the codebook, denoted as $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^K$, where $\mathbf{c}_i$ is the i -th prototype. Note that, the order of the prototypes is fixed in the codebook.

Step II: Histogram computation. For a dataset to be encoded $\mathcal{D}_e = \{\mathbf{x}_i^e\}_{i=1}^{N_e}$ where N_e is the number of images, we project it onto codebook $\mathcal{C}$ of size K to compute its BoP representation. Specifically, after extracting image features $\mathcal{F}_e := \{\mathbf{F}(\mathbf{x}_i^e)\}_{i=1}^{N_e}$ from $\mathcal{D}_e$, for each image feature, we compute its distance with each the K prototypes in the codebook, yielding K distances $d_1, \dots, d_K$, where d_i is the distance between an image feature and the i -th prototype. An image feature is quantized to prototype

	BoW in NLP	BoW in CV	BoP
Encoded objects	Documents	Images	Datasets (a set of images)
Codewords in codebook	Words	Cluster centers of local descriptors	Prototypes of image features
Clustering?	No	Yes	Yes
Codewords semantics	Clear	Little	Sensible
Codebook size	$> 10^3$	$10^3 - 10^6$	$\sim 10^2$ (dataset dependent)

Table 7.1: **Comparing BoP with BoW model in natural language processing (NLP) and computer vision (CV).** The objective of BoW in NLP and CV is encoding texts and images respectively, while BoP is proposed to represent datasets.

c_i if d_i is the lowest among $d_1, \dots, d_k$. Following the quantization, we generate a K -dimensional one-hot encoding where the i -th entry is 1 and all the others are 0. Having computed the one-hot vectors for all the image features, we sum them which is then normalized by N_e , the number of images in D_e . This gives the histogram representation $\mathbf{h}_e$, or **BoP representation**, for D_e where the i -th entry indicates the density of features in D_e belonging to prototype c_i .

7.2.3 Measuring Dataset-to-Dataset Similarity

Similar to image / document retrieval where BoW vectors of instances are used for similarity comparison [Sivic and Zisserman, 2003; Philbin et al., 2007; Csurka et al., 2004; Nister and Stewenius, 2006; Fei-Fei and Perona, 2005], this work uses the BoP representation to calculate dataset-to-dataset similarity. Specifically, given BoP representations $\mathbf{h}_x$ and $\mathbf{h}_y$ of two datasets $\mathcal{D}_x$ and $\mathcal{D}_y$, we simply define their similarity $S_{x,y}$ using Jensen-Shannon divergence (JS divergence), which is designed for histogram-based similarity measurement [Dagan et al., 1997; Manning and Schütze, 1999].

Task-oriented similarity measure. We can build a *universal* codebook on a large-scale dataset following BoW [Csurka et al., 2004; Zheng et al., 2017]. By doing so, the resulting BoP representations are generic. We can also build a *task-oriented* codebook on a reference dataset from a specific task to consider more task-oriented information. The latter is more suitable for the two dataset-level tasks considered. For the task of training set suitability assessment, we use the target dataset as the reference for codebook generation to fully consider the semantic information. As a result, the JS divergence between BoP representations of the training set and the target dataset can well capture how a training set is similar to the target set. Similarly, for the task of test set difficulty assessment, we build codebook on the training set. This practice can effectively measure how an unlabeled test is similar to a given training set.

7.2.4 Discussion

Working mechanism of BoP. Codebook generation of BoP can be viewed as Centroidal Voronoi Tessellations [Du et al., 1999]. Specifically, the prototypes of codebook tessellate the feature space into Voronoi cells. Then, histogram computation approximates a probability distribution function in the same way as the *nonparametric* histogram [Freedman and Diaconis, 1981; Boots et al., 2009; Polianskii et al., 2022]. That is, BoP representation reflects the distribution of a dataset in the feature space.

As shown in Fig. 7.1, the prototypes of reference dataset tessellate feature space into Voronoi cells. Based on this, we quantify the histogram of the reference dataset to represent its distribution. Given a new dataset, we conduct the same histogram calculation procedure and correspondingly capture its dataset distribution with the histogram. Then, we measure discrepancy of the two datasets by calculating JS divergence between their histograms. Compared with common measures of dataset distance (e.g., FD [Fréchet, 1957], KID [Bińkowski et al., 2018] and MMD [Gretton et al., 2006]) that only reflect global structure (e.g., first few moments) of dataset distributions, BoP, combined with JS divergence, considers more local structures.

Training set suitability vs. transferability estimation. The two tasks are related but differ significantly: 1) Given an unlabeled target dataset and a pool of training datasets, the former aims to select the most suitable training set for the target. The latter assumes a labeled target dataset and a pool of models pretrained on a source dataset, with the goal of selecting the most suitable source model for the target without fine-tuning them all [Pándy et al., 2022; Agostinelli et al., 2022a,b]; 2) Datasets in training set suitability are used for the same classification problem. In contrast, in transferability estimation, the problem in the target dataset (e.g., CIFAR-10 classification) is different from that of the source dataset (e.g. ImageNet classification).

Analysis of the number of prototypes in a codebook. The codebook size is a critical factor influencing the usefulness of BoP. A small codebook means a coarser partition of feature space, where similar features will likely be in the same cluster, but dissimilar features may also be in the same cluster. A large codebook provides a finer description of the space, where dissimilar features are quantized to different prototypes and more semantics are explored. According to our experiment results in Fig. 7.2 and Fig. 7.5, we find, reassuringly, BoP is robust against codebook size: prototype number can deviate within a wide range around the true classes number (e.g., 345 for DomainNet [Peng et al., 2019]) without significantly affecting performance.

Application scope and future directions. BoP is proposed to study the two dataset-level tasks, and the datasets considered in each task share the same label space. We may encounter some situations where we need to compare datasets with different label spaces (e.g., pre-training datasets selection [Achille et al., 2019]). In this case, one potential way is to build a universal codebook on a large-scale and representative dataset similar to BoW models [Csurka et al., 2004; Zheng et al., 2017]. By doing so, the resulting BoP representations can encode diverse and sufficient semantics for comparing datasets across various label spaces. We view our BoP as a starting point to encode datasets. It would be interesting to study other dataset

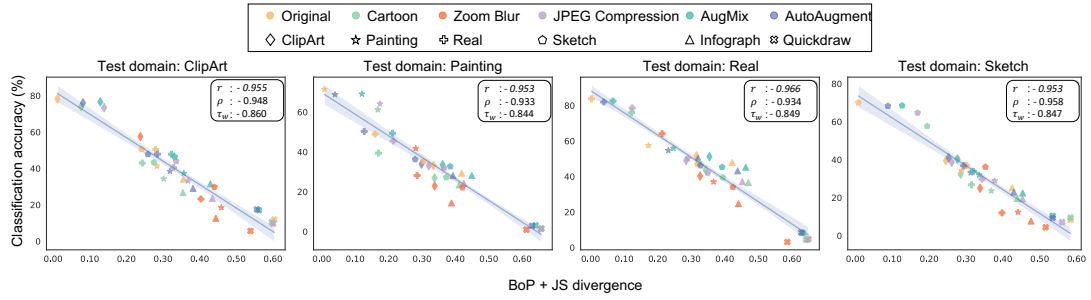


Figure 7.2: **Correlation study for training suitability of datasets.** We report the correlation strength between BoP + JS and model classification accuracy on four test domains of DomainNet: ClipArt, Painting, Real and Sketch. The model architecture is ResNet-101. Each dot denotes a model trained on a training set of DomainNet. We mark training domains (e.g., ClipArt) by different shapes and transformation operations (e.g., AugMix) by different colors. The straight lines are fit with robust linear regression [Huber, 2011]. We consistently observe high correlation results: Pearson’s correlation ($|r| > 0.95$), Spearman’s correlation ($|\rho| > 0.93$) and weighted Kendall’s correlation ($|\tau_w| > 0.84$). This suggests that BoP + JS is predictive of the suitability of a training set.

vectorization schemes and dataset-level tasks.

7.3 Comparing Training Suitability of Datasets

This task studies dataset valuation where multiple training sets are provided by different data contributors. The goal is to select the most suitable training set (ideally without training) whose trained model performs the best on a target test set. In this section, we first validate that BoP, collaborated with JS divergence (BoP + JS), is predictive of dataset suitability for the target test set. Then, we show that BoP is robust when using a wide range of codebook sizes and different networks.

7.3.1 Experimental Settings

Correlation study under DomainNet setup. We use domain generalization benchmark DomainNet Peng et al. [2019], which consists of 6 domains: Painting, Real, Infograph, Quickdraw, Sketch and ClipArt, where the tasks are 345-way object classification. Each domain has its training and test splits. We conduct the correlation study in a *leave-one-out* manner, leading to 6 groups of correlation studies, with each group using the test split of one domain as the target test set. Additionally, we apply image transformations to the training split of six original domains. Specifically, we employ ‘Cartoon’ [Jung, 2018], ‘Zoom Blur’ and ‘JPEG Compression’ [Hendrycks and Dietterich, 2019b] to convert domains’ style to be one specific type. We also use ‘AugMix’ [Hendrycks et al., 2020] and ‘AutoAugment’ [Cubuk et al., 2019], which transform images with various operations to generate domains with mixed styles. This process synthesizes 30 new datasets, so we have 36 training sets in total.

Method	ResNet-34			ResNet-101		
	r	ρ	τ_w	r	ρ	τ_w
FD [Fréchet, 1957]	-0.860	-0.926	-0.828	-0.903	-0.902	-0.802
MMD [Gretton et al., 2006]	-0.817	-0.801	-0.691	-0.821	-0.817	-0.704
KID [Bińkowski et al., 2018]	-0.773	-0.904	-0.804	-0.876	-0.896	-0.800
BoP + JS	-0.960	-0.927	-0.840	-0.961	-0.929	-0.840

Table 7.2: Compare averaged Pearson’s correlation (r), Spearman’s correlation (ρ) and weighted Kendall’s correlation (τ_w) of Fréchet distance (FD), maximum mean discrepancy (MMD), kernel inception distance (KID) and BoP + JS (codebook size 1000) on six test sets in DomainNet. We report two groups of results using ResNet-34 (Left) and ResNet-101 (Right). We show BoP + JS is more effective in assessing training set suitability than others.

We follow the training scheme provided by TLlib [Jiang et al., 2020a] to train ResNet-101 model [He et al., 2016b], whose weights are pretrained on ImageNet [Deng et al., 2009], yielding 36 models. Moreover, the penultimate output of the pre-trained ResNet-101 is used as image feature. On the test set, we generate a codebook of size 1000. Then, for each training set, we compute its BoP histogram, BoP + JS from the test set, and the accuracy of its trained model on the test set. After this, we calculate correlation strength between BoP + JS and model accuracy to evaluate whether BoP is indicative of training-set suitability.

Evaluation metric. We use Pearson’s correlation r and Spearman’s rank correlation ρ to show linearity and monotonicity between BoP-based dataset distance and model accuracy, respectively. Both metrics range in $[-1, 1]$. If $|r|$ or $|\rho|$ is close to 1, the linearity or monotonicity is strong, and vice versa. In addition to these two metrics, we also use a weighted variant of Kendall’s correlation (τ_w) [Vigna, 2015]. It is shown to be useful when selection of the best-ranked item is of interest [Shieh, 1998], while a major application of BoP + JS is to select the training dataset leading to the best performance on a test set. This metric has the same range where a number closer to -1 or 1 indicates stronger negative or positive correlation, respectively, and 0 means no correlation.

7.3.2 Evaluation

Strong correlation: A training set is more suitable for a given test set if it has small BoP + JS. Fig. 7.2 shows correlation study on ClipArt, Painting, Real and Sketch. We notice that there are strong Pearson’s correlations ($|r| > 0.95$), Spearman’s rank correlations ($|\rho| > 0.93$) and relatively high weighted Kendall’s correlations ($|\tau_w| > 0.84$) on four test sets. This suggests that BoP + JS is stable and useful across test sets. Table 7.2 compares average correlation strength of BoP + JS with Fréchet distance (FD) [Fréchet, 1957], maximum mean discrepancy (MMD) [Gretton et al., 2006] and kernel inception distance (KID) [Bińkowski et al., 2018]. They use that

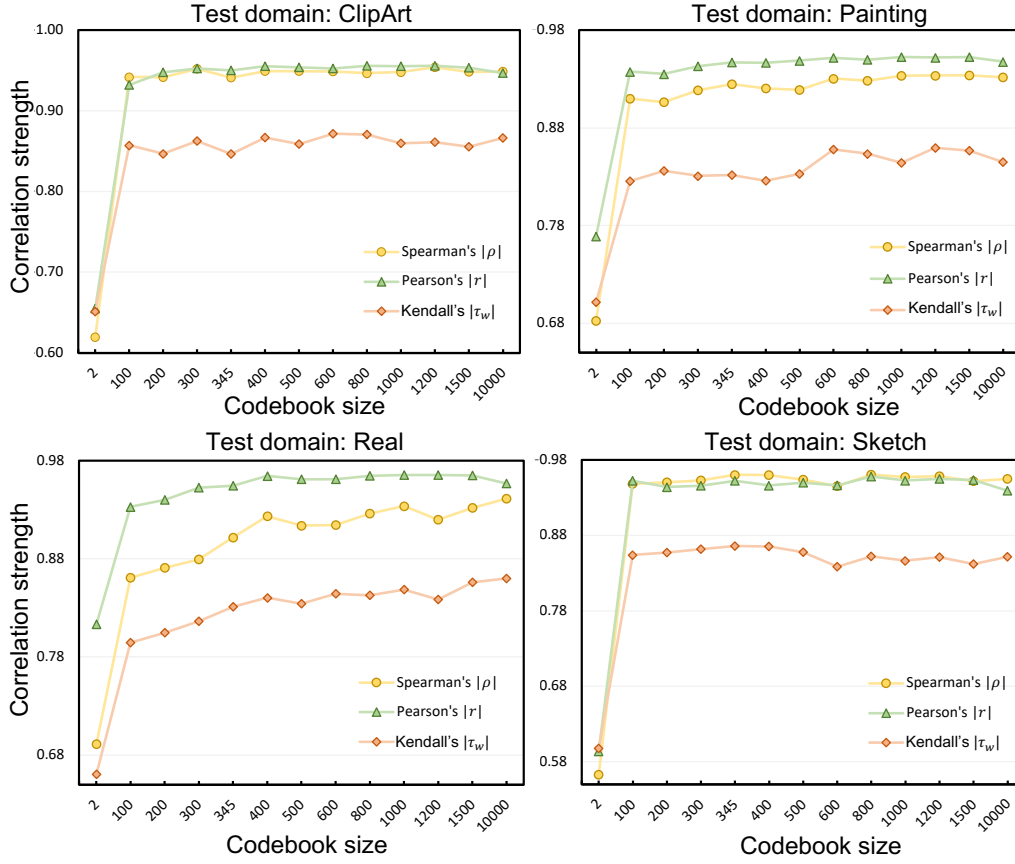


Figure 7.3: **Impact of codebook size on correlation strength for ResNet-101 on four test domains: ClipArt, Painting, Real and Sketch.** For example, on Real domain, correlation scores $|\rho|$, $|r|$ and $|\tau_w|$ are relatively low under a small size and remain stably high when the size is greater than 400.

same image features as BoP. According to their formulae, mean and covariance of these features are used for distance computation. We see that BoP + JS has the highest average correlation scores on six test sets ($|r| = 0.961$, $|\rho| = 0.929$ and $|\tau_w| = 0.840$). On average, BoP + JS is superior in depicting training sets suitability for a test set without any training procedure.

Impact of codebook size is shown in the Fig. 7.3. We construct codebooks with different size within approximately one order of magnitude around 345. We find that the three correlation scores increase and then become stable when codebook size becomes larger. This indicates that the performance BoP + JS is overall consistent.

Correlation study with a different model architecture. We additionally validate the robustness of BoP for ResNet-34 with codebook size 1000. As shown in Table 7.2, we compare the average correlation scores of BoP + JS, FD, MMD and KID. We see that BoP + JS has consistent performance on two models and remains preferable to characterize training suitability.

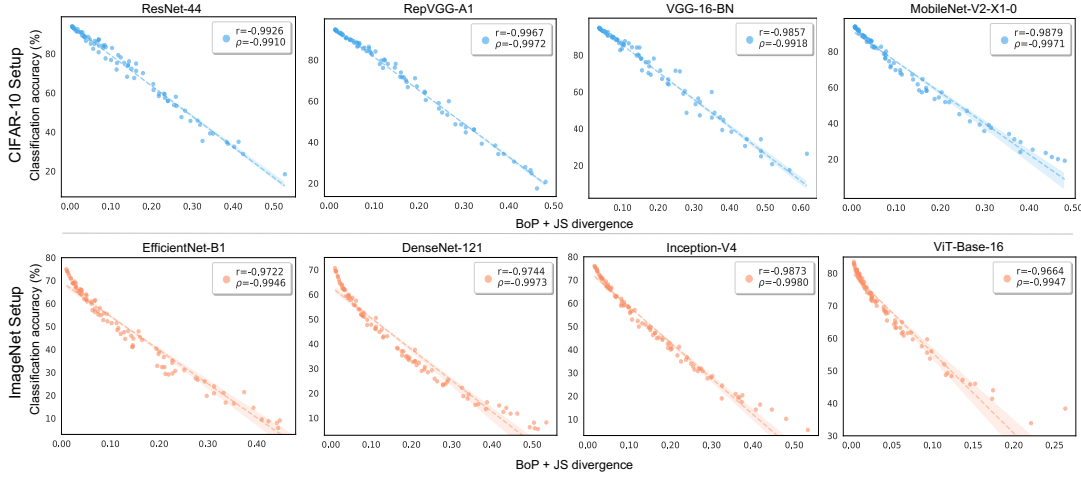


Figure 7.4: **Correlation between train-test distance measured by BoP + JS and model accuracy.** **Top:** Correlation study under CIFAR-10 setup using ResNet-44, RepVGG-A1, VGG-16-BN and MobileNet-V2. Each data point denotes a dataset from CIFAR-10-C. **Bottom:** Correlation study under ImageNet setup using EfficientNet-B1, DenseNet-121, Inception-V4 and ViT-Base-16. ImageNet-C datasets are used as test sets. The straight lines are fit with robust linear regression [Huber, 2011]. Under both setups, we observe the strong Spearman’s rank correlation ($|\rho| > 0.98$) between BoP + JS and model accuracy.

7.4 Assessing Test Set Difficulty without Labels

In the task of test set difficulty assessment, we are provided with a labeled training set and a set of unlabeled datasets for testing. Given a classifier trained on the training set, the goal is to estimate the model accuracy on these test sets without any data annotations. In this section, we first show dataset distance measured by BoP + JS exhibits strong negative correlation with classifier accuracy. We then demonstrate an accuracy predictor based on the BoP representation gives accurate performance estimates compared to state-of-the-art methods.

7.4.1 Experimental Settings

Correlation study under CIFAR-10 setup. We conduct a correlation study by comparing BoP + JS with classifier accuracy. Following the same setup in [Deng and Zheng, 2021c], we use a series of datasets sharing the same label space (but usually with distribution shift) with CIFAR-10 [Krizhevsky et al., 2009]. Specifically, we train ResNet-44 classifier [He et al., 2016b] on the training set of CIFAR-10, which consists of 50,000 images from 10 classes. Here, we use the CIFAR-10-C benchmark [Hendrycks and Dietterich, 2019a] for correlation study, which contains different types of corruptions with 5 levels of severity including per-pixel noise, blurring, synthetic weather effects, and digital transforms. Then, for each dataset, we compute its BoP vector, its BoP + JS from CIFAR-10 training set and the classifier accuracy. In addition to ResNet-44, we also study the RepVGG-A1 [Ding et al., 2021], VGG-16-BN

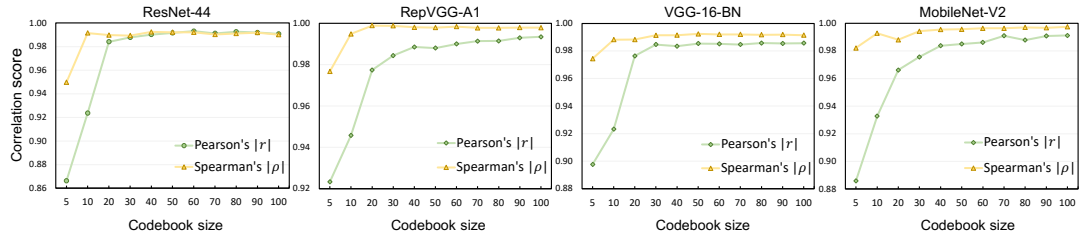


Figure 7.5: **Impact of codebook size on correlation strength on CIFAR-10-C.** Correlation scores $|\rho|$ and $|r|$ are relatively low under a small size and become stably high when the size is greater than 20 for all four model architectures.

[Simonyan and Zisserman, 2014] and MobileNet-V2 [Sandler et al., 2018] .

Predicting classification accuracy under CIFAR-10 setup. We train a regressor that takes as input the BoP representation and outputs classification accuracy. The regressor is a neural network with 3 fully connected layers and trained on CIFAR-10-C (regression training set). We evaluate accuracy prediction on CIFAR-10.1 [Recht et al., 2018b], CIFAR-10.2 [Recht et al., 2018b] and CIFAR-10.2- $\tilde{C}$ [Mintun et al., 2021]. The former two are real-world datasets with natural shift, while the latter one is manually corrupted by the same synthetic shift as [Mintun et al., 2021]. Specifically, we add 10 types of unseen and unique corruptions such as warps, blurs, color distortions and noise additions, with 5 severity levels to CIFAR-10.2. Note that, these corruptions have no overlap with those in CIFAR-10-C [Mintun et al., 2021].

For the above, we extract image features (output of penultimate layer of ResNet-44) from CIFAR-10 training set. We construct a codebook by dividing the features into 80 clusters with k-means.

Correlation study under ImageNet setup. We use DenseNet-121 [Huang et al., 2017] classifier trained on ImageNet training set. We employ a series of datasets from the ImageNet-C benchmark [Hendrycks and Dietterich, 2019b] where the classifier is tested. ImageNet-C uses the same types of corruptions as CIFAR-10-C. We construct a codebook of size 1000 on the ImageNet training set from which images features are extracted by the penultimate layer of DenseNet-121. We project each dataset in ImageNet-C onto the codebook and obtain their BoP representations. When exhibiting linear correlations, we calculate BoP + JS between each ImageNet-C dataset and the training set, and compute classification accuracy. We also use EfficientNet-B1 [Tan and Le, 2019], Inception-V4 [Szegedy et al., 2017] and ViT-Base-16 [Dosovitskiy et al., 2021b] to repeat above procedure for correlation study.

Evaluation metric. Same as Section 7.3.1, we use Pearson’s correlation r and Spearman’s rank correlation ρ to show linearity and monotonicity between BoP based dataset distance and model accuracy, respectively. To evaluate the effectiveness of accuracy estimation, we use root mean squared error (RMSE) by calculating the difference between estimated accuracy and ground truth before taking the mean across all the test sets. A larger RMSE means a less accurate prediction, and vice versa.

Compared methods. We compare our system with four existing ones. 1) *Prediction score* estimates model accuracy using the maximum of Softmax output (*i.e.*, confidence score). An image with a confidence score greater than a predefined threshold $\tau \in [0, 1]$ is considered correctly predicted. We select two thresholds ($\tau = 0.8$ and

Table 7.3: **Method comparison in predicting classifier accuracy under CIFAR-10 setup.** We compare four methods: predicted score-based method with hard threshold τ , neural network regression based on feature statistics ($\mu + \sigma + FD$) [Deng and Zheng, 2021a], average thresholded confidence with maximum confidence score function (ATC-MC) [Garg et al., 2022a] and difference of confidences (DoC) [Guillory et al., 2021]. We use CIFAR-10.1 and CIFAR-10.2 (both real-world) and CIFAR-10.2- $\bar{C}$ (manually corrupted) as unseen test sets for accuracy prediction. We use RMSE (%) to indicate precision of estimates. In each column, we compare our method with the best of the competing ones. We report results by average of five runs.

Method	CIFAR-10.1	CIFAR-10.2	CIFAR-10.2- $\bar{C}$ (50)					Overall
			S 1	S 2	S 3	S 4	S 5	
Prediction score ($\tau = 0.8$)	4.899	4.800	10.127	12.869	16.809	21.427	24.371	17.910
Prediction score ($\tau = 0.9$)	0.297	0.550	3.638	5.078	8.048	11.804	14.108	9.404
ATC-MC	2.650	2.672	3.080	4.306	7.108	11.015	13.040	8.601
DoC	0.490	0.263	2.247	2.916	5.117	9.012	6.637	5.745
$\mu + \sigma + FD$	0.455	0.561	5.875	5.823	4.724	4.908	6.486	5.602
BoP ($K = 80$)	0.218	0.122	2.458	2.818	3.730	5.836	6.451	4.551
BoP ($K = 100$)	0.186	0.124	2.849	2.808	3.548	4.025	4.777	3.677

0.9). 2) *Difference of confidence (DoC)* [Guillory et al., 2021] trains a linear regressor mapping average confidence to classifier accuracy; 3) *Average thresholded confidence with maximum confidence score function (ATC-MC)* [Garg et al., 2022b] calculates a threshold on CIFAR-10 validation set and regards an image with a confidence score higher than the threshold as correctly classified; 4) *Network regression ($\mu + \sigma + FD$)* [Deng and Zheng, 2021c] trains a neural network that takes as input the feature mean, covariance and Fréchet distance between a set of interest and training set and outputs model accuracy. All methods, if applicable, are compared under the same conditions as our system, *e.g.*, classification training set and regression training set.

7.4.2 Evaluation

Strong correlation: A test set is difficult (low accuracy) if it is dissimilar to the training set using BoP + JS. Fig. 7.4 presents the correlation study of two setups and various classifiers. We observe a very high Spearman’s rank correlation ($|\rho| > 0.99$) in all the scenarios. It indicates that classification accuracy is highly correlated with JS divergence between BoPs of training and test sets. That is, test accuracy drops proportionally to the distance between the given training set and a test set. The results demonstrate BoP + JS between training and test sets is an effective indicator of classification accuracy. More studies are presented in the supplementary materials.

Effectiveness of the BoP representation in predicting classification accuracy on various unseen test sets. After performing correlation study, we train a neural network regressor on CIFAR-10-C and test it on a series of test sets. Results are sum-

marized in Table 7.3. We have the following observations. **First**, BoP representation achieves the best accuracy prediction performance, evidenced by the lowest RMSE across all test scenarios. For example, on the test sets of CIFAR-10.2- $\bar{C}$, the RMSE of our method is 4.551, which is 1.051 lower than the second best method [Deng and Zheng, 2021c]. This clearly validates the effectiveness of BoP representation. **Second**, we observe that the “Prediction score” method is unstable. While it has good results under $\tau = 0.9$ on CIFAR-10.1 and CIFAR-10.2 datasets, it is generally inferior to the competing methods in other test scenarios. Our observation is similar to [Deng and Zheng, 2021c], suggesting that a more robust threshold selection method is needed for this method. **Third**, although BoP has slightly higher RMSE than DoC on Severity 1 of CIFAR-10.2- $\bar{C}$ (2.458 *v.s.*, 2.247), we stress that BoP is overall more stable and effective on real-world datasets and other severity levels of synthetic datasets.

Impact of codebook size is summarized in Fig. 7.5 under CIFAR-10 setup. We conduct a correlation study using different sizes on four classifiers. We observe correlation scores first increase and then become stable when codebook size is larger than 20. These results provide validation and support the use of competitive and stable codebook sizes in Table 7.3.

7.5 Summary

This chapter introduces a *bag-of-prototypes* (BoP) dataset representation to vectorize visual datasets. It first computes a codebook composed of clustering prototypes and then a prototype histogram for a dataset. The BoP vector considers the underlying local semantic distribution of a dataset and is thus more discriminative than global dataset statistics. When used in conjunction with JS divergence, the proposed descriptor better captures the underlying relationship across datasets. This advantage is validated by its promising results in two dataset-level tasks: assessing training set suitability and test set difficulty. This chapter establishes the baseline usage of the BoP scheme, and more investigations and applications will be made in future work.

Conclusion and Future Directions

8.1 Contributions and Summary

This thesis investigates the trustworthiness of modern deep learning models, particularly vision-language models (VLMs) and large multimodal models (LMMs), under the realistic conditions encountered during deployment. Classical evaluation assumes that training and test data are identically distributed and that ground-truth labels are available for assessment. In practice, neither assumption holds: models routinely encounter distribution shifts, and the cost of obtaining labels makes conventional supervised evaluation impractical. This thesis addresses both challenges through two complementary lines of inquiry: analyzing model reliability under distribution shifts, and developing label-free methods for assessing model performance.

Understanding model reliability under distribution shifts. The first part of this thesis analyzes the reliability of deep and vision-language models under distribution shifts. Chapter 3 presents a comprehensive empirical study of the robustness of contrastive language-image pre-training (CLIP) models. Instead of focusing solely on overall accuracy, the study examines multiple aspects of model behavior, including robustness to visual factors, out-of-distribution detection, calibration of prediction confidence, cross-modal retrieval performance, and sensitivity to three-dimensional perturbations. The results reveal several systematic patterns in how vision-language models behave under distribution shifts. Building on these observations, Chapter 4 studies uncertainty estimation and calibration in vision-language models. The results show that strong zero-shot performance does not necessarily imply reliable confidence estimates, and that many models are not inherently well calibrated. Nevertheless, simple post-hoc calibration methods can substantially improve uncertainty estimates, even under distribution shifts and across different label spaces, using only a small amount of calibration data.

Label-free performance assessment. The second part of this thesis considers the practical problem of assessing model performance when labels from the target environment are unavailable. Chapter 5 shows that prediction probabilities contain useful information about model generalization. Beyond conventional confidence-based statistics, the proposed SoftmaxCorr measure jointly characterizes prediction certainty and diversity and provides an informative signal for ranking classifiers across a wide range of datasets and distribution shifts. Chapter 6 extends label-free model

ranking from conventional classifiers to modern LMMs. By studying uncertainty at the token and sequence levels, it shows that signals derived directly from model generation, including negative log-likelihood and consistency across stochastic outputs, can indicate relative model performance on unlabeled multimodal benchmarks. Finally, Chapter 7 introduces a complementary data-centric perspective. Rather than relying solely on model outputs, the proposed bag-of-prototypes representation summarizes the semantic structure of an unlabeled dataset and enables dataset-level reasoning about training set suitability, dataset similarity, and test set difficulty.

8.1.1 A Unified Framework for Trustworthy Model Assessment

Although the individual contributions address different evaluation problems, they can be understood as complementary components of a unified framework for trustworthy model assessment. At a high level, reliable deployment requires answering two related questions: *how does a model fail when deployment conditions change?*, and *how can such changes in reliability be detected when representative labels are unavailable?* The two parts of this thesis address these questions from different but closely connected perspectives under realistic deployment conditions.

The reliability analyses in Chapters 3 and 4 first establish that model performance cannot be adequately characterized by a single in-distribution accuracy value. Reliability depends on multiple aspects of model behavior, including sensitivity to visual variations, robustness to distribution shifts, the ability to detect unfamiliar inputs, and the extent to which predictive confidence reflects actual correctness. These analyses therefore identify both the dimensions along which reliability should be evaluated and the types of failure that can arise after deployment.

Chapters 5 and 6 then ask whether some of these reliability properties can be inferred without repeatedly obtaining new annotations. Their results show that model outputs themselves provide useful evidence about performance. For classifiers, the distribution and structure of prediction probabilities can reveal information about generalization. For LMMs, token-level uncertainty and consistency across generated responses provide analogous signals. Thus, uncertainty is not only a reliability property but also a useful signal for model evaluation.

Chapter 7 complements this model-centric view with a data-centric one. Changes in model reliability do not arise solely from the model; they are also driven by changes in the deployment environment. Dataset-level representations provide a way to characterize these changes directly by measuring how the semantic structure of new data differs from training or reference data. Model-derived and data-derived signals therefore provide complementary evidence: the former indicate how the model is responding to its current environment, while the latter describe how that environment differs from previously observed data.

Taken together, these contributions suggest a broader evaluation process for trustworthy machine learning. Robustness analysis identifies relevant failure modes; uncertainty estimation evaluates whether a model can faithfully express its confidence; label-free model assessment uses prediction- and generation-based signals

to estimate performance when annotations are unavailable; and dataset-level analysis provides additional evidence about changes in the underlying data distribution. These components could ultimately support continuous model assessment in which changes in both model behaviour and deployment data are monitored after deployment rather than relying only on a fixed benchmark evaluated before deployment.

Importantly, this thesis does not propose a single end-to-end algorithm that integrates all of these components. Instead, it establishes a unified evaluation perspective: trustworthy machine learning requires moving beyond static accuracy-based evaluation towards complementary model-centric and data-centric signals that remain informative under distribution shift and limited supervision. From this perspective, robustness evaluation, uncertainty estimation, label-free model ranking, and dataset characterization address different parts of the same underlying problem—determining whether a model remains reliable when the conditions under which it operates differ from those under which it was originally evaluated.

8.2 Limitations

While the proposed label-free assessment methods reduce the dependence on ground-truth annotations, they also introduce assumptions and practical costs that should be considered when applying them to broader deployment settings.

Computational cost and scalability. The proposed methods have different computational bottlenecks. SoftmaxCorr in Chapter 5 requires constructing a class–class correlation matrix whose size grows quadratically with the number of classes, which may become costly for tasks with very large output spaces. For the uncertainty-based methods in Chapter 6, the computational cost depends on the uncertainty signal being used. Measures that operate on the full token probability distribution, such as entropy, become more expensive as vocabulary size increases. In contrast, sample-based consistency measures avoid requiring access to internal probability distributions, but require multiple stochastic generations for each input and therefore introduce additional inference cost. These issues become increasingly relevant as multimodal foundation models adopt larger vocabularies and more expensive generation procedures. More computationally efficient uncertainty measures and fewer-sample consistency estimators would therefore improve the scalability of label-free model assessment.

Dependence on model outputs and task assumptions. The proposed ranking methods also depend on the information exposed by the evaluated models. SoftmaxCorr assumes access to prediction probability vectors and is primarily designed for classification problems with a predefined output space. Similarly, probability-based LMM ranking requires access to token-level likelihoods, which may not be available for proprietary models accessed only through black-box APIs. Sample-based consistency provides an alternative in such settings, but its effectiveness depends on decoding choices such as sampling temperature and the number of generated responses. In addition, SoftmaxCorr relies on an estimate of the target class distribution, and its performance can therefore be influenced by the quality of this estimate.

Generality across evolving models and tasks. The empirical validation of the proposed methods necessarily reflects the model families and tasks available at the time of this research. SoftmaxCorr is evaluated primarily on image classification models, while the methods in Chapter 6 focus on autoregressive LMMs whose outputs are generated token by token. As multimodal foundation models continue to evolve, future architectures may use different multimodal fusion mechanisms, output representations, or generation paradigms, for which the same uncertainty measures may not remain equally informative. Similarly, the BoP representation in Chapter 7 is evaluated mainly on visual datasets with related semantic spaces, and its applicability to substantially different modalities or heterogeneous multimodal datasets remains to be established. These limitations suggest that the specific measures proposed in this thesis should be continually re-evaluated as model architectures and application settings evolve, even though the broader principle of using model- and data-derived signals for label-free assessment remains applicable.

8.3 Open Problems and Future Directions

8.3.1 Future Directions in Model Reliability Analysis

Understanding the reliability of deep learning models remains an important challenge, particularly as modern systems become larger and are deployed in increasingly complex environments. Building on the contributions of this thesis, we identify several directions that could further extend the reliability analysis of deep models.

Evaluation under adversarial perturbations. This thesis primarily focuses on natural distribution shifts and does not consider adversarially perturbed inputs. Adversarial examples represent a carefully constructed form of distribution shift in which imperceptible perturbations cause models to produce incorrect or attacker-controlled predictions. For VLMs specifically, adversarial inputs across visual and textual modalities may interact in complex ways, potentially amplifying model failures beyond what either modality alone would produce. These properties suggest that existing robustness evaluation and uncertainty estimation methods may not transfer directly to adversarial settings. Future work could investigate how these techniques behave under adversarial conditions and develop principled methods for detecting and rejecting such inputs to improve overall model reliability.

Reliability of Spatially Grounded AI Systems. Recent advances in large multimodal models [Liu et al., 2024c] and emerging world models [Ha and Schmidhuber, 2018; Hafner et al., 2025] introduce new reliability challenges that differ meaningfully from those studied in this thesis. Beyond static image understanding, modern AI systems are increasingly expected to reason about dynamic, three-dimensional environments, navigating physical space, predicting future states, and making sequential decisions grounded in spatial context [Driess et al., 2023; Yang et al., 2024a, 2025]. Recent benchmarking efforts have revealed that while multimodal models exhibit competitive spatial reasoning ability, their performance remains substantially below human level, with spatial reasoning identified as the primary bottleneck rather than

visual perception or language understanding [Yang et al., 2024a]. In these settings, failures may manifest as spatially inconsistent predictions, breakdowns in geometric reasoning, or compounding errors as an agent interacts with a changing environment over time [Chen et al., 2024a; Hong et al., 2023]. Understanding how uncertainty, calibration, and robustness behave when a model must reason about depth, layout, and object relationships in the physical world remains a largely open problem. Future research should develop reliability frameworks that move beyond flat image benchmarks and address the structured, sequential, and spatially grounded outputs that characterize next-generation multimodal and embodied AI systems.

Towards automated reliability evaluation pipelines. As machine learning systems are deployed at increasing scale, manual reliability evaluation becomes difficult to maintain in practice. Automated pipelines that combine robustness evaluation, uncertainty estimation, and label-free model assessment could continuously monitor model behavior as data distributions shift over time. For example, such a pipeline might flag performance degradation by tracking changes in prediction entropy or agreement across model ensembles, without requiring any labeled data. Integrating automated reliability analysis into model development and deployment workflows could help practitioners identify potential failure modes early and maintain model performance in dynamic real-world environments.

8.3.2 Future Directions in Label-Free Model Assessment

Evaluating model performance without labeled data is an increasingly important problem as machine learning systems are deployed at scale. This thesis demonstrates that signals derived from model predictions, such as probability distributions and uncertainty estimates, can provide informative indicators for ranking models on unlabeled data. However, label-free model assessment remains a relatively early research direction with many open questions.

Beyond image classification. Much existing work on label-free evaluation focuses on image classification tasks, where prediction probabilities can be directly related to model accuracy. However, modern machine learning systems increasingly address more complex tasks such as visual question answering, multimodal reasoning, generative modeling, and sequential decision-making. In these settings, model outputs are often structured, open-ended, or multi-step, making evaluation without labels significantly more challenging. Future research could explore how label-free assessment methods can be extended to these tasks, for example by leveraging token-level uncertainty, consistency across generated outputs, or agreement between multiple prompts or modalities.

Self-refinement. Another promising direction is to explore whether evaluation signals that correlate with model performance can also be used to improve models. If a metric consistently reflects model generalization ability, it may serve as a useful training or selection signal even without labeled supervision. For instance, such metrics could be incorporated into training objectives, regularization schemes, or model selection strategies to encourage models to exhibit more robust and re-

liable behaviors. Developing methods that integrate label-free evaluation signals into model optimization could help bridge the gap between model assessment and model improvement, ultimately enabling systems that become more reliable as they encounter new environments. Label-free model assessment sits at the intersection of practical deployment constraints and fundamental questions about what makes a model trustworthy. Advancing this area requires developing evaluation signals that remain robust across tasks, architectures, and deployment environments, and the payoff extends beyond convenience: in a world where labeled evaluation data are scarce or unavailable, reliable label-free methods may become a cornerstone of responsible machine learning deployment.

8.4 Final Remarks

Trustworthy machine learning cannot be established by predictive accuracy alone. Models may achieve strong benchmark performance while remaining poorly calibrated, fragile to distribution shifts, or difficult to evaluate once deployed in environments where representative labels are unavailable. The central theme of this thesis is that understanding and assessing these behaviours requires looking beyond conventional supervised evaluation.

The studies presented in this thesis approach this problem from complementary directions. Robustness analysis reveals where models fail; uncertainty estimation investigates whether confidence faithfully reflects those failures; model-derived signals enable performance comparison without target-domain annotations; and dataset representations provide a complementary view of changes in the deployment environment. Together, these contributions establish a unified perspective on model assessment in which reliability is treated as a property that must be continually examined as models, tasks, and data distributions change.

The methods developed here do not resolve all aspects of trustworthy machine learning, and their assumptions, computational requirements, and applicability must be reconsidered as foundation models continue to evolve. Nevertheless, they demonstrate that meaningful evidence about model reliability can be extracted from both model behaviour and unlabeled deployment data. Developing such evidence into scalable, architecture-agnostic, and continuously operating evaluation mechanisms will be increasingly important as multimodal models move from static benchmarks into complex real-world systems.

Appendix: Models and Datasets

A.1 Models

A.1.1 Zero-shot CLIP

We use zero-shot CLIP models accessible by OpenCLIP library [Ilharco et al., 2021], presented in the format (architecture, source). Training details for OpenAI’s models are available at <https://github.com/openai/CLIP/blob/d50d76daa670286dd6caf3bcd80b5e4823fc8e1/model-card.md>. For OpenCLIP models, detailed training information can be found in their respective model cards on Hugging Face. Below, we provide an example for (ViT-B-32, laion400m_e31):

(ViT-B-32, laion400m_e31)

Visual Encoder:	ViT-B-32
Pretraining Dataset:	laion400m
Number of Parameters:	151M
Training Samples:	400M
Batch Size:	79K
Epochs:	31

(RN50, yfcc15m)
 (RN50, cc12m)
 (RN101, yfcc15m)
 (RN50, openai)
 (RN50-quickgelu, openai)
 (RN101, openai)
 (RN101-quickgelu, openai)
 (RN50×4, openai)
 (RN50×16, openai)
 (RN50×64, openai)
 (ViT-B-32, openai)
 (ViT-B-32-quickgelu, openai)
 (ViT-B-16, openai)
 (ViT-L-14, openai)
 (ViT-L-14-336, openai)

(ViT-B-32, laion400m_e31)
 (ViT-B-32, laion400m_e32)
 (ViT-B-32, laion2b_e16)
 (ViT-B-32, laion2b_s34b_b79k)
 (ViT-B-32-quickgelu, laion400m_e31)
 (ViT-B-32-quickgelu, laion400m_e32)
 (ViT-B-16, laion400m_e31)
 (ViT-B-16, laion400m_e32)
 (ViT-B-16, laion2b_s34b_b88k)
 (ViT-B-16-plus-240, laion400m_e31)
 (ViT-B-16-plus-240, laion400m_e32)
 (ViT-L-14, laion400m_e31)
 (ViT-L-14, laion400m_e32)
 (ViT-L-14, laion2b_s32b_b82k)
 (ViT-H-14, laion2b_s32b_b79k)
 (ViT-g-14, laion2b_s12b_b42k)
 (ViT-g-14, laion2b_s34b_b88k)
 (ViT-bigG-14, laion2b_s39b_b160k)
 (convnext_base, laion400m_s13b_b51k)
 (convnext_base_w, laion2b_s13b_b82k)
 (convnext_base_w, laion2b_s13b_b82_augreg)
 (convnext_base_w, laion_aesthetic_s13b_b82k)
 (convnext_base_w_320, laion_aesthetic_s13b_b82k)
 (convnext_base_w_320, laion_aesthetic_s13b_b82k_augreg)
 (convnext_large_d, laion2b_s26b_b102k_augreg)
 (convnext_large_d_320, laion2b_s29b_b131k_ft)
 (convnext_large_d_320, laion2b_s29b_b131k_ft_soup)
 (convnext_xxlarge, laion2b_s34b_b82k_augreg)
 (convnext_xxlarge, laion2b_s34b_b82k_augreg_rewind)
 (convnext_xxlarge, laion2b_s34b_b82k_augreg_soup)

Nine models provided by [Cherti et al., 2023] <https://github.com/LAION-AI/scaling-laws-openclip>. The information regarding training cost is included in https://github.com/LAION-AI/scaling-laws-openclip/blob/master/trained_models_info.csv.

(ViT-B-32, Model-B-32_Data-80M_Samples-3B_lr-5e-4_bs-32k.pt)
 (ViT-B-32, Model-B-32_Data-80M_Samples-13B_lr-5e-4_bs-32k.pt)
 (ViT-B-32, Model-B-32_Data-80M_Samples-34B_lr-1e-3_bs-88k.pt)
 (ViT-B-16, Model-B-16_Data-80M_Samples-3B_lr-1e-3_bs-88k.pt)
 (ViT-B-16, Model-B-16_Data-80M_Samples-13B_lr-1e-3_bs-88k.pt)
 (ViT-B-16, Model-B-16_Data-80M_Samples-34B_lr-1e-3_bs-88k.pt)
 (ViT-L-14, Model-L-14_Data-80M_Samples-3B_lr-1e-3_bs-88k.pt)
 (ViT-L-14, Model-L-14_Data-80M_Samples-13B_lr-1e-3_bs-88k.pt)
 (ViT-L-14, Model-L-14_Data-80M_Samples-34B_lr-1e-3_bs-88k.pt)

We also evaluate models trained on DataComp and CommonPool [Gadre et al., 2023] provided by OpenCLIP:

(ViT-B-32, datacomp_m_s128m_b4k)
 (ViT-B-32, commonpool_m_clip_s128m_b4k)


```
(ViT-B-32, commonpool_m_laion_s128m_b4k)
(ViT-B-32, commonpool_m_image_s128m_b4k)
(ViT-B-32, commonpool_m_text_s128m_b4k)
(ViT-B-32, commonpool_m_basic_s128m_b4k)
(ViT-B-32, commonpool_m_s128m_b4k)
(ViT-B-32, datacomp_s_s13m_b4k)
(ViT-B-32, commonpool_s_clip_s13m_b4k)
(ViT-B-32, commonpool_s_laion_s13m_b4k)
(ViT-B-32, commonpool_s_image_s13m_b4k)
(ViT-B-32, commonpool_s_text_s13m_b4k)
(ViT-B-32, commonpool_s_basic_s13m_b4k)
(ViT-B-32, commonpool_s_s13m_b4k)
(ViT-B-16, datacomp_l_s1b_b8k)
(ViT-B-16, commonpool_l_clip_s1b_b8k)
(ViT-B-16, commonpool_l_laion_s1b_b8k)
(ViT-B-16, commonpool_l_image_s1b_b8k)
(ViT-B-16, commonpool_l_text_s1b_b8k)
(ViT-B-16, commonpool_l_basic_s1b_b8k)
(ViT-B-16, commonpool_l_s1b_b8k)
(ViT-L-14, datacomp_xl_s13b_b90k)
(ViT-L-14, commonpool_xl_clip_s13b_b90k)
(ViT-L-14, commonpool_xl_laion_s13b_b90k)
(ViT-L-14, commonpool_xl_s13b_b90k)
```

We also include EVA-CLIP, MetaCLIP, SigLIP and CLIPA and CLIP trained using DFN:

```
(EVA01-g-14, laion400m_s11b_b41k)
(EVA02-E-14, laion2b_s4b_b115k)
(EVA02-E-14-plus, laion2b_s9b_b144k)
(ViT-H-14-CLIPA-336, laion2b)
(ViT-B-16, dfn2b)
(ViT-B-16-quickgelu, metaclip_400m)
(ViT-L-14-quickgelu, dfn2b)
(ViT-L-14-quickgelu, metaclip_400m)
(ViT-S0400M-14-SigLIP-384, webli)
```

A.1.2 Fine-tuned CLIP on ImageNet

We consider the fine-tuned CLIP models provided in TIMM [Wightman \[2019\]](#). Detailed training information can be found in their respective model cards on Hugging Face. We use `vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k` as an example:

```
vit_huge_patch14_clip_336.
laion2b_ft_in12k_in1k
```

Architecture:	ViT-H-14
Pretraining Dataset:	LAION, ImageNet-12K
Dataset:	ImageNet-1K
Number of Parameters:	632M
GMACs:	363.7

```

vit_huge_patch14_clip_336.laion2b_ft_in12k_in1k
vit_huge_patch14_clip_224.laion2b_ft_in12k_in1k
vit_large_patch14_clip_336.laion2b_ft_in12k_in1k
vit_large_patch14_clip_224.laion2b_ft_in12k_in1k
vit_base_patch16_clip_384.laion2b_ft_in12k_in1k
vit_base_patch16_clip_224.laion2b_ft_in12k_in1k
vit_base_patch32_clip_448.laion2b_ft_in12k_in1k
vit_base_patch32_clip_384.laion2b_ft_in12k_in1k
vit_base_patch32_clip_224.laion2b_ft_in12k_in1k
vit_huge_patch14_clip_224.laion2b_ft_in1k
vit_large_patch14_clip_336.laion2b_ft_in1k
vit_large_patch14_clip_224.laion2b_ft_in1k
vit_base_patch16_clip_384.laion2b_ft_in1k
vit_base_patch16_clip_224.laion2b_ft_in1k
vit_base_patch32_clip_224.laion2b_ft_in1k
vit_large_patch14_clip_336.openai_ft_in12k_in1k
vit_large_patch14_clip_224.openai_ft_in12k_in1k
vit_base_patch16_clip_384.openai_ft_in12k_in1k
vit_base_patch16_clip_224.openai_ft_in12k_in1k
vit_base_patch32_clip_384.openai_ft_in12k_in1k
vit_large_patch14_clip_224.openai_ft_in1k
vit_base_patch16_clip_384.openai_ft_in1k
vit_base_patch16_clip_224.openai_ft_in1k
vit_base_patch32_clip_224.openai_ft_in1k
convnext_base.clip_laion2b_augreg_ft_in1k
convnext_base.clip_laion2b_augreg_ft_in1k_384
convnext_large_mlp.clip_laion2b_augreg_ft_in1k
convnext_large_mlp.clip_laion2b_augreg_ft_in1k_384
convnext_large_mlp.clip_laion2b_soup_ft_in12k_in1k_384
convnext_large_mlp.clip_laion2b_soup_ft_in12k_in1k_320
convnext_base.clip_laion2b_augreg_ft_in12k_in1k_384
convnext_base.clip_laion2b_augreg_ft_in12k_in1k

```

A.1.3 Other ImageNet Models

The standard models and those which are pre-trained on more data are publicly available on TIMM as well. We refer to the supplementary of [Taori et al., 2020] and [Deng et al., 2022b] for more details. For models trained by contrastive learning, they are provided by [Geirhos et al., 2021] in <https://github.com/bethgelab/model-vs-human>. They are:

```

PIRL [Misra and Maaten, 2020]
InsDis [Wu et al., 2018]
MoCo [He et al., 2020]
MoCoV2 [Chen et al., 2020b]
InfoMin [Tian et al., 2020]
simclr_resnet50×1, simclr_resnet50×2, simclr_resnet50×4 [Chen et al., 2020a]

```

For OOD detection, we consider following models to compare with CLIP:

```

vit_base_patch16_384,

```

```

vit_base_patch16_224_21kpre,
swinv2-1k,
swinv2-22k,
deit3-384-1k,
deit3_base_patch16_224,
xcit_medium_24_p16_224,
xcit_medium_24_p16_224_dist,
convnext_base,
convnext_base_in22ft1k,
BiT_s,
BiT_m,
tf_efficientnetv2_m,
tf_efficientnetv2_m_in21ft1k,
resnet50,
resnet18.tv_in1k,
resnet34.tv_in1k',
efficientnet_b0,
tf_efficientnet_b1.in1k,
tf_efficientnet_b7,
repvgg_b0.rvgg_in1k,
inception_v4.tf_in1k,
mobilenetv2_050.lamb_in1k,
deit3-384-22k,
deit3_base_patch16_224_in21ft1k

```

We evaluate 6 LLaVA models provided by [Karamcheti et al., 2024]:

```

prism-mistral-instruct-v0.2+7b+clip
prism-mistral-instruct-v0.2+7b+siglip
prism-llama2-chat+7b+clip
prism-llama2-chat+7b+siglip
clip-336px-resize-naive+7b
siglip-384px-resize-naive+7b

```

ImageNet models are publicly available via TIMM [Wightman, 2019]. Models pre-trained by contrastive learning can be used from model-vs-human project [Geirhos et al., 2021]. Some MAE models [He et al., 2022] are accessible from (<https://github.com/facebookresearch/mae>).

```

PIRL
InsDis
MoCo
MoCoV2
InfoMin
simclr_resnet50x4
simclr_resnet50x1
simclr_resnet50x2
resnet50_l2_eps1
resnet50_l2_eps0_5
resnet50_l2_eps0_25

```

resnet50_l2_eps0_03
resnet50_l2_eps0_01
ens_adv_inception_resnet_v2
adv_inception_v3
tf_efficientnet_b0_ap
resnet50_trained_on_SIN_and_IN
resnet50_trained_on_SIN_and_IN_then_finetuned_on_IN
resnet50_trained_on_SIN
resnet50_augmix
resnet50_cutmix
resnet50_deepaugument
resnet50_deepaugument_and_augmix
resnet50_feature_cutmix
resnet50_pixmix
convnext_xlarge_384_in22ft1k
convnext_xlarge_in22ft1k
resnetv2_152x2_bitm
resnetv2_50x1_bitm
convnext_base_384_in22ft1k
convnext_base_in22ft1k
resmlp_big_24_224_in22ft1k
resmlp_big_24_distilled_224
tf_efficientnetv2_s_in21ft1k
tf_efficientnetv2_m_in21ft1k
tf_efficientnetv2_l_in21ft1k
tf_efficientnetv2_xl_in21ft1k
vit_large_patch16_384
swin_large_patch4_window12_384
beit_large_patch16_512
beit_large_patch16_384
beit_large_patch16_224
beit_base_patch16_384
vit_base_patch16_384
vit_small_r26_s32_384
vit_tiny_patch16_384
vit_large_r50_s32_384
mixer_b16_224_miil
resmlp_big_24_224
resnetv2_50x1_bit_distilled
ig_resnext101_32x16d
ig_resnext101_32x32d
ig_resnext101_32x8d
ig_resnext101_32x48d
resnext101_32x16d_wsl
tf_efficientnet_l2_ns_475
tf_efficientnet_b7_ns
tf_efficientnet_b6_ns
tf_efficientnet_b5_ns
ssl_resnext101_32x8d
ssl_resnext101_32x16d

swsl_resnext101_32x8d
swsl_resnext101_32x16d
ssl_resnext101_32x4d
ssl_resnext50_32x4d
ssl_resnet50
swsl_resnext101_32x4d
swsl_resnext50_32x4d
swsl_resnet50
swin_small_patch4_window7_224
swin_base_patch4_window12_384
deit_base_patch16_224
deit_small_distilled_patch16_224
resmlp_36_224
cait_s36_384
cait_s24_224
convit_base
convit_tiny
twins_pcpvt_base
eca_nfnet_l1
xcit_tiny_24_p8_384_dist
efficientnet_b1
efficientnet_b3
efficientnet_b4
tf_efficientnet_b2
tf_efficientnet_lite1
convnext_base
convnext_small
resnetrs350
pit_xs_distilled_224
crossvit_small_240
botnet26t_256
tinynet_e
tinynet_d
repvgg_b2g4
mnasnet_small
dla46x_c
lcnet_050
tv_resnet34
tv_resnet50
tv_resnet101
tv_resnet152
densenet121
inception_v4
resnet26d
mobilenetv2_140
hrnet_w40
xception
xception41
resnet18
resnet34

seresnet50
mobilenetv2_050
seresnet33ts
wide_resnet50_2
wide_resnet101_2
resnet18d
hrnet_w18_small
gluon_resnet152_v1d
hrnet_w48
hrnet_w44
repvgg_b2
densenet201
hrnet_w18_small
resnet101d
gluon_resnet101_v1d
gluon_resnet101_v1s
gluon_xception65
gluon_seresnext50_32x4d
gluon_senet154
gluon_inception_v3
gluon_resnet101_v1c
tf_inception_v3
tv_densenet121
tv_resnext50_32x4d
repvgg_b1g4
resnext26ts
ghostnet_100
crossvit_9_240
deit_base_patch16_384
rexnet_150
rexnet_130
resnetrs50
resnet50d
resnet50
resnetv2_50
resnetrs152
resnetrs101
resnet50_aa
resnet50_fastaa
resnet50_randaa
vgg19_bn
vgg16_bn
vgg13_bn
vgg11_bn
vgg11
vgg11_bn
vgg16
vgg19
resnet10t
resnet14t

```
darknet53
cs3darknet_m
cs3darknet_focus_m
cs3darknet_l
cs3darknet_focus_l
regnety_040
regnety_064
regnety_080
regnetv_040
regnetv_064
regnetz_040
regnetz_040h
```

CIFAR-10 models can be downloaded through (<https://github.com/chenyaofu/pytorch-cifar-models>). The training script for training trajectory experiment borrows from (<https://github.com/kuangliu/pytorch-cifar>).

```
VGG16
VGG13
ResNet34
ResNet50
ResNet101
ResNet152
ShuffleNetG2
ShuffleNetG3
PreActResNet18
PreActResNet34
PreActResNet50
PreActResNet101
densenet_cifar
DenseNet121
DenseNet169
DenseNet201
DenseNet161
ResNeXt29_8x64d
ResNeXt29_32x4d
MobileNetV2
RegNetX_200MF
DLA
DPN
PNASNetB
RegNetX_400MF
RegNetY_400MF
MobileNet
ResNet18
VGG11
SimpleDLA
ResNeXt29_4x64d
ResNeXt29_2x64d
EfficientNetB0
```

SENet18
 ShuffleNetV2
 GoogLeNet
 DPN92
 ResNet34-160
 ResNet34-170
 ShuffleNetG2-170
 ShuffleNetG2-180
 ShuffleNetG2-190
 SimpleDLA-90
 SimpleDLA-105
 SimpleDLA-120
 SimpleDLA-135
 SimpleDLA-150
 SENet18-60
 SENet18-75
 SENet18-90
 SENet18-120
 SENet18-135
 SENet18-150
 SENet18-165
 ShuffleNetG2-170
 ShuffleNetG2-180
 ShuffleNetG2-190
 densenet_cifar-135
 densenet_cifar-150
 densenet_cifar-165
 MobileNetV2-135
 MobileNetV2-150
 MobileNetV2-165
 LeNet
 PreActResNet152

WILDS models (Camelyon17 and DomainNet) are trained by github (<https://github.com/p-lambda/wilds>).

Large multimodal models are as follows:

LLaVA-V1.5

llava_v1.5_7b
 llava_v1.5_13b

LLaVA-NeXT

llava_next_mistral_7b
 llava_next_vicuna_7b
 llava_next_vicuna_13b

LLaVA-NeXT-Interleave

llava_next_interleave_7b
llava_next_interleave_7b_dpo

LLaVA-OneVision

llava_onevision_qwen2_0.5b_ov
llava_onevision_qwen2_7b_ov
llava_onevision_qwen2_0.5b_si
llava_onevision_qwen2_7b_si

ShareGPT4V

sharegpt4v_7b
sharegpt4v_13b

InstructBLIP

InstructBLIP_7b
InstructBLIP_13b

Eagle

Eagle-X5-7B
Eagle-X5-13B
Eagle-X5-13B-Chat

InternVL

Mini-InternVL-Chat-2B-V1-5
Mini-InternVL-Chat-4B-V1-5
InternVL2-1B
InternVL2-2B
InternVL2-4B
InternVL2-8B

PaliGemma

paligemma-3b-mix-224
paligemma-3b-mix-448

Mantis

Mantis-8B-Defics2
Mantis-8B-clip-llama3
Mantis-8B-siglip-llama3

mPLUG-Owl2
mPLUG-Owl2
Qwen2-VL
Qwen2-VL-2B-Instruct
Qwen-VL
deepseek_vl2_tiny
LLaVA Prismatic
reproduction-llava-v15+7b
one-stage+7b
full-ft-multi-stage+7b
full-ft-one-stage+7b
in1k-224px+7b
dinov2-224px+7b
clip-224px+7b
siglip-224px+7b
clip-336px-resize-crop+7b
clip-336px-resize-naive+7b
siglip-384px-letterbox+7b
llama2-no-cotraining+7b
llava-lvis4v+7b
llava-lrv+7b
llava-lvis4v-lrv+7b

A.2 Datasets

We carefully check the licenses of all datasets and list the open sources to them.

Datasets used in factor-level robustness:

ImageNet [Deng et al., 2009] (<https://www.image-net.org>);

ImageNet-X [Idrissi et al., 2022] (<https://github.com/facebookresearch/imagenetx>);

Datasets used in out-of-distribution detection:

iNaturalist [Van Horn et al., 2018] (<https://www.kaggle.com/c/inaturalist-2021>);

SUN [Hendrycks et al., 2021b] (<https://groups.csail.mit.edu/vision/SUN/hierarchy.html>);

TEXTURE [Cimpoi et al., 2014] (<https://www.robots.ox.ac.uk/~vgg/data/dtd/>);

PLACES [Zhou et al., 2017] (<https://github.com/CSAILVision/places365>);

ImageNet-0 [Hendrycks et al., 2021b] (<https://github.com/hendrycks/natural-adv-examples>);

Datasets used in predictive uncertainty:

ImageNet-V2 [Shankar et al., 2021] (<https://github.com/modestyachts/ImageNetV2>);

ImageNet-Adversarial [Hendrycks et al., 2021b] (<https://github.com/hendrycks/natural-adv-examples>);

Factor	Number of samples
Pose	16080
Background	15441
Pattern	6571
Color	6476
Smaller	1473
Shape	696
Partial View	678
Subcategory	614
Texture	286
Larger	157
Darker	125
Object Blocking	78
Person Blocking	61
Style	45
Brighter	45
Multiple Objects	40
Overall	48868

Table A.1: Statistics of ImageNet-X of each factor.

ImageNet-Rendition [Hendrycks et al., 2021a] (<https://github.com/hendrycks/robustness>);
 ImageNet-Sketch [Wang et al., 2019] (<https://github.com/HaohanWang/ImageNet-Sketch>);
 ObjectNet [Barbu et al., 2019] (<https://objectnet.dev/download.html>);
 ImageNet-Vid-Robust [Shankar et al., 2021] (<https://github.com/modestyachts/imagenet-vid-robust-example>);

Datasets used in retrieval:

Flick30K [Young et al., 2014] (<https://huggingface.co/datasets/nlphuji/flickr30k>);
 MSCOCO [Chen et al., 2015] (<https://cocodataset.org/#home>);

Datasets used in 3D awareness:

ScanNet Dai et al. [2017] (<http://www.scan-net.org/>);
 NAVI [Jampani et al., 2023] (<https://github.com/google/navi>);
 SPair-71K [Min et al., 2019] (<https://cvlab.postech.ac.kr/research/SPair-71k/>);
 3DCC [Zhang et al., 2024a] (<https://3dcommoncorruptions.epfl.ch/>);

Datasets used in interaction between image/text encoders:

ImageNet-D [Zhang et al., 2024a] (https://github.com/chenshuang-zhang/imagenet_d).

Additional datasets for model ranking:

CIFAR-10 [Krizhevsky et al., 2009]:(<https://www.cs.toronto.edu/~kriz/cifar.html>);
 CIFAR-10-C [Hendrycks and Dietterich, 2019a]: (<https://github.com/hendrycks/robustness>);
 CIFAR-10.2 [Recht et al., 2018b]:(<https://github.com/modestyachts/CIFAR-10.1>);
 CINIC [Darlow et al., 2018]: (<https://www.v7labs.com/open-datasets/cinic-10>);

Camelyon17 [Bandi et al., 2018]: (<https://camelyon17.grand-challenge.org/>);
DomainNet [Peng et al., 2019]: (<http://ai.bu.edu/M3SDA/>);

Additional datasets for LMM ranking: VLMEvalKit [Duan et al., 2024]. We consider 8 datasets and their corresponding links to download TSV files via the toolkit:

ScienceQA [Lu et al., 2022] (https://opencompass.openxlab.space/utils/VLMEval/ScienceQA_TEST.tsv);

AI2D [Hiippala et al., 2021] (https://opencompass.openxlab.space/utils/VLMEval/AI2D_TEST.tsv);

ChartQA [Masry et al., 2022] (https://opencompass.openxlab.space/utils/VLMEval/ChartQA_TEST.tsv);

OCRVQA [Mishra et al., 2019] (https://opencompass.openxlab.space/utils/VLMEval/OCRVQA_TESTCORE.tsv);

TextVQA [Singh et al., 2019] (https://opencompass.openxlab.space/utils/VLMEval/TextVQA_VAL.tsv);

DocVQA [Mathew et al., 2021] (https://opencompass.openxlab.space/utils/VLMEval/DocVQA_VAL.tsv);

RealWorldQA [x.ai, 2024] (<https://opencompass.openxlab.space/utils/VLMEval/RealWorld.tsv>);

MMMU [Yue et al., 2024] (https://opencompass.openxlab.space/utils/VLMEval/MMMU_DEV_VAL.tsv);

GQA [Ainslie et al., 2023] (https://opencompass.openxlab.space/utils/VLMEval/GQA_TestDev_Balanced.tsv)

Bibliography

- ACHIAM, J.; ADLER, S.; AGARWAL, S.; AHMAD, L.; AKKAYA, I.; ALEMAN, F. L.; ALMEIDA, D.; ALTENSCHMIDT, J.; ALTMAN, S.; ANADKAT, S.; ET AL., 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, (2023). (cited on page 32)
- ACHILLE, A.; LAM, M.; TEWARI, R.; RAVICHANDRAN, A.; MAJI, S.; FOWLKES, C. C.; SOATTO, S.; AND PERONA, P., 2019. Task2vec: Task embedding for meta-learning. In *International Conference on Computer Vision*, 6430–6439. (cited on pages 82, 83, and 86)
- AGOSTINELLI, A.; PÁNDY, M.; UIJLINGS, J.; MENSINK, T.; AND FERRARI, V., 2022a. How stable are transferability metrics evaluations? In *European Conference on Computer Vision*, 303–321. Springer. (cited on page 86)
- AGOSTINELLI, A.; UIJLINGS, J.; MENSINK, T.; AND FERRARI, V., 2022b. Transferability metrics for selecting source model ensembles. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7936–7946. (cited on page 86)
- AINSLIE, J.; LEE-THORP, J.; DE JONG, M.; ZEMLYANSKIY, Y.; LEBRÓN, F.; AND SANGHAI, S., 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Conference on Empirical Methods in Natural Language Processing*. (cited on pages 75 and 114)
- ARORA, S.; GE, R.; NEYSHABUR, B.; AND ZHANG, Y., 2018a. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, 254–263. (cited on page 9)
- ARORA, S.; GE, R.; NEYSHABUR, B.; AND ZHANG, Y., 2018b. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, 254–263. (cited on page 54)
- AROYO, L.; LEASE, M.; PARITOSH, P.; AND SCHAEKERMANN, M., 2022. Data excellence for ai: why should you care? *Interactions*, 29, 2 (2022), 66–69. (cited on page 81)
- ASANO, Y. M.; RUPPRECHT, C.; AND VEDALDI, A., 2019. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations*. (cited on page 58)
- AZARIA, A. AND MITCHELL, T., 2023. The internal state of an LLM knows when it’s lying. In *Conference on Empirical Methods in Natural Language Processing*. <https://openreview.net/forum?id=y2V6YgLaW7>. (cited on page 80)
- BAEK, C.; JIANG, Y.; RAGHUNATHAN, A.; AND KOLTER, J. Z., 2022. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In *Advances in Neural Information Processing Systems*, 19274–19289. (cited on page 55)

-
- BALDOCK, R.; MAENNEL, H.; AND NEYSHABUR, B., 2021. Deep learning through the lens of example difficulty. In *Advances in Neural Information Processing Systems*, 10876–10889. (cited on pages 57 and 59)
- BANDI, P.; GEESINK, O.; MANSON, Q.; VAN DIJK, M.; BALKENHOL, M.; HERMSEN, M.; BENJORDI, B. E.; LEE, B.; PAENG, K.; ZHONG, A.; ET AL., 2018. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, (2018). (cited on pages 60 and 114)
- BARBU, A.; MAYO, D.; ALVERIO, J.; LUO, W.; WANG, C.; GUTFREUND, D.; TENENBAUM, J.; AND KATZ, B., 2019. Objectnet: a large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*. (cited on pages 1, 7, 11, 15, 42, 59, and 113)
- BARTLETT, P. L.; FOSTER, D. J.; AND TELGARSKY, M., 2017. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, 6241–6250. (cited on pages 9 and 53)
- BAY, H.; TUYTELAARS, T.; AND GOOL, L. V., 2006. Surf: Speeded up robust features. In *European Conference on Computer Vision*, 404–417. Springer. (cited on page 84)
- BELLO, I.; FEDUS, W.; DU, X.; CUBUK, E. D.; SRINIVAS, A.; LIN, T.-Y.; SHLENS, J.; AND ZOPH, B., 2021. Revisiting resnets: Improved training and scaling strategies. In *Advances in Neural Information Processing Systems*. (cited on pages 20 and 42)
- BEN-DAVID, S.; BLITZER, J.; CRAMMER, K.; AND PEREIRA, F., 2006a. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 137–144. (cited on page 7)
- BEN-DAVID, S.; BLITZER, J.; CRAMMER, K.; AND PEREIRA, F., 2006b. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*. (cited on page 54)
- BENDALE, A. AND BOULT, T. E., 2016. Towards open set deep networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1563–1572. (cited on page 8)
- BEYER, L.; STEINER, A.; PINTO, A. S.; KOLESNIKOV, A.; WANG, X.; SALZ, D.; NEUMANN, M.; AL-ABDULMOHSIN, I.; TSCHANNEN, M.; BUGLIARELLO, E.; ET AL., 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, (2024). (cited on page 75)
- BIŃKOWSKI, M.; SUTHERLAND, D. J.; ARBEL, M.; AND GRETTON, A., 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, (2018). (cited on pages 86 and 88)
- BITTERWOLF, J.; MÜLLER, M.; AND HEIN, M., 2023. In or out? Fixing ImageNet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, 2471–2506. (cited on pages 15, 22, and 37)
- BOOTS, B.; SUGIHARA, K.; CHIU, S. N.; AND OKABE, A., 2009. Spatial tessellations: concepts and applications of voronoi diagrams. (2009). (cited on page 86)
- BROWN, T.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J. D.; DHARIWAL, P.; NEELAKANTAN, A.; SHYAM, P.; SASTRY, G.; ASKELL, A.; ET AL., 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*. (cited on page 48)

-
- CHEN, B.; XU, Z.; KIRMANI, S.; ICHTER, B.; DRIESS, D.; FLORENCE, P.; SADIGH, D.; GUIBAS, L.; AND XIA, F., 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14455–14465. (cited on page 99)
- CHEN, J.; LIU, F.; AVCI, B.; WU, X.; LIANG, Y.; AND JHA, S., 2021. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. In *Advances in Neural Information Processing Systems*, 14980–14992. (cited on page 8)
- CHEN, J.; YU, Q.; SHEN, X.; YUILLE, A.; AND CHEN, L.-C., 2024b. Vitamin: Designing scalable vision models in the vision-language era. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 35)
- CHEN, L.; CHEN, H.; WEI, Z.; JIN, X.; TAN, X.; JIN, Y.; AND CHEN, E., 2022. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 57)
- CHEN, L.; LI, J.; DONG, X.; ZHANG, P.; HE, C.; WANG, J.; ZHAO, F.; AND LIN, D., 2023a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*. (cited on page 75)
- CHEN, T.; KORNBLITH, S.; NOROUZI, M.; AND HINTON, G., 2020a. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. (cited on pages 14 and 104)
- CHEN, X.; FAN, H.; GIRSHICK, R.; AND HE, K., 2020b. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, (2020). (cited on page 104)
- CHEN, X.; FANG, H.; LIN, T.-Y.; VEDANTAM, R.; GUPTA, S.; DOLLÁR, P.; AND ZITNICK, C. L., 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, (2015). (cited on pages 15, 25, and 113)
- CHEN, X.; LIN, M.; SCHÄRLI, N.; AND ZHOU, D., 2023b. Teaching large language models to self-debug. In *International Conference on Learning Representations*. (cited on page 73)
- CHEN, Z.; WU, J.; WANG, W.; SU, W.; CHEN, G.; XING, S.; ZHONG, M.; ZHANG, Q.; ZHU, X.; LU, L.; ET AL., 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198. (cited on page 75)
- CHERTI, M.; BEAUMONT, R.; WIGHTMAN, R.; WORTSMAN, M.; ILHARCO, G.; GORDON, C.; SCHUHMANN, C.; SCHMIDT, L.; AND JITSEV, J., 2023. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 12, 39, and 102)
- CHUANG, C.-Y.; MROUEH, Y.; GREENEWALD, K.; TORRALBA, A.; AND JEGELKA, S., 2021. Measuring generalization with optimal transport. In *Advances in Neural Information Processing Systems*, 8294–8306. (cited on pages 9 and 53)
- CHUANG, C.-Y.; TORRALBA, A.; AND JEGELKA, S., 2020. Estimating generalization under distribution shifts via domain-invariant representations. In *International Conference on Machine Learning*. (cited on page 9)

-
- CIMPOI, M.; MAJI, S.; KOKKINOS, I.; MOHAMED, S.; AND VEDALDI, A., 2014. Describing textures in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 15 and 112)
- COBBE, K.; KOSARAJU, V.; BAVARIAN, M.; CHEN, M.; JUN, H.; KAISER, L.; PLAPPERT, M.; TWOREK, J.; HILTON, J.; NAKANO, R.; HESSE, C.; AND SCHULMAN, J., 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, (2021). (cited on page 73)
- CSURKA, G.; DANCE, C.; FAN, L.; WILLAMOWSKI, J.; AND BRAY, C., 2004. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, vol. 1, 1–2. (cited on pages 85 and 86)
- CUBUK, E. D.; ZOPH, B.; MANE, D.; VASUDEVAN, V.; AND LE, Q. V., 2019. Autoaugment: Learning augmentation policies from data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 87)
- CUI, S.; WANG, S.; ZHUO, J.; LI, L.; HUANG, Q.; AND TIAN, Q., 2020. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3941–3950. (cited on page 57)
- DAGAN, I.; LEE, L.; AND PEREIRA, F., 1997. Similarity-based methods for word sense disambiguation. In *Annual Meeting of the Association for Computational Linguistics*, 56–63. (cited on page 85)
- DAI, A.; CHANG, A. X.; SAVVA, M.; HALBER, M.; FUNKHOUSER, T.; AND NIESSNER, M., 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5828–5839. (cited on pages 15, 26, and 113)
- DAI, W.; LI, J.; LI, D.; TIONG, A. M. H.; ZHAO, J.; WANG, W.; LI, B.; FUNG, P.; AND HOI, S., 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*. (cited on pages 67, 69, and 75)
- DARLOW, L. N.; CROWLEY, E. J.; ANTONIOU, A.; AND STORKEY, A. J., 2018. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, (2018). (cited on pages 42, 59, and 113)
- DENG, A.; LI, S.; XIONG, M.; CHEN, Z.; AND HOOI, B., 2022a. Trust, but verify: Using self-supervised probing to improve trustworthiness. In *European Conference on Computer Vision*. (cited on page 54)
- DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; AND FEI-FEI, L., 2009. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 14, 40, 42, 59, 81, 88, and 112)
- DENG, W.; GOULD, S.; AND ZHENG, L., 2022b. On the strong correlation between model invariance and generalization. In *Advances in Neural Information Processing Systems*. (cited on page 104)
- DENG, W.; SUH, Y.; GOULD, S.; AND ZHENG, L., 2023. Confidence and dispersity speak: Characterizing prediction matrix for unsupervised accuracy estimation. In *International Conference on Machine Learning*, 7658–7674. (cited on pages 8 and 9)
- DENG, W. AND ZHENG, L., 2021a. Are labels always necessary for classifier accuracy evaluation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15069–15078. (cited on pages 9, 82, and 92)

-
- DENG, W. AND ZHENG, L., 2021b. Are labels always necessary for classifier accuracy evaluation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15069–15078. (cited on page 70)
- DENG, W. AND ZHENG, L., 2021c. Are labels always necessary for classifier accuracy evaluation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (2021), 1–1. doi: 10.1109/TPAMI.2021.3136244. (cited on pages 90, 92, and 93)
- DING, X.; CHEN, H.; ZHANG, X.; HAN, J.; AND DING, G., 2022. Repmlpnet: Hierarchical vision mlp with re-parameterized locality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 578–587. (cited on page 14)
- DING, X.; ZHANG, X.; MA, N.; HAN, J.; DING, G.; AND SUN, J., 2021. Repvgg: Making vgg-style convnets great again. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733–13742. (cited on page 90)
- DINH, L.; PASCANU, R.; BENGIO, S.; AND BENGIO, Y., 2017. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 1019–1028. (cited on pages 9 and 54)
- DJOLONGA, J.; YUNG, J.; TSCHANNEN, M.; ROMIJNDERS, R.; BEYER, L.; KOLESNIKOV, A.; PUIGSERVER, J.; MINDERER, M.; D’AMOUR, A.; MOLDOVAN, D.; ET AL., 2021. On robustness and transferability of convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16458–16468. (cited on page 7)
- DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; AND HOULSBY, N., 2021a. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=YicbFdNTTy>. (cited on pages 14, 41, 42, and 64)
- DOSOVITSKIY, A.; BEYER, L.; KOLESNIKOV, A.; WEISSENBORN, D.; ZHAI, X.; UNTERTHINER, T.; DEHGHANI, M.; MINDERER, M.; HEIGOLD, G.; GELLY, S.; USZKOREIT, J.; AND HOULSBY, N., 2021b. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. (cited on page 91)
- DRIESS, D.; XIA, F.; SAJJADI, M. S. M.; LYNCH, C.; CHOWDHERY, A.; ICHTER, B.; WAHID, A.; TOMPSON, J.; VUONG, Q.; YU, T.; HUANG, W.; CHEBOTAR, Y.; SERMANET, P.; DUCKWORTH, D.; LEVINE, S.; VANHOUCHE, V.; HAUSMAN, K.; TOUSSAINT, M.; GREFF, K.; ZENG, A.; MORDATCH, I.; AND FLORENCE, P., 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning (ICML)*. (cited on page 98)
- DU, Q.; FABER, V.; AND GUNZBURGER, M., 1999. Centroidal voronoi tessellations: Applications and algorithms. *SIAM review*, 41, 4 (1999), 637–676. (cited on page 86)
- DUAN, H.; YANG, J.; QIAO, Y.; FANG, X.; CHEN, L.; LIU, Y.; DONG, X.; ZANG, Y.; ZHANG, P.; WANG, J.; LIN, D.; AND CHEN, K., 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. <https://arxiv.org/abs/2407.11691>. (cited on page 114)
- DZIUGAITE, G. K.; DROUIN, A.; NEAL, B.; RAJKUMAR, N.; CABALLERO, E.; WANG, L.; MITLIAGKAS, I.; AND ROY, D. M., 2020. In search of robust measures of generalization. In *Advances in Neural Information Processing Systems*. (cited on page 9)

-
- DZIUGAITE, G. K. AND ROY, D. M., 2017a. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence*. (cited on page 9)
- DZIUGAITE, G. K. AND ROY, D. M., 2017b. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Conference on Uncertainty in Artificial Intelligence*. (cited on page 54)
- EL BANANI, M.; RAJ, A.; MANINIS, K.-K.; KAR, A.; LI, Y.; RUBINSTEIN, M.; SUN, D.; GUIBAS, L.; JOHNSON, J.; AND JAMPANI, V., 2024. Probing the 3d awareness of visual foundation models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21795–21806. (cited on pages 12, 25, 26, and 27)
- ENGSTROM, L.; ILYAS, A.; SANTURKAR, S.; TSIPRAS, D.; STEINHARDT, J.; AND MADRY, A., 2020. Identifying statistical bias in dataset replication. In *International Conference on Machine Learning*, 2922–2932. PMLR. (cited on page 55)
- EYUBOGLU, S.; KARLAŠ, B.; RÉ, C.; ZHANG, C.; AND ZOU, J., 2022. dcbench: A benchmark for data-centric ai systems. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning*. (cited on page 81)
- FANG, A.; ILHARCO, G.; WORTSMAN, M.; WAN, Y.; SHANKAR, V.; DAVE, A.; AND SCHMIDT, L., 2022. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*. (cited on pages 12, 16, and 39)
- FANG, A.; JOSE, A. M.; JAIN, A.; SCHMIDT, L.; TOSHEV, A. T.; AND SHANKAR, V., 2024. Data filtering networks. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=KAk6ngZ09F>. (cited on pages 14 and 35)
- FEI-FEI, L. AND PERONA, P., 2005. A bayesian hierarchical model for learning natural scene categories. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 524–531. (cited on page 85)
- FORT, S.; REN, J.; AND LAKSHMINARAYANAN, B., 2021. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*. (cited on page 21)
- FRÉCHET, M., 1957. Sur la distance de deux lois de probabilité. *Comptes Rendus Hebdomadaires des Seances de L Academie des Sciences*, 244, 6 (1957), 689–692. (cited on pages 79, 83, 86, and 88)
- FREEDMAN, D. AND DIACONIS, P., 1981. On the histogram as a density estimator: L 2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 57, 4 (1981), 453–476. (cited on page 86)
- FU, J.; LI, S.; JIANG, Y.; LIN, K.-Y.; QIAN, C.; LOY, C. C.; WU, W.; AND LIU, Z., 2022. Stylegan-human: A data-centric odyssey of human generation. In *European Conference on Computer Vision*, 1–19. Springer. (cited on page 81)
- GADRE, S. Y.; ILHARCO, G.; FANG, A.; HAYASE, J.; SMYRNIS, G.; NGUYEN, T.; MARTEN, R.; WORTSMAN, M.; GHOSH, D.; ZHANG, J.; ET AL., 2023. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, 27092–27112. (cited on pages 14, 35, 37, 41, and 102)

-
- GALIL, I.; DABBAH, M.; AND EL-YANIV, R., 2023. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers. In *International Conference on Learning Representations*. (cited on page 40)
- GANIN, U.; AJAKAN, G.; LAROCHELLE, L.; MARCHAND; AND LEMPITSKY, 2016. Domain-adversarial training of neural networks. In *Journal of Machine Learning Research*. (cited on page 64)
- GARG, S.; BALAKRISHNAN, S.; LIPTON, Z. C.; NEYSHABUR, B.; AND SEDGHI, H., 2022a. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*. (cited on pages 8, 54, 56, 57, 59, 73, and 92)
- GARG, S.; BALAKRISHNAN, S.; LIPTON, Z. C.; NEYSHABUR, B.; AND SEDGHI, H., 2022b. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*. (cited on page 92)
- GARG, S.; WU, Y.; BALAKRISHNAN, S.; AND LIPTON, Z., 2020. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, 3290–3300. (cited on pages 62 and 65)
- GEIRHOS, R.; JACOBSEN, J.-H.; MICHAELIS, C.; ZEMEL, R.; BRENDDEL, W.; BETHGE, M.; AND WICHMANN, F. A., 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 11 (2020), 665–673. (cited on page 18)
- GEIRHOS, R.; NARAYANAPPA, K.; MITZKUS, B.; THIERINGER, T.; BETHGE, M.; WICHMANN, F. A.; AND BRENDDEL, W., 2021. Partial success in closing the gap between human and machine vision. In *Advances in Neural Information Processing Systems*. (cited on pages 104 and 105)
- GEIRHOS, R.; RUBISCH, P.; MICHAELIS, C.; BETHGE, M.; WICHMANN, F. A.; AND BRENDDEL, W., 2019a. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bygh9j09KX>. (cited on page 7)
- GEIRHOS, R.; RUBISCH, P.; MICHAELIS, C.; BETHGE, M.; WICHMANN, F. A.; AND BRENDDEL, W., 2019b. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=Bygh9j09KX>. (cited on pages 15, 18, and 20)
- GEIRHOS, R.; RUBISCH, P.; RAUBER, J.; TEMME, C. R. M.; MICHAELIS, C.; BRENDDEL, W.; BETHGE, M.; AND WICHMANN, F. A., 2019c. Inducing a human-like shape bias leads to emergent human-level distortion robustness in cnns. *Journal of Vision*, 19, 10 (2019), 209c–209c. (cited on page 20)
- GHORBANI, A. AND ZOU, J., 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, 2242–2251. PMLR. (cited on page 81)
- GOTTESMAN, D. AND GEVA, M., 2024. Estimating knowledge in large language models without generating a single token. In *Conference on Empirical Methods in Natural Language Processing*, 3994–4019. Association for Computational Linguistics, Miami, Florida, USA. [doi:10.18653/v1/2024.emnlp-main.232](https://doi.org/10.18653/v1/2024.emnlp-main.232). <https://aclanthology.org/2024.emnlp-main.232/>. (cited on page 80)

-
- GOYAL, S.; KUMAR, A.; GARG, S.; KOLTER, Z.; AND RAGHUNATHAN, A., 2023. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19338–19347. (cited on pages 19, 20, 33, and 39)
- GRETTON, A.; BORGWARDT, K.; RASCH, M.; SCHÖLKOPF, B.; AND SMOLA, A., 2006. A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*. (cited on pages 83, 86, and 88)
- GU, X.; LIN, T.-Y.; KUO, W.; AND CUI, Y., 2021. Open-vocabulary object detection via vision and language knowledge distillation. (2021). (cited on page 39)
- GUILLORY, D.; SHANKAR, V.; EBRAHIMI, S.; DARRELL, T.; AND SCHMIDT, L., 2021. Predicting with confidence on unseen distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1134–1144. (cited on pages 8, 9, 54, 56, and 92)
- GUO, C.; PLEISS, G.; SUN, Y.; AND WEINBERGER, K. Q., 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330. (cited on pages 8, 23, 41, 42, 56, and 77)
- GUPTA, K.; RAHIMI, A.; AJANTHAN, T.; MENSINK, T.; SMINCHISESCU, C.; AND HARTLEY, R., 2021. Calibration of neural networks using splines. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=eQe8DEWNN2W>. (cited on pages 23, 42, and 48)
- HA, D. AND SCHMIDHUBER, J., 2018. World models. *arXiv preprint arXiv:1803.10122*, (2018). (cited on page 98)
- HAFNER, D.; PASUKONIS, J.; BA, J.; ET AL., 2025. Mastering diverse control tasks through world models. *Nature*, 640 (2025), 647–653. doi:10.1038/s41586-025-08744-2. (cited on page 98)
- HAN, B. AND LIU, T., 2025. *Trustworthy Machine Learning under Imperfect Data*. Springer, Singapore, 1 edn. ISBN 978-981-96-9396-2. doi:10.1007/978-981-96-9396-2. (cited on page 1)
- HE, K.; CHEN, X.; XIE, S.; LI, Y.; DOLLÁR, P.; AND GIRSHICK, R., 2022. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 105)
- HE, K.; FAN, H.; WU, Y.; XIE, S.; AND GIRSHICK, R., 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 14 and 104)
- HE, K.; ZHANG, X.; REN, S.; AND SUN, J., 2016a. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 14 and 42)
- HE, K.; ZHANG, X.; REN, S.; AND SUN, J., 2016b. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 88 and 90)
- HEIN, M.; ANDRIUSHCHENKO, M.; AND BITTERWOLF, J., 2019. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 41–50. (cited on page 56)

-
- HENDRYCKS, D.; BASART, S.; MU, N.; KADAVATH, S.; WANG, F.; DORUNDO, E.; DESAI, R.; ZHU, T.; PARAJULI, S.; GUO, M.; ET AL., 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *International Conference on Computer Vision*, 8340–8349. (cited on pages 1, 7, 11, 15, 42, 54, 59, and 113)
- HENDRYCKS, D. AND DIETTERICH, T., 2019a. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*. (cited on pages 7, 29, 59, 64, 90, and 113)
- HENDRYCKS, D. AND DIETTERICH, T., 2019b. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*. (cited on pages 87 and 91)
- HENDRYCKS, D. AND GIMPEL, K., 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*. (cited on pages 8, 21, 54, 56, and 59)
- HENDRYCKS, D.; LEE, K.; AND MAZEIKA, M., 2019. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, 2712–2721. (cited on page 8)
- HENDRYCKS, D.; MU, N.; CUBUK, E. D.; ZOPH, B.; GILMER, J.; AND LAKSHMINARAYANAN, B., 2020. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*. (cited on page 87)
- HENDRYCKS, D.; ZHAO, K.; BASART, S.; STEINHARDT, J.; AND SONG, D., 2021b. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15262–15271. (cited on pages 7, 11, 15, and 112)
- HERMANN, K.; CHEN, T.; AND KORNBLITH, S., 2020. The origins and prevalence of texture bias in convolutional neural networks. In *Advances in Neural Information Processing Systems*, 19000–19015. (cited on page 20)
- HEUSEL, M.; RAMSAUER, H.; UNTERTHINER, T.; NESSLER, B.; AND HOCHREITER, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. (cited on pages 46 and 47)
- HIIPPALA, T.; ALIKHANI, M.; HAVERINEN, J.; KALLIOKOSKI, T.; LOGACHEVA, E.; OREKHOVA, S.; TUOMAINEN, A.; STONE, M.; AND BATEMAN, J. A., 2021. Ai2d-rst: A multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*, 55 (2021), 661–688. (cited on pages 67, 75, and 114)
- HONG, Y.; ZHEN, H.; CHEN, P.; ZHENG, S.; DU, Y.; CHEN, Z.; AND GAN, C., 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36 (2023), 20482–20494. (cited on page 99)
- HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; AND ADAM, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, (2017). (cited on page 64)
- HSU, Y.-C.; SHEN, Y.; JIN, H.; AND KIRA, Z., 2020. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10951–10960. (cited on page 8)

-
- HU, D.; LUO, M.; LIANG, J.; AND FOO, C.-S., 2024. Towards reliable model selection for unsupervised domain adaptation: An empirical study and a certified baseline. In *Advances in Neural Information Processing Systems*. <https://openreview.net/forum?id=rI7kbFTSpr>. (cited on page 71)
- HUANG, G.; LIU, Z.; VAN DER MAATEN, L.; AND WEINBERGER, K. Q., 2017. Densely connected convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4700–4708. (cited on pages 64 and 91)
- HUANG, Y.; SONG, J.; WANG, Z.; ZHAO, S.; CHEN, H.; JUEFEI-XU, F.; AND MA, L., 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. (cited on pages 71 and 73)
- HUANG, Z.; LIU, C.; DONG, Y.; SU, H.; ZHENG, S.; AND LIU, T., 2024. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *International Conference on Machine Learning*. (cited on page 67)
- HUANG, Z.; NIU, G.; HAN, B.; SUGIYAMA, M.; AND LIU, T., 2025. Towards out-of-modal generalization without instance-level modal correspondence. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=LuVulfPgZN>. (cited on page 67)
- HUBER, P. J., 2011. Robust statistics. In *International Encyclopedia of Statistical Science*, 1248–1251. Springer. (cited on pages 17, 21, 33, 34, 35, 56, 60, 69, 87, and 90)
- IDRISSI, B. Y.; BOUCHACOURT, D.; BALESTRIERO, R.; EVTIMOV, I.; HAZIRBAS, C.; BALLAS, N.; VINCENT, P.; DROZDZAL, M.; LOPEZ-PAZ, D.; AND IBRAHIM, M., 2022. Imagenet-x: Understanding model mistakes with factor of variation annotations. In *International Conference on Learning Representations*. (cited on pages 7, 12, 15, and 112)
- ILHARCO, G.; WORTSMAN, M.; WIGHTMAN, R.; GORDON, C.; CARLINI, N.; TAORI, R.; DAVE, A.; SHANKAR, V.; NAMKOONG, H.; MILLER, J.; HAJISHIRZI, H.; FARHADI, A.; AND SCHMIDT, L., 2021. Openclip. [doi:10.5281/zenodo.5143773](https://doi.org/10.5281/zenodo.5143773). <https://doi.org/10.5281/zenodo.5143773>. If you use this software, please cite it as below. (cited on pages 14, 42, and 101)
- JAMPANI, V.; MANINIS, K.-K.; ENGELHARDT, A.; KARPUR, A.; TRUONG, K.; SARGENT, K.; POPOV, S.; ARAUJO, A.; MARTIN BRUALLA, R.; PATEL, K.; ET AL., 2023. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. In *Advances in Neural Information Processing Systems Dataset and Benchmark Track*. (cited on pages 15, 26, and 113)
- JÉGOU, H. AND CHUM, O., 2012. Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening. In *European Conference on Computer Vision*, 774–787. (cited on page 84)
- JÉGOU, H.; DOUZE, M.; SCHMID, C.; AND PÉREZ, P., 2010. Aggregating local descriptors into a compact image representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3304–3311. (cited on page 84)
- JELINEK, F.; MERCER, R. L.; BAHL, L. R.; AND BAKER, J. M., 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62 (1977). <https://api.semanticscholar.org/CorpusID:121680873>. (cited on page 72)

-
- JIA, C.; YANG, Y.; XIA, Y.; CHEN, Y.-T.; PAREKH, Z.; PHAM, H.; LE, Q.; SUNG, Y.-H.; LI, Z.; AND DUEIRIG, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. (cited on pages 11 and 39)
- JIANG, A. Q.; SABLAYROLLES, A.; MENSCH, A.; BAMFORD, C.; CHAPLOT, D. S.; CASAS, D. D. L.; BRESSAND, F.; LENGUEL, G.; LAMPLE, G.; SAULNIER, L.; ET AL., 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*, (2023). (cited on pages 14, 30, and 77)
- JIANG, D.; HE, X.; ZENG, H.; WEI, C.; KU, M.; LIU, Q.; AND CHEN, W., 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*, (2024). (cited on page 75)
- JIANG, J.; CHEN, B.; FU, B.; AND LONG, M., 2020a. Transfer-learning-library. <https://github.com/thuml/Transfer-Learning-Library>. (cited on page 88)
- JIANG, Y.; FORET, P.; YAK, S.; ROY, D. M.; MOBAHI, H.; DZIUGAITE, G. K.; BENGIO, S.; GUNASEKAR, S.; GUYON, I.; AND NEYSHABUR, B., 2020b. Neurips 2020 competition: Predicting generalization in deep learning. *arXiv preprint arXiv:2012.07976*, (2020). (cited on page 53)
- JIANG, Y.; NAGARAJAN, V.; BAEK, C.; AND KOLTER, J. Z., 2022. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=WvOGCEAQhxl>. (cited on page 9)
- JIANG, Y.; NATEKAR, P.; SHARMA, M.; AITHAL, S. K.; KASHYAP, D.; SUBRAMANYAM, N.; LASSANCE, C.; ROY, D. M.; DZIUGAITE, G. K.; GUNASEKAR, S.; GUYON, I.; FORET, P.; YAK, S.; MOBAHI, H.; NEYSHABUR, B.; AND BENGIO, S., 2021. Methods and analysis of the first competition in predicting generalization of deep learning. In *NeurIPS Competition and Demonstration Track*, 170–190. (cited on page 9)
- JIANG, Y.; NEYSHABUR, B.; MOBAHI, H.; KRISHNAN, D.; AND BENGIO, S., 2020c. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=SJgIPJBFvH>. (cited on pages 9 and 54)
- JIN, Y.; WANG, X.; LONG, M.; AND WANG, J., 2020. Minimum class confusion for versatile domain adaptation. In *European Conference on Computer Vision*, 464–480. (cited on page 57)
- JOACHIMS, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning*, 137–142. (cited on page 83)
- JOACHIMS, T.; FREITAG, D.; MITCHELL, T.; ET AL., 1997. Webwatcher: A tour guide for the world wide web. In *International Joint Conference on Artificial Intelligence*, 770–777. Citeseer. (cited on pages 83 and 84)
- JUNG, A. B., 2018. imgaug. <https://github.com/aleju/imgaug>. [Online; accessed 30-Oct-2018]. (cited on page 87)
- KAR, O. F.; TONIONI, A.; POKLUKAR, P.; KULSHRESTHA, A.; ZAMIR, A.; AND TOMBARI, F., 2024. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, 113–132. (cited on page 29)
- KAR, O. F.; YEO, T.; ATANOV, A.; AND ZAMIR, A., 2022. 3d common corruptions and data augmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18963–18974. (cited on pages 12 and 29)

-
- KARAMCHETI, S.; NAIR, S.; BALAKRISHNA, A.; LIANG, P.; KOLLAR, T.; AND SADIGH, D., 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *International Conference on Machine Learning*. (cited on pages 15, 77, and 105)
- KARPATHY, A. AND FEI-FEI, L., 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 15)
- KENDALL, M. G., 1948. Rank correlation methods. (1948). (cited on pages 42, 55, and 70)
- KESKAR, N. S.; MUDIGERE, D.; NOCEDAL, J.; SMELYANSKIY, M.; AND TANG, P. T. P., 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=H1oyRlYgg>. (cited on pages 9 and 53)
- KHATTAK, M. U.; RASHEED, H.; MAAZ, M.; KHAN, S.; AND KHAN, F. S., 2023. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 33)
- KHATTAK, M. U.; WASIM, S. T.; NASEER, M.; KHAN, S.; YANG, M.-H.; AND KHAN, F. S., 2023. Self-regulating prompts: Foundational model adaptation without forgetting. In *International Conference on Computer Vision*, 15190–15200. (cited on page 33)
- KIRICHENKO, P.; IZMAILOV, P.; AND WILSON, A. G., 2020. Why normalizing flows fail to detect out-of-distribution data. In *Advances in Neural Information Processing Systems*, 20578–20589. (cited on page 8)
- KIRSCH, A. AND GAL, Y., 2022. A note on "assessing generalization of sgd via disagreement". *Transactions on Machine Learning Research*, (2022). (cited on page 7)
- KOH, P. W.; SAGAWA, S.; MARKLUND, H.; XIE, S. M.; ZHANG, M.; BALSUBRAMANI, A.; HU, W.; YASUNAGA, M.; PHILLIPS, R. L.; GAO, I.; ET AL., 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 5637–5664. (cited on pages 7 and 60)
- KOUW, W. M. AND LOOG, M., 2019. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 43, 3 (2019), 766–785. (cited on page 63)
- KRIZHEVSKY, A.; HINTON, G.; ET AL., 2009. Learning multiple layers of features from tiny images. (2009). (cited on pages 40, 42, 59, 73, 83, 90, and 113)
- KUMAR, A.; RAGHUNATHAN, A.; JONES, R.; MA, T.; AND LIANG, P., 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=UYneFzXSJWh>. (cited on page 33)
- LAKSHMINARAYANAN, B.; PRITZEL, A.; AND BLUNDELL, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*. (cited on page 8)
- LEE, K.; LEE, K.; LEE, H.; AND SHIN, J., 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*. (cited on page 8)

-
- LEWIS, D. D. AND GALE, W. A., 1994. A sequential algorithm for training text classifiers. In *Conference on Research and Development in Information Retrieval*, 3–12. (cited on pages 83 and 84)
- LI, B.; ZHANG, Y.; GUO, D.; ZHANG, R.; LI, F.; ZHANG, H.; ZHANG, K.; LI, Y.; LIU, Z.; AND LI, C., 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, (2024). (cited on page 75)
- LI, C.; XU, H.; TIAN, J.; WANG, W.; YAN, M.; BI, B.; YE, J.; CHEN, H.; XU, G.; CAO, Z.; ET AL., 2022a. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, (2022). (cited on page 75)
- LI, F.; ZHANG, R.; ZHANG, H.; ZHANG, Y.; LI, B.; LI, W.; MA, Z.; AND LI, C., 2024b. Llava-next: Tackling multi-image, video, and 3d in large multimodal models. <https://llava-vl.github.io/blog/2024-06-16-llava-next-interleave/>. (cited on page 75)
- LI, J.; LI, D.; SAVARESE, S.; AND HOI, S., 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742. PMLR. (cited on pages 34 and 60)
- LI, J.; LI, D.; XIONG, C.; AND HOI, S., 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. (cited on pages 34, 40, 41, 55, and 60)
- LI, S.; LV, F.; XIE, B.; LIU, C. H.; LIANG, J.; AND QIN, C., 2021. Bi-classifier determinacy maximization for unsupervised domain adaptation. In *Annual AAAI Conference on Artificial Intelligence*, 8455–8464. (cited on page 57)
- LI, T.; WEN, Z.; LI, Y.; AND LEE, T. S., 2024c. Emergence of shape bias in convolutional neural networks through activation sparsity. In *Advances in Neural Information Processing Systems*. (cited on page 20)
- LI, X.; WANG, Z.; AND XIE, C., 2023b. An inverse scaling law for clip training. In *Advances in Neural Information Processing Systems*. (cited on page 41)
- LIANG, S.; LI, Y.; AND SRIKANT, R., 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*. (cited on page 8)
- LIANG, S.; LI, Y.; AND SRIKANT, R., 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*. (cited on page 54)
- LIANG, T.; POGGIO, T.; RAKHLIN, A.; AND STOKES, J., 2019. Fisher-rao metric, geometry, and complexity of neural networks. In *International Conference on Artificial Intelligence and Statistics*, 888–896. (cited on page 9)
- LIANG, W.; TADESSE, G. A.; HO, D.; FEI-FEI, L.; ZAHARIA, M.; ZHANG, C.; AND ZOU, J., 2022. Advances, challenges and opportunities in creating data for trustworthy ai. *Nature Machine Intelligence*, 4, 8 (2022), 669–677. (cited on page 81)
- LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; AND ZITNICK, C. L., 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755. Springer. (cited on page 81)

-
- LIPTON, Z.; WANG, Y.-X.; AND SMOLA, A., 2018. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, 3122–3130. (cited on pages 62 and 65)
- LIU, H.; LI, C.; LI, Y.; AND LEE, Y. J., 2024a. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306. (cited on pages 67, 69, and 75)
- LIU, H.; LI, C.; LI, Y.; LI, B.; ZHANG, Y.; SHEN, S.; AND LEE, Y. J., 2024b. Llava-next: Improved reasoning, ocr, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>. (cited on page 75)
- LIU, H.; LI, C.; WU, Q.; AND LEE, Y. J., 2024c. Visual instruction tuning. In *Advances in Neural Information Processing Systems*. (cited on pages 12, 14, 29, 30, 67, and 98)
- LIU, W.; WANG, X.; OWENS, J.; AND LI, Y., 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 21464–21475. (cited on page 8)
- LIU, Z.; LIN, Y.; CAO, Y.; HU, H.; WEI, Y.; ZHANG, Z.; LIN, S.; AND GUO, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*. (cited on pages 14 and 42)
- LIU, Z.; MAO, H.; WU, C.-Y.; FEICHTENHOFER, C.; DARRELL, T.; AND XIE, S., 2022. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 14, 40, 41, and 42)
- LOWE, D., 1999. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1150–1157. (cited on page 84)
- LU, P.; MISHRA, S.; XIA, T.; QIU, L.; CHANG, K.-W.; ZHU, S.-C.; TAFJORD, O.; CLARK, P.; AND KALYAN, A., 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2507–2521. (cited on pages 67, 75, and 114)
- LU, Y.; QIN, Y.; ZHAI, R.; SHEN, A.; CHEN, K.; WANG, Z.; KOLOURI, S.; STEPPUTTIS, S.; CAMPBELL, J.; AND SYCARA, K., 2023. Characterizing out-of-distribution error via optimal transport. In *Advances in Neural Information Processing Systems*, 17602–17622. (cited on page 9)
- MACQUEEN, J. ET AL., 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 281–297. Oakland, CA, USA. (cited on page 84)
- MADANI, O.; PENNOCK, D. M.; AND FLAKE, G. W., 2004. Co-validation: using model disagreement on unlabeled data to validate classification algorithms. In *Advances in Neural Information Processing Systems*, 873–880. (cited on page 9)
- MALININ, A. AND GALES, M., 2020. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*. (cited on page 72)
- MANAKUL, P.; LIUSIE, A.; AND GALES, M. J., 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Annual Meeting of the Association for Computational Linguistics*. (cited on page 71)

-
- MANIPARAMBIL, M.; VORSTER, C.; MOLLOY, D.; MURPHY, N.; MCGUINNESS, K.; AND O'CONNOR, N. E., 2023. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *International Conference on Computer Vision*, 262–271. (cited on page 32)
- MANNING, C. D. AND SCHÜTZE, H., 1999. *Foundations of Statistical Natural Language Processing*. MIT press. (cited on pages 72 and 85)
- MANSOUR, Y.; MOHRI, M.; AND ROSTAMIZADEH, A., 2009. Domain adaptation: Learning bounds and algorithms. In *Conference on Learning Theory*. (cited on page 7)
- MASRY, A.; LONG, D. X.; TAN, J. Q.; JOTY, S.; AND HOQUE, E., 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Annual Meeting of the Association for Computational Linguistics*. (cited on pages 67, 68, 75, and 114)
- MATHEW, M.; KARATZAS, D.; AND JAWAHAR, C., 2021. Docvqa: A dataset for vqa on document images. 2200–2209. (cited on pages 67, 68, 75, and 114)
- MAZUMDER, M.; BANBURY, C.; YAO, X.; KARLAŠ, B.; ROJAS, W. G.; DIAMOS, S.; DIAMOS, G.; HE, L.; KIELA, D.; JURADO, D.; ET AL., 2022. Dataperf: Benchmarks for data-centric ai development. *arXiv preprint arXiv:2207.10062*, (2022). (cited on page 81)
- MCCALLUM, A.; NIGAM, K.; ET AL., 1998. A comparison of event models for naive bayes text classification. In *AAAI Workshop on Learning for Text Categorization*, 41–48. (cited on page 83)
- MILLER, J. P.; TAORI, R.; RAGHUNATHAN, A.; SAGAWA, S.; KOH, P. W.; SHANKAR, V.; LIANG, P.; CARMON, Y.; AND SCHMIDT, L., 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 7721–7735. (cited on pages 1, 16, 54, 55, 59, and 71)
- MIN, J.; LEE, J.; PONCE, J.; AND CHO, M., 2019. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, (2019). (cited on pages 15, 27, and 113)
- MINDERER, M.; DJOLONGA, J.; ROMIJNDERS, R.; HUBIS, F.; ZHAI, X.; HOULSBY, N.; TRAN, D.; AND LUCIC, M., 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, 15682–15694. (cited on pages 8, 22, 23, 40, and 56)
- MING, Y.; CAI, Z.; GU, J.; SUN, Y.; LI, W.; AND LI, Y., 2022. Delving into out-of-distribution detection with vision-language representations. In *Advances in Neural Information Processing Systems*, 35087–35102. (cited on pages 8, 12, and 21)
- MINTUN, E.; KIRILLOV, A.; AND XIE, S., 2021. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, 3571–3583. (cited on pages 7 and 91)
- MIRZA, M. J.; KARLINSKY, L.; LIN, W.; DOVEH, S.; MICOREK, J.; KOZINSKI, M.; KUEHNE, H.; AND POSSEGER, H., 2024. Meta-prompting for automating zero-shot visual recognition with llms. In *European Conference on Computer Vision*, 370–387. (cited on page 32)
- MISHRA, A.; SHEKHAR, S.; SINGH, A. K.; AND CHAKRABORTY, A., 2019. Ocr-vqa: Visual question answering by reading text in images. 947–952. doi:10.1109/ICDAR.2019.00156. (cited on pages 67, 75, and 114)

-
- MISRA, I. AND MAATEN, L. v. D., 2020. Self-supervised learning of pretext-invariant representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 104)
- MOKADY, R.; HERTZ, A.; AND BERMANO, A. H., 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, (2021). (cited on page 39)
- MUCSÁNYI, B.; KIRCHHOF, M.; NGUYEN, E.; RUBINSTEIN, A.; AND OH, S. J., 2023. Trustworthy machine learning. *arXiv preprint arXiv:2310.08215*, (2023). (cited on page 1)
- MURRAY, K. AND CHIANG, D., 2018. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, 212–223. Association for Computational Linguistics, Brussels, Belgium. doi:10.18653/v1/W18-6322. <https://aclanthology.org/W18-6322>. (cited on page 72)
- NAEINI, M. P.; COOPER, G.; AND HAUSKRECHT, M., 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. (cited on page 15)
- NAGARAJAN, V. AND KOLTER, J. Z., 2019. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, (2019). (cited on pages 9 and 53)
- NAGELKERKE, N. J. ET AL., 1991. A note on a general definition of the coefficient of determination. *biometrika*, 78, 3 (1991), 691–692. (cited on page 42)
- NALISNICK, E.; MATSUKAWA, A.; TEH, Y. W.; GORUR, D.; AND LAKSHMINARAYANAN, B., 2018. Do deep generative models know what they don’t know? In *International Conference on Learning Representations*. (cited on page 8)
- NASEER, M. M.; RANASINGHE, K.; KHAN, S. H.; HAYAT, M.; SHAHBAZ KHAN, F.; AND YANG, M.-H., 2021. Intriguing properties of vision transformers. In *Advances in Neural Information Processing Systems*. (cited on page 19)
- NEAL, L.; OLSON, M.; FERN, X.; WONG, W.-K.; AND LI, F., 2018. Open set learning with counterfactual images. In *European Conference on Computer Vision*, 620–635. (cited on page 8)
- NEYSHABUR, B.; BHOJANAPALLI, S.; MCALLESTER, D.; AND SREBRO, N., 2017a. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*. (cited on page 53)
- NEYSHABUR, B.; BHOJANAPALLI, S.; AND SREBRO, N., 2017b. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*. (cited on page 9)
- NEYSHABUR, B.; TOMIOKA, R.; AND SREBRO, N., 2015. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, 1376–1401. (cited on pages 9 and 53)
- NGUYEN, K. AND O’CONNOR, B., 2015. Posterior calibration and exploratory analysis for natural language processing models. In *Conference on Empirical Methods in Natural Language Processing*, 1587–1598. (cited on page 8)
- NGUYEN, T.; ILHARCO, G.; WORDSMAN, M.; OH, S.; AND SCHMIDT, L., 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. In *Advances in Neural Information Processing Systems*. (cited on pages 11, 16, and 39)

-
- NISTER, D. AND STEWENIUS, H., 2006. Scalable recognition with a vocabulary tree. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2161–2168. (cited on page 85)
- NOVACK, Z.; MCAULEY, J.; LIPTON, Z. C.; AND GARG, S., 2023. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*. (cited on page 44)
- OQUAB, M.; DARCET, T.; MOUTAKANNI, T.; VO, H. V.; SZAFRANIEC, M.; KHALIDOV, V.; FERNANDEZ, P.; HAZIZA, D.; MASSA, F.; EL-NOUBY, A.; ASSRAN, M.; BALLAS, N.; GALUBA, W.; HOWES, R.; HUANG, P.-Y.; LI, S.-W.; MISRA, I.; RABBAT, M.; SHARMA, V.; SYNNAEVE, G.; XU, H.; JEGOU, H.; MAIRAL, J.; LABATUT, P.; JOULIN, A.; AND BOJANOWSKI, P., 2024. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, (2024). <https://openreview.net/forum?id=a68SUt6zFt>. (cited on pages 27 and 77)
- OVADIA, Y.; FERTIG, E.; REN, J.; NADO, Z.; SCULLEY, D.; NOWOZIN, S.; DILLON, J.; LAKSHMINARAYANAN, B.; AND SNOEK, J., 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*. (cited on pages 8, 25, 43, 46, 56, and 65)
- OZA, P. AND PATEL, V. M., 2019. C2ae: Class conditioned auto-encoder for open-set recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2307–2316. (cited on page 8)
- PÁNDY, M.; AGOSTINELLI, A.; UIJLINGS, J.; FERRARI, V.; AND MENSINK, T., 2022. Transferability estimation using bhattacharyya class separability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9172–9182. (cited on page 86)
- PAPINENI, K.; ROUKOS, S.; WARD, T.; AND ZHU, W.-J., 2002. Bleu: A method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 311–318. Association for Computational Linguistics. (cited on page 73)
- PARASHAR, S.; LIN, Z.; LIU, T.; DONG, X.; LI, Y.; RAMANAN, D.; CAVERLEE, J.; AND KONG, S., 2024. The neglected tails in vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12988–12997. (cited on page 32)
- PAULLADA, A.; RAJI, I. D.; BENDER, E.; DENTON, E.; AND HANNA, A., 2021. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, (2021). (cited on page 81)
- PENG, X.; BAI, Q.; XIA, X.; HUANG, Z.; SAENKO, K.; AND WANG, B., 2019. Moment matching for multi-source domain adaptation. In *International Conference on Computer Vision*. (cited on pages 40, 42, 63, 82, 86, 87, and 114)
- PENG, X.; LI, Y.; AND SAENKO, K., 2020. Domain2vec: Domain embedding for unsupervised domain adaptation. In *European Conference on Computer Vision*, 756–774. (cited on pages 82 and 83)
- PERRONNIN, F. AND DANCE, C., 2007. Fisher kernels on visual vocabularies for image categorization. In *International Conference on Computer Vision*, 1–8. (cited on page 84)
- PERRONNIN, F.; SÁNCHEZ, J.; AND MENSINK, T., 2010. Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, 143–156. (cited on page 84)

-
- PHILBIN, J.; CHUM, O.; ISARD, M.; SIVIC, J.; AND ZISSERMAN, A., 2007. Object retrieval with large vocabularies and fast spatial matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1–8. (cited on page 85)
- PLATANIOS, E.; POON, H.; MITCHELL, T. M.; AND HORVITZ, E. J., 2017. Estimating accuracy from unlabeled data: A probabilistic logic approach. In *Advances in Neural Information Processing Systems*. (cited on page 9)
- PLATANIOS, E. A.; DUBEY, A.; AND MITCHELL, T., 2016. Estimating accuracy from unlabeled data: a bayesian approach. In *International Conference on Machine Learning*, 1416–1425. (cited on page 9)
- PLATT, J. ET AL., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10, 3 (1999), 61–74. (cited on page 40)
- POLIANSKII, V.; MARCHETTI, G. L.; KRAVBERG, A.; VARAVA, A.; POKORNY, F. T.; AND KRAGIC, D., 2022. Voronoi density estimator for high-dimensional data: Computation, compactification and convergence. In *Conference on Uncertainty in Artificial Intelligence*, 1644–1653. PMLR. (cited on page 86)
- POLO, F. M.; WEBER, L.; CHOSHEN, L.; SUN, Y.; XU, G.; AND YUROCHKIN, M., 2024. tinybenchmarks: evaluating llms with fewer examples. *arXiv preprint arXiv:2402.14992*, (2024). (cited on page 79)
- PRATT, S.; COVERT, I.; LIU, R.; AND FARHADI, A., 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *International Conference on Computer Vision*. (cited on pages 32, 47, and 48)
- RADFORD, A.; KIM, J. W.; HALLACY, C.; RAMESH, A.; GOH, G.; AGARWAL, S.; SASTRY, G.; ASKELL, A.; MISHKIN, P.; CLARK, J.; ET AL., 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. (cited on pages 11, 14, 15, 30, 32, 37, 39, 40, 41, 55, 60, 64, 69, and 78)
- RAJARAMAN, A. AND ULLMAN, J. D., 2011. *Mining of massive datasets*. Cambridge University Press. (cited on page 84)
- RECHT, B.; ROELOFS, R.; SCHMIDT, L.; AND SHANKAR, V., 2018a. Do cifar-10 classifiers generalize to cifar-10? (2018). <https://arxiv.org/abs/1806.00451>. (cited on pages 42 and 59)
- RECHT, B.; ROELOFS, R.; SCHMIDT, L.; AND SHANKAR, V., 2018b. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, (2018). (cited on pages 91 and 113)
- RECHT, B.; ROELOFS, R.; SCHMIDT, L.; AND SHANKAR, V., 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. (cited on pages 1, 7, 11, 15, 40, 42, 54, and 59)
- REN, J.; LIU, P. J.; FERTIG, E.; SNOEK, J.; POPLIN, R.; DEPRISTO, M.; DILLON, J.; AND LAKSHMINARAYANAN, B., 2019. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*. (cited on page 8)
- REN, Z.; SU, Y.; AND LIU, X., 2023. Chatgpt-powered hierarchical comparisons for image classification. In *Advances in Neural Information Processing Systems*. (cited on page 44)

-
- RIDNIK, T.; BEN-BARUCH, E.; NOY, A.; AND ZELNIK-MANOR, L., 2021. Imagenet-21k pretraining for the masses. (cited on page 42)
- ROMBACH, R.; BLATTMANN, A.; LORENZ, D.; ESSER, P.; AND OMMER, B., 2022a. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. (cited on page 37)
- ROMBACH, R.; BLATTMANN, A.; LORENZ, D.; ESSER, P.; AND OMMER, B., 2022b. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 48)
- SAGAWA, S.; KOH, P. W.; LEE, T.; GAO, I.; XIE, S. M.; SHEN, K.; KUMAR, A.; HU, W.; YASUNAGA, M.; MARKLUND, H.; ET AL., 2021. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, (2021). (cited on page 63)
- SANDLER, M.; HOWARD, A.; ZHU, M.; ZHMOGINOV, A.; AND CHEN, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520. (cited on page 91)
- SANTURKAR, S.; TSIPRAS, D.; AND MADRY, A., 2020. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*. (cited on pages 45 and 49)
- SCHIFF, Y.; QUANZ, B.; DAS, P.; AND CHEN, P.-Y., 2021. Predicting deep neural network generalization with perturbation response curves. In *Advances in Neural Information Processing Systems*. (cited on page 9)
- SCHUHMAN, C. AND BEAUMONT, R., 2022. Laion-aesthetics. <https://laion.ai/blog/laion-aesthetics/>. Accessed: 2025-04-25. (cited on page 35)
- SCHUHMAN, C.; BEAUMONT, R.; VENCU, R.; GORDON, C. W.; WIGHTMAN, R.; CHERTI, M.; COOMBES, T.; KATTA, A.; MULLIS, C.; WORTSMAN, M.; SCHRAMOWSKI, P.; KUNDURTHY, S. R.; CROWSON, K.; SCHMIDT, L.; KACZMARCZYK, R.; AND JITSEV, J., 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*. <https://openreview.net/forum?id=M3Y74vmsMcY>. (cited on page 14)
- SERRÀ, J.; ÁLVAREZ, D.; GÓMEZ, V.; SLIZOVSKAIA, O.; NÚÑEZ, J. F.; AND LUQUE, J., 2019. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations*. (cited on page 8)
- SHANKAR, V.; DAVE, A.; ROELOFS, R.; RAMANAN, D.; RECHT, B.; AND SCHMIDT, L., 2021. Do image classifiers generalize across time? In *International Conference on Computer Vision*. (cited on pages 15, 40, 112, and 113)
- SHARMA, P.; DING, N.; GOODMAN, S.; AND SORICUT, R., 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*, 2556–2565. (cited on page 14)
- SHI, M.; LIU, F.; WANG, S.; LIAO, S.; RADHAKRISHNAN, S.; HUANG, D.-A.; YIN, H.; SAPRA, K.; YACOOB, Y.; SHI, H.; CATANZARO, B.; TAO, A.; KAUTZ, J.; YU, Z.; AND LIU, G., 2024. Eagle: Exploring the design space for multimodal llms with mixture of encoders. *arXiv:2408.15998*, (2024). (cited on page 75)

-
- SHI, Z.; CARLINI, N.; BALASHANKAR, A.; SCHMIDT, L.; HSIEH, C.-J.; BEUTEL, A.; AND QIN, Y., 2023. Effective robustness against natural distribution shifts for models with different training data. In *Advances in Neural Information Processing Systems*, 73543–73558. (cited on page 12)
- SHIEH, G. S., 1998. A weighted kendall’s tau statistic. *Statistics & Probability Letters*, 39, 1 (1998), 17–24. (cited on pages 70 and 88)
- SIMONYAN, K. AND ZISSERMAN, A., 2014. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*. (cited on pages 64 and 91)
- SINGH, A.; HU, R.; GOSWAMI, V.; COUAIRON, G.; GALUBA, W.; ROHRBACH, M.; AND KIELA, D., 2022. Flava: A foundational language and vision alignment model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 41 and 60)
- SINGH, A.; NATARAJAN, V.; SHAH, M.; JIANG, Y.; CHEN, X.; BATRA, D.; PARIKH, D.; AND ROHRBACH, M., 2019. Towards vqa models that can read. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8317–8326. (cited on pages 68, 75, and 114)
- SIVIC AND ZISSERMAN, 2003. Video google: a text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, 1470–1477. (cited on pages 84 and 85)
- SMITH, S. L. AND LE, Q. V., 2017. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*. (cited on pages 9 and 54)
- SUN, B.; FENG, J.; AND SAENKO, K., 2016. Return of frustratingly easy domain adaptation. In *Annual AAAI Conference on Artificial Intelligence*. (cited on page 82)
- SUN, B. AND SAENKO, K., 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 443–450. (cited on pages 64 and 82)
- SUN, Q.; FANG, Y.; WU, L.; WANG, X.; AND CAO, Y., 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, (2023). (cited on page 14)
- SUN, T.; LU, C.; AND LING, H., 2022. Prior knowledge guided unsupervised domain adaptation. In *European Conference on Computer Vision*. (cited on pages 62 and 65)
- SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; AND ALEMI, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Annual AAAI Conference on Artificial Intelligence*. (cited on page 91)
- TAN, M. AND LE, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. (cited on pages 20, 64, and 91)
- TAN, M. AND LE, Q., 2021. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*. (cited on page 20)
- TAORI, R.; DAVE, A.; SHANKAR, V.; CARLINI, N.; RECHT, B.; AND SCHMIDT, L., 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*, 18583–18599. (cited on pages 1, 14, 16, 17, 21, and 104)

-
- TEAM, V., 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. <https://lmsys.org/blog/2023-03-30-vicuna/>. *Based on evaluations comparing Vicuna’s responses to those of ChatGPT. (cited on page 69)
- TENEY, D.; LIN, Y.; OH, S. J.; AND ABBASNEJAD, E., 2023. Id and ood performance are sometimes inversely correlated on real-world datasets. In *Advances in Neural Information Processing Systems*, 71703–71722. (cited on page 1)
- TENEY, D.; LIN, Y.; OH, S. J.; AND ABBASNEJAD, E., 2024. Id and ood performance are sometimes inversely correlated on real-world datasets. In *Advances in Neural Information Processing Systems*. (cited on pages 54 and 55)
- TIAN, Y.; SUN, C.; POOLE, B.; KRISHNAN, D.; SCHMID, C.; AND ISOLA, P., 2020. What makes for good views for contrastive learning? In *Advances in Neural Information Processing Systems*. (cited on page 104)
- TOLSTIKHIN, I. O.; HOULSBY, N.; KOLESNIKOV, A.; BEYER, L.; ZHAI, X.; UNTERTHINER, T.; YUNG, J.; STEINER, A.; KEYSERS, D.; USZKOREIT, J.; ET AL., 2021. Mlp-mixer: An all-mlp architecture for vision. In *Advances in Neural Information Processing Systems*. (cited on page 14)
- TOMANI, C.; WASEDA, F. K.; SHEN, Y.; AND CREMERS, D., 2023. Beyond in-domain scenarios: robust density-aware calibration. In *International Conference on Machine Learning*, 34344–34368. PMLR. (cited on page 43)
- TONG, S.; LIU, Z.; ZHAI, Y.; MA, Y.; LECUN, Y.; AND XIE, S., 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9568–9578. (cited on page 67)
- TOUVRON, H.; CORD, M.; AND JÉGOU, H., 2022. Deit iii: Revenge of the vit. In *European Conference on Computer Vision*, 516–533. Springer. (cited on page 27)
- TOUVRON, H.; MARTIN, L.; STONE, K.; ALBERT, P.; ALMAHAIRI, A.; BABAEI, Y.; BASHLYKOV, N.; BATRA, S.; BHARGAVA, P.; BHOSALE, S.; ET AL., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, (2023). (cited on pages 14, 30, and 69)
- TU, W.; DENG, W.; AND GEDEON, T., 2023a. A closer look at the robustness of contrastive language-image pre-training (clip). In *Advances in Neural Information Processing Systems*, 13678–13691. (cited on pages 8, 39, and 40)
- TU, W.; DENG, W.; GEDEON, T.; AND ZHENG, L., 2023b. A bag-of-prototypes representation for dataset-level applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2881–2892. (cited on page 9)
- TU, W.; DENG, W.; GEDEON, T.; AND ZHENG, L., 2023c. A bag-of-prototypes representation for dataset-level applications. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2881–2892. (cited on page 70)
- TU, W.; DENG, W.; ZHENG, L.; AND GEDEON, T., 2024. What does softmax probability tell us about classifiers ranking across diverse test conditions? *Transactions on Machine Learning Research*, (2024). <https://openreview.net/forum?id=vtiDUgGjyx>. (cited on page 71)
- VAN HORN, G.; MAC AODHA, O.; SONG, Y.; CUI, Y.; SUN, C.; SHEPARD, A.; ADAM, H.; PERONA, P.; AND BELONGIE, S., 2018. The inaturalist species classification and detection dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 15 and 112)

-
- VANSCHOREN, J., 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, (2018). (cited on page 82)
- VIGNA, S., 2015. A weighted correlation index for rankings with ties. In *Proceedings of International Conference on World Wide Web*, 1166–1176. (cited on page 88)
- WANG, H.; GE, S.; LIPTON, Z.; AND XING, E. P., 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*. (cited on pages 1, 7, 11, 15, 40, 42, 59, and 113)
- WANG, H.; HUANG, Z.; LIN, Z.; AND LIU, T., 2024a. NoiseGPT: Label noise detection and rectification through probability curvature. In *Advances in Neural Information Processing Systems*. <https://openreview.net/forum?id=VRRvJnxgQe>. (cited on page 67)
- WANG, P.; BAI, S.; TAN, S.; WANG, S.; FAN, Z.; BAI, J.; CHEN, K.; LIU, X.; WANG, J.; GE, W.; ET AL., 2024b. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, (2024). (cited on page 75)
- WANG, T. AND ISOLA, P., 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 9929–9939. (cited on page 58)
- WEI, J.; WANG, X.; SCHUURMANS, D.; BOSMA, M.; XIA, F.; CHI, E.; LE, Q. V.; ZHOU, D.; ET AL., 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 24824–24837. (cited on page 32)
- WENZEL, F.; DITTADI, A.; GEHLER, P. V.; SIMON-GABRIEL, C.-J.; HORN, M.; ZIETLOW, D.; KERNERT, D.; RUSSELL, C.; BROX, T.; SCHIELE, B.; ET AL., 2022. Assaying out-of-distribution generalization in transfer learning. In *Advances in Neural Information Processing Systems*. (cited on pages 55 and 61)
- WIGHTMAN, R., 2017. Train cifar10 with pytorch. <https://github.com/kuangliu/pytorch-cifar>. (cited on page 59)
- WIGHTMAN, R., 2019. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>. doi:10.5281/zenodo.4414861. (cited on pages 14, 42, 59, 103, and 105)
- WOLF, T.; DEBUT, L.; SANH, V.; CHAUMOND, J.; DELANGUE, C.; MOI, A.; CISTAC, P.; RAULT, T.; LOUF, R.; FUNTOWICZ, M.; AND BREW, J., 2020. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing*, 38–45. (cited on page 75)
- WORTSMAN, M.; ILHARCO, G.; KIM, J. W.; LI, M.; KORNBLITH, S.; ROELOFS, R.; LOPES, R. G.; HAJISHIRZI, H.; FARHADI, A.; NAMKOONG, H.; ET AL., 2022. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 12, 33, and 39)
- WU, Z.; CHEN, X.; PAN, Z.; LIU, X.; LIU, W.; DAI, D.; GAO, H.; MA, Y.; WU, C.; WANG, B.; XIE, Z.; WU, Y.; HU, K.; WANG, J.; SUN, Y.; LI, Y.; PIAO, Y.; GUAN, K.; LIU, A.; XIE, X.; YOU, Y.; DONG, K.; YU, X.; ZHANG, H.; ZHAO, L.; WANG, Y.; AND RUAN, C., 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. <https://arxiv.org/abs/2412.10302>. (cited on page 75)

-
- WU, Z.; XIONG, Y.; YU, S. X.; AND LIN, D., 2018. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on pages 14 and 104)
- X.AI, 2024. Grok 1.5v: The future of ai models. <https://x.ai/blog/grok-1.5v>. (cited on pages 67, 68, 75, and 114)
- XIAO, J.; HAYS, J.; EHINGER, K. A.; OLIVA, A.; AND TORRALBA, A., 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 15)
- XIAO, Y.; TANG, Z.; WEI, P.; LIU, C.; AND LIN, L., 2023. Masked images are counterfactual samples for robust fine-tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20301–20310. (cited on page 18)
- XIAO, Z.; YAN, Q.; AND AMIT, Y., 2020. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *Advances in Neural Information Processing Systems*, 20685–20696. (cited on page 8)
- XIE, R.; WEI, H.; FENG, L.; CAO, Y.; AND AN, B., 2023. On the importance of feature separability in predicting out-of-distribution error. In *Advances in Neural Information Processing Systems*, 27783–27800. (cited on page 9)
- XU, H.; XIE, S.; TAN, X. E.; HUANG, P.-Y.; HOWES, R.; SHARMA, V.; LI, S.-W.; GHOSH, G.; ZETTEMAYER, L.; AND FEICHTENHOFER, C., 2024. Demystifying clip data. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=5BCFlnfE1g>. (cited on pages 14, 35, and 41)
- XUE, L.; SHU, M.; AWADALLA, A.; WANG, J.; YAN, A.; PURUSHWALKAM, S.; ZHOU, H.; PRABHU, V.; DAI, Y.; RYOO, M. S.; ET AL., 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, (2024). (cited on page 37)
- YANG, J.; YANG, S.; GUPTA, A. W.; HAN, R.; LI, F.-F.; AND XIE, S., 2024a. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*, (2024). (cited on pages 98 and 99)
- YANG, J.; ZHOU, K.; LI, Y.; AND LIU, Z., 2024b. Generalized out-of-distribution detection: A survey. *International Journal on Computer Vision*, 132, 12 (2024), 5635–5662. (cited on page 8)
- YANG, S.; VAN DE WEIJER, J.; HERRANZ, L.; JUI, S.; ET AL., 2021. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Advances in Neural Information Processing Systems*, 29393–29405. (cited on page 58)
- YANG, S.; YANG, J.; HUANG, P.; BROWN, E.; YANG, Z.; YU, Y.; TONG, S.; ZHENG, Z.; XU, Y.; WANG, M.; LU, D.; FERGUS, R.; LECUN, Y.; FEI-FEI, L.; AND XIE, S., 2025. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, (2025). (cited on page 98)
- YOU, K.; LIU, Y.; WANG, J.; AND LONG, M., 2021. Logme: Practical assessment of pre-trained models for transfer learning. In *International Conference on Machine Learning*. (cited on pages 55 and 70)

-
- YOUNG, P.; LAI, A.; HODOSH, M.; AND HOCKENMAIER, J., 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2 (2014), 67–78. doi:10.1162/tac1_a_00166. <https://aclanthology.org/Q14-1006>. (cited on pages 15, 25, and 113)
- YU, J.; WANG, Z.; VASUDEVAN, V.; YEUNG, L.; SEYEDHOSSEINI, M.; AND WU, Y., 2022a. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, (2022). (cited on page 39)
- YU, Y.; BATES, S.; MA, Y.; AND JORDAN, M., 2022b. Robust calibration with multi-domain temperature scaling. In *Advances in Neural Information Processing Systems*, 27510–27523. (cited on page 43)
- YU, Y.; YANG, Z.; WEI, A.; MA, Y.; AND STEINHARDT, J., 2022c. Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, 25721–25746. (cited on page 9)
- YUE, X.; NI, Y.; ZHANG, K.; ZHENG, T.; LIU, R.; ZHANG, G.; STEVENS, S.; JIANG, D.; REN, W.; SUN, Y.; ET AL., 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9556–9567. (cited on pages 68, 75, and 114)
- YUKSEKGONUL, M.; BIANCHI, F.; KALLURI, P.; JURAFSKY, D.; AND ZOU, J., 2022. When and why vision-language models behave like bag-of-words models, and what to do about it? In *International Conference on Learning Representations*. (cited on page 39)
- ZENG, A.; ATTARIAN, M.; ICHTER, B.; CHOROMANSKI, K.; WONG, A.; WELKER, S.; TOMBARI, F.; PUROHIT, A.; RYOO, M.; SINDHWANI, V.; ET AL., 2022. Socratic models: Composing zero-shot multimodal reasoning with language. In *International Conference on Learning Representations*. (cited on page 39)
- ZHAI, X.; MUSTAFA, B.; KOLESNIKOV, A.; AND BEYER, L., 2023. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision*, 11975–11986. (cited on pages 14, 30, 35, 41, 60, and 69)
- ZHANG, C.; PAN, F.; KIM, J.; KWEON, I. S.; AND MAO, C., 2024a. Imagenet-d: Benchmarking neural network robustness on diffusion synthetic object. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21752–21762. (cited on pages 7, 12, 15, 30, and 113)
- ZHANG, C.; ZHANG, M.; ZHANG, S.; JIN, D.; ZHOU, Q.; CAI, Z.; ZHAO, H.; LIU, X.; AND LIU, Z., 2022a. Delving deep into the generalization of vision transformers under distribution shifts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. (cited on page 19)
- ZHANG, J.; HERRMANN, C.; HUR, J.; CHEN, E.; JAMPANI, V.; SUN, D.; AND YANG, M.-H., 2024b. Telling left from right: Identifying geometry-aware semantic correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3076–3085. (cited on page 25)
- ZHANG, J.; HERRMANN, C.; HUR, J.; POLANIA CABRERA, L.; JAMPANI, V.; SUN, D.; AND YANG, M.-H., 2024c. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems*. (cited on page 25)

-
- ZHANG, R.; ZHANG, W.; FANG, R.; GAO, P.; LI, K.; DAI, J.; QIAO, Y.; AND LI, H., 2022b. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*. (cited on pages 14, 19, 20, and 33)
- ZHANG, T.; DAS, R.; SIMA'AN, K.; AND SUTSKEVER, I., 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. (cited on page 73)
- ZHAO, Y.; PANG, T.; DU, C.; YANG, X.; LI, C.; CHEUNG, N.-M. M.; AND LIN, M., 2023. On evaluating adversarial robustness of large vision-language models. In *Advances in Neural Information Processing Systems*. (cited on pages 12 and 37)
- ZHENG, J.; SHEN, J.; YAO, Y.; WANG, M.; YANG, Y.; WANG, D.; AND LIU, T., 2025. Chain-of-focus prompting: Leveraging sequential visual cues to prompt large autoregressive vision models. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=noidywkBba>. (cited on page 67)
- ZHENG, L.; YANG, Y.; AND TIAN, Q., 2017. Sift meets cnn: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 40, 5 (2017), 1224–1244. (cited on pages 85 and 86)
- ZHOU, B.; LAPEDRIZA, A.; KHOSLA, A.; OLIVA, A.; AND TORRALBA, A., 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 40, 6 (2017), 1452–1464. (cited on pages 15 and 112)
- ZHOU, K.; LIU, Z.; QIAO, Y.; XIANG, T.; AND LOY, C. C., 2022a. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (2022). (cited on page 63)
- ZHOU, K.; YANG, J.; LOY, C. C.; AND LIU, Z., 2022b. Learning to prompt for vision-language models. *International Journal on Computer Vision*, 130, 9 (2022), 2337–2348. (cited on pages 14, 19, 20, 33, and 47)
- ZHOU, X.; GIRDHAR, R.; JOULIN, A.; KRÄHENBÜHL, P.; AND MISRA, I., 2022c. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, 350–368. (cited on page 39)
- ZOU, Y.; DENG, W.; AND ZHENG, L., 2023. Adaptive calibrator ensemble: Navigating test set difficulty in out-of-distribution scenarios. In *International Conference on Computer Vision*, 19333–19342. (cited on page 43)