An Empirical Study Into What Matters for Calibrating Vision–Language Models

Weijie Tu [1] **Weijian Deng** [1] **Dylan Campbell** [1] **Stephen Gould** [1] **Tom Gedeon** [1 2 3]

Abstract

Vision–Language Models (VLMs) have emerged as the dominant approach for zero-shot recognition, adept at handling diverse scenarios and significant distribution changes. However, their deployment in risk-sensitive areas requires a deep understanding of their uncertainty estimation capabilities, a relatively uncharted area. In this study, we explore the calibration properties of VLMs across different architectures, datasets, and training strategies. In particular, we analyze the uncertainty estimation performance of VLMs when calibrated in one domain, label set or hierarchy level, and tested in a different one. Our findings reveal that while VLMs are not inherently calibrated for uncertainty, temperature scaling significantly and consistently improves calibration, even across shifts in distribution and changes in label set. Moreover, VLMs can be calibrated with a very small set of examples. Through detailed experimentation, we highlight the potential applications and importance of our insights, aiming for more reliable and effective use of VLMs in critical, real-world scenarios.

1. Introduction

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., 2022b; 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., 2023), 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., 2023).

However, their application in risk-sensitive domains necessitates a more rigorous understanding of their uncertainty estimation capabilities, an area that remains largely underexplored. 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. (2023) 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.

We investigate which factors affect the calibration of VLMs. Starting from prior research that demonstrates zero-shot CLIP can be well-calibrated with simple temperature scaling under distribution shifts (Tu et al., 2023), 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., 2022). 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

[1]The Australian National University [2]Curtin University [3]University of Óbuda. Correspondence to: Weijie Tu <weijie.tu@anu.edu.au>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

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:

- After calibrating all models with temperature scaling (Platt et al., 1999), vision–language models are better calibrated than the other models in this study, which is not necessarily the case prior to calibration;
- 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 a few samples for calibration. For example, VLMs can be calibrated using temperature scaling, spline fitting (Gupta et al., 2021) or histogram binning (Zadrozny & Elkan, 2001) with less than 100 samples;
- 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.

2. Related Work

Vision–language models have demonstrated strong capabilities by leveraging web-scale datasets and language supervision to learn joint image–text representations (Bommasani et al., 2021; Radford et al., 2021; Jia et al., 2021). A seminal work by Radford et al. (2021) introduced CLIP, a large VLM trained on 400 million filtered web-crawled image–text pairs, which exhibits unprecedented zero-shot ability on numerous downstream visual tasks. Inspired by CLIP, various algorithms have been created to enhance the performance of the model (Singh et al., 2022; Li et al., 2022; Zhai et al., 2023; Li et al., 2023b). For example, Li et al. (2022) propose BLIP, a new pre-training framework, that bootstraps captions to effectively leverage noisy web data. Zhai et al. (2023) designs a simple pairwise sigmoid loss which solely operates on the image–text pairs without requiring the global view of pairwise similarities for normalization.

Encouraged by the strong generalizability of VLMs, researchers have explored their properties from diverse perspectives, such as robustness and bias. For instance, Schiappa et al. (2022) and Qiu et al. (2022) investigate their robustness through perturbations, while Fang et al. (2022) attribute the remarkable robustness of CLIP to its diverse training distribution. Additionally, Yuksekgonul et al. (2022) and Thrush et al. (2022) assess the capability of VLMs to encode compositional information. Liang et al. (2022) study the modality gap from the perspectives of model initialization and contrastive learning optimization. While Tu et al. (2023) focus on the calibration of CLIP models using temperature scaling on ImageNet, our study extends this to a diverse array of VLMs and various calibration methods (*e.g.*, Spline, histogram binning, and vector scaling). We provide an in-depth analysis of calibration factors, examine the uncertainty estimates of VLMs, and consider different aspects that may influence calibration performance. Furthermore, we demonstrate the practical utility of our findings in a real-world problem setup.

Confidence calibration aims to calibrate models so that their prediction probabilities align with the empirical frequency of correctness (Nguyen & O'Connor, 2015; Guo et al., 2017). Much research effort has been made in 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 calibration of modern neural networks (Guo et al., 2017; Ovadia et al., 2019; Minderer et al., 2021; Tu et al., 2023). 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. (2023) show that zero-shot CLIP models are well-calibrated with temperature scaling. This paper builds on this prior research by studying a more comprehensive suite of factors that influence the uncertainty estimation performance of VLMs.

3. Definition and Notation

Let $\mathcal{Y} = 1, \ldots, K$ 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 $\mathbf{f} : \mathcal{X} \to \Delta^k$. 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: $\mathbf{f} =: \sigma \circ \mathbf{g}$, where $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^k$ is a non-probabilistic k-way classifier, and $\sigma : \mathbb{R}^k \to \Delta^k$ is the softmax operator $\sigma_i(\mathbf{z}) = \frac{\exp(\mathbf{z}_i)}{\sum_{j=1}^k \exp(\mathbf{z}_j)}$ 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 $\hat{y} =: \arg\max_i \mathbf{f}_i(\mathbf{x})$ and the corresponding confidence score $\hat{p} =: \max_i \mathbf{f}i(\mathbf{x})$.

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 confi-

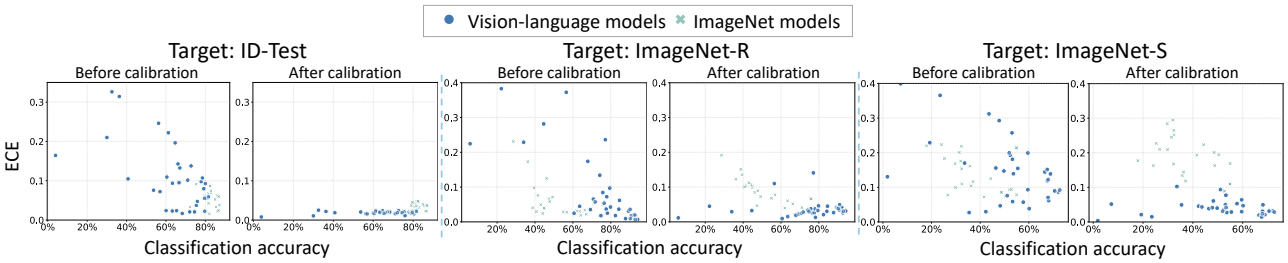

Figure 1. **Comparing the calibration performance of ImageNet-trained models and VLMs**. We report the results on the in-distribution test set (ID-Test) and two out-of-distribution (OOD) test sets: ImageNet-R and ImageNet-S. We plot the expected calibration error (ECE) before and after temperature scaling 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.

dence 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. Experimental Setup

Compared models: 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., 2022). In particular, we evaluate zero-shot CLIP models 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., 2023). CLIP variants with different image encoders are also assessed, including ViT (Dosovitskiy et al., 2020) 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., 2023c). Unless specified, for each model, we use their default prompt sets from Radford et al. (2021).

Compared models: 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., 2016) and ConvNeXt (Liu et al., 2022), and vision transformers, exemplified by ViT (Dosovitskiy et al., 2020) 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 the modelsmentioned are publicly accessible through 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. OOD test sets are ImageNet-V2-A (Recht et al., 2019), ImageNet-R(endition) (Hendrycks et al., 2021), 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., 2018b) 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

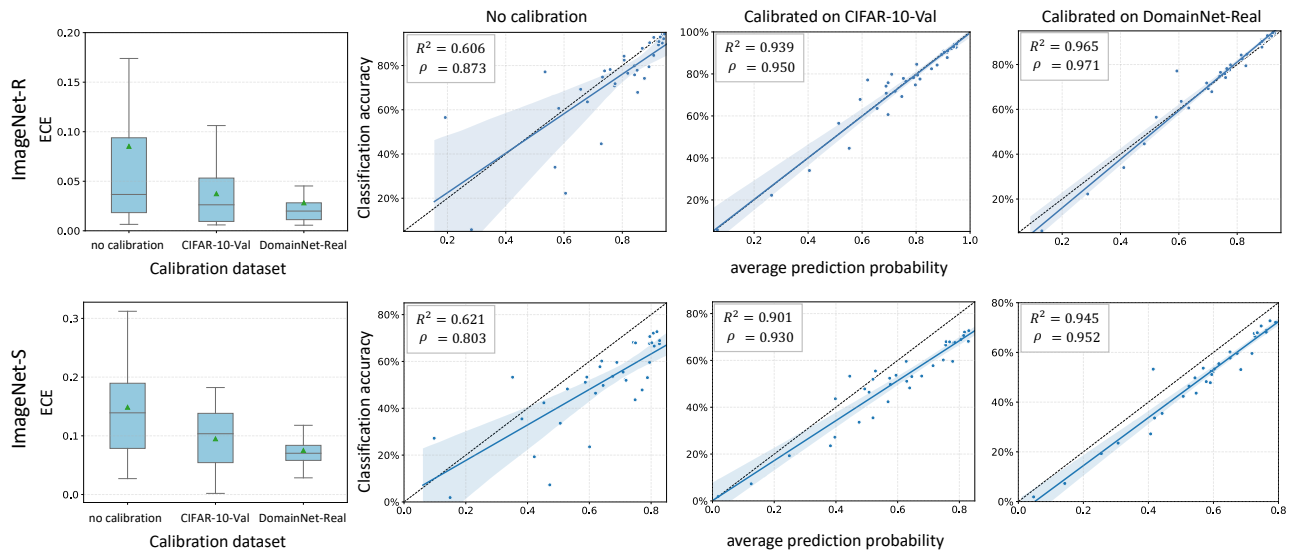

Figure 2. **Adaptability of VLMs to different calibration label sets. Left: Calibration error reduction.** Here, we observe a significant decrease in the expected calibration error for VLMs following cross-label-set calibration, as opposed to when no calibration is applied. **Right: Correlation between VLM prediction probability and classification accuracy.** This graph illustrates the 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.

spans $[-1, 1]$, where a value closer to 1 (or -1) indicates a better ranking index, while 0 indicates no correlation.

5. 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 are designed to scrutinize the uncertainty estimation of VLMs under various conditions, including changes in (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. Figure 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 indicates that VLMs benefit more significantly 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.

5.1. 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 its potential for cross-label-set calibration. To evaluate the effect of calibration label sets, we conduct experiments calibrating VLMs on datasets with different label sets. The findings, depicted in Figure 2, demonstrate an im-

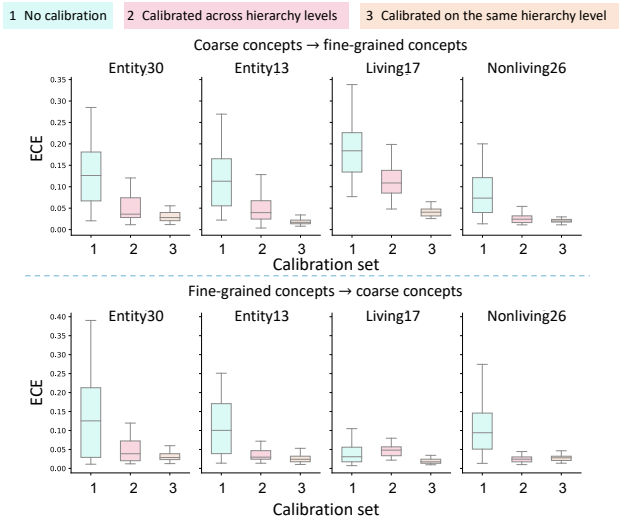

Figure 3. **Robustness of VLM calibration to label hierarchy levels.** This figure presents box plots summarizing the calibration errors (ECEs) of VLMs calibrated with label hierarchies differing in granularity from the target dataset (ImageNet-S). The top row shows calibration at a coarser level, and 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.

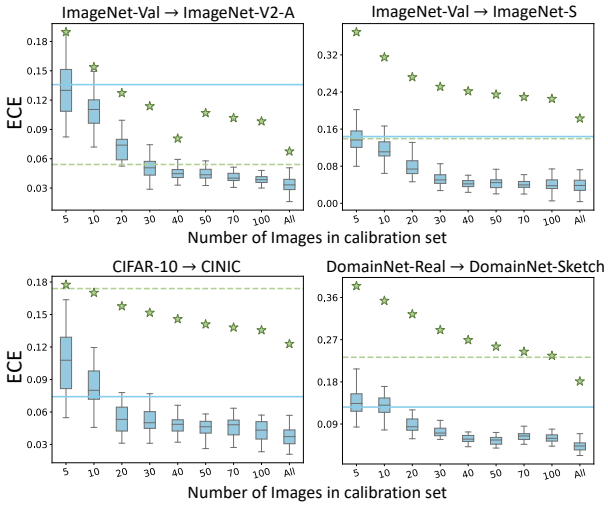

Figure 4. **Data-efficiency of VLM calibration across diverse datasets.** This figure displays the ECE of VLMs as a function of the calibration set size across four datasets: ImageNet-V2-A, ImageNet-S, CINIC, and DomainNet. The green stars are the average ECE of calibrated models trained on the dataset. The blue solid and green dashed horizontal lines represent the average ECE before calibration of VLMs and non-VLMs, respectively. The ECE values, averaged over ten random seeds, plateau after including merely 40–50 images in the calibration set, chosen at random, 10 times. The results closely approximate 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. These results highlight the data-efficiency of VLM calibration.

provement 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.

5.2. 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 hierarchy. 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. Figure 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.

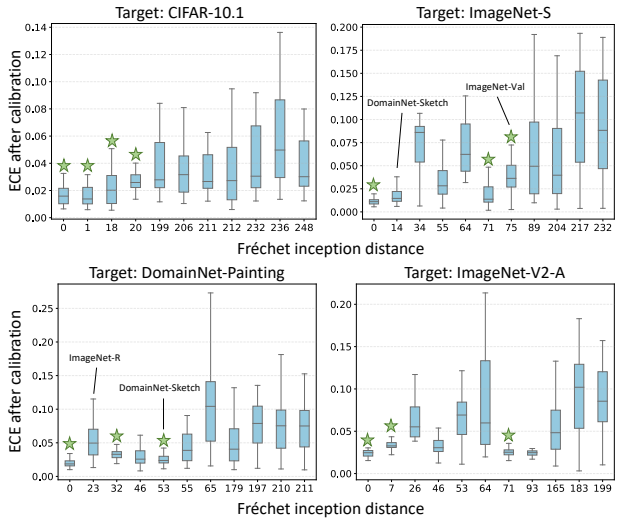

Figure 5. **Impact of the distance between calibration and target set distributions on VLM uncertainty estimates.** The distance between the calibration dataset and target dataset is computed by Fréchet inception distance (Heusel et al., 2017). A green star indicates that the dataset for this column 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.

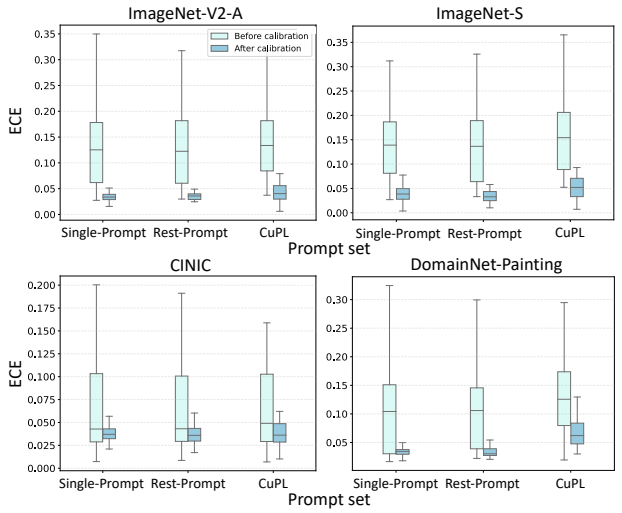

Figure 6. **Efficient calibration with single prompts for VLMs.** The figure illustrates the transferability of calibration effectiveness when using a single prompt for VLMs. The temperature scalar estimated using one random prompt, denoted as "Single-Prompt" remains highly effective when applied to 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.

5.3. 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. 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. Figure 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, and that it is a low-dimensional problem.

5.4. 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.

Figure 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.

5.5. Transferability of Calibration Across Prompts

VLMs classify images by comparing image features with class weights computed by the text encoder, which takes as input the textual prompts describing each class of interest. This suggests that given the same textual class names, a better prompt set may improve 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., 2022a). In the previous experiment, 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.

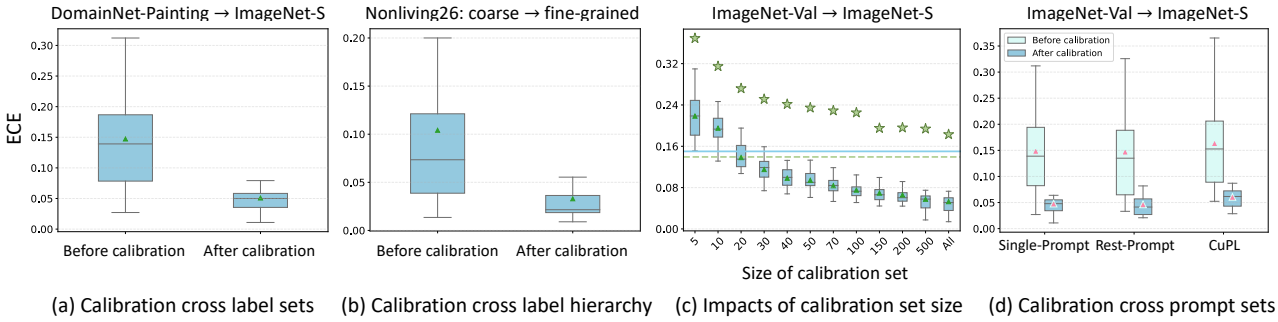

Figure 7. **Uncertainty estimation performance of VLMs with spline calibration.** Performance evaluation of VLMs using the spline method (Gupta et al., 2021) across four factors: (a) label sets, (b) label hierarchy, (c) calibration set size, and (d) prompt sets. Our results demonstrate that the phenomena 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.

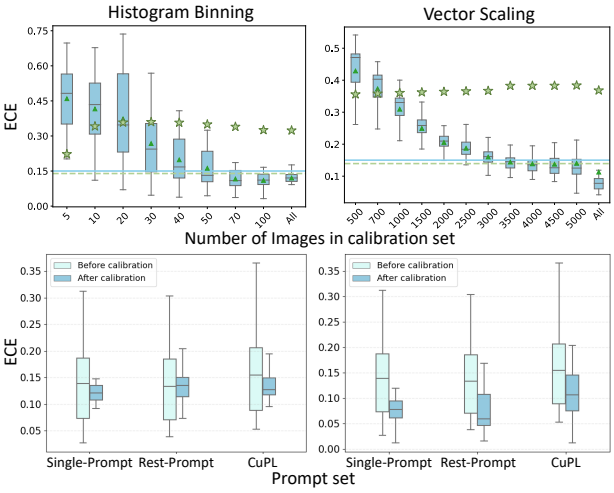

Figure 8. **Uncertainty estimation performance of VLMs with histogram binning and vector scaling** across two factors: calibration set size (up) and prompt sets (bottom). The target domain is ImageNet-S and the calibration set is ImageNet-Validation.

To answer this question, we conducted 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 the human-designed prompts (Rest-Prompt) and machine-generated prompts (CuPL), as depicted in Figure 6. This discovery highlights a key insight: 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.

5.6. Impact of the Calibration Method

In addition to temperature scaling, we undertake an evaluation of three post-hoc calibration methods: spline fitting (Gupta et al., 2021), histogram binning (Zadrozny & Elkan, 2001) and vector scaling (Guo et al., 2017). Spline calibration involves deriving a function through spline fitting (SF), directly aligning classifier outputs with calibrated probabilities. Histogram binning (HB) and vector scaling (VS) both use class-specific parameters and assume an identical label set between calibration set and target set. Hence, HB and VS do not support calibration across label sets and label hierarchy levels. Additionally, as suggested by (Guo et al., 2017), VS requires the presence of samples for *each* class, meaning that VS demands more images for calibration. This section investigates whether the observed calibration properties of VLMs hold using SF, HB and VS.

In Figure 7 and Figure 8, we present our findings on SF, HB, and VS calibration methods. First, we observe that VLMs achieve lower ECE than calibrated non-VLMs using the three calibration methods under a distribution shift. Second, both SF and HB effectively calibrate VLMs using a small number of samples (*i.e.* 100). In contrast, VS requires a larger number of images (10% of $50,000$ samples) for effective calibration, consistent with the findings in Guo et al. (2017). Third, similar to temperature scaling, both SF and VS can calibrate VLMs using a single prompt, which can differ from the test prompt set. While HB effectively reduces the overall average ECE, it may not benefit all outliers. Lastly, when calibrating VLMs using datasets with different label sets or hierarchy levels compared to the target test set, spline calibration exhibits trends and outcomes similar to those of temperature scaling.

6. Application: Calibration-by-Synthesis

In this section, we take some of the calibration robustness findings from Section 5 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

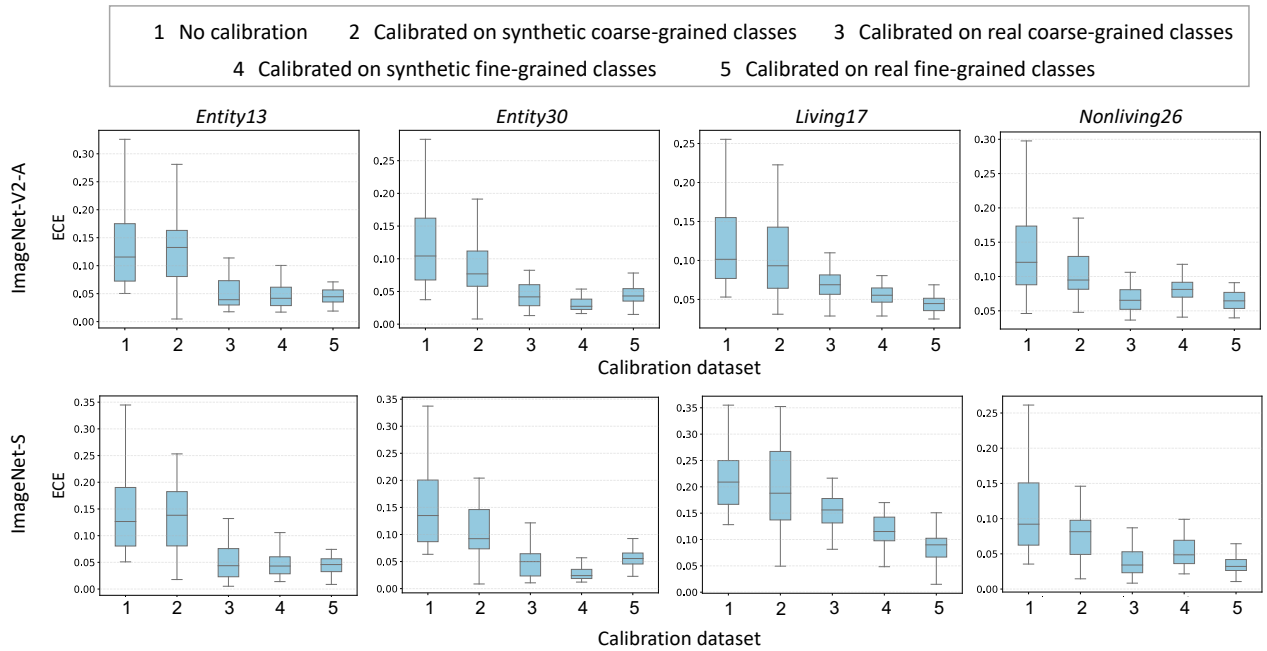

Figure 9. **Can VLMs be calibrated without labeled data? A calibration-by-synthesis approach.** Each column represents a pre-defined coarse-fine hierarchical mapping, and each row corresponds to a test set formed according to this class hierarchy. The uncertainty estimation performance of VLMs is plotted 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.

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., 2022), 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 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 Figure 9, we plot the ECE, evaluated on the target dataset with fine-grained class labels, for a set of VLMs grouped by

	No calibration			Synthetic calibration		
Label set	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 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 the model accuracy, and Spearman's rank correlation (ρ ↑). We see that by calibrating on a synthetic set with fine-grained classes, prediction probabilities are informative of the model's performance and its ranking.

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 out-

performing calibration 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 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.

7. Conclusion and Discussion

In this study, we have 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 holds promise for tasks that demand precise uncertainty information. Furthermore, we have demonstrated VLMs' remarkable adaptability—they can be effectively calibrated with datasets of 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 synthetic 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.

This work leaves open many interesting directions for future research. We primarily focus on calibrating classification confidence estimates for vanilla/backbone VLM networks. It would be interesting to explore the calibration of regression confidence estimates. Some more parameter-efficient fine-tuning methods for prompts can be included to study the transferability of calibration across prompts. Additionally, a broader class of large language model-based VLMs, such as LLaVA (Liu et al., 2023), can be further evaluated and examined. These models could potential present a unique calibration property for the lack of inherent capability to generate classification confidence scores.

Lastly, our scope is the calibration of classification confidence estimates for vanilla/backbone VLM networks, rather than the calibration of regression confidence estimates used for other tasks (*e.g.*, object detection and segmentation). In the context of VLMs, these tasks require compound networks with additional modules. We acknowledge that the findings in this study may not transfer to these other tasks, and that this would be interesting to study in future work.

Acknowledgements

We sincerely thank all the anonymous reviewers and area chairs for their constructive comments and valuable suggestions, which have greatly helped in enhancing this paper.

Impact Statement

Poorly calibrated models make under- or over-confident predictions on average, which can have significant negative societal consequences when these models are deployed in the real world. For example, vehicle safety systems that ascribe a high probability of "sky" to a region that actually contains the back of a truck can lead to accidents. This work investigates what factors might contribute to these calibration errors after temperature scaling, and shows that VLMs are surprisingly robust to a variety of factors when constructing the calibration set. This increases our confidence in the use of calibrated VLMs in safety-critical sectors like healthcare and autonomous systems. Nonetheless, cautious deployment is advisable, since there are likely to be other factors that have not been studied here that may contribute to how well-calibrated a VLM is in its target domain.

References

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems*, 2019.

Bello, I., Fedus, W., Du, X., Cubuk, E. D., Srinivas, A., Lin, T.-Y., Shlens, J., and Zoph, B. Revisiting resnets: Improved training and scaling strategies. In *Advances in Neural Information Processing Systems*, pp. 22614–22627, 2021.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R.,

Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

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

Darlow, L. N., Crowley, E. J., Antoniou, A., and Storkey, A. J. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

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

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

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.

Galil, I., Dabbah, M., and El-Yaniv, R. What can we learn from the selective prediction and uncertainty estimation performance of 523 imagenet classifiers. In *International Conference on Learning Representations*, 2023.

Goyal, S., Kumar, A., Garg, S., Kolter, Z., and Raghunathan, A. Finetune like you pretrain: Improved finetuning of zero-shot vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Gupta, K., Rahimi, A., Ajanthan, T., Mensink, T., Sminchisescu, C., and Hartley, R. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Helber, P., Bischke, B., Dengel, A., and Borth, D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721. PMLR, 2019.

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

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

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. Openclip, July 2021. URL https://doi.org/10.5281/zenodo.5143773. If you use this software, please cite it as below.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Kendall, M. G. Rank correlation methods. 1948.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Li, D., Li, J., Le, H., Wang, G., Savarese, S., and Hoi, S. C. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023a.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, pp. 9694–9705, 2021.

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

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023b.

Li, X., Wang, Z., and Xie, C. An inverse scaling law for clip training. In *Advances on Neural Information Processing Systems*, 2023c.

Liang, V. W., Zhang, Y., Kwon, Y., Yeung, S., and Zou, J. Y. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35: 17612–17625, 2022.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023.

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

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

Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., and Lucic, M. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, pp. 15682–15694, 2021.

Mokady, R., Hertz, A., and Bermano, A. H. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Nagelkerke, N. J. et al. A note on a general definition of the coefficient of determination. *biometrika*, 78(3):691–692, 1991.

Nguyen, K. and O'Connor, B. Posterior calibration and exploratory analysis for natural language processing models. In *Conference on Empirical Methods in Natural Language Processing*, 2015.

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. Quality not quantity: On the interaction between dataset design and robustness of clip. In *Advances in Neural Information Processing Systems*, 2022.

Novack, Z., McAuley, J., Lipton, Z. C., and Garg, S. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, 2023.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74, 1999.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Qiu, J., Zhu, Y., Shi, X., Wenzel, F., Tang, Z., Zhao, D., Li, B., and Li, M. Are multimodal models robust to image and text perturbations? *Journal of Data-centric Machine Learning Research*, 2022.

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

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018a.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018b. https://arxiv.org/abs/1806.00451.

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

Ren, Z., Su, Y., and Liu, X. Chatgpt-powered hierarchical comparisons for image classification. In *Advances in Neural Information Processing Systems*, 2023.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Santurkar, S., Tsipras, D., and Madry, A. Breeds: Benchmarks for subpopulation shift. In *International Conference on Learning Representations*, 2020.

Schiappa, M. C., Rawat, Y. S., Vyas, S., Vineet, V., and Palangi, H. Multi-modal robustness analysis against language and visual perturbations. In *Advances on Neural Information Processing Systems Datasets and Benchmark Track*, 2022.

Shankar, V., Dave, A., Roelofs, R., Ramanan, D., Recht, B., and Schmidt, L. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9661–9669, 2021.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Tomani, C., Waseda, F. K., Shen, Y., and Cremers, D. Beyond in-domain scenarios: robust density-aware calibration. In *International Conference on Machine Learning*, pp. 34344–34368. PMLR, 2023.

Tu, W., Deng, W., and Gedeon, T. A closer look at the robustness of contrastive language-image pre-training (clip). In *Advances on Neural Information Processing Systems*, 2023.

Wang, H., Ge, S., Lipton, Z., and Xing, E. P. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

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

Xu, H., Xie, S., Tan, X. E., Huang, P.-Y., Howes, R., Sharma, V., Li, S.-W., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022a.

Yu, Y., Bates, S., Ma, Y., and Jordan, M. Robust calibration with multi-domain temperature scaling. In *Advances in Neural Information Processing Systems*, pp. 27510–27523, 2022b.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. When and why vision-language models behave like bag-of-words models, and what to do about it? In *International Conference on Learning Representations*, 2022.

Zadrozny, B. and Elkan, C. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, 2001.

Zeng, A., Attarian, M., Ichter, B., Choromanski, K., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M., Sindhwani, V., et al. Socratic models: Composing zero-shot

multimodal reasoning with language. In *International Conference on Learning Representations*, 2022.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022a.

Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., and Misra, I. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368, 2022b.

Zou, Y., Deng, W., and Zheng, L. Adaptive calibrator ensemble: Navigating test set difficulty in out-of-distribution scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19333–19342, October 2023.

A. Appendix.

A.1. Datasets:

`ImageNet` (Deng et al., 2009) (https://www.image-net.org/);
`ImageNet-V2` (Shankar et al., 2021) (https://github.com/modestyachts/ImageNetV2);
`ImageNet-Rendition` (Hendrycks et al., 2021) (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);
`CIFAR-10` (Krizhevsky et al., 2009): (https://www.cs.toronto.edu/ kriz/cifar.html);
`CIFAR-10.1` (Recht et al., 2018a): (https://github.com/modestyachts/CIFAR-10.1);
`CIFAR-10.2` (Recht et al., 2018a): (https://github.com/modestyachts/CIFAR-10.1);
`CINIC` (Darlow et al., 2018): (https://www.v7labs.com/open-datasets/cinic-10);
`DomainNet` (Peng et al., 2019): (http://ai.bu.edu/M3SDA/);

A.2. Models Included in Experiments

(1) Vision–language models:

(1) we use the zero-shot CLIP models provided in OpenCLIP (Ilharco et al., 2021). They are listed as follows in the pattern (`architecture`, `source`):

(`RN50`, `openai`)

(`RN50`, `yfcc15m`)

(`RN50`, `cc12m`)

(`ViT-B-32`, `openai`)

(`ViT-B-32`, `laion400m_e32`)

(`ViT-B-32`, `laion2b_s34b_b79k`)

(`ViT-B-16`, `openai`)

(`ViT-B-16`, `laion400m_e32`)

(`ViT-L-14`, `openai`)

(`ViT-L-14`, `laion400m_e32`)

(`ViT-H-14`, `laion2b_s32b_b79k`)

(`ViT-g-14`, `laion2b_s34b_b88k`)

(`ViT-bigG-14`, `laion2b_s39b_b160k`)

(`convnext_base`, `laion400m_s13b_b51k`)

(`convnext_base_w`, `laion_aesthetic_s13b_b82k`)

(`convnext_xxlarge`, `laion2b_s34b_b82k_augreg`)

(`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-34B_lr-1e-3_bs-88k.pt`)

(`ViT-L-14`, `Model-L-14_Data-80M_Samples-34B_lr-1e-3_bs-88k.pt`)

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

(`ViT-B-32`, `datacomp_s_s13m_b4k`)

(`ViT-B-16`, `datacomp_l_s1b_b8k`)

(ViT-L-14, datacomp_xl_s13b_b90k)

(2) Other vision–language models:

(EVA02-B-16, merged2b_s8b_b131k)

(EVA02-L-14, merged2b_s8b_b131k)

(ViT-B-32-quickgelu, metaclip_400m)

(ViT-B-16-quickgelu, metaclip_fullcc)

(ViT-B-16-SigLIP, webli)

(ViT-B-16-SigLIP-256, webli)

(ViT-H-14-CLIPA, datacomp1b)

(ViT-L-14-CLIPA, datacomp1b)

Flava can be derived from Hugging Face, Transformer module (Wolf et al., 2020).

BLIP (Li et al., 2022), BLIP-2 (Li et al., 2023b) and ALBEF (Li et al., 2021) models are publicly available through LAVIS package (Li et al., 2023a).

(2) ImageNet-trained models:

we compare VLMs' uncertainty estimate performance with following models from TIMM (Wightman, 2019)

Trained from scratch on ImageNet:

swin_base_patch4_window12_384

deit_small_distilled_patch16_224

deit_base_patch16_224

swin_small_patch4_window7_224

deit_base_patch16_384

wide_resnet50_2

convnext_base

tv_resnet50

densenet121

inception_v4

resmlp_36_224

xception

vgg19

Pre-trained on a larger dataset then fine-tuned on ImageNet training split:

vit_base_patch16_224.orig_in21k_ft_in1k

vit_base_patch32_224

vit_large_r50_s32_384

beit_large_patch16_224

vit_large_patch16_384

swin_large_patch4_window12_384

`convnext_small.in12k_ft_in1k`

`ig_resnext101_32x16d`

`mixer_b16_224_miil`

`resnetv2_50x1_bitm`

`ig_resnext101_32x8d`

A.3. Prompt Template

In this section, we provide the prompt set used in our experiments. For experiments included in Section 5.1, Section 5.2, Section 5.4 and Section 5.6, we also use default prompt sets provided by Radford et al. (2021). For experiments in Section 5.5, the calibration prompt set is "*a photo of a <class>*". In Section 6, the calibration prompt set is "*a photo of a <class>*", while the test prompt set is generated by CuPL.

A.4. Computation Resources

PyTorch version is 1.10.0+cu111 and timm version is 0.8.21dev0. All experiment is run on one 3090 and the CPU AMD EPYC 7343 16-Core Processor.

A.5. Experiments on Specialized Domain: Satellite Images

We further conduct experiments on whether our findings maintain on more specialized domains: Satellite images. We use EuroSAT (Helber et al., 2019) as the target set. Two different calibration datasets are considered: 1) ImageNet validation set; 2) A generated dataset which consists of 68 images from finer-grained classes of Living17 hierarchy. The calibration method is temperature scaling. We see that our observations persist on such a specialized domain that VLMs can be calibrated across different label sets using a few samples.

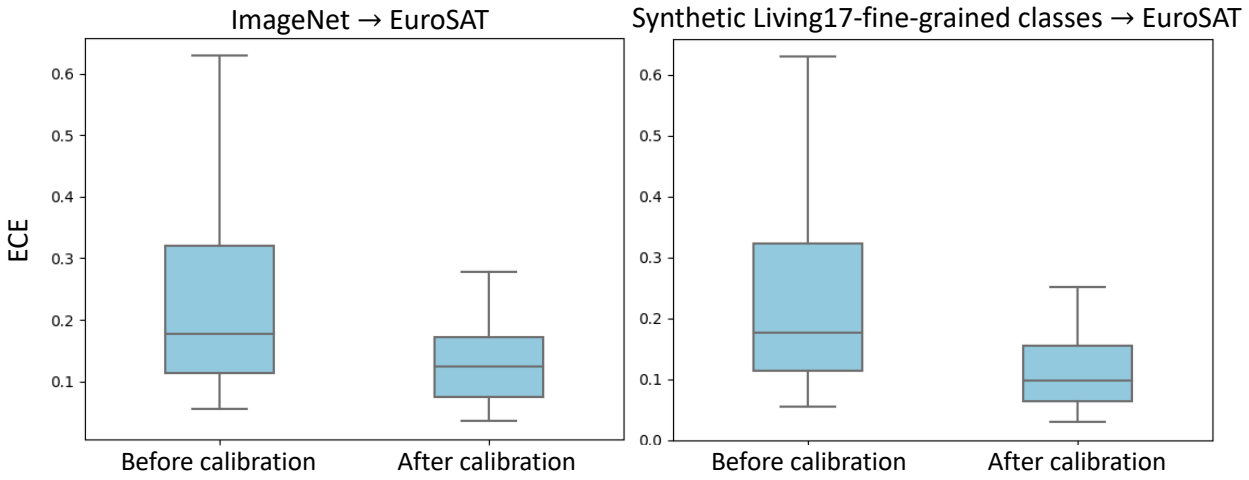

Figure 10. **Uncertainty estimate performance of VLMs when the target domain contains Satellite images.** Left: The calibration set is ImageNet-Validation set. Right: The calibration set is a synthesized dataset containing fine-grained classes of Living17.

A.6. Additional Results for Adaptability to Different Calibration Label Sets

Following the same practice as Section 5.1, we additionally validate that the findings on ImageNet-S and ImageNet-R persist when the target datasets change to ImageNet-A or ObjectNet.

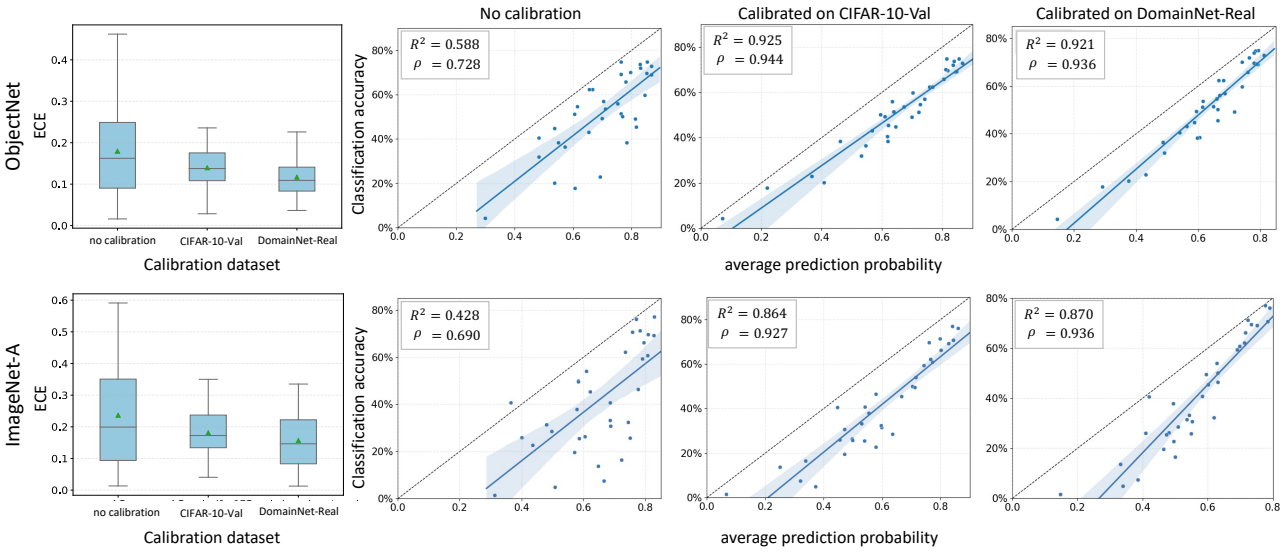

Figure 11. **Adaptability of VLMs to different calibration label sets. Left: Calibration error reduction.** Here, we observe a significant decrease in the expected calibration error for VLMs following cross-label-set calibration, as opposed to when no calibration is applied. **Right: Correlation between VLM prediction probability and classification accuracy.** This graph illustrates the classification accuracy of VLMs on ObejctNet and ImageNet-A 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.

A.7. Additional Results for Calibration Across Semantic Hierarchy Levels

We also show the ECE of VLMs calibrated with label hierarchies differing in granularity from the target dataset (ImageNet-V2-A). The results validate that VLMs can be calibrated across label hierarchies.

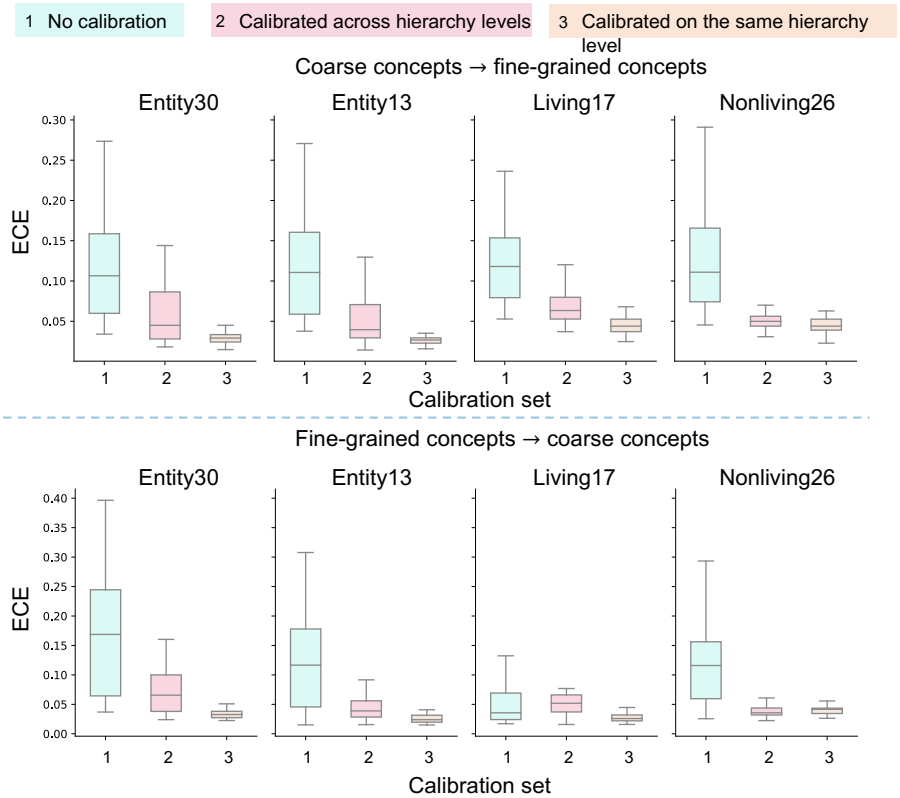

Figure 12. **Robustness of VLM calibration to label hierarchy levels.** This figure presents box plots summarizing the calibration errors (ECEs) of VLMs calibrated with label hierarchies differing in granularity from the target dataset (ImageNet-V2-A). The top row shows calibration at a coarser level, 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.