ZERO-SHOT CALIBRATION OF LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Calibration of deep learning models is crucial to their trustworthiness and safe usage, and as such, has been extensively studied in supervised classification models, with methods crafted to decrease miscalibration. However, there has yet to be a comprehensive study of the calibration of vision-language models that are used for zero-shot inference, like CLIP. We measure calibration across relevant variables like prompt, dataset, and architecture, and find that zero-shot inference with CLIP is miscalibrated. Furthermore, we propose a modified version of temperature scaling that is aligned with the common use cases of CLIP as a zero-shot model, and show that a single learned temperature generalizes for each specific CLIP model across inference dataset and prompt choice.

1 INTRODUCTION

Interpretability is one of the main hurdles in the trust, safety, and reliability of deep learning models. One specific area of concern is the miscalibration of these models, where the confidence of the predictions given do not reflect the probability of being correct. There exists many studies on the calibration and corresponding interpretability of classification models(Guo et al. (2017); Kull et al. (2019); Rajendran & LeVine (2019)) that are trained and tested on in a traditional fashion - a given dataset of one modality (like images) is split into a train, validation, and test set, with a known, fixed number of classes. The model is trained on the train set, tuned with the validation set, and evaluated on the test set. However, the use of vision-language models for zero-shot inference, like CLIP Radford et al. (2021), is becoming increasingly popular. In this setting, the dataset is multimodal, and the inference paradigm allows for zero-shot inference, where the class being predicted was not explicitly defined as a class of interest during training.

There has yet to be either an extensive study of the calibration of CLIP as a zero-shot model or applications of calibration methods to CLIP's zero-shot setting. We therefore propose "Zero-Shot Temperature Scaling," a method based on Temperature Scaling but trained in a zero-shot way similar to CLIP. Our main **contributions** are

- An extensive analytical study of the calibration of CLIP stratified by architecture, dataset (pre-training and inference), and input prompt. There exists literature that briefly mentions the calibration of CLIP Minderer et al. (2021), but not one that studies the calibration across any of the previously mentioned experimental variables.

- The novel application of Temperature Scaling to CLIP in a zero-shot way with an exposition on its robustness to changes in inference dataset and prompt. We note that the only modification to inference is to perform Temperature Scaling on the text-image similarity with temperature T. We show that this parameter varies solely with changes in underlying architecture and pre-training dataset (identical to the axes of variation allowed in the parameters of CLIP), and thus can be used at inference time for any arbitrary set of prompts or inference datasets. Therefore, to perform inference on a given dataset of interest requires no training, tuning, or calibration - meaning our method matches the zero-shot paradigm as used in CLIP.

2 PRELIMINARIES

2.1 PROBLEM SETUP

Let $X : \Omega \to \mathcal{X} \subset \mathbb{R}^D$ be the input (random) variable and let $Y : \Omega \to \{1, 2, ..., C-1, C\} \subset \mathbb{N}$ (where C is the number of output classes) be the response (random) variable. Typically X has some information about Y and we'd like to make inferences about Y given X. A common situation is trying to compute $\underset{c}{\operatorname{argmax}}\, \mathbb{E}(Y = c|X = x)$ with a model $\hat{f} : \mathbb{R}^d \to [0, 1]^C$. That is, $\hat{f}$ predicts the most likely class among the output classes as $\underset{c}{\operatorname{argmax}} \hat{f}_c(x_i)$. Typically, f uses an intermediate logit function $L : \mathbb{R}^D \to \mathbb{R}^C$. That is, the logit function returns a real number per class. For a given class index c, the logit corresponding to that class $L_c(x_i)$ ideally increases as the resemblance increases between the inference example x_i and the training inputs of class c (formally $\{x_j \text{ s.t. } y_j = c\}_{j=1}^N$).

2.2 CLIP

Introduced in Radford et al. (2021), CLIP is trained to align image/text pairs. This enables zero-shot inference, where the output classes of Y are chosen at inference time based on natural language prompts. CLIP accomplishes this by calculating its logit function as the cosine similarity between the embedding of a given input image x_i - which we denote as $E_{\text{im}}(x_i)$ - and the language embeddings of the natural language corresponding to the output classes. We denote the language embedding of the natural language corresponding to a given output class y_c as $E_{\text{lang}}(y_c)$. More formally, the logit function of CLIP is

$$L_c^{\text{CLIP}}(x_i) = 100 * \frac{E_{\text{im}}(x_i) \cdot E_{\text{lang}}(y_c)}{|E_{\text{im}}(x_i)||E_{\text{lang}}(y_c)|}$$

For CLIP and non-CLIP models, the softmax function is typically used to convert these logits into class probabilities. That is

$$\hat{f}_c(x_i) = \frac{e^{L_c^{\text{CLIP}}(x_i)}}{\sum_{j=1}^{C} e^{L_j^{\text{CLIP}}(x_i)}}$$

2.3 CALIBRATION

Ideally, we would like the confidence estimate $\hat{p}(x_i, \hat{f}) = \max_c \hat{f}_c(x_i)$ to be in alignment with the accuracy of $\hat{f}$ on x_i and points with confidences similar to that of x_i. As an example noted in Guo et al. (2017), given a set of 100 predictions with confidences of 0.8, we would hope that 80 of these predictions would be correctly classified. If so, we would consider the model to be *calibrated*. Let $D^{test} = \{(x_i, y_i)\}_{i=1}^N$ be the test set on which $\hat{f}$ is evaluated, where each x_i and y_i are examples drawn from (X, Y) (or a subset thereof), respectively. Further let $D_p^{test} = \{(x_i, y_i) \in D^{test} \text{ s.t. } \hat{p}(x_i, \hat{f}) = p\}$. Formally, a model is calibrated if

$$\text{acc}(\hat{f}, D_p^{test}) = p \;\forall\; p \in [0, 1] \tag{1}$$

We further note as in Guo et al. (2017) that the probability in 1 cannot be computed on a single sample, since an accuracy is computed on a set of examples rather than a single sample. Thus, there is a need for empirical approximations that can capture the essence of 1, which we describe below.

2.3.1 VISUALIZING MISCALIBRATION VIA RELIABILITY DIAGRAMS

We visualize the calibration of our estimator through **reliability diagrams** DeGroot & Fienberg (1983); Niculescu-Mizil & Caruana (2005). These diagrams group points by their predicted confidence scores into M equally spaced bins, and then compute the true and estimated accuracies in each bin as follows: let B_m be the (x_i, y_i) test samples whose confidence (i.e. estimated accuracy) falls into the interval $I_m = \big((m-1)/M, m/M\big]$, for $m = 2, \ldots, M$, and $I_1 = [0, \frac{1}{M}]$. The true

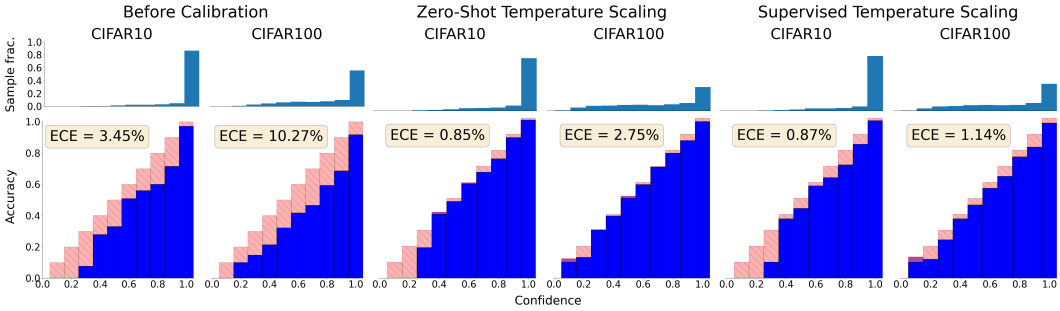

Figure 1: Confidence histograms (top) and reliability diagrams (bottom) of **Left 2 Columns:** CLIP without any calibration, **Center 2 Columns:** CLIP with our method, Zero-Shot Temperature Scaling, and **Right 2 Columns:** CLIP with Supervised Temperature Scaling. Additionally, the Expected Calibration Error (ECE) for each evaluation is given, where a lower ECE is better. We note that our method is as calibrated and almost as calibrated as Supervised Temperature Scaling on CIFAR10 and CIFAR100, as measured by ECE, respectively, even though our method is zero-shot. Here, we use ViT-B-16 (Dosovitskiy et al. (2020)) pretrained on laion400m_e31 (Schuhmann et al. (2021)) as our model, and "a photo of {*CLASS NAME*}" as our prompt.

accuracy is $\text{acc}(\hat{f}, B_m)$ and the average *estimated* accuracy (i.e. average confidence) within B_m is $1/|B_m| \sum_{(x_i,y_i)\in B_m} \hat{p}(x_i, \hat{f})$, which we write in short-hand as $\hat{p}(\hat{f}, B_m)$. The reliability diagram plots the difference between true accuracy and estimated accuracy for all M bins, and deviations from the line $f(x) = x$ represent miscalibrations: areas where there is a significant difference between the estimated and true accuracy. In Figure 1, the pink bars represent a theoretically perfectly calibrated model, while the blue bars represent the calibration of each model. For all experiments, we let $M = 10$, as is standard (Guo et al., 2017; Rajendran & LeVine, 2019; Kull et al., 2019).

2.3.2 Quantifying Miscalibration Via Expected Calibration Error (*ECE*)

We can quanitfy this miscalibration with the **Expected Calibration Error (*ECE*)** introduced in Naeini et al. (2015). *ECE*, aimed at summarizing the miscalibration visualized in reliability diagrams, is calculated as

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{|D|} \left| \hat{p}(\hat{f}, B_m) - \text{acc}(\hat{f}, B_m) \right| \tag{2}$$

In the leftmost columns of Figure 1, we show the calibration of CLIP, as it is regularly used, via reliability diagrams.

2.4 Calibrating With Labels Via Temperature Scaling

Typically used to reduce miscalibration, Temperature Scaling Guo et al. (2017) geometrically decreases the logit function L by a scalar T. That is, $\hat{f}$ has a logit function that employs Temperature Scaling as $L_c^{\text{calibrated}}(x_i; T) = L_c(x_i)/T$ (or in the case of CLIP, $L_c^{\text{calibrated}}(x_i; T) = L_c^{\text{CLIP}}(x_i)/T$). Temperature Scaling typically calibrates a frozen network post-training using a dataset $D_{\text{calibration}} = \{(x_i, y_i)\}_{i=1}^{N}$ by minimizing the cross-entropy loss function:

$$\mathcal{L}_{CE}(\hat{f}, D_{\text{calibration}}) = -\sum_{(x_i,y_i)\in D_{\text{calibration}}} \log \hat{f}_{y_i}(x_i)$$

In the rightmost two rows of Figure 1, we present reliability diagrams of CLIP once calibrated via Temperature Scaling. In Appendix Section A.1, we present reliability diagrams in this evaluation setting using additional supervised calibration methods Isotonic Regression introduced in Zadrozny & Elkan (2002), and Histogram Binning introduced in Zadrozny & Elkan (2001), as well as Unsupervised Temperature Scaling introduced in Mozafari et al. (2019). These are solely for context, since the main benefits of using CLIP are to be able to perform inference on a task of interest both

Method		CLIP	CLIP + 0-Shot TS	CLIP + TS
Architecture	Pre-Train Data			
ViT-B-16	laion400m	6.34	2.22	0.91
	laion2b	4.65	2.96	0.98
ViT-L-14	laion400m	6.68	1.36	0.72
	laion2b	3.17	2.38	0.85
ViT-B-32	laion400m	4.69	3.06	1.66
	laion2b	3.88	2.69	0.80
ViT-H-14	laion2b	3.67	2.47	0.88
ResNet-50	yfcc15m	26.69	7.60	2.61
	cc12m	26.56	6.18	3.31

Table 1: *ECE* results on a variety of prompts, architectures, inference datasets, and pre-training datasets. Each results is the mean across 3 different datasets with different prompts for each dataset (per the original CLIP paper from Minderer et al. (2021)) which are enumerated in Appendix Section A.2. For citations and links to the architectures and pre-training datasets, please see Appendix Section A.3. All numbers are percentages.

without having any training labels on that task (which are required by the supervised methods) and without a calibration dataset that is specific for each inference dataset (which are required by all of the previously mentioned methods).

3 Our Method: Zero-Shot Temperature Scaling

To address this gap of the inability to calibrate CLIP without a calibration dataset, we propose Zero-Shot Temperature Scaling. For a given architecture and pre-training dataset of CLIP, we simply train a temperature on an auxiliary dataset via Temperature Scaling. We then use this temperature on all downstream inferences of this model regardless of prompt or inference dataset. For all experiments, we use ImageNet-1k Huang & Li (2021) as our auxiliary dataset with "a photo of {}" as the prompt in the supervised training of the temperature ultimately used in our Zero-Shot Temperature Scaling. Once trained, this model can be used in a zero-shot fashion - it does not require any re-training or tuning to be used on any given inference dataset with any given prompt.

We note that the training of Zero-Shot Temperature Scaling *does* require a dataset on which to perform *a* training process. However, this method is zero-shot in an identical sense to CLIP: CLIP trains parameters on an auxiliary dataset, where the training of the parameters on a given architecture and a given pre-training dataset is independent of the training of the parameters of different architectures and pre-training datasets. Therefore, we allow a different temperature for each architecture/pre-training dataset pair, matching the CLIP paradigm. We will make the singular parameter T learned in this process for each such pair publicly available if accepted, although we do note that users can train this temperature T themselves on an auxiliary dataset of their choice. Given this single parameter T associated with the CLIP architecture and pre-training dataset, a user can utilize our method by simply diving the CLIP logits by T, without any training, tuning, or calibration processes necessary.

In the middle two columns of Figure 1, we present reliability diagrams of a CLIP-based model once updated via Zero-Shot Temperature Scaling.

4 Results

4.1 Comparison to Vanilla CLIP and CLIP + Temperature Scaling

In Table 1, we show the Expected Calibration Error results of our method compared to CLIP and CLIP calibrated via supervised Temperature Scaling on a variety of prompts, architectures, inference datasets, and pre-training datasets. For detailed results across prompts and datasets for a single pre-training dataset and architecture (selected arbitrarily), please see Figure 3. We note that our results are superior to CLIP without any calibration in all settings, but that our method results in models that are still significantly less calibrated than those calibrated via the supervised variant of Temperature

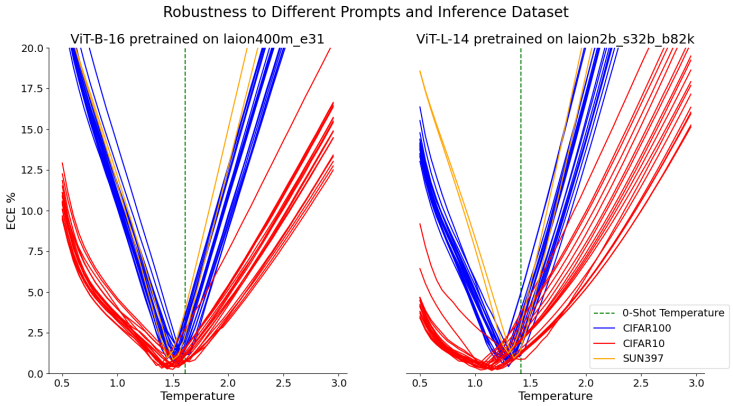

Figure 2: A plot of the effect of temperature on *ECE* for various prompts and inference datasets (enumerated in Appendix Section A.2). Different lines of the same color represent different prompts. Note that the optimal temperature T per inference dataset and prompt is very close to the one learned in a zero-shot way and is approximately the same across inference datasets and prompts.

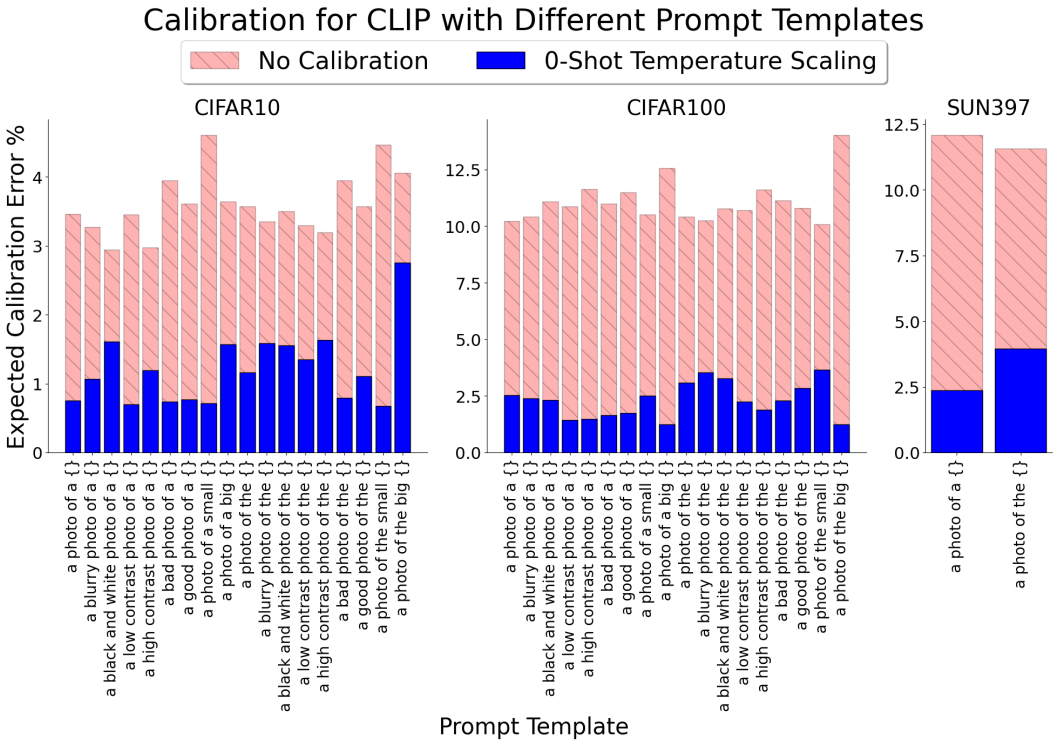

Figure 3: Differences in *ECE* with and without Zero-Shot Temperature Scaling for various prompts and datasets. Regions where Zero-Shot Temperature Scaling lowered *ECE* (i.e. improved calibration) can be visualized in the pink portion of the bars. Here, we use ViT-B-16 (Dosovitskiy et al. (2020)) pretrained on laion400m_e31 (Schuhmann et al. (2021)) as our model.

Scaling - thus, leaving room for future zero-shot CLIP calibration methods that improve upon our method.

4.2 Robustness to Changes in Prompt and Inference Dataset

In Figure 2, we show that, for a single architecture and pre-training dataset, the optimal temperature T across prompts and inference datasets are approximately the same. This is remarkable considering each of the inference and auxiliary datasets has a different number of classes from each other (since CIFAR10 has 10 classes, CIFAR100 has 100 classes, SUN397 has 397 classes, and ImageNet-1k has 1000 classes), as well as different distributions. We do note that this temperature T needs to be trained per architecute/pre-training dataset pair, as different pairs have (slightly) different optimal T's, as can be visualized by comparing the approximate optimal T's in the left plot (around $T = 1.6$) and in the right plot (around $T = 1.4$) of Figure 2.

5 Conclusion and Future Work

The miscalibration of supervised classification models has been extensively studied and improved via many calibration methods. Yet, prior to this paper, that has not been the case for vision-language models (like CLIP) with a different evaluation setup than traditional deep learning models. In this paper, we have shown that CLIP out-of-the-box is generally miscalibrated for a variety of experimental parameters. Lastly, to address this miscalibration, we have also presented a zero-shot calibration method for CLIP that modifies inference with a single learned temperature that is aligned with the CLIP zero-shot paradigm. Future work will extend additional supervised calibration methods to CLIP's zero-shot setting and provide improvements to our method to close the gap in calibration between our Zero-Shot Temperature Scaling and the supervised variant of Temperature Scaling.

References

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Morris H DeGroot and Stephen E Fienberg. Comparing probability forecasters: Basic binary concepts and multivariate extensions. Technical report, CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF STATISTICS, 1983.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8710–8719, 2021.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, and Christian Gagné. Unsupervised temperature scaling: An unsupervised post-processing calibration method of deep networks. *arXiv preprint arXiv:1905.00174*, 2019.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pp. 2901–2907. AAAI Press, 2015. ISBN 0262511290.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Vickram Rajendran and William LeVine. Accurate layerwise interpretable competence estimation. *Advances in Neural Information Processing Systems*, 32, 2019.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 609–616, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1558607781.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pp. 694–699, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 158113567X. doi: 10.1145/775047.775151. URL https://doi.org/10.1145/775047.775151.

A Appendix

A.1 Calibration Results of Supervised Methods

In Figure 4 we present reliability diagrams for CLIP using the following supervised calibration methods: Isotonic Regression Zadrozny & Elkan (2002) and Histogram Binning Zadrozny & Elkan (2001), as well as Unsupervised Temperature Scaling Mozafari et al. (2019). All three of these methods perform well and significantly reduce miscalibration. However, as mentioned in Section 2.4, the use of these methods is inconsistent with how CLIP is often used, and is therefore impractical for wide-scale adoption.

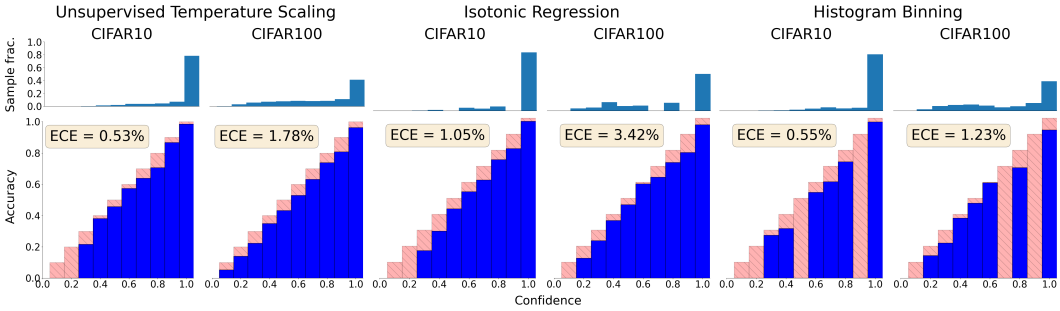

Figure 4: Reliability diagrams of **Left 2 Columns:** CLIP with Unsupervised Temperature Scaling Mozafari et al. (2019), **Center 2 Columns:** CLIP with Isotonic Regression Zadrozny & Elkan (2002), Zero-Shot Temperature Scaling, and **Right 2 Columns:** CLIP with Histogram Binning Zadrozny & Elkan (2001). Here, we use ViT-B-16 (Dosovitskiy et al. (2020)) pretrained on laion400m_e31 (Schuhmann et al. (2021)) as our model, and "a photo of {*CLASS NAME*}" as our prompt.

A.2 Prompts and Datasets Used in Figure 1

Our *ECE* results in Figure 1 are averaged over the following datasets and prompts:

1. SUN397 Yu et al. (2015) with the following prompts:
   (a) "a photo of {}"
   (b) "a photo of the {}"
2. CIFAR10 and CIFAR100 Krizhevsky (2009) with the following prompts:
   (a) "a photo of a {}"
   (b) "a blurry photo of a {}"
   (c) "a black and white photo of a {}"
   (d) "a low contrast photo of a {}"
   (e) "a high contrast photo of a {}"
   (f) "a bad photo of a {}"
   (g) "a good photo of a {}"
   (h) "a photo of a small {}"
   (i) "a photo of a big {}"
   (j) "a photo of the {}"
   (k) "a blurry photo of the {}"
   (l) "a black and white photo of the {}"
   (m) "a low contrast photo of the {}"
   (n) "a high contrast photo of the {}"
   (o) "a bad photo of the {}"
   (p) "a good photo of the {}"
   (q) "a photo of the small {}"
   (r) "a photo of the big {}"

A.3 Architectures and Pretraining Datasets Used in Figure 1

All models in Figure 1 are from OpenCLIP Ilharco et al. (2021). All ViT models used are pretrained on either LAION-400M Schuhmann et al. (2021) or LAION-2B Schuhmann et al. (2022). The ResNet models are pretrained on YFCC15M, a subset of YFCC100M Thomee et al. (2016), or the Conceptual Captions Dataset Changpinyo et al. (2021)