This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Intra-Class Probabilistic Embeddings for Uncertainty Estimation in Vision-Language Models

Zhenxiang Lin Maryam Haghighat Will Browne Dimity Miller
Queensland University of Technology, Brisbane, Australia
{z25.lin, maryam.haghighat, will.browne, d24.miller}@qut.edu.au

Abstract

Vision-language models (VLMs), such as CLIP, have gained popularity for their strong open vocabulary classification performance, but they are prone to assigning high confidence scores to misclassifications, limiting their reliability in safety-critical applications. We introduce a training-free, post-hoc uncertainty estimation method for contrastive VLMs that can be used to detect erroneous predictions. The key to our approach is to measure visual feature consistency within a class, using feature projection combined with multivariate Gaussians to create class-specific probabilistic embeddings. Our method is VLM-agnostic, requires no fine-tuning, demonstrates robustness to distribution shift, and works effectively with as few as 10 training images per class. Extensive experiments on ImageNet, Flowers102, Food101, EuroSAT and DTD show state-of-the-art error detection performance, significantly outperforming both deterministic and probabilistic VLM baselines. Code is available at https://github.com/zhenxianglin/ICPE.

1. Introduction

Contrastive vision-Language Models (VLMs), such as CLIP [40] and SigLIP [56], have achieved remarkable performance across a wide range of visual recognition tasks by learning from internet-scale datasets to align images and textual concepts in a shared embedding space. This has led to rapid adoption for open-vocabulary perception in safety-critical applications, such as medical diagnostics [46, 58], robot navigation [14, 24] and robot action planning [9, 26]. However, with real-world deployment, model errors have the potential for harmful consequences, and so models must also output a measure of confidence in their predictions.

Contrastive VLMs predict class labels by computing the cosine similarity between an image embedding and the textual prompt for each class [16, 27, 40, 55, 56, 59]. This cosine similarity, or its softmax transformation, is often used to represent prediction confidence [11, 12]. Recent work has shown that this is not a reliable measure of prediction confidence, as VLMs can produce high cosine similarity scores even when an input is misclassified as an incorrect semantic class [31, 36, 47, 53]. In this paper, we address the task of error detection for contrastive VLMs, introducing a technique to produce a new confidence score that can be used to identify errors in classification.

Importantly, prior work has shown that contrastive VLMs suffer from a modality gap where image and text embeddings occupy disjoint regions of the feature space [33]. This suggests that inter-modal similarity measures may not reliably capture prediction uncertainty for identifying errors, as distances between image and text features may be dominated by the modality gap itself rather than by true semantic uncertainty. Building on this observation, Dong et al. [11] demonstrated that using deterministic visual prototypes for image–image similarity offers stronger error detection than image–text cosine similarity, highlighting the benefit of extracting uncertainty within the visual modality.

Other recent works have introduced uncertainty estimation methods for VLMs by learning probabilistic embeddings that model inter-modal uncertainty (i.e. image-text alignment) [2, 5–7, 25, 50]. While this improves alignment in the embedding space and improves classification accuracy, it does not necessarily capture uncertainty for error detection (see Sec. 4.2). Additionally, these techniques often rely on large-scale pretraining or fine-tuning of VLMs, making them either computationally expensive [7, 25] or restricting their generalization beyond the fine-tuning dataset [5, 6, 50].

Motivated by this, we propose a technique for training-free, post-hoc uncertainty estimation for error detection in VLMs. Rather than modeling inter-modal uncertainty, we introduce probabilistic embeddings that capture intra-class feature distributions in the visual embedding space (see Fig. 1). Leveraging training data, we fit a multivariate Gaussian for each class to represent the range of typical visual

The authors acknowledge continued support from the Queensland University of Technology (QUT) through the Centre for Robotics.

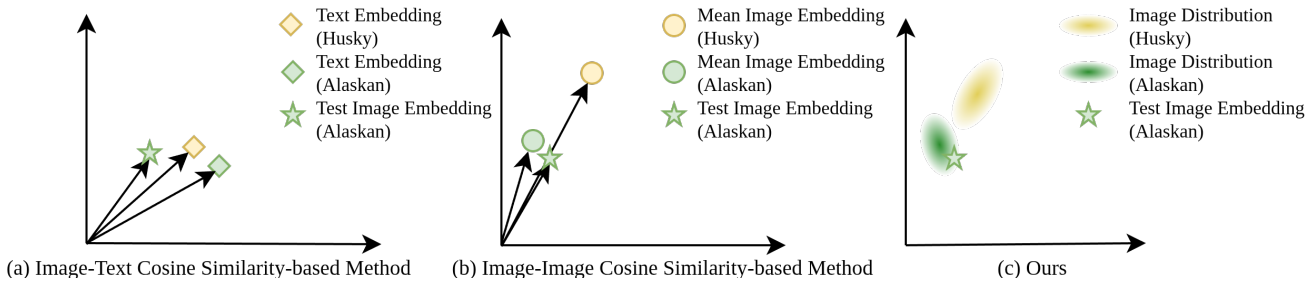


Figure 1. Illustration of different uncertainty estimation paradigms for vision-language models (VLMs). (a) The standard image–text cosine similarity approach assigns high confidence to the Alaskan image (green star) due to its high similarity with the "Husky" text embedding, despite being a misclassification. (b) Image–image similarity methods estimate uncertainty based on proximity to the class mean (yellow circle), but ignore the feature distribution, leading to unreliable scores. (c) Our method models intra-class distributions using image features and assigns higher uncertainty when a test sample deviates from the class distribution, enabling more accurate uncertainty estimation.

features. At inference, we retrieve the class distributions and compute the log-probability of the test image under each Guassian, producing an uncertainty score that reflects how well the visual features agree with known intra-class structure.

We make the following claims:

- Compared to deterministic [11, 12] and inter-modal uncertainty estimation techniques [2, 5, 7, 11], our intra-class probabilistic embeddings achieve state-of-the-art performance for VLM error detection (see Sec. 4.2). Our approach is post-hoc and training-free, and can obtain state-of-the-art performance with multiple VLM backbones with as few as 10 images per class (see Sec. 4.3).
- We identify that PCA-based feature projection is crucial for modeling each class distribution effectively (see Sec. 4.4), positing that this mitigates ill-conditioned covariance matrices in the VLM feature space.
- Our approach exhibits robustness to both label shift (semantic mismatches between query labels and dictionary labels) and feature shift (distributional differences between training and test images) (see Sec. 4.5).

2. Related Works

2.1. Uncertainty Estimation for Vision Models

Detecting misclassifications in image classifiers is a key challenge in computer vision. A common method is the maximum softmax probability (MSP), which use the model's most confident prediction as a proxy for correctness [22]. However, deep networks are prone to overconfidence [47, 53], often assigning high confidence to incorrect predictions.

To address this, several works [18, 32] apply temperature scaling to calibrate confidence. Differently, Bayesian approximations-based methods [3, 15, 17, 34, 43] treat model parameters as probability distributions and introduce uncertainty directly into the model weights. Ensemble-based methods [1, 29, 51, 54] train multiple independent models and combine the variance from their predictions to quantify uncertainty. Nonetheless, these methods still require high computational costs due to the training and testing of multiple models. Single network deterministic methods [35, 41, 44, 52] incorporate uncertainty estimation in a single deterministic forward pass and predict the distribution parameters of the output.

These methods have achieved strong performance in confidence calibration and misclassification detection across a variety of image classification tasks. However, they have been primarily developed for conventional vision-only architectures, such as CNNs and Transformers, where prediction is made solely based on visual inputs. For VLMs, like CLIP [40] and SigLIP [56], which perform classification by computing similarities between image and text embeddings, these techniques are not directly applicable.

2.2. Uncertainty Estimation for VLMs

VLMs [16, 27, 40, 55, 56, 59] have demonstrated impressive performance on a variety of visual recognition tasks through their ability to align images and textual concepts in a shared embedding space. In classification, VLMs typically assign labels by computing the cosine similarity between the image embedding and each class's textual prompt, and then applying softmax over the similarity scores. However, unlike traditional vision-only models, which rely on an explicit classifier, this similarity-based prediction mechanism raises fundamental challenges for estimating uncertainty.

To address this, Zero [12] uses the test-time adaptation strategy to calibrate confidence. However, it relies on multiple forwards at inference, causing a low inference speed. PCME [6], PCME++ [5] and ProbVLM [50] use probabilistic adapters to model distributions over text and image embeddings, but they require finetuning on a dataset

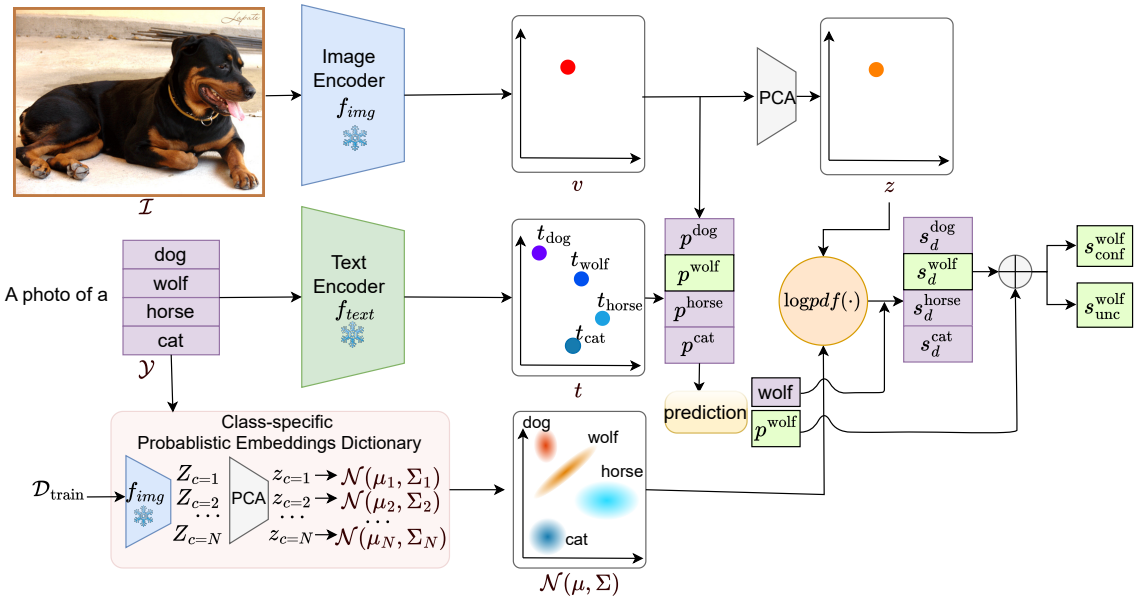


Figure 2. Overview of the proposed retrieval-augmented uncertainty estimation pipeline, which utilises a dictionary of intra-class probabilistic distributions to estimate uncertainty.

for a specific task. MAP [25] and ProLIP [7] proposed pre-training models for probabilistic representation that can benefit many downstream tasks. However, these methods need to cost a large number of computation sources and rely on an Internet-scale dataset. BayesVLM [2] leverages a posterior Laplace approximation to quantify uncertainties over cosine similarities, but its dependence on Hessian-based estimation requires access to the underlying loss function and model gradients, which may not always be accessible for all VLMs. TrustVLM [11], which is a concurrent work to ours, takes visual feature consistency into consideration and explore the error detection task, but they do not model a distribution and instead have a single prototype per class. Our method leverages class-wise distribution modeling over frozen VLM features to assess the distributional typicality of predictions, achieving effective uncertainty estimation without retraining or architectural modification.

3. Methodology

3.1. Preliminaries of VLM Classification

Contrastive VLMs, such as CLIP [40], encode both images and natural language into a shared feature space using modality-specific encoders. VLMs can be used for zero-shot classification by comparing the similarity between the encoded image and a set of class-descriptive text prompts.

Let $\mathcal{I}$ denote an input image and $\mathcal{Y} = \{y_1, \ldots, y_K\}$ the set of K class labels. Each image $\mathcal{I}$ is encoded into a visual feature vector $v = f_{\text{img}}(\mathcal{I}) \in \mathbb{R}^d$ using a visual encoder f_{img}. Each class label y_k is formatted into a natural language prompt (e.g., "a photo of a `<label>`"), and then encoded into a text feature $t_k = f_{\text{text}}(y_k)$ using a text encoder f_{text}.

Classification is performed by computing the cosine similarity between the image embedding v and each text embedding t_k:

$$s_k = \frac{v \cdot t_k}{\|v\| \, \|t_k\|}, \quad \text{for } k = 1, \ldots, K.$$

The predicted class $\hat{y}$ corresponds to the label with the highest similarity:

$$\hat{y} = \arg\max_k \, s_k.$$

The similarity scores $\{s_k\}_{k=1}^K$ can also be interpreted as a measure of confidence or uncertainty. Typically, these scores are normalized using a softmax function:

$$p_k = \frac{\exp(s_k)}{\sum_{j=1}^K \exp(s_j)},$$

and the maximum probability or the entropy of the distribution can also serve as a confidence or uncertainty estimate. However, prior work has shown that VLM's can assign high cosine similarity even when an input is misclassified [11, 12], making these cosine-based uncertainty measures suboptimal for error detection.

3.2. Method Pipeline

An overview of our approach is shown in Figure 2. Utilizing existing training datasets, our approach creates a dictionary of class-specific probabilistic distributions of feature

embeddings for all labeled classes. This captures class-level structure in the VLM feature space, which is absent in the standard cosine similarity scores.

At inference, given a test image, we first extract its deterministic visual feature and obtain its predicted class label (following the standard approach introduced in Sec. 3.1). We then project the feature to a low dimension space and compute the log-probabilities of the projected embedding under the Gaussian distribution of all queried classes in our dictionary. The softmax-normalized log-probability of the predicted class is combined with the image-text softmax score and used to represent prediction uncertainty, where predictions with high uncertainty are flagged as likely errors. Notably, our method does not alter the model's original prediction and introduces no training. It operates entirely on frozen features, providing a post-hoc uncertainty estimate that is agnostic to VLM backbone.

3.3. A Dictionary of Class-specific Probabilistic Embeddings

We assume access to a dataset $\mathcal{D}_{\text{train}} = \{(\mathcal{I}_i, y_i)\}_{i=1}^{N}$ consisting of N images $\mathcal{I}_i$ paired with class labels $y_i \in \mathcal{Y}_{\text{train}}$. Importantly, the visual features from $\mathcal{D}_{\text{train}}$ should be independent and identically distributed (i.i.d.) with the visual features of test images. In practice, this can be achieved by using the training subset for any given test dataset. While our method assumes i.i.d. visual features, it does not require $\mathcal{Y}_{\text{train}}$ to identically match the test label set. In Section 4.5, we show how our approach can handle moderate class label shift and feature shift.

To construct our dictionary of class-specific probabilistic distributions, we iterate over each class label $c \in \mathcal{Y}_{\text{train}}$. For each class c, we retrieve all training images from $\mathcal{D}_{\text{train}}$ that are labeled with c, and pass them through the frozen visual encoder of the VLM to obtain a set of high-dimensional feature vectors. We denote this set of visual embeddings as $Z_c = \{\mathbf{v_i} \,|\, y_i = c\}$. As motivated in Section 4.4, we then apply principal component analysis (PCA) to reduce the dimensionality of all embeddings before fitting class-conditional distributions.

We treat the PCA-transformed embeddings in Z_c as i.i.d. samples drawn from a multivariate distribution over the reduced feature space. Each class c becomes a key in our dictionary, associated with a multivariate Gaussian probabilistic embedding, parameterized by the mean μ_c and covariance matrix Σ_c:

$$\mu_c = \frac{1}{|Z_c|} \sum_{\mathbf{v} \in Z_c} \mathbf{v}, \quad \Sigma_c = \frac{1}{|Z_c| - 1} \sum_{v \in Z_c} (\mathbf{v} - \mu_c)(\mathbf{v} - \mu_c)^\top.$$

3.4. Feature Projection for Mitigating Ill-conditioned Covariance Matrices

Although VLMs' image embeddings are powerful for semantic alignment, their distribution has been shown to be anisotropic and highly correlated [30, 49]. Specifically, the covariance matrix $\mathbf{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{v_i} - \boldsymbol{\mu})(\mathbf{v_i} - \boldsymbol{\mu})^\top$ of the feature embeddings $\mathbf{v_i}$ contains significant off-diagonal entries, indicating strong linear correlations among feature dimensions [30]. In addition, the eigenvalue spectrum of $\mathbf{\Sigma}$ is highly non-uniform, suggesting extreme variance disparities across directions [30].

Such a covariance structure is ill-conditioned, characterised by rank deficiency and a high condition number, i.e. $\kappa(\mathbf{\Sigma}) = \frac{\lambda_1}{\lambda_d}$ is very large. This makes class-conditional modeling unstable and often numerically infeasible, especially when the number of class samples is limited relative to embedding dimension.

To mitigate this, we apply PCA to obtain an orthogonal basis $\mathbf{U} \in \mathbb{R}^{d \times d}$ such that $\mathbf{\Sigma} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$. We retain only the top-k principal components, and project embeddings to $\mathbf{z}_i = \mathbf{U_k}^\top (\mathbf{v_i} - \boldsymbol{\mu})$, where $\mathbf{U} \in \mathbb{R}^{d \times k}$, and $k \ll d$. The resulting covariance matrix in this projected space is approximately diagonal and significantly better conditioned with a reduced condition number $\kappa(\mathbf{\Sigma}) = \frac{\lambda_1}{\lambda_k} \ll \kappa(\mathbf{\Sigma})$. The distribution of the projected embeddings can then be approximated as $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}_k)$ enabling multivariate gaussian modeling. Depending on the amount of data available for Gaussian parameter estimation, our method can either adopt a diagonal covariance approximation or a full covariance approximation [37]. We report results for both variations and explore the trade-off between data availability and the covariance approximation in Sec. 4.3.

3.5. Uncertainty Estimation during Inference

With class-wise Gaussian distributions modeled in a stabilized feature space, we introduce an uncertainty scoring method that reflects how well a test image aligns with the feature distribution of its predicted class. Given a test image $\mathcal{I}$, we obtain its visual embedding $\mathbf{v} = f_{\text{img}}(\mathcal{I})$ using the frozen VLM image encoder. The predicted class $\hat{c}$ is then selected by computing cosine similarity between $\mathbf{v}$ and the class text embeddings $\{t_k\}_{k=1}^{K}$:

$$\hat{c} = \arg\max_k \langle \mathbf{v}, t_k \rangle.$$

To compute uncertainty, we first project $\mathbf{v}$ into the PCA-transformed space:

$$\mathbf{z} = \mathbf{U_k}^\top (\mathbf{v} - \boldsymbol{\mu}). \tag{1}$$

Using the text label for our predicted class $\hat{c}$, we then extract the relevant probabilistic embedding from our dictionary and compute the log-likelihood of $\mathbf{z}$ under the Gaussian distribution of the predicted class:

$$s_{\hat{c}} = \log pdf(\mathbf{z} \mid \mu_{\hat{c}}, \Sigma_{\hat{c}}), \tag{2}$$

where $pdf(\cdot)$ is the multivariate Gaussian probability density function. This score reflects how well the test image

aligns with the predicted class's feature distribution: higher values indicate typicality, while lower values suggest the prediction may be overconfident or erroneous.

To enable uncertainty comparisons across examples, we normalize the log-likelihood scores across all classes using a softmax function to find an intra-class uncertainty score s_d. To create our final uncertainty score, we combine this with the VLM inter-modal similarity

$$s_{\text{unc}} = 1 - \frac{p_{\max} + s_d}{2}. \tag{3}$$

where $p_{\max}$ denotes the maximum softmax probability obtained from image–text similarities. This combination allows our uncertainty score to represent both inter-modal similarity (which VLMs are extensively trained to capture) as well as the intra-class feature consistency.

The uncertainty score s_{unc} can be used for error detection by applying a simple thresholding strategy. If $s_{\textbf{unc}} > \tau$, where τ is a user-defined threshold, the model rejects the prediction as a likely misclassification. Our evaluation metrics test over a range of τ, and in Sec. 4.6, we explore the sensitivity of our method to this threshold.

3.6. Scalability

An important practical aspect of the proposed method is its scalability to large label spaces. While our dictionary construction only needs to occur once before deployment, the computational cost of this scales linearly with the number of classes and number of image samples per class. The most costly operation is the feature extraction for each image, which requires a forward pass through the VLM visual encoder. This cost can be alleviated by parallelization and GPU-accelerated inference, and in Sec. 4.3 we show that our method can achieve state-of-the-art performance with at least ten labeled images per class. In combination, this enables our method to feasibly scale to large-scale classification scenarios if such compute is available. We note, however, that our approach may be intractable for applications where the dictionary must be computed on a compute-limited edge devices.

4. Experimental Results

4.1. Experimental Setup

Metrics: We treat the error detection task as a binary classification problem (i.e. correct predictions to retain versus error to reject) and adopt three standardized metrics used for this task [11, 12, 21, 39, 45]: Area Under the Receiver Operating Characteristic Curve (AuROC), the False Positive Rate at 95% True Positive Rate (FPR95) and the Area Under the Precision-Recall Curve (AuPR). We additionally report top-1 accuracy on the multi-class classification task in the Supplementary Material, noting that this is a secondary metric and it is not our goal to improve accuracy.

Datasets: We evaluate our method on five standard image classification benchmarks with different domains and scales: ImageNet-1K [10] (1000 categories, generic objects and animals), Flowers-102 [38] (102 categories, flower species), Food-101 [4] (101 categories, food types), EuroSAT [19] (10 categories, satellite imagery and land types) and DTD [8] (47 categories, textures). For each dataset, we use the official training sets to model our probabilistic embeddings.

Baselines: We compare to four standard baselines for uncertainty estimation: maximum cosine similarity, maximum softmax score, entropy of the softmax distribution and temperature scaling [48]. We additionally compare to 6 state-of-the-art uncertainty estimation methods designed for contrastive VLMs, testing their performance for error detection. This includes methods that model probabilistic embeddings – ProbVLM [50], PCME++ [5], ProLIP [7] and BayesVLM [2] – as well as non-probabilistic approaches Zero [12] and TrustVLM [11]. Most similar to our approach is TrustVLM [11], which was also designed for the task of error detection and considers intra-class similarity, but does not consider a probability distribution.

Implementation Details: We use the CLIP ViT-B/32 (OpenAI), CLIP ViT-B/16 (OpenAI) and SigLIP (google/siglip-base-patch16-224) as backbones. When applying PCA to feature embeddings, we reduce the feature dimensionality to 128. PCA is fit once on the full set of training embeddings for each dataset and reused across all test images regardless of class. For TrustVLM [11], we use their public code and run on ImageNet directly. Due to different training and test data split on Food101 [4], Flowers102 [38], EuroSAT [19] and DTD [8], we use our data split to keep a fair comparison.

4.2. Uncertainty for Error Detection

In Table 1, we show that our approach achieves state-of-the-art performance for error detection across all 5 datasets. In particular, we observe between a 4-23% increase in AuROC and a 11-31% reduction in FPR95 when compared to the next best method. Notably, our approach consistently outperforms the baselines for both the CLIP [40] and SigLIP [56] VLMs. These results validate one of the core claims of our paper – state-of-the-art performance for VLM error detection that is agnostic to the underlying VLM.

Our strong performance can be attributed to two key insights underpinning our method design: (1) capturing intra-class visual consistency, with (2) the use of probabilistic embeddings. Comparatively, the other baselines that leverage probabilistic embeddings [2, 5, 7, 50] perform poorly for the task of error detection, often outperformed by the simple MaxSoftmax baseline – these approaches were designed for improving top-1 accuracy or confidence calibration tasks, and instead use cross-modal comparison

Table 1. Our approach outperforms the uncertainty estimation baselines for the CLIP and SigLIP backbones and the task of error detection across all datasets. *TrustVLM additionally uses DinoV2 features alongside VLM features. Ours-D indicates when our method uses a diagonal covariance approximation for the class-wise Gaussian distributions. Best results are shown in bold.

	Method	ImageNet			Flowers102			Food101			EuroSAT			DTD		
		AuROC (↑)	AuPR (↑)	FPR95 (↓)	AuROC (↑)	AuPR (↑)	FPR95 (↓)	AuROC (↑)	AuPR (↑)	FPR95 (↓)	AuROC (↑)	AuPR (↑)	FPR95 (↓)	AuROC (↑)	AuPR (↑)	FPR95 (↓)
CLIP ViT-B/32	MaxCosine	65.79	74.58	86.40	72.71	82.61	80.85	74.60	92.42	81.65	45.72	35.14	97.84	66.16	58.61	86.56
	MaxSoftmax	80.71	87.43	71.38	86.26	92.21	61.11	87.05	96.70	62.96	63.25	51.62	87.97	76.98	74.32	78.49
	Entropy	78.27	86.01	76.75	85.22	91.56	64.70	85.56	96.33	67.26	61.03	50.28	92.07	76.43	73.92	80.14
	TempScaling [48]	80.63	87.37	71.64	86.15	92.11	61.93	87.06	96.70	62.31	63.25	51.62	87.95	77.01	74.36	77.56
	ProbVLM [50]	50.18	61.79	95.31	43.05	53.74	97.20	52.87	82.99	95.16	42.72	30.23	95.53	46.57	39.52	97.26
	PCME++ [5]	77.31	90.68	71.82	80.05	93.89	70.11	87.60	97.84	53.44	89.31	**99.80**	42.57	82.06	**93.15**	75.65
	BayesVLM [2]	79.41	86.05	72.83	85.65	91.88	91.37	86.12	96.32	64.67	61.46	41.49	90.56	73.77	58.74	77.32
	Ours-D	85.80	90.53	59.02	**96.96**	**98.43**	14.79	**92.01**	**997.93**	37.04	96.56	92.57	24.82	**91.56**	89.44	**40.23**
	Ours	**87.52**	**91.61**	**55.26**	97.59	98.68	**4.86**	92.58	98.08	**33.98**	**96.37**	94.14	**16.39**	91.13	89.09	46.43
CLIP ViT-B/16	MaxCosine	70.20	87.04	82.31	79.98	94.32	64.23	77.61	96.30	75.11	59.24	48.18	87.98	69.69	64.96	85.67
	MaxSoftmax	84.21	**94.10**	65.30	93.49	98.64	40.18	91.43	98.92	49.98	76.21	73.69	81.15	79.96	78.54	76.33
	Entropy	83.46	93.87	68.64	93.09	98.55	36.77	90.79	98.84	54.79	71.97	67.99	83.73	79.15	77.54	76.86
	TempScaling [48]	81.28	89.87	72.16	85.41	92.60	60.58	89.21	98.12	54.61	76.22	73.70	81.28	79.94	78.48	76.33
	Zero [12]	78.87	90.84	73.95	83.73	92.30	62.98	85.13	97.84	56.62	66.73	61.81	87.23	76.64	77.68	81.79
	ProLIP [7]	70.57	85.77	81.23	80.67	92.82	59.95	77.05	96.81	76.04	62.42	60.64	89.44	78.96	86.70	72.74
	TrustVLM* [11]	83.39	90.43	75.67	89.18	**99.86**	34.43	89.37	98.19	56.04	73.72	91.84	73.66	77.17	88.98	74.49
	Ours-D	86.80	92.57	56.57	97.28	98.79	10.62	93.40	98.82	29.03	97.32	95.91	7.39	**93.00**	**91.55**	**33.86**
	Ours	**88.01**	93.23	**54.10**	**98.25**	99.15	**2.47**	**94.02**	**98.94**	**27.08**	**97.64**	**96.16**	**3.35**	92.61	91.31	36.94
SigLIP ViT-B/16	MaxCosine	70.20	87.04	82.31	79.98	94.32	64.23	77.61	96.30	75.11	60.47	54.20	98.93	68.82	78.99	85.19
	MaxSoftmax	84.21	94.10	65.30	93.49	98.64	40.18	91.43	98.92	49.98	74.91	69.24	80.30	82.05	89.21	82.29
	Entropy	83.46	93.87	68.64	93.09	98.55	36.77	90.79	98.84	54.79	74.41	69.04	82.51	80.30	88.18	71.18
	TempScaling [48]	84.12	94.06	66.06	93.64	98.66	38.38	91.43	98.92	49.71	74.40	68.62	80.85	81.48	88.84	72.41
	Zero [12]	79.43	93.55	68.65	86.30	96.84	53.09	85.23	97.96	55.76	66.63	55.79	87.95	75.65	86.51	74.22
	BayesVLM [2]	70.06	87.84	83.01	70.59	90.49	85.19	78.17	96.58	74.28	70.54	62.52	82.52	80.06	78.40	61.16
	TrustVLM* [11]	81.66	90.85	81.65	93.65	**99.93**	29.82	90.08	98.50	53.15	74.97	92.16	74.20	82.84	92.78	79.90
	Ours-D	86.17	94.80	56.88	**99.38**	99.88	**1.40**	**95.06**	**99.37**	28.13	96.93	96.95	12.67	**94.08**	**96.22**	**24.88**
	Ours	**88.57**	**95.85**	**48.87**	98.81	99.77	1.60	94.96	99.35	**24.50**	**98.10**	**98.03**	**2.93**	89.44	94.27	68.58



Figure 3. Visualization of our uncertainty on ImageNet [10]. Green indicates our method correctly distinguished between correct and error, and red indicates an incorrect distinction. The uncertainty threshold for error rejection is 0.5.

(image-text) rather than our proposed intra-class comparison (image-images). The most competitive baseline is TrustVLM [11]. Similar to our approach, TrustVLM utilizes intra-class visual consistency to estimate uncertainty, however it does not leverage probabilistic embeddings and only considers a mean embedding for each class – our substantial performance gains emphasize the importance of modeling the entire distribution for each class to reliably capture how typical a test image is with respect to the visual feature distribution of its predicted class.

Errors can arise from both epistemic uncertainty (due to out of distribution inputs [28]) and aleatoric uncertainty (due to ambiguous or noisy inputs [28]). Figure 3 shows a random sample of qualitative examples of our uncertainty estimates on ImageNet (additional included in the Supp. Material). Our method primarily appears to capture aleatoric uncertainty, e.g. the noisy data label example of the cowboy boot/hat image and the ambiguous siamese cat image. Our method also shows some evidence of identifying long-tail or out-of-distribution inputs arising from epistemic uncertainty, e.g. the atypical wooden spoon image, which has in fact been classified correctly despite being as-

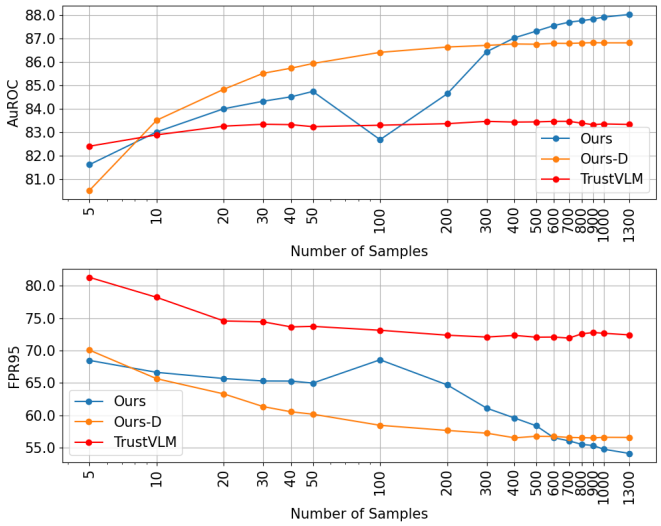


Figure 4. Testing on ImageNet with a CLIP ViT-B/16, our method achieves SOTA with 10 labeled images per class.

signed high uncertainty.

4.3. Data Dependency of Probabilistic Embeddings

To assess the robustness of our method to data availability, we examine how the number of labeled training images per class (used for calculating our dictionary) affects uncertainty estimation performance. As shown in Figure 4, both variants of our method achieves SOTA performance with only 10 labeled samples per class. As expected, both variants achieve optimal performance with more data availability (as a greater range of class features can be captured by the class distributions). In the presence of less than 400 labeled images per class, our method using a diagonal covariance matrix for each class Gaussian is preferable to the full covariance Gaussian variant. These findings show that our approach remains competitive even in low-data settings.

4.4. Importance of Feature Projection for Probabilistic Modeling

In Table 2, we evaluate the effect of feature projection in our method by comparing our intra-class uncertainty with and without the PCA feature projection (note that here we do not include the inter-modal softmax). As introduced in Sec. 3.4, PCA is applied to reduce the dimensionality of CLIP image features and mitigate ill-conditioned covariance matrices. After incorporating PCA, performance on ImageNet, Food101 and EuroSAT in particular are improved by 2-8% AuROC and 15-24% FPR95.

In addition, in Figure 5, we show that the use of PCA on the feature embeddings substantially reduces the condition number of the covariance matrices for each class probabilistic embedding in ImageNet, Food101 and EUROSAT (see Supp. Material for Flowers102 and DTD). Higher condition numbers indicate ill-conditioned covariance matrices that are prone to numerical instability. This result, combined with the substantial performance increase on these datasets when applying PCA, supports one of the core claims of our paper – PCA-based feature projection is crucial for allowing each class distribution to be modeled effectively as it mitigates ill-conditioned covariance matrices in the VLM feature space.

Table 2. Error detection performance of our method with (w) and without (w/o) PCA.

Metric	ImageNet		Flowers102		Food101		EuroSAT		DTD	
	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w
AuROC	82.46	87.71	95.62	95.39	84.78	93.18	95.69	98.44	88.61	86.95
FPR95	74.70	54.98	3.19	3.64	59.80	35.58	20.17	4.26	100.00	100.00

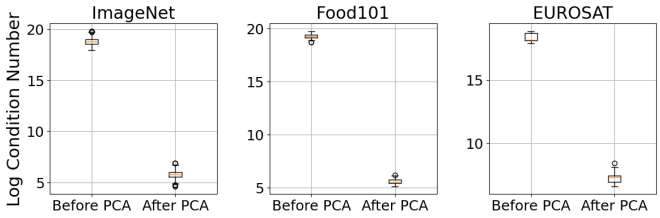


Figure 5. For our probabilistic embeddings, application of PCA mitigates ill-conditioned covariance matrices, measured by the log condition number of the covariance matrices for each class.

4.5. Robustness to Distribution Shift

We explore the robustness of our method under two types of data distribution shift [57]: feature shift (also referred to as covariate shift) and class label shift.

Feature shift: We create our dictionary using the ImageNet training data, and then evaluate performance on three ImageNet variants with different levels of feature shift: ImageNetV2 [42] (image source distribution shift through newly collected real images), ImageNet-C [20] (synthetic distribution shift via image corruptions) , and ImageNet-A [23] (adversarial natural distribution shift via naturally occurring hard examples).

As shown in Table 3, our method achieves the best AuROC and FPR95 performance on both ImageNetV2 and ImageNet-C, demonstrating robustness to feature shift introduced by image sources and synthetic image corruptions. However, ImageNet-A poses a significant challenge for both our approach and TrustVLM [11], where error detection on natural adversarial images is tested – in this case, the simple baseline of MaxSoftmax achieves best performance. As noted in [23], the image distribution shift between ImageNet and ImageNet-A is significant, highlighting the limits of error detection methods that rely on data

Table 3. Error detection performance under feature shift, where training data is from ImageNet and methods are evaluated on ImageNet variants with CLIP ViT-16.

Method	ImageNet-V2			ImageNet-C			ImageNet-A		
	AuROC (↑)	AuPR (↑)	FPR95 (↓)	AuROC (↑)	AuPR (↑)	FPR95 (↓)	AuROC (↑)	AuPR (↑)	FPR95 (↓)
MaxCosine	67.84	79.59	83.55	65.99	59.88	87.41	64.29	60.95	86.67
MaxSoftmax	81.30	89.83	71.00	81.48	78.34	68.32	**76.14**	74.63	77.74
Entropy	78.78	88.62	78.08	78.84	75.72	72.86	74.82	73.40	**77.66**
TempScaling [48]	81.53	89.95	70.43	81.65	78.57	68.14	76.12	**74.84**	78.20
TrustVLM [11]	83.03	89.68	78.76	82.12	**83.73**	71.99	64.28	49.53	92.92
Ours-D	85.32	91.91	62.27	**83.51**	80.12	**64.09**	70.34	67.53	77.99
Ours	**86.19**	**92.42**	**59.00**	82.45	78.41	65.06	68.83	64.74	78.38

for uncertainty estimation.

Class label shift: We test our performance when the VLM query labels do not exactly match the keys in our class probabilistic embedding dictionary, simulating a coarse-to-fine mismatch setting on the ImageNet-1k dataset [10]. We assume that every query class is represented in the dictionary, but that the text input may not exactly match, utilizing the WordNet hierarchy [13] to group the 1,000 fine-grained ImageNet labels into coarser categories. We construct two levels: the 1st-level superclass containing 548 labels and the 2nd-level (grand) superclass containing 299 labels. The test set labels are replaced with their corresponding superclasses, while the dictionary's class labels remain the original fine-grained labels. To adapt our method to this setting, we retrieve the labels from the dictionary which are among the top-K most semantically similar classes with the test class, where similarity is measured by cosine similarity in the CLIP text embedding space. The images with these labels are then used to construct probabilistic embeddings for the corresponding superclass. The number K is selected based on the ratio between the number of classes in the dictionary and the number of test-time labels, calculated by

$$K = \left\lfloor \frac{N_{\text{retrieval}}}{N_{\text{test}}} \right\rceil . \tag{4}$$

Table 4 shows our method outperforms all baselines in all metrics under both label shift levels. This suggests that our class-specific probabilistic embeddings capture uncertainty more reliably than similarity-based baselines. Compared to TrustVLM-LS, which also attempts to estimate uncertainty under label shift using mean embeddings of top-K classes, our method demonstrates improved robustness by modeling the full feature distribution rather than only relying on mean feature. This result supports the final core claim of our paper that our method exhibits robustness to label shift, and can continue to exhibit SOTA performance when the probabilistic embedding labels have some discrepancy with the test query labels.

Table 4. Results of label shift on the 1st and 2nd level superclasses. For 1st-level, we set $K = 2$. For 2nd-level, we set $K = 3$.

Method	1st level				2nd level			
	AuROC	AuPR	FPR95	Acc	AuROC	AuPR	FPR95	Acc
MaxCosine	62.58	43.87	87.72	**32.53**	63.53	27.90	87.06	**19.15**
MaxSoftmax	71.78	55.46	81.05	**32.53**	67.89	31.86	82.73	**19.15**
Entropy	69.78	53.69	83.76	**32.53**	67.43	31.51	83.55	**19.15**
TempScaling [48]	71.76	55.56	81.09	32.51	67.58	31.68	82.81	19.14
TrustVLM-LS [11]	74.00	57.42	87.40	26.72	68.06	28.60	91.99	13.81
Ours-D-LS	77.15	62.39	75.44	**32.53**	70.47	34.05	**79.86**	**19.15**
Ours-LS	**77.34**	**62.69**	**75.18**	**32.53**	**70.64**	**34.28**	80.83	**19.15**

4.6. Sensitivity to Uncertainty Threshold

To further examine the robustness of our approach, we explore the sensitivity of error detection performance to the rejection threshold τ, which determines when a prediction is flagged as an expected error. As shown in Figure 6, across all five datasets, the F1 score of our method remains at a stable peak between 0.4 and 0.6 for CLIP backbones and between 0.1 and 0.4 for the SigLIP backbone. This indicates that our method does not require precise tuning of τ, and a value of 0.4 can be used for both backbones to consistently achieve strong performance for error detection.

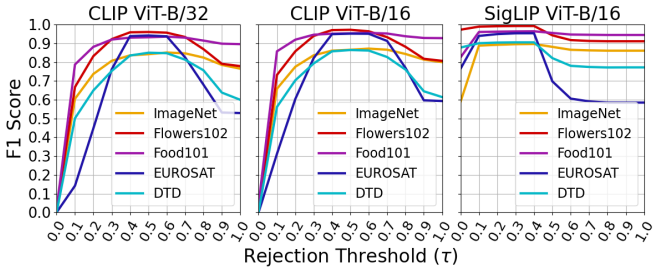


Figure 6. Sensitivity of our approach for error detection across a range of uncertainty rejection thresholds.

5. Conclusion

In this paper, we present a training-free, post-hoc method for uncertainty estimation in vision-language models, focusing on error detection. Our method estimates uncertainty by combining inter-modal uncertainty with a measure of intra-class feature consistency captured by distributions in the visual embedding space. Although the proposed method achieves state-of-the-art performance across multiple benchmarks, its reliance on data for building the class-wise probabilistic distributions introduces two challenges that should be explored in future work: (1) performance degrades under feature shift on adversarially curated natural examples, indicating susceptibility in boundary-adjacent, hard regions of the input space; and (2) this method may become intractable when the distributions must be built on compute-constrained edge devices.

References

[1] Omer Achrack, Raizy Kellerman, and Ouriel Barzilay. Multi-loss sub-ensembles for accurate classification with uncertainty estimation. *arXiv preprint arXiv:2010.01917*, 2020. 2

[2] Anton Baumann, Rui Li, Marcus Klasson, Santeri Mentu, Shyamgopal Karthik, Zeynep Akata, Arno Solin, and Martin Trapp. Post-hoc probabilistic vision-language models. *arXiv preprint arXiv:2412.06014*, 2024. 1, 2, 3, 5, 6

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015. 2

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pages 446–461. Springer, 2014. 5

[5] Sanghyuk Chun. Improved probabilistic image-text representations. In *International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 5, 6

[6] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[7] Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 2, 3, 5, 6

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5

[9] Xuzhe Dang, Stefan Edelkamp, and Nicolas Ribault. Clipmotion: Learning reward functions for robotic actions using consecutive observations. *arXiv preprint arXiv:2311.03485*, 2023. 1

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 6, 8

[11] Hao Dong, Moru Liu, Jian Liang, Eleni Chatzi, and Olga Fink. To trust or not to trust your vision-language model's prediction. *arXiv preprint arXiv:2505.23745*, 2025. 1, 2, 3, 5, 6, 7, 8

[12] Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024. 1, 2, 3, 5, 6

[13] Christiane Fellbaum. *WordNet: An electronic lexical database*. MIT press, 1998. 8

[14] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023. 1

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 1, 2

[17] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011. 2

[18] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 2

[19] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. 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. 5

[20] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 7

[21] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 5

[22] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *CVPR*, 2021. 7

[24] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. *arXiv preprint arXiv:2210.05714*, 2022. 1

[25] Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, and Yujiu Yang. Map: Multimodal uncertainty-aware vision-language pre-training model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23262–23271, 2023. 1, 3

[26] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734, 2025. 1

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2

[28] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 6

[29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[30] Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of clip. *arXiv preprint arXiv:2411.14517*, 2024. 4

[31] Will LeVine, Benjamin Pikus, Pranav Vishnu Raja, and Fernando Amat. Enabling calibration in the zero-shot inference of large vision-language models. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 1

[32] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 2

[33] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1

[34] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. *Advances in neural information processing systems*, 30, 2017. 2

[35] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018. 2

[36] Dimity Miller, Niko Sünderhauf, Alex Kenna, and Keita Mason. Open-set recognition in the age of vision-language models. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1

[37] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. 4

[38] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5

[39] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1686–1695, 2023. 5

[40] 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*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5

[41] Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *International conference on machine learning*, pages 5281–5290. PMLR, 2019. 2

[42] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 7

[43] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 2

[44] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 2

[45] Qiaozhi Tan, Long Bai, Guankun Wang, Mobarakol Islam, and Hongliang Ren. Endoood: Uncertainty-aware out-of-distribution detection in capsule endoscopy diagnosis. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024. 5

[46] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12): 1399–1406, 2022. 1

[47] Weijie Tu, Weijian Deng, and Tom Gedeon. A closer look at the robustness of contrastive language-image pre-training (clip). *Advances in Neural Information Processing Systems*, 36:13678–13691, 2023. 1, 2

[48] Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *International Conference on Machine Learning*, pages 48791–48808. PMLR, 2024. 5, 6, 8

[49] Kirill Tyshchuk, Polina Karpikova, Andrew Spiridonov, Anastasiia Prutianova, Anton Razzhigaev, and Alexander Panchenko. On isotropy of multimodal embeddings. *Information*, 14(7):392, 2023. 4

[50] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023. 1, 2, 5, 6

[51] Matias Valdenegro-Toro. Deep sub-ensembles for fast uncertainty estimation in image classification. *arXiv preprint arXiv:1910.08168*, 2019. 2

[52] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020. 2

[53] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1802–1812, 2023. 1, 2

[54] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020. 2

[55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 1, 2

[56] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, 2023. 1, 2, 5

[57] Aston Zhang, Zachary C. Lipton, Mu Li, and Alexander J. Smola. *Dive into Deep Learning*. Cambridge University Press, 2023. https://D2L.ai. 7

[58] Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 458–468. Springer, 2024. 1

[59] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2