This ICCV 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.

# ViLU: Learning Vision-Language Uncertainties for Failure Prediction

Marc Lafon[⋆,1] Yannis Karmim[⋆,1] Julio Silva-Rodríguez[3] Paul Couairon[2]
Clément Rambour[1,2] Raphaël Fournier-Sniehotta[1,2] Ismail Ben Ayed[3] Jose Dolz[3]
Nicolas Thome[2]
[1]Conservatoire National des Arts et Métiers, CEDRIC, F-75141 Paris, France
[2]Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
[3]ETS Montreal, Canada

## Abstract

*Reliable Uncertainty Quantification (UQ) and failure prediction remain open challenges for Vision-Language Models (VLMs). We introduce* ***ViLU****, a new* ***Vi****sion-***L****anguage* ***U****ncertainty quantification framework that contextualizes uncertainty estimates by leveraging all task-relevant textual representations. ViLU constructs an uncertainty-aware multi-modal representation by integrating the visual embedding, the predicted textual embedding, and an image-conditioned textual representation via cross-attention. Unlike traditional UQ methods based on loss prediction, ViLU trains an uncertainty predictor as a binary classifier to distinguish correct from incorrect predictions using a weighted binary cross-entropy loss, making it loss-agnostic. In particular, our proposed approach is well-suited for post-hoc settings, where only vision and text embeddings are available without direct access to the model itself. Extensive experiments on diverse datasets show the significant gains of our method compared to state-of-the-art failure prediction methods. We apply our method to standard classification datasets, such as ImageNet-1k, as well as large-scale image-caption datasets like CC12M and LAION-400M. Ablation studies highlight the critical role of our architecture and training in achieving effective uncertainty quantification. Our code is publicly available and can be found here: ViLU Repository.*

## 1. Introduction

Vision Language Models (VLMs) [9, 25, 27, 36, 50] are highly popular foundation models pre-trained on large-scale image-text datasets, *e.g.*, LAION [37]. They possess the appealing capability of performing zero-shot image classification, meaning they can classify samples into classes that were not seen during training.

Uncertainty quantification (UQ) is a fundamental challenge in deep learning and involves estimating the confidence of a model's predictions. This paper addresses the problem of reliable UQ for VLMs to detect their potential failures in downstream tasks. Reliable UQ is crucial in safety-critical domains and offers significant opportunities for various applications involving VLMs, such as failure prediction [17], out-of-distribution detection [46], active learning [47], and reinforcement learning [12], among others.

The vanilla method for UQ with VLMs is the Maximum Concept Matching (MCM) score [30], a direct extension of Maximum Class Probability (MCP) [17] for classification. Although MCM is a strong baseline, it suffers from fundamental drawbacks: by design, it assigns high confidence to failures and struggles with fine-grained concepts. As illustrated in Fig. 1, the VLM misclassifies the "`Eskimo dog`" image as a "`Siberian husky`", and the high MCM score prevents the detection of the error.

In classification, learning-based methods have also been explored for failure prediction. In particular, a few deep learning models designed to predict the classifier's loss and dedicated to learning visual uncertainties (LVU) have been proposed [8, 19, 47]. However, when applied to VLMs, these methods do not model the relationships between downstream concepts, intrinsically limiting their failure prediction performance. The example in Fig. 1 still receives a high confidence score with LVU methods. Furthermore, although recent methods have proposed specific calibration techniques for VLMs [31, 34, 42, 44, 48], fewer works have focused on UQ solutions for VLMs in general, with the exception of the recent [2].

This paper introduces ViLU, a post-hoc framework for learning **Vi**sion-**L**anguage **U**ncertainties and detecting VLMs' failure. The core idea in ViLU is to define a confidence score depending both on the visual input but also on the set of concepts that defines the downstream task,

⋆ Equal contribution
Corresponding author: marc.lafon@lecnam.net

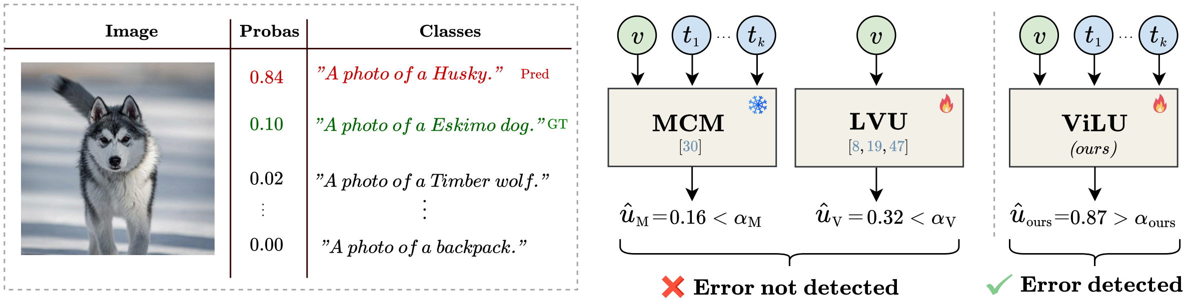



Figure 1. **Motivation of ViLU.** In zero-shot classification with VLMs, uncertainty may arise from both the image and target concept definitions. Given an ambiguity between several concepts, the vanilla Maximum Concept Maxing (MCM) [30], can assign a high confidence to a wrong prediction, *e.g.* a "`Eskimo dog`" wrongly classified as a "`Siberian husky`". Previous methods based on Learning Visual Uncertainties (LVU) [8, 19, 47] do not account for ambiguity between concepts and often fail, whereas ViLU captures a fine-grained uncertainty by contextualizing the image within the spectrum of possible textual concepts.

*e.g.* classification or caption matching. By finely modeling the interaction between the image and the target concepts, ViLU assigns a low confidence score to the misclassified example – see Fig. 1.

We summarize our key contributions as follows:

- ViLU introduces a novel multi-modal uncertainty representation that integrates visual embeddings, predicted textual embeddings, and image-conditioned textual representations via a cross-attention module. This formulation enables the model to capture fine-grained ambiguities between the input image and candidate concepts, significantly improving failure prediction.
- We propose a dedicated uncertainty predictor that operates on this enriched representation and is trained to discriminate between correct and incorrect predictions using weighted binary cross-entropy (BCE). Unlike conventional loss-prediction-based UQ methods, ViLU is fully loss-agnostic, making it particularly well-suited for post-hoc uncertainty estimation in black-box VLMs.

We conduct an extensive experimental validation of ViLU on various downstream classification datasets, as well as image-caption datasets such as CC12M and LAION-400M. First, we highlight that state-of-the-art methods struggle to outperform the MCM baseline, illustrating the difficulty of the task. In contrast, ViLU delivers significant and consistent improvements over multiple baselines, including recent VLM-specific methods [2]. Thorough ablation studies further validate our architectural choices and training design for optimal performance.

## 2. Related Work

**Vision-language models (VLMs)** have gained popularity for aligning image and text representations [9, 25, 27, 36, 50], achieving unprecedented zero-shot performance by jointly learning a shared vision-text embedding space on large-scale web datasets [9, 24, 37]. They have been applied in a plethora of domains including image classification [22, 52, 54], open-vocabulary segmentation [23, 43, 49, 51] and cross-modal retrieval [1, 18]. Nevertheless, VLMs align deterministic text and image representations without accounting for uncertainty in zero-shot predictions [2, 42]. While simple uncertainty quantification methods exist, *e.g.* MCM [30], a robust approach for failure prediction remains lacking. We address this gap with an effective UQ method for failure prediction of VLMs.

**Uncertainty quantification (UQ) for failure prediction.** In classification, MCP [17] is the vanilla method for UQ. However, by only considering the maximum over the predicted probabilities, MCP tends to overestimate confidence in failure cases [8]. Another common approach is to use the Shannon entropy of the predicted softmax distribution [38], but its invariance to label permutations [14, 38] limits its effectiveness for failure detection. Recently, Doctor [14] was proposed to refine the Shannon entropy [38] using Rényi entropy, while Rel-U [13] further extended it by incorporating a learned distance matrix to model class relationships from classifier predictions. Such methods, however, suffer from a limited expressiveness in their UQ models, and struggle to capture finer-grained uncertainties. Another line of work is to learn the classifier's loss using deep neural networks [8, 19, 47]. [47] estimates model loss to improve active learning, while [8] learns the cross-entropy loss for classification and segmentation. More recently, [19] applies this approach for large-scale UQ with vision transformers. However, these methods are limited to learning visual uncertainties (LVU), and generalize poorly to VLMs as they do not account for ambiguity in the language modality. To overcome this limitation, we propose an efficient UQ method for frozen, pre-trained VLMs and show that directly predicting failures outperforms loss prediction.

**VLMs' UQ.** MCP can be easily adapted for UQ of VLMs by leveraging their zero-shot probabilities, leading to the Maximum Concept Matching (MCM) method [30]. However, MCM inherits MCP's limitations for failure prediction, especially its overconfidence for errors. However, VLMs' overconfidence is generally less pronounced than for standard classifiers [29, 41]. As a result, when the VLM's downstream accuracy is sufficiently high, MCM remains a strong baseline for failure prediction. Several works extend VLMs for UQ in cross-modal retrieval tasks [5, 6, 26, 32], often by learning probabilistic embeddings for each modality. However, most require retraining both visual and textual backbones, limiting their practicality [5, 6]. To address this, ProbVLM [42] learns distribution parameters over embeddings via adapters, but overlooks cross-modal similarity scores, which limits its effectiveness for failure prediction. BayesVLM [2] applies a Laplace approximation to model uncertainty over similarities post-hoc. While both capture vision-language uncertainty, their objectives are not tailored for failure prediction, which limits their performance compared to our approach.

# 3. Background

## 3.1. Contrastive vision-language models

**Zero-shot predictions.** Consider a particular dataset $\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{N}$, where each image $\boldsymbol{x}_i \in \mathcal{X}$ is paired with a class name or caption $\boldsymbol{t}_i \in \mathcal{T}$. Contrastive VLMs are composed of a pre-trained vision encoder $f_{\mathcal{V}}(\cdot)$ that maps an input image $\boldsymbol{x}_i$ to a visual embedding $\boldsymbol{z}_{v_i} = f_{\mathcal{V}}(\boldsymbol{x}_i) \in \mathbb{R}^d$, and a text encoder $f_{\mathcal{T}}(\cdot)$ that embeds a textual input $\boldsymbol{t}_j$ into a representation $\boldsymbol{z}_{t_j} = f_{\mathcal{T}}(\boldsymbol{t}_j) \in \mathbb{R}^d$. This multi-modal joint embedding space is learned during pre-training by aligning paired concepts. Thus, it enables zero-shot image classification, by computing the probability $p(\boldsymbol{t}_j|\boldsymbol{x}_i)$ of image $\boldsymbol{x}_i$ being associated to caption $\boldsymbol{t}_j$ using the softmax of the similarities between the visual representations $\boldsymbol{z}_{v_i}$ and the embeddings $\boldsymbol{z}_{t_j}$ of a set of $K$ candidate textual concepts:

$$p(\boldsymbol{t}_j|\boldsymbol{x}_i) = \frac{\exp\left(\boldsymbol{z}_{v_i}^{\top}\boldsymbol{z}_{t_j}/\tau\right)}{\sum_k \exp\left(\boldsymbol{z}_{v_i}^{\top}\boldsymbol{z}_{t_k}/\tau\right)} \tag{1}$$

where $\tau$ is a temperature parameter optimized during pre-training, and $\boldsymbol{z}_{v_i}$ and $\boldsymbol{z}_{t_j}$ are $\ell_2$-normalized embeddings.

## 3.2. VLMs' failure prediction with MCM

We aim to determine whether VLMs can recognize when their predictions are unreliable. This is captured by an uncertainty scoring function, $u(\boldsymbol{x})$, where higher values indicate a greater likelihood of misclassification.

**Maximum Concept Matching (MCM) [30]** corresponds to the probability of the predicted caption for image $\boldsymbol{x}_i$, *i.e.*, the one with the highest probability:

$$u_{\text{MCM}}(\boldsymbol{x}_i, \boldsymbol{t}_1, ..., \boldsymbol{t}_k) = 1 - \max_j p(\boldsymbol{t}_j|\boldsymbol{x}_i). \tag{2}$$

**Limitations.** Despite being a reasonable UQ method in coarse-grained classification tasks, $u_{\text{MCM}}$ has fundamental limitations. Firstly, the max operation in Eq. (2) by design assigns an overestimated UQ for incorrect predictions. Therefore, MCM performances drop when the zero-accuracy of the classifier is low, as shown in the experiments. Also, in fine-grained classification or open-vocabulary settings, visual and textual alignments become more dispersed, reducing MCM's reliability. This limitation is particularly problematic for modern VLMs, which are trained on large-scale datasets for open-vocabulary tasks.

# 4. ViLU model

This section presents our ViLU framework for multi-modal failure prediction on VLMs, as illustrated in Fig. 2. We first introduce our general post-hoc methodology in Sec. 4.1, which enables UQ on VLMs without modifying their internal parameters. We then describe our architecture in Sec. 4.2, detailing the design of ViLU's embedding using image-text cross-attention, and of our failure classification head. Finally, we outline ViLU's training procedure in Sec. 4.3, including our weighted cross-entropy loss that directly aligns with the failure prediction task.

## 4.1. Methodology for UQ on VLMs

**Post-hoc Setting.** We aim to design a discriminative and reliable UQ measure for pre-trained VLMs. To achieve this, we adopt a post-hoc approach, relying solely on visual and textual representations, *i.e.*, $\mathcal{D} = \{(\boldsymbol{z}_{v_i}, \boldsymbol{z}_{t_i})\}_{i=1}^{N}$. This allows the UQ model to be easily integrated on top of a pre-trained VLM, providing uncertainty estimates without modifying internal representations, requiring fine-tuning, or depending on the loss function used during training.

**Learning vision-language uncertainties.** We propose leveraging the interactions between the *visual modality* and the *set of candidate concepts* to estimate uncertainty for failure prediction. Uncertainty in model predictions can arise from *visual patterns* (*e.g.*, low image quality, ambiguous features) or *textual patterns*, which define concept distinctions. Additionally, it is shaped by cross-modal interactions. To model these interactions, we learn a global uncertainty representation $u_\theta(\cdot)$ that captures visual-textual interactions and their inherent uncertainty. We adopt a *data-driven approach*, training the uncertainty module to predict VLMs' misclassifications. Specifically, we frame our objective as a *binary classification task*, where $u_\theta$ predicts whether an input will be misclassified by the VLM (see Sec. 4.3). Formally, let us define our uncertainty module as:

$$\begin{array}{rccc} u_\theta : & \mathbb{R}^d \times \mathbb{R}^{K \times d} & \rightarrow & [0,1] \\ & (\boldsymbol{z}_v, \boldsymbol{z}_{t_1}, ..., \boldsymbol{z}_{t_K}) & \rightarrow & u_\theta(\boldsymbol{z}_v, \boldsymbol{z}_{t_1}, ..., \boldsymbol{z}_{t_K}) \end{array} \tag{3}$$

Note that our uncertainty function in Eq. (3) can handle varying values of $K$, as the number of candidate concepts may vary during inference.

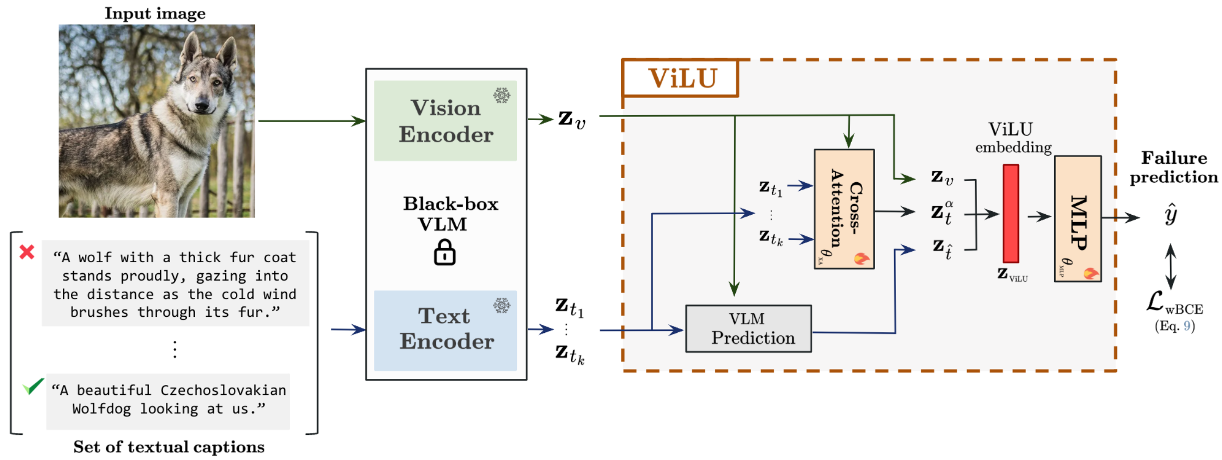



Figure 2. **Overview of ViLU:** A learning-based **Vi**sion-**L**anguage **U**ncertainty quantification framework for VLM failure prediction. The key strength of ViLU lies in its ability to contextualize uncertainty estimates by leveraging all textual representations relevant to the task (see Sec. 4.1). It constructs an uncertainty-aware representation by combining the visual embedding $\boldsymbol{z}_v$, the predicted textual embedding $\boldsymbol{z}_{\hat{t}}$, and an image-conditioned textual representation $\boldsymbol{z}_t^\alpha$ obtained via cross-attention (see Sec. 4.2). Instead of relying on loss prediction, ViLU trains an uncertainty predictor as a binary classifier to distinguish between correct and incorrect predictions (see Sec. 4.3). ViLU is a post-hoc approach that can be efficiently deployed on top of any pre-trained VLM without accessing its weights (black box) and supports both image-caption and image-label tasks.

## 4.2. ViLU's architecture

This section details the architectural components of our uncertainty module $u_\theta$ and how a visual feature, $\boldsymbol{z}_v$, for a given sample and textual embeddings of $K$ target concepts, $Z_t = \{\boldsymbol{z}_{t_j}\}_{1\leq j\leq K}$, are combined to provide an expressive representation for failure detection.

**Vision-textual cross-attention.** To enable flexible representations of the visual and textual modalities, we employ a *cross-attention module* $h_{\theta_{\text{XA}}}$ that produces an image-specific textual representation $\boldsymbol{z}_t^\alpha$ allowing to efficiently capture inherent cross-modal uncertainty:

$$\boldsymbol{z}_t^\alpha = h_{\theta_{\text{XA}}}(\boldsymbol{z}_v, \boldsymbol{z}_{t_1}, ..., \boldsymbol{z}_{t_K}) \tag{4}$$

More specifically, the *query* is the visual representation $\boldsymbol{z}_v$, and *keys and values* are the $K$ textual embeddings $Z_t$. The cross-attention's output is obtained as:

$$\begin{aligned} \boldsymbol{\alpha} &= \text{softmax}\left((W_Q\boldsymbol{z}_v)^\top (W_K Z_t)/\sqrt{d}\right) \\ \boldsymbol{z}_t^\alpha &= \sum_{j=1}^{K} \boldsymbol{\alpha}_j (W_V \boldsymbol{z}_{t_j}) \end{aligned} \tag{5}$$

where $\boldsymbol{\alpha}$ are the attention weights, and $W_Q$, $W_K$, and $W_V$ denote the projection matrices for the queries, keys, and values, respectively. This weighted textual embedding refines the textual context based on the model's predicted distribution over the candidate captions, allowing for a more accurate uncertainty assessment. This contextualized representation can be computed for any number of concepts, ensuring a generic uncertainty module that remains well-defined across different concept sets.

**ViLU embeddings.** To resolve vision-language ambiguity for failure prediction, our model must capture key information to distinguish correct from incorrect predictions. At a minimum, this requires the visual representation $\boldsymbol{z}_v$ and the textual embedding of the predicted caption $\boldsymbol{z}_{\hat{t}}$, allowing the model to approximate MCM's uncertainty estimator. However, this limited representation overlooks ambiguous alternatives, making error detection unreliable. To address this, we construct a rich vision-language uncertainty embedding $\boldsymbol{z}_{\text{ViLU}} = (\boldsymbol{z}_v, \boldsymbol{z}_{\hat{t}}, \boldsymbol{z}_t^\alpha)$ by concatenating $\boldsymbol{z}_v$, $\boldsymbol{z}_{\hat{t}}$, and the cross-attention output $\boldsymbol{z}_t^\alpha$.

**Learning complex patterns for failure prediction.** Detecting misclassifications among numerous fine-grained concepts is challenging, as it requires capturing complex cross-modal relationships. To overcome this, we apply a non-linear transformation on our ViLU embeddings via a multi-layer perceptron (MLP), $g_{\theta_{\text{MLP}}}$, which enhances feature expressiveness and produces a scalar uncertainty estimate. Formally, uncertainty quantification is expressed as the predicted failure score $\hat{y}_i$:

$$\hat{y}_i = \sigma\big(g_{\theta_{\text{MLP}}}(\boldsymbol{z}_{\text{ViLU}})\big), \tag{6}$$

where $\sigma$ denotes the sigmoid function, ensuring that $\hat{y}_i \to 1$ when the predicted zero-shot label is likely incorrect. Importantly, $g_{\theta_{\text{MLP}}}$ in Eq. (6) boils down to the unnormalized MCM score when using a bilinear form on $\boldsymbol{z}_{\text{ViLU}}$ (see Appendix B). Our failure predictor is thus a *consistent generalization of MCM*, incorporating its prior and learning to refine it for finer multi-modal UQ.

### 4.3. Training procedure

ViLU accommodates both image-caption and image-label tasks during training and inference.

**1) Image-Label Tasks:** Image-label classification considers a predefined set of $K$ target categories, leading to a *batch-independent* predictive pipeline. Here, the textual representations of categories are obtained using *text templates* (*e.g.*, "`A photo of a [CLASS]`"), resulting in a fixed set of textual captions $\{\boldsymbol{t}_j\}_{j\in\{1,\ldots,K\}}$. The predicted concept for an image is then determined as:

$$\hat{\boldsymbol{t}}_i = \underset{j\in\{1,\ldots,K\}}{\arg\max}\; p(\boldsymbol{t}_j|\boldsymbol{x}_i). \tag{7}$$

This setting is batch-independent, making it suitable for standard classification datasets with predefined labels.

**2) Image-Caption Tasks:** When a set of textual descriptions is available from an open-vocabulary domain, the goal is to assign the most similar caption to a given input image. It is typically validated using batches of paired images and text descriptions. Given a batch $\mathcal{B}$ of images with associated text descriptions, $\{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i\in\mathcal{B}}$, the predicted concept for an image is determined as:

$$\hat{\boldsymbol{t}}_i = \underset{j\in\mathcal{B}}{\arg\max}\, p(\boldsymbol{t}_j|\boldsymbol{x}_i). \tag{8}$$

Here, the performance is *batch-dependent*, with larger batch sizes increasing task complexity.

**Training objective.** Our uncertainty module $u_\theta$ is trained as a binary classifier to predict VLM zero-shot misclassifications. The parameters $\theta = \{\theta_{\text{XA}}, \theta_{\text{MLP}}\}$ of our uncertainty model are trained by mini-batch gradient descent to minimize the following weighted binary cross-entropy loss:

$$\mathcal{L}_{\text{wBCE}} = -\frac{1}{B}\sum_i \left[w y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right] \tag{9}$$

where $y_i = \mathbb{1}_{\{\hat{\boldsymbol{t}}_i \neq \boldsymbol{t}_i\}}$ is the target label, and $w$ is a weighting factor that mitigates the potential class imbalance between correctly and incorrectly classified examples. This weight is dynamically adjusted based on the empirical classification accuracy of the VLM within each mini-batch:

$$w = \log\left(1 + \frac{\sum_{i=1}^{B}(1-y_i)}{\sum_{i=1}^{B} y_i}\right) \tag{10}$$

Previous UQ methods predict the model's training loss [8, 19, 47], assuming a high loss value indicates misclassification. Applying this approach to VLMs would require predicting CLIP's [36] contrastive loss or SigLIP's [50] sigmoid loss, which is impractical in a post-hoc setting where the pre-training loss is unknown. In contrast, ViLU is loss-agnostic, relying only on whether a prediction is correct. As shown in the experiments Tab. 4, training it as a binary classifier consistently outperforms loss prediction for VLMs, naturally aligning with standard error detection metrics.

## 5. Experiments

This section presents experimental results validating our multi-modal uncertainty model. We first focus on predicting VLM failures in zero-shot classification across two setups: ***i)*** standard image-label classification datasets and ***ii)*** large-scale image-caption datasets. Next, we experimentally analyze the different components of our model and training procedure, conducting various assessments to evaluate its behavior. We provide qualitative results and visualizations illustrating the ability of ViLU to quantify uncertainty.

### 5.1. Experimental setup

**Datasets.** We conduct a benchmark across 16 commonly used datasets to evaluate the ability of VLMs to detect misclassification in image classification. As outlined in Sec. 4.3, we consider two settings: ***i)*** image-label datasets and ***ii)*** image-caption datasets. Image-label datasets are standard datasets used in CLIP's transfer learning [22, 52], covering general object recognition, fine-grained classification, and specialized domains (see Appendix A.1). We use official or CoOp [53] data splits for evaluation. Image-caption datasets, including CC3M [39], CC12M [4], and LAION-400M [37], contain free-text descriptions for each sample. From these datasets, we randomly hold out 1% of the data for testing.

**Implementation details.** We use CLIP ViT-B/32 as the default backbone in all experiments. Full implementation details, along with additional results using other backbones, are provided in Appendix A.2 and C.4, respectively.

**Baselines.** We explore relevant baselines for uncertainty estimation in vision, as well as recent approaches for VLMs. Further details on these baselines and implementation specifics can be found in Appendix A.3. **1) Measures of output distribution:** MCM [30], described in Sec. 3, along with its calibrated variant [15], provide robust uncertainty estimates without additional training. Similarly, Entropy [38] and Doctor [14] are commonly used as baselines for uncertainty estimation. **2) Data-driven predictors:** We implement the recent Rel-U [13], which incorporates cross-label uncertainties in the logit space. However, since Rel-U relies on label-based information, it does not apply to image-caption datasets. Additionally, we assess methods that leverage embedding representations to learn patterns related to uncertainty. Specifically, we compare against baselines that predict classifier loss, such as ConfidNet [8] and other vision-only estimators [19, 47], collectively referred to as 'Learning Visual Uncertainty' (LVU). Finally,

| | CIFAR-10 88.3% | | CIFAR-100 68.6% | | Caltech101 91.4% | | Flowers102 64.0% | | OxfordPets 85.1% | | Food101 78.9% | | ImageNet-1k 62.0% | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ |
| MCM [30] | 89.9 | 52.1 | 82.7 | 67.3 | 88.1 | 68.7 | 86.6 | 68.0 | 87.2 | 59.9 | 86.4 | 63.3 | 80.8 | 71.3 |
| TS [15] + MCM [30] | 89.9 | 51.5 | 83.9 | 68.4 | 90.4 | 55.7 | 86.9 | 66.0 | 89.1 | 55.5 | 86.9 | 62.8 | 80.7 | 71.5 |
| Entropy [38] | 88.7 | 59.9 | 79.8 | 71.9 | 86.1 | 78.8 | 85.5 | 65.0 | 88.0 | 60.0 | 86.1 | 65.0 | 78.3 | 76.8 |
| Doctor [14] | 89.5 | 56.5 | 82.3 | 69.7 | 88.7 | 66.5 | 86.7 | 63.9 | 88.9 | 56.6 | 86.8 | 63.4 | 80.3 | 72.9 |
| Rel-U [13] | 86.2 | 54.4 | 81.0 | 68.2 | 90.2 | 58.5 | 90.0 | 47.3 | 83.5 | 59.3 | 81.8 | 73.4 | 75.1 | 85.0 |
| LVU [8, 19, 47] | 96.6 | 21.2 | 80.3 | 68.5 | 89.8 | 50.9 | 90.5 | 38.3 | 84.1 | 55.7 | 82.7 | 69.9 | 78.7 | 77.0 |
| BayesVLM [2] | 92.6 | 44.9 | 87.0 | 60.3 | 94.0 | 37.4 | 87.3 | 62.4 | 89.5 | 60.3 | 87.8 | 60.3 | 81.5 | 70.3 |
| **ViLU (*Ours*)** | **98.3** | **7.7** | **91.5** | **35.4** | **96.7** | **18.2** | **98.7** | **5.1** | **94.4** | **24.5** | **94.8** | **28.5** | **89.5** | **47.4** |

| | FGVCAircraft 18.1% | | EuroSAT 35.8% | | StanfordCars 60.1% | | DTD 43.0% | | SUN397 62.1% | | UCF101 61.6% | | **Average** 62.7% | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ |
| MCM [30] | 75.7 | 82.9 | 64.1 | 87.6 | 81.4 | 73.4 | 77.4 | 77.9 | 78.8 | 75.9 | 84.1 | 68.9 | 81.8 | 70.6 |
| TS [15] + MCM [30] | 74.9 | 82.9 | 63.0 | 88.1 | 81.6 | 71.9 | 76.9 | 78.3 | 79.0 | 75.5 | 84.5 | 69.7 | 82.0 | 69.5 |
| Entropy [38] | 74.1 | 83.6 | 61.0 | 92.1 | 79.4 | 77.5 | 76.4 | 80.2 | 75.8 | 78.6 | 83.5 | 72.9 | 80.2 | 74.0 |
| Doctor [14] | 74.8 | 82.9 | 62.4 | 89.5 | 80.9 | 73.1 | 76.9 | 82.6 | 78.2 | 77.2 | 84.5 | 70.2 | 81.6 | 71.2 |
| Rel-U [13] | 68.6 | 82.5 | 76.3 | 72.1 | 75.5 | 78.9 | 81.4 | 69.7 | 75.2 | 81.3 | 84.1 | 61.1 | 80.7 | 68.6 |
| LVU [8, 19, 47] | 74.8 | 83.5 | 95.9 | 19.3 | 78.4 | 75.4 | 87.2 | 55.9 | 76.7 | 76.9 | 88.6 | 53.6 | 85.0 | 57.4 |
| BayesVLM [2] | 70.9 | 84.3 | 74.3 | 86.6 | 87.7 | 63.4 | 77.6 | 77.0 | 80.3 | 73.4 | 84.6 | 66.2 | 84.2 | 65.1 |
| **ViLU (*Ours*)** | **81.0** | **71.7** | **98.8** | **4.4** | **90.1** | **46.8** | **93.8** | **28.8** | **88.6** | **50.3** | **95.9** | **20.9** | **93.2** | **29.9** |

Table 1. **Misclassification detection on image-label datasets**. The evaluation leverages each dataset's labeled classes as textual queries, which are fixed for each batch. Values below dataset names denote CLIP zero-shot accuracy on the respective dataset.

we evaluate the most recent post-hoc UQ probabilistic modeling approach designed for VLMs, BayesVLM [2].

**Metrics.** To assess the performance of our uncertainty model, we rely on two standard metrics: **1)** False Positive Rate at $95\%$ True Positive Rate (**FPR95**) and **2)** Area Under the receiver-operating characteristic Curve (**AUC**). These are commonly used in failure prediction to quantify the model's ability to detect incorrect classifications [8, 13, 14].

### 5.2. Main results

**Image classification datasets.** In Tab. 1, we evaluate misclassification detection on CLIP's zero-shot predictions across 13 standard image-label datasets. Notably, MCM emerges as a strong baseline in this setting, even outperforming more complex methods such as Doctor and Rel-U. This advantage arises because these methods rely on task-specific vision classifiers, which are generally less well-calibrated than large-scale pre-trained models. Consequently, MCM achieves superior performance without requiring post-processing steps, such as temperature scaling [15], for failure prediction. On the other hand, expressive data-driven methods such as LVU, BayesVLM, and our proposed ViLU enable more effective failure detection. Our proposed ViLU ranks first across all datasets and metrics. Its novel architectural design leverages the class-semantic information embedded in MCM while also capturing uncertainty patterns specific to each sample's ambiguities. As a clear demonstration of its effectiveness, ViLU achieves remarkable improvements in FPR95, surpassing MCM, BayesVLM, and LVU by margins of $-40.7$, $-35.2$, and $-27.5$, respectively (Tab. 1, Average column).

CLIP's zero-shot accuracy for each dataset is reported in Tab. 1. MCM and BayesVLM performances closely follow CLIP's accuracy, struggling in low zero-shot accuracy settings: they achieve only $64.1\%$ and $74.3\%$ AUC in EuroSAT, a dataset on which CLIP's accuracy is $35.8\%$. This limitation makes them unreliable when zero-shot accuracy is low – an unpredictable scenario in real-world settings. In contrast, ViLU remains effective across all accuracy regimes. See Appendix C for further analysis.

**Large-scale image-caption datasets.** We now evaluate the ability of state-of-the-art methods to detect failures in open-vocabulary settings, where tasks vary based on the inference batch and its specific captions. Tab. 2 presents results for applicable baselines in this challenging setting, along with the proposed ViLU. Unlike in previous experiments on image-label datasets, LVU fails to outperform even the MCM baseline, underscoring the importance of considering target objectives alongside sample-related uncertainties in open-vocabulary scenarios. As a result, methods specifically designed for vision-language models, such as BayesVLM, achieve superior performance across all three evaluated datasets. Our proposed ViLU achieves the best performance, consistently surpassing BayesVLM with FPR95 improvements of $-21.5$, $-28.1$, and $-5.2$ on CC3M, CC12M, and LAION-400M, respectively. This advantage stems from our explicit modeling of misclassifi-

| | CC3M 58.8% | | CC12M 73.5% | | LAION-400M 90.5% | |
|---|---|---|---|---|---|---|
| | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ |
| MCM [30] | 83.9 | 69.0 | 88.8 | 58.8 | 91.7 | 50.2 |
| Entropy [38] | 82.5 | 73.3 | 87.7 | 63.0 | 89.4 | 62.5 |
| Doctor [14] | 83.7 | 70.1 | 88.6 | 59.9 | 91.2 | 54.5 |
| LVU [8, 19, 47] | 69.3 | 82.5 | 74.4 | 76.5 | 80.2 | 72.3 |
| BayesVLM [2] | <u>87.1</u> | <u>62.6</u> | <u>90.9</u> | <u>53.3</u> | <u>95.1</u> | <u>26.4</u> |
| **ViLU (*Ours*)** | **91.4** | **41.1** | **95.2** | **25.2** | **97.3** | **21.2** |

Table 2. **Misclassification detection on image-caption datasets**. The evaluation uses the captions of each randomly-retrieved batch as textual queries. Hence, the textual queries vary for each batch. Results reported with a batch size of 1024 samples for inference.

cation errors, whereas BayesVLM focuses on the uncertainty of similarities between individual embeddings, which is more implicit.

### 5.3. Ablation studies

**Architectural design of ViLU.** Tab. 3 highlights the contributions of different components in our model, namely the cross-attention mechanism and the inclusion of the predicted caption as input. As observed in [8, 19], using **only visual information** (first row) is effective on CIFAR-10 but struggles on larger datasets with fine-grained or semantically similar concepts. For instance, on ImageNet-1k, visual information alone fails to outperform MCM. Incorporating the **predicted class textual embedding** (second row) significantly improves performance across all datasets, yielding a +10 AUC gain on ImageNet-1k and +14 AUC on CC12M. This additional input helps the model to handle class ambiguity, which is crucial for datasets where categories are easily confused. For example, as shown in Fig. 3, the class `container ship` is often misclassified as `ocean liner`, another type of boat. Finally, the **cross-attention module** (third row) enables the model to integrate contextual information from all candidate classes or captions. By re-contextualizing predictions among available textual inputs, this mechanism proves particularly beneficial for CC12M, where captions change dynamically across batches. Unlike CIFAR-10 and ImageNet-1k, where class sets are fixed, omitting cross-attention in CC12M results in an AUC plateau at 88.9, only slightly above MCM. Incorporating this mechanism enhances the model's ability to detect ambiguities and assess potential errors.

**Loss function design for failure prediction.** In Sec. 4.3, we discussed our optimization target and choice of loss function. Unlike prior works [8, 19, 47], which treat the problem as a regression task by directly predicting the test-time loss and optimizing it with MSE, we instead frame it as a binary classification task, distinguishing between errors and correct predictions. As shown in Tab. 4, this approach consistently outperforms MSE-based loss approximations,

| Visual embed. | Cross attention | Predicted caption | CIFAR-10 | | ImageNet-1k | | CC12M | |
|---|---|---|---|---|---|---|---|---|
| | | | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ |
| ✓ | ✗ | ✗ | 96.4 | 21.8 | 78.7 | 77.0 | 74.0 | 76.5 |
| ✓ | ✗ | ✓ | 97.9 | 10.8 | 88.8 | 50.1 | 88.9 | 48.9 |
| ✓ | ✓ | ✗ | 97.7 | 11.4 | 86.1 | 63.5 | 93.6 | 37.0 |
| ✓ | ✓ | ✓ | **98.3** | **8.2** | **89.5** | **47.4** | **95.2** | **25.2** |
| | MCM [30] | | 89.9 | 52.1 | 80.8 | 71.3 | 88.8 | 58.8 |

Table 3. **Ablation on different components of ViLU**.

**GT**: *"container ship"* **Pred**: *"ocean liner"*
**GT**: *"mailbox"* **Pred**: *"birdhouse"*

Figure 3. **Two qualitative examples.** Misclassifications detected by ViLU, but not MCM [30] and visual-only baselines [8, 19].

yielding a +3 AUC improvement on ImageNet-1k while reducing FPR95 by 16 points. Additionally, our automatic weighting of the two classes (misclassified and correctly classified) further enhances the performance, resulting in a +1 AUC gain on ImageNet-1k and a 2-point reduction in FPR95 on CIFAR-10.

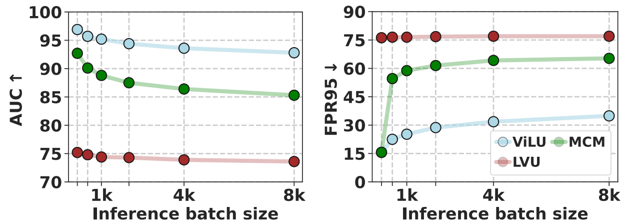
| | CIFAR-10 | | ImageNet-1k | |
|---|---|---|---|---|
| | AUC↑ | FPR95↓ | AUC↑ | FPR95↓ |
| MSE [8, 19] | 95.1 | 30.4 | 84.8 | 64.4 |
| w/ Weighting | 95.6 | 29.6 | 85.7 | 63.2 |
| BCE | 97.7 | 10.5 | 88.6 | 48.4 |
| **w/ Weighting (*Ours*)** | **98.3** | **8.2** | **89.5** | **47.4** |

Table 4. **Role of the proposed weighted loss**. Effect of loss function choice and adaptive weighting in Eq. (10) for failure detection.

### 5.4. In-depth analysis

**Classification complexity on image-caption datasets.** Fig. 4 explores the performance of ViLU in challenging image-caption classification tasks, where difficulty is determined by the number of concepts to be distinguished simultaneously, *i.e.*, batch size used for inference. Naturally, increasing the batch size makes the classification more complex, making errors harder to detect for methods that rely on semantic relationships among query concepts, such as ViLU or MCM. However, ViLU consistently outperforms MCM across all batch sizes. Notably, this configuration does not affect prior LVU methods, which estimate vision-only uncertainty. However, they fall short in performance, particularly compared to the proposed ViLU.

**Requirements on training data.** As a data-driven uncer-



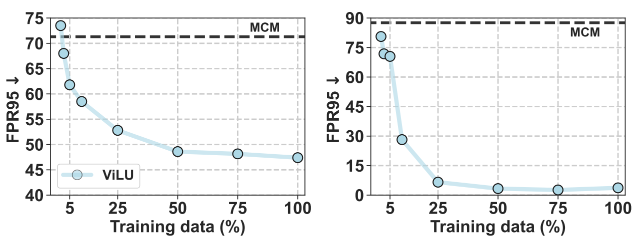

AUC FPR95

Figure 4. **Robustness to image-caption task complexity.** Inference batch size effect in failure detection for ViLU (CC12M).

| Dataset | MCM | LVU | **ViLU** |
|---|---|---|---|
| CIFAR-10 | **52.1** | 77.2 | 54.2 |
| CIFAR-100 | 67.3 | 83.8 | **59.9** |
| Caltech101 | 68.7 | 82.5 | **48.8** |
| Flowers102 | 68.0 | 96.8 | **67.4** |
| OxfordPets | 59.9 | 93.1 | **58.1** |
| Food101 | **63.3** | 87.2 | 67.4 |
| FGVCAircraft | 82.9 | 94.5 | **82.3** |
| EuroSAT | 87.6 | 88.2 | **85.7** |
| DTD | **77.9** | 93.1 | 78.2 |
| SUN397 | 75.9 | 90.1 | **72.7** |
| StanfordCars | **73.4** | 92.6 | 84.1 |
| UCF101 | 68.9 | 90.4 | **63.8** |
| **Average** | 70.5 | 89.1 | **68.6** |

Table 5. FPR95↓ across datasets. Zero-shot performance when pre-trained on CC12M.

tainty quantifier, the proposed ViLU requires a subset of image-caption or image-label examples. Fig. 5 illustrates the data-efficiency of the proposed approach for uncertainty quantification. The results showcase the efficiency of ViLU, which *requires only a small amount of data to surpass the strong baseline MCM*, *e.g*., 2.5% for ImageNet and even less for specialized datasets such as EuroSAT.




**(a) ImageNet-1k** **(b) EuroSAT**

Figure 5. **Data-efficiency.** Performance, in terms of FPR95↓, w.r.t. the available data for training ViLU.

**Cross-datasets generalization.** To evaluate ViLU's robustness to dataset shifts and its ability to generalize without per-dataset tuning, we report in Tab. 5 its zero-shot transfer performance when pre-trained on the large-scale CC12M dataset. ViLU consistently outperforms MCM across all 12 datasets in this pure transfer setting. The table also highlights ViLU's clear advantage over LVU (also trained on CC12M), underscoring the benefit of explicitly modeling vision-language uncertainty for zero-shot generalization. While these results demonstrate strong transfer capabilities, generalization can still be improved. In Appendix C.6, we explore how smarter batch sampling strategies aimed at improving concept coverage during pre-training could further enhance performance.

**Qualitative assessment.** In Fig. 3, we present two qualitative examples from ImageNet where the VLM misclassified the input images. On the left, the model predicted `ocean liner` instead of `container ship`, likely due to their shared visual features as large vessels. On the right, a `mailbox` was misclassified as a `birdhouse`, possibly influenced by the surrounding vegetation and the mailbox's opening, which resemble typical birdhouse features. ViLU effectively captures both sources of ambiguity: semantic confusion between visually similar classes and misinterpretations driven by contextual cues.

**Uncertainty distribution score.** Fig. 6 illustrates the effectiveness of different uncertainty estimation methods in distinguishing between correctly and incorrectly classified samples on ImageNet. In MCM and LVU, the uncertainty distributions for success and failure overlap significantly. In contrast, ViLU produces a more distinct separation. Indeed, misclassified samples receive high uncertainty (peaking near 1.0), while correct predictions have low uncertainty (peaking near 0.0). Its higher density at the extremes indicates a well-calibrated uncertainty measure.

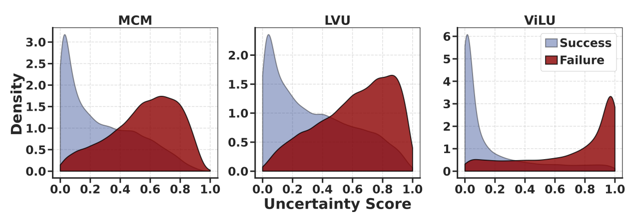



Figure 6. **Uncertainty score distribution.** Predictions for correctly and incorrectly classified samples on ImageNet.

## 6. Conclusion

We have presented our ViLU method, a new uncertainty quantification approach for failure prediction of pre-trained VLMs. ViLU learns an appropriate uncertainty embedding space including fine interactions between visual and concepts ambiguities to learn an effective binary failure predictor on the downstream task. Extensive experiments on several datasets show the significant gain of our method compared to state-of-the-art UQ methods for failure prediction. In addition, ablation studies clearly validate our architectural choices and training design. We also show that ViLU remains effective even in the low-performance regime of VLMs. Future works include adjusting ViLU in the context of domain adaptation and test-time adaptation of VLMs.

## Acknowledgment

This work was supported by the French National Research Agency (Agence Nationale de la Recherche, ANR) under grants from project DIAMELEX (ANR-20-CE45-0026) and project RODEO (ANR-24-CE23-5886). It was granted access to the HPC resources of IDRIS under the allocation AD011013370R2 made by GENCI.

## References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A CLIP-hitchhiker's guide to long video retrieval. *arXiv preprint arXiv:2205.08508*, 2022. 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, 6, 7, 13

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014. 12

[4] 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 *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE, 2021. 5

[5] Sanghyuk Chun. Improved probabilistic image-text representations. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[6] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8415–8424, 2021. 3

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014. 12

[8] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 5, 6, 7, 12, 13, 14, 15

[9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024. 1, 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 12

[11] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Worskshops (CVPRW)*, pages 178–178, 2004. 12

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, page 1050–1059. JMLR.org, 2016. 1

[13] Eduardo Dadalto Câmara Gomes, Marco Romanelli, Georg Pichler, and Pablo Piantanida. A data-driven measure of relative uncertainty for misclassification detection. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 5, 6, 12

[14] Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. Doctor: A simple method for detecting misclassification errors. *Advances in Neural Information Processing Systems*, 34:5669–5681, 2021. 2, 5, 6, 7, 12

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017. 5, 6, 12

[16] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3606–3613, 2018. 12

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *The Fifth International Conference on Learning Representations*. OpenReview.net, 2017. 1, 2

[18] Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, and Donglin Wang. Vop: Text-video co-operative prompt tuning for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 6565–6574, 2023. 2

[19] Michael Kirchhof, Mark Collier, Seong Joon Oh, and Enkelejda Kasneci. Pretrained visual uncertainties. *arXiv preprint arXiv:2402.16569*, 2024. 1, 2, 5, 6, 7, 12, 13, 14

[20] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3498–3505, 2012. 12

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 12

[22] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2024. 2, 5

[23] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 2

[24] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[25] Hugo Laurençon, Leo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2

[26] Hao Li, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Haonan Zhang, and Gongfu Li. A differentiable semantic metric approximation in probabilistic embedding for cross-modal retrieval. In *Advances in Neural Information Processing Systems*, 2022. 3

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 2

[28] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. In *ArXiv Preprint*, 2013. 12

[29] 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. 3

[30] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35:35087–35102, 2022. 1, 2, 3, 5, 6, 7, 12, 13, 14

[31] Balamurali Murugesan, Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *European Conference on Computer Vision*, pages 147–165, 2024. 1

[32] Andrei Neculai, Yanbei Chen, and Zeynep Akata. Probabilistic compositional embeddings for multimodal image retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4546–4556, 2022. 3

[33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 12

[34] Changdae Oh, Hyesu Lim, Mijoo Kim, Dongyoon Han, Sangdoo Yun, Jaegul Choo, Alexander Hauptmann, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. *Advances in Neural Information Processing Systems*, 37:12677–12707, 2024. 1

[35] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3498–3505, 2012. 12

[36] 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, 5, 14

[37] 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. *CoRR*, abs/2111.02114, 2021. 1, 2, 5

[38] Murat Sensoy, Lance M. Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 3183–3193, 2018. 2, 5, 6, 7, 12

[39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics, 2018. 5

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *ArXiv Preprint*, 2012. 12

[41] 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. 3

[42] 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, 3, 12

[43] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 2

[44] Shuoyuan Wang, Jindong Wang, Guoqing Wang, Bob Zhang, Kaiyang Zhou, and Hongxin Wei. Open-vocabulary calibration for fine-tuned CLIP. In *Proceedings of the 41st International Conference on Machine Learning*, pages 51734–51754, 2024. 1

[45] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010. 12

[46] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey, 2024. 1

[47] Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019. 1, 2, 5, 6, 7, 12, 13, 14

[48] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-TPT:

Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[49] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 2

[50] 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*, pages 11975–11986, 2023. 1, 2, 5, 12, 14

[51] Zhuowen Tu Zheng Ding, Jieke Wang. Open-vocabulary universal image segmentation with maskclip. In *International Conference on Machine Learning*, 2023. 2

[52] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16816–16825, 2022. 2, 5

[53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 5

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2