ViLU: Learning Vision-Language Uncertainties for Failure Prediction

Supplementary Material

A. Additional details on experimental setup

A.1. Datasets

In this section, we provide additional details on the image–label datasets used in the main experiments presented in the paper. These span general object recognition, fine-grained classification, and specialized domains. The datasets include ImageNet-1k [10], CIFAR-10 [21], CIFAR-100 [21], SUN397 [45], FGVCAircraft [28], EuroSAT [16], StanfordCars [20], Food101 [3], OxfordPets [35], Flowers102 [33], Caltech101 [11], DTD [7], and UCF101 [40].

A.2. Implementation details

As mentioned in the main paper, we use CLIP ViT-B/32 backbone in all our experiments. We provide additional results using other CLIP backbones and architectures, such as SigLIP [50] in Appendix C.4. Regarding the proposed ViLU, the MLP layer for misclassification prediction follows a four-layer architecture with dimensions $[512, 256, 128, 1]$ and ReLU activations. Training is performed using SGD as the optimizer, with the cross-attention layers remaining frozen during the first epoch. We select the learning rate through a grid search over $\{10^{-1}, 10^{-2}, 10^{-3}\}$ and explore batch sizes among $\{128, 256, 512, 1024\}$.

A.3. Baselines & Implementation

- **MCM [30]:** Maximum Concept Matching (MCM) estimates uncertainty in VLMs by leveraging the softmax probability distribution over all classes or captions. It selects the most likely caption for an image based on the highest probability score, providing a natural measure of confidence in the model's predictions. No additional training is required since MCM directly uses the model's output probabilities.
- **MCM + TS [15]:** This method extends MCM by applying Temperature Scaling (TS) to adjust the softmax probabilities better. TS optimizes the temperature parameter to refine the confidence scores, leading to more calibrated uncertainty estimates. The multiplicative temperature parameter is learned using the whole training dataset to minimize expected calibration error, using LBFGS optimizer.
- **Entropy** [38]: This method quantifies uncertainty in neural network predictions by calculating the Shannon entropy of the output probability distribution. High entropy indicates more significant uncertainty, as the model assigns similar probabilities across multiple classes, reflecting ambiguity in its prediction. On the other hand, low entropy signifies confidence, with the model favoring a specific class. Entropy-based uncertainty estimation does not require additional training.
- **DOCTOR** [14]: DOCTOR quantifies uncertainty by analyzing the confidence distribution of the model's predictions. It computes the Rényi entropy of order two, a measure based on the squared probabilities assigned to each class, emphasizing how concentrated or dispersed the probability mass is. A prediction with one dominant probability value will yield a low uncertainty score, while a more evenly spread distribution results in higher uncertainty. This method does not require additional training and operates directly on the model's softmax outputs.
- **Rel-U** [13]: Rel-U is a data-driven method that incorporates cross-label uncertainties directly in the logit space. Learning relationships between class logits provides a refined estimation of uncertainty beyond traditional confidence scores. Due to its reliance on a cross-label cost penalty matrix, Rel-U does not apply to image-text datasets where labels are absent. Rel-U's hyperparameters are fixed to $\lambda = 0.15$ and $T = 0.5$ greedily, since they provided the best performance.
- **Learning Visual Uncertainties (LVU) [8, 19, 47]:** LVU refers to a class of models designed to predict the loss of a visual backbone as a means to estimate potential errors. ConfidNet [8] established that accurately predicting uncertainty is equivalent to estimating the model's loss—if a model can predict the loss of its visual backbone, it inherently quantifies its error. Another approach, Pretrained Visual Uncertainties [19], follows a similar principle by learning to predict backbone loss, leveraging pretraining on ImageNet-21k.

  Implementation: To evaluate the LVU baseline, we use the same MLP architecture as our model but restrict the input to the visual token only. Additionally, following [8, 19], this baseline is trained with an MSE loss, in contrast to our method, which uses a BCE loss.
- **ProbVLM [42]:** ProbVLM introduces a probabilistic adapter that estimates probability distributions for embeddings of pre-trained VLMs. This is achieved through inter- and intra-modal alignment in a post-hoc manner. The goal is to capture the inherent ambiguity in embeddings, reflecting the fact that multiple samples can represent the same concept in the physical world. This method enhances the calibration of embedding uncertainties in retrieval tasks and benefits downstream applications like active learning and model selection.

  Implementation: ProbVLM models probability distributions over the embeddings of image and text modalities. However, it does not explicitly model the uncertainty in their interaction via cosine similarity. As a result, directly adapting the method for image classification is not straightforward. We attempted to include ProbVLM in our baseline comparison by using its proposed visual aleatoric uncertainty metric, but it resulted in nearly random failure prediction performance. Additionally, we explored using its cross-modal loss as an uncertainty logit, applying a softmax transformation, but this approach also proved ineffective. In contrast, BayesVLM addresses this limitation by modeling the uncertainty over the similarity computation, enabling a more principled approach to downstream tasks like image classification.

- **BayesVLM [2]:** BayesVLM is a training-free method for estimating predictive uncertainty. It employs a post-hoc approximation of the Bayesian posterior, allowing for analytic computation of uncertainty propagation through the VLM. By approximating the Bayesian posterior over model parameters, BayesVLM captures uncertainties inherent to the model itself (image and text encoders). These model uncertainties are then propagated through the VLM to produce uncertainty estimates for predictions. **Implementation:** To evaluate BayesVLM, we follow the implementation provided in its official Github repository https://github.com/AaltoML/BayesVLM

B. Additional details on ViLU

B.1. Bilinear interpretation of MCM

In Sec. 4.2 of the main paper, we mentioned that ViLU is a consistent generalization of MCM. More precisely, the uncertainty module g_θ can model the unnormalized MCM score by approximating the following bilinear form on $\boldsymbol{z}_{\text{ViLU}} = (\boldsymbol{z}_v, \boldsymbol{z}_{\hat{t}}, \boldsymbol{z}_t^{\alpha})$:

$$g_\theta(\boldsymbol{z}_{\text{ViLU}}) = \frac{1}{2}\, \boldsymbol{z}_{\text{ViLU}}^{T}\, A\, \boldsymbol{z}_{\text{ViLU}} = \boldsymbol{z}_v^{\top} \boldsymbol{z}_{\hat{t}}, \qquad (11)$$

with $$A = \begin{pmatrix} \mathbf{0} & \mathrm{I}_d & \mathbf{0} \\ \mathrm{I}_d & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{3d \times 3d}.$$

B.2. ViLU variant for generalization experiments

For the generalization experiments presented in the main paper (cross-dataset transfer) and the supplementary material (domain generalization and concept coverage), we used a slightly modified version of ViLU. Specifically, the MCM score was explicitly provided as an additional input to the uncertainty module g_θ, alongside the visual and textual embeddings. While the original design of ViLU allows g_θ to model this behavior implicitly through interactions between the modalities, we found that explicitly including the MCM score improves uncertainty generalization.

C. Additional experimental results

C.1. Impact of MLP depth on performance

The results in Fig. 7 show that ViLU is relatively robust to MLP depth variations, particularly on ImageNet, where performance remains stable across different configurations. Across all tested datasets, a depth of 4 layers consistently achieved strong results, suggesting that this architecture provides a good balance between expressiveness and generalization for failure prediction.

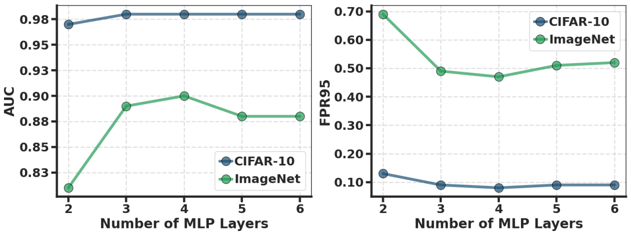

Figure 7. **Impact of MLP depth.** Performance of ViLU on ImageNet and CIFAR-10 for different MLP depths.

C.2. Robustness to image-text task complexity

We analyze in Tab. 6 how inference-time batch size affects failure detection performance for MCM, LVU, and ViLU on the CC12M dataset. As batch size increases, the number of candidate captions used during inference grows, introducing more semantic competition and making the task more complex. Despite this, ViLU consistently outperforms both MCM and LVU across all tested settings. Notably, even under very large batch sizes—16,384 and 32,768—ViLU maintains strong performance, with only moderate degradation in AUC and FPR95. These results confirm the robustness of our method to increased image-text ambiguity at test time.

	MCM [30]		LVU [8, 19, 47]		ViLU	
Batch Size	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓
128	92.7	15.6	75.2	76.2	**96.9**	**15.6**
512	90.1	54.6	74.8	76.5	**95.7**	**22.5**
1024	88.8	58.8	74.4	76.5	**95.2**	**25.2**
2048	87.5	61.5	74.3	76.8	**94.4**	**28.7**
4096	86.4	64.2	73.9	77.0	**93.6**	**31.8**
8192	85.3	65.3	73.6	77.0	**92.8**	**34.9**
16384	84.5	66.4	73.3	76.7	**91.9**	**37.8**
32768	83.7	67.0	73.2	76.6	**91.1**	**39.9**

Table 6. Numerical results corresponding to Fig. 4, showing the effect of inference batch size on failure detection for ViLU (on CC12M).

C.3. Reliability of misclassification detection

Fig. 8 illustrates the relationship between misclassification detection performance and the zero-shot accuracy of the vision-language model for each baseline. Each dot at a given x-coordinate represents the classification performance of different baselines on the same dataset. The results emphasize the superior reliability of the uncertainty estimates provided by our method, particularly in low zero-shot accuracy settings. Notably, the tendency curves indicate a strong correlation between model performance and uncertainty metrics for both MCM and BayesVLM. Specifically, as zero-shot accuracy decreases, these two methods exhibit the worst performance. This suggests that they are only reliable when the model's zero-shot accuracy is high—an unpredictable scenario in real-world settings, where ground-truth labels are unavailable.

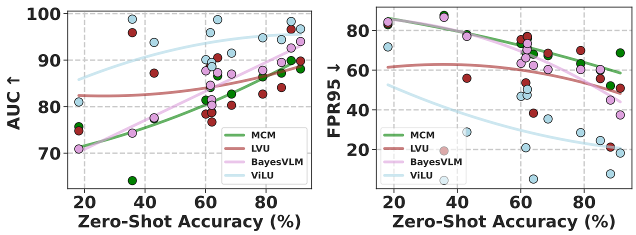

Figure 8. **Reliability of misclassification detection.** Our method, ViLU, enhances misclassification detection by providing more reliable uncertainty estimates, particularly when zero-shot accuracy is low.

C.4. Extension to different VLMs.

Tab. 7 presents the performance of ViLU when applied to different zero-shot vision-language backbones, including CLIP [36] and SigLIP [50], with both ViT-B and ViT-L variants. Across all settings, ViLU consistently outperforms MCM by a large margin in both AUC and FPR95, demonstrating strong and reliable failure detection. On CIFAR-10, the improvements are particularly pronounced: for example, using CLIP ViT-L/14, ViLU achieves an AUC of 99.0 compared to 93.6 for MCM, and reduces FPR95 from 31.5 to just 4.1. On ImageNet-1k, the gains remain substantial, with up to 30-point reductions in FPR95. Unlike LVU-based methods [8, 19, 47], which require access to the model's pre-training loss, ViLU is trained solely from classification correctness, making it applicable to a broad range of frozen or proprietary VLMs. Overall, the consistent results across architectures confirm that ViLU generalizes effectively with minimal assumptions.

	Backbone	Method	CIFAR-10		ImageNet-1k	
			AUC↑	FPR95↓	AUC↑	FPR95↓
CLIP [36]	ViT-B/16	MCM [30]	90.9	47.3	81.0	73.0
		ViLU	**98.4**	**8.0**	**90.3**	**44.2**
	ViT-L/14	MCM [30]	93.6	31.5	82.9	68.9
		ViLU	**99.0**	**4.1**	**91.2**	**39.2**
SigLIP [50]	ViT-B/16	MCM [30]	92.8	46.1	84.2	65.8
		ViLU	**97.6**	**13.3**	**90.7**	**44.4**
	ViT-L/16	MCM [30]	95.6	29.1	86.8	64.1
		ViLU	**98.4**	**7.8**	**91.3**	**41.4**

Table 7. **Generalization across backbones.** ViLU shows consistent performance gains on several VLMs compared to MCM.

C.5. Domain generalization on ImageNet variants

To evaluate ViLU's robustness under distribution shift, we consider a domain generalization setup in which ViLU is trained on the original ImageNet dataset and evaluated on two domain-shifted variants: ImageNet-V2 (IN-V2) and ImageNet-Sketch (IN-S). We first assess ViLU's uncertainty estimates in a zero-shot transfer setting, where the model is applied directly to each variant without any adaptation. As shown in Tab. 8, ViLU achieves competitive performance, notably outperforming LVU on IN-S (FPR95 of 73.1 vs. 86.6) and remaining close to MCM (70.9). On IN-V2, ViLU performs even better, reaching an FPR95 of 54.8 compared to 71.7 for MCM and 77.3 for LVU. These results confirm that ViLU retains reliable uncertainty estimates even when evaluated on unseen domains.

We then explore a few-shot adaptation scenario, where ViLU is fine-tuned using only five labeled images per class from the target domain (IN-V2 or IN-S). On IN-S, this minimal supervision significantly reduces FPR95 from 73.1 to 54.4, outperforming both MCM (70.9) and LVU (72.3). On IN-V2, ViLU achieves similarly strong improvements, lowering FPR95 from 54.8 to 52.7, once again surpassing MCM (71.7) and LVU (68.7). These results highlight ViLU's strong adaptability in low-data regimes and confirm that even minimal adaptation of the uncertainty head can lead to substantial gains in reliability under distribution shifts.

Dataset	**MCM**	**LVU** (*zero-shot*)	**ViLU** (*zero-shot*)	**LVU** (*5-shot*)	**ViLU** (*5-shot*)
IN-V2	71.7	77.3	**54.8**	68.7	**52.7**
IN-S	70.9	86.6	73.1	72.3	**54.4**

Table 8. FPR95↓ on ViLU's domain generalization from ImageNet to ImageNet-V2 and ImageNet-Sketch.

C.6. Impact of concept coverage in pre-training

In the main paper, we evaluated the zero-shot generalization ability of ViLU when pre-trained on CC12M and tested on 12 downstream datasets spanning various domains. In this section, we conduct a controlled experiment to assess whether better coverage of target concepts during pre-training improves zero-shot transfer. To this end, we construct a synthetic multi-dataset by combining the training sets of the 12 downstream datasets. Each image is paired with a pseudo-caption of the form “This is a photo of a ”, allowing us to train ViLU in the same image-caption setting as for CC12M. As shown in Tab. 9, this targeted pre-training leads to a substantial reduction in FPR95 across most datasets, with an average of 63.1 compared to 68.6 for the CC12M variant and 70.5 for MCM. These results confirm that more explicit coverage of the target classes during pre-training can significantly improve the quality of uncertainty estimates in zero-shot settings.

		CC12M	Multi-datasets
Dataset	**MCM**	**ViLU**	**ViLU**
CIFAR-10	52.1	54.2	**31.9**
CIFAR-100	67.3	59.9	**50.3**
Caltech101	68.7	**48.8**	70.8
Flowers102	68.0	67.4	**45.9**
OxfordPets	59.9	**58.1**	72.1
Food101	63.3	67.4	**36.2**
FGVCAircraft	82.9	82.3	**80.3**
EuroSAT	87.6	**85.7**	91.3
DTD	77.9	78.2	**75.2**
SUN397	75.9	**72.7**	81.0
StanfordCars	**73.4**	84.1	76.4
UCF101	68.9	63.8	**45.4**
Average	70.5	68.6	**63.1**

Table 9. FPR95↓ across datasets. Zero-shot performance when pre-trained on a Multi-datasets *vs*. CC12M.

C.7. Qualitative results

We provide additional visualizations on eight datasets in Fig. 9 and Fig. 10, illustrating the distribution of uncertainty scores for correctly and incorrectly classified validation samples. Our results demonstrate the consistency of ViLU in assigning high uncertainty scores to misclassified samples (red) and low uncertainty scores to correctly classified ones (blue).

Unlike visual uncertainty models such as ConfidNet [8], which rely solely on image features, our multimodal architecture leverages both visual and textual information to provide more reliable uncertainty estimates. Learning a cross-attention mechanism between image and text allows ViLU to better capture ambiguities in class definitions, leading to improved uncertainty calibration across diverse datasets.

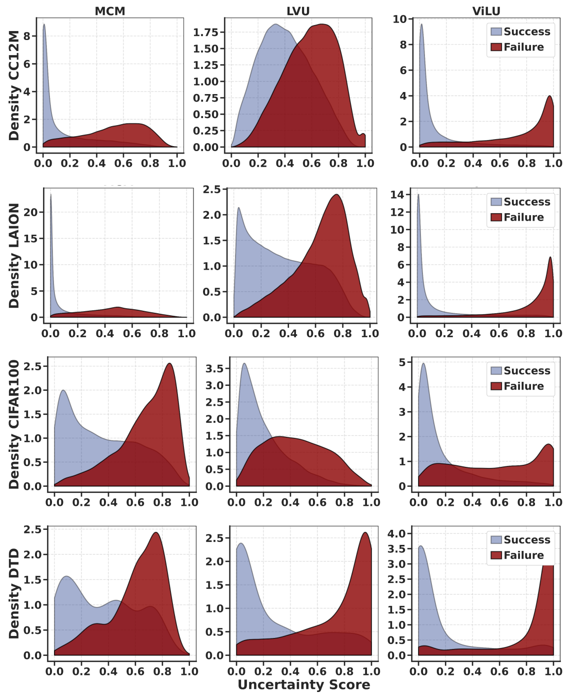

Figure 9. **Uncertainty score distribution.** Prediction for correctly and incorrectly classified samples on CC12M, LAION400M, CIFAR100 and DTD.

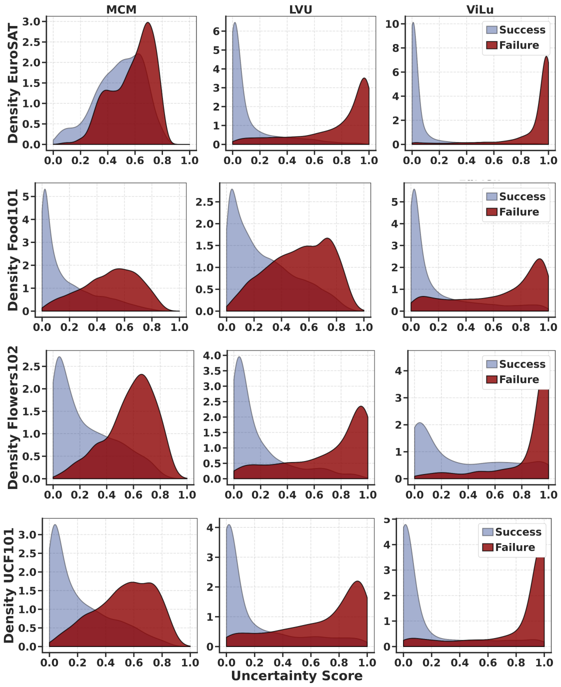

Figure 10. **Uncertainty score distribution.** Prediction for correctly and incorrectly classified samples on EuroSAT, Food101, Flowers102 and UCF101.