Over the Top-1: Uncertainty-Aware Cross-Modal Retrieval with CLIP

Lluis Gomez[1]

[1] Computer Vision Center, Universitat Autònoma de Barcelona.

Abstract

State-of-the-art vision-language models, such as CLIP, achieve remarkable performance in cross-modal retrieval tasks, yet estimating their predictive uncertainty remains an open challenge. While recent works have explored probabilistic embeddings to quantify retrieval uncertainty, these approaches often require model retraining or fine-tuning adapters, making them computationally expensive and dataset-dependent. In this work, we propose a training-free framework for uncertainty estimation in cross-modal retrieval. We start with a simple yet effective baseline that uses the cosine similarity between a query and its top-ranked match as an uncertainty measure. Building on this, we introduce a method that estimates uncertainty by analyzing the consistency of the top-1 retrieved item across samples drawn from the posterior predictive distribution using Monte Carlo Dropout (MCD) or Deep Ensembles. Finally, we propose an adversarial perturbation-based method, where the minimal perturbation required to alter the top-1 retrieval serves as a robust indicator of uncertainty. Empirical results in two standard cross-modal retrieval benchmarks demonstrate that our approach achieves superior calibration compared to learned probabilistic methods, all while incurring zero additional training cost.

1 INTRODUCTION

Cross-modal retrieval systems enable the retrieval of information across different modalities, such as using a text query to find matching images or vice versa. This capability has gained significant momentum in recent years, driven by the growing need to efficiently search and manage information in increasingly large multimodal databases. The ability to bridge the gap between different modalities has become a cornerstone of modern artificial intelligence applications, ranging from image search engines to multimodal content understanding.

State-of-the-art deep learning models for cross-modal retrieval, such as CLIP (Contrastive Language-Image Pre-training) (Radford et al. [2021]), ALIGN (Li et al. [2021]), Flamingo (Alayrac et al. [2022]), and BLIP (Li et al. [2023]), have demonstrated remarkable performance in tasks like image retrieval, text-to-image synthesis, and multimodal reasoning. These models are typically trained by maximizing the agreement between image and text representations, leveraging large-scale datasets to learn rich, semantically aligned embeddings. Their impressive performance on benchmark datasets, such as MSCOCO (Lin et al. [2014]) and Flickr30K (Young et al. [2014]), underscores their effectiveness in bridging visual and textual information.

However, a critical aspect often overlooked in the evaluation of cross-modal retrieval models is the uncertainty associated with their predictions. Standard evaluation metrics, such as Recall@K, focus solely on the accuracy of the retrieval results, providing no insight into how confident the model is in its predictions. In real-world applications, where retrieval errors can have significant consequences – such as in medical imaging, autonomous systems, or content moderation – understanding the model's uncertainty is as important as its accuracy.

Reliable measures of predictive uncertainty are essential for distinguishing between confident, trustworthy predictions and those where the model may be uncertain or even erroneous. Quantifying uncertainty in cross-modal retrieval is particularly challenging due to the complex interactions between modalities, where ambiguity can arise not only from the model but also from the data, the task, and the inherent variability in semantic alignment across different modalities.

In this work, we focus on *epistemic uncertainty* – lack of knowledge about the correct mapping from inputs to outputs.

In Bayesian deep learning, epistemic uncertainty is captured by placing a distribution over model weights and observing the spread in predictions this induces. The variance of predictions across posterior samples directly quantifies the model's epistemic uncertainty. Monte Carlo Dropout (Gal and Ghahramani [2016]) and Deep Ensembles (Lakshminarayanan et al. [2017]) are practical posterior-sampling techniques that make prediction variability observable, providing a Bayesian estimate of model uncertainty and helping to mitigate the overconfidence of a single deterministic network. Intuitively, if the model's belief about the best prediction is unstable across different sampled weight configurations, it signals a lack of knowledge. In classification tasks, epistemic uncertainty can be measured by the consistency of model's predictions: high posterior confidence implies the model returns the same top class across nearly all samples, whereas variation in predictions indicates significant epistemic uncertainty.

We argue that the same principles apply to retrieval and ranking tasks, including cross-modal retrieval. A Monte Carlo Dropout (MCD) or Deep Ensemble ranker outputs a predictive relevance distribution rather than a single score, and the dispersion of this predictive distribution corresponds to the model's uncertainty in the ranking. By leveraging posterior sampling (via dropout or ensembles) and tracking changes in the top retrieval result, we obtain a principled and quantifiable indicator of epistemic uncertainty in retrieval outcomes, consistent with established definitions of model uncertainty in classification and regression tasks.

In this paper, we propose several training-free approaches to quantify uncertainty in cross-modal retrieval, with a focus on pre-trained CLIP models. We explore simple yet effective baselines, such as cosine similarity-based uncertainty scores, as well as more sophisticated methods for predictive uncertainty estimation, including Monte Carlo Dropout (MCD), Deep Ensembles, and adversarial perturbation-based uncertainty estimation. Our evaluation, conducted on standard datasets (MSCOCO and Flickr30K), demonstrates that the proposed uncertainty measures not only correlate well with retrieval performance but also help to identify unreliable rankings, improve retrieval robustness, and enhance the overall trustworthiness of cross-modal retrieval systems.

2 RELATED WORK

Uncertainty estimation has been extensively studied in unimodal tasks such as image classification, natural language processing, and time series forecasting. Methods in Bayesian Neural Networks (Blundell et al. [2015], Neal [2012]), including variational inference and Hamiltonian Monte Carlo, provide principled approaches to estimate predictive uncertainty but are often computationally expensive, limiting their scalability to large models. To address these limitations, more scalable techniques have been developed, such as Monte Carlo Dropout (Gal and Ghahramani [2016]) and Deep Ensembles (Lakshminarayanan et al. [2017]), which approximate Bayesian inference through stochastic regularization and model diversity, respectively. Additionally, post-hoc calibration methods like temperature scaling (Guo et al. [2017]) have been proposed to adjust confidence estimates without modifying the underlying model.

In the multimodal domain, uncertainty estimation remains less explored. Probabilistic embedding methods (Chun et al. [2021], Li et al. [2022], Neculai et al. [2022], Ji et al. [2023]) model cross-modal retrieval as a probabilistic matching task, learning uncertainty-aware representations via probabilistic contrastive losses. However, these methods often require retraining models from scratch, which limits their scalability to large pre-trained vision-language models (VLMs).

To reduce computational overhead, adapter-based approaches (Chun, Upadhyay et al. [2023]) have been proposed. For example, ProbVLM (Upadhyay et al. [2023]) introduces a probabilistic adapter trained post hoc to estimate uncertainty distributions from frozen VLM embeddings. While ProbVLM achieves strong calibration without modifying the base model, it still relies on additional training and dataset-specific fine-tuning.

In contrast, our work focuses on training-free uncertainty estimation methods for cross-modal retrieval. We systematically investigate approaches such as top-1 similarity-based uncertainty, Monte Carlo Dropout, Deep Ensembles, and adversarial perturbation-based techniques, all of which can be directly applied to pre-trained models like CLIP without additional fine-tuning. Our goal is to provide practical, computationally efficient predictive uncertainty estimates that improve retrieval robustness and help identify unreliable predictions in cross-modal retrieval.

3 UNCERTAINTY ESTIMATION IN CROSS-MODAL RETRIEVAL MODELS

In this section, we introduce our framework for estimating uncertainty in cross-modal retrieval models. We first discuss a simple baseline using similarity scores as an uncertainty measure before exploring probabilistic techniques such as Monte Carlo Dropout and Deep Ensembles. Finally, we propose an adversarial perturbation-based approach that quantifies uncertainty based on retrieval robustness.

3.1 BACKGROUND

State-of-the-art cross-modal retrieval models such as CLIP learn embedding functions ϕ_q for text queries and ϕ_I for images (and vice versa). These functions project their respective inputs into a shared embedding space, aiming to position the representation of a text query $\phi_q(q)$ and an image $\phi_I(I)$ closely together if the image I is relevant to

the query q. The similarity between embeddings, quantified using a similarity metric $\mathrm{sim}(\phi_q(q), \phi_I(I))$, guides the retrieval of relevant images in response to a given text query. Let $R(q, \mathcal{I})$ denote the retrieval ranking for a query q obtained as:

$$R(q, \mathcal{I}) = \mathrm{argsort}_{I \in \mathcal{I}}[\mathrm{sim}(\phi_q(q), \phi_I(I))] \tag{1}$$

where $\mathrm{argsort}$ sorts images in in the retrieval set $\mathcal{I}$ in descending order by similarity score.

The standard evaluation metric for cross-modal retrieval models is *Recall at Rank K* (R@K), which measures the proportion of queries where a relevant item appears in the top-K results. This metric is particularly preferred in datasets such as MSCOCO and Flickr30k, where each text query has only a single relevant image, and each image query has only five relevant captions. This sparsity in ground-truth relevance makes other retrieval metrics like Mean Average Precision (mAP) less suitable, as they assume multiple relevant items per query. R@K is also more suitable for practical retrieval tasks where users primarily interact with top-ranked results.

In this scenario, a natural approach for estimating retrieval confidence is to use the distance to the top-1 retrieved item as a proxy for the uncertainty for a given query ranking; similar to using the max value of a softmax as a proxy for predictive uncertainty of classification models (Guo et al. [2017]). In metric spaces, confidence and uncertainty are inherently linked to the density of relevant items. In high-confidence cases, queries should be embedded close to their correct match, yielding high similarity scores, whereas ambiguous queries exhibit lower similarity due to embedding uncertainty. Therefore, we define a simple confidence measure as:

$$C(q) = \mathrm{sim}(\phi_q(q), \phi_I(I^*)), \tag{2}$$

where $I^* = \mathrm{argmax}_{I \in \mathcal{I}}[\mathrm{sim}(\phi_q(q), \phi_I(I))]$ is the top-1 retrieved image for query q. In CLIP, the similarity function sim is cosine similarity, ensuring that the confidence score $C(q)$ is bounded in the range $[0, 1]$. This bounded range makes it an interpretable and normalized proxy for confidence estimation. In the experimental section, we will analyze how this simple confidence measure demonstrates strong calibration with retrieval performance (in terms of R@K), establishing it as an effective baseline for uncertainty estimation ($U(q) = 1 - C(q)$) in cross-modal retrieval.

3.2 MONTE CARLO DROPOUT

Monte Carlo Dropout (Gal and Ghahramani [2016]) provides an approximation to Bayesian inference in deep neural networks by enabling dropout (Srivastava et al. [2014]) at inference time, effectively sampling from the approximate posterior distribution. More formally, given a neural network with weights W, we introduce stochasticity through a dropout mask $z \sim \mathrm{Bernoulli}(p)$ applied independently to each layer during each forward pass:

$$y^*(x, W, z) = f(x; W, z), \tag{3}$$

where $y^*(x, W, z)$ is the output given input x. The Bayesian posterior predictive mean is approximated using M stochastic forward passes:

$$\mathbb{E}_{\hat{p}(y^*|x^*)}[y^*] \approx \frac{1}{M} \sum_{m=1}^{M} y_m^*, \tag{4}$$

where $y_m^* = f(x^*; W, z_m)$ is the output from the m-th stochastic forward pass. Similarly, the predictive variance is given by:

$$\begin{aligned} \mathrm{Var}_{\hat{p}(y^*|x^*)}(y^*) \approx & \tau^{-1} I + \frac{1}{M} \sum_{m=1}^{M} y_m^* y_m^{*T} \\ & - \mathbb{E}_{\hat{p}(y^*|x^*)}[y^*] \mathbb{E}_{\hat{p}(y^*|x^*)}[y^*]^T \end{aligned} \tag{5}$$

where $\tau^{-1} I_D$ represents the observation noise variance, accounting for aleatoric uncertainty; the second term captures epistemic uncertainty by averaging the variance of multiple stochastic forward passes; while the final term ensures proper centering of the variance estimate around the predictive mean. This formulation enables uncertainty estimation by leveraging variability across multiple stochastic forward passes.

Gal and Ghahramani [2016] demonstrated the effectiveness of Monte Carlo Dropout (MCD) for regression and classification tasks, showing that dropout can serve as an efficient Bayesian approximation. Although MCD has been widely applied in unimodal settings such as image classification (Gustafsson et al. [2020]) and natural language processing (Xiao and Wang [2019]), its application to cross-modal retrieval remains largely unexplored.

A key challenge in applying MCD to retrieval models is that uncertainty estimation in retrieval is fundamentally different from both classification and regression tasks. In classification, uncertainty is estimated over discrete class probabilities, while in regression, it is captured by the variance of scalar outputs. However, in retrieval models, outputs are rankings derived from distances in a high-dimensional embedding space. In this context, the embedding functions ϕ_q and ϕ_I can be seen as high-dimensional regressors, mapping input text queries and images (and vice versa) into a shared space where semantic similarity is measured.

Unlike traditional regression tasks, where uncertainty is directly estimated on a continuous output variable, retrieval

uncertainty must be inferred from the variability in ranked similarity scores across stochastic forward passes. Therefore, applying MCD in retrieval requires analyzing the variance of retrieval rankings rather than direct output distributions.

Given a retrieval query q and image gallery $\mathcal{I}$, we obtain M stochastic forward passes of the embedding functions ϕ_q and ϕ_I, resulting in a set of retrieval rankings $\{R^m(q,\mathcal{I})\}_{m=1}^M$. From a Bayesian perspective, these retrieval rankings represent samples from the posterior distribution over rankings, induced by the model's uncertainty in embedding representations under dropout.

To quantify retrieval uncertainty, we propose to measure the consistency of the top-1 retrieval outcome across posterior samples:

$$U_{\text{MCD}}(q) = 1 - \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}[R^m(q,\mathcal{I})_1 = R^*(q,\mathcal{I})_1], \quad (6)$$

where $R^*(q,\mathcal{I})_1$ is the most frequently retrieved top-1 item across all Monte Carlo samples. This formulation reflects epistemic uncertainty, as greater variability in top-1 retrievals suggests higher model uncertainty in ranking stability.

Intuitively, if the same top-1 item appears consistently across stochastic passes ($U_{\text{MCD}}(q) \approx 0$), the model is confident in its retrieval decision. Conversely, if the retrieved top-1 item varies significantly across posterior samples ($U_{\text{MCD}}(q) \approx 1$), the model exhibits high epistemic uncertainty, signaling potential ambiguity in the ranking.

In Appendix A we analyze the sensitivity of MCD uncertainty estimation to key hyperparameters: dropout rate and the number of samples. Our experiments indicate robustness across typical dropout values, with 0.2 providing optimal calibration performance. Moreover, increasing the number of samples improves stability in uncertainty estimates, with 20 samples offering a good trade-off between computational efficiency and performance, although our default choice of 50 ensures more robust results.

3.3 DEEP ENSEMBLES

Deep Ensembles Lakshminarayanan et al. [2017] provide a robust approach for predictive uncertainty estimation by training multiple independent neural networks with different random initializations. Although originally introduced as a non-Bayesian technique, Deep Ensembles have been shown to approximate Bayesian inference Hoffmann and Elster [2021], where each model in the ensemble represents a sample from a multimodal posterior distribution over the model parameters.

Formally, consider an ensemble of K independently trained retrieval models $\{\mathcal{M}_k\}_{k=1}^K$, each parameterized by weights θ_k. The posterior predictive distribution for a new input x^* is approximated as a uniformly-weighted mixture:

$$\hat{p}(y^*|x^*) = \frac{1}{K}\sum_{k=1}^{K} p_{\theta_k}(y^*|x^*,\theta_k), \quad (7)$$

where $p_{\theta_k}(y^*|x^*,\theta_k)$ is the predictive distribution of the k-th model. This formulation aligns to Bayesian model averaging, where the ensemble acts as an approximation to the true posterior by representing it as a mixture of delta functions centered at the maximum a posteriori (MAP) estimates of each model's parameters.

For regression tasks, this mixture can be approximated by a Gaussian distribution, with the posterior predictive mean and variance given by:

$$\mathbb{E}_{\hat{p}(y^*|x^*)}[y^*] \approx \frac{1}{K}\sum_{k=1}^{K} \mu_{\theta_k}(x^*) \quad (8)$$

$$\begin{aligned}\text{Var}_{\hat{p}(y^*|x^*)}(y^*) \approx \frac{1}{K}\sum_{k=1}^{K} \left(\sigma^2_{\theta_k}(x^*) + \mu^2_{\theta_k}(x^*)\right) \\ - \left(\mathbb{E}_{\hat{p}(y^*|x^*)}[y^*]\right)^2\end{aligned} \quad (9)$$

where $\mu_{\theta_k}(x^*)$ and $\sigma^2_{\theta_k}(x^*)$ represent the mean and variance predicted by the k-th model.

Applying Deep Ensembles to cross-modal retrieval introduces challenges similar to those encountered with Monte Carlo Dropout. Specifically, uncertainty must be inferred from variability in retrieval rankings rather than scalar outputs or probability distributions over classes. Given a query q and an image gallery $\mathcal{I}$, we obtain K retrieval rankings $\{R^k(q,\mathcal{I})\}_{k=1}^K$ from each ensemble member.

To quantify retrieval uncertainty, we propose measuring the consistency of the top-1 retrieval across ensemble members:

$$U_{\text{Ens}}(q) = 1 - \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}[R^k(q,\mathcal{I})_1 = R^*(q,\mathcal{I})_1], \quad (10)$$

where $R^*(q,\mathcal{I})_1$ is the most frequently retrieved top-1 item across all ensemble models. This metric captures epistemic uncertainty, as greater variability among ensemble predictions indicates less confidence in the retrieval outcome.

3.4 ADVERSARIAL PERTURBATIONS FOR UNCERTAINTY ESTIMATION

Building on the confidence scores based on top-1 distance (our baseline from section 3.1) and top-1 consistency (sections 3.2 and 3.3), we propose an uncertainty estimation framework based on adversarial perturbations. The core idea is that robustness to small perturbations in the embedding

space can serve as an indicator of model uncertainty: confident rankings should remain stable under minor changes of the query embedding, while uncertain predictions are more susceptible to fluctuations.

Formally, given an input query q and its corresponding embedding $\phi_q(q)$, an adversarial perturbation δ is defined as the minimal perturbation required to alter the model's output, in our case, the top-1 retrieved item. This can be expressed as the following optimization problem:

$$\delta^* = \min\{\delta \mid R(\phi_q(q) + \delta, \mathcal{I})_1 \neq R(\phi_q(q), \mathcal{I})_1\} \quad (11)$$

This formulation seeks the smallest perturbation δ^* that changes the top-1 retrieval result. Eq. 11 is solved via Projected Gradient Descent (PGD) Madry et al. [2018]:

$$\phi_q(q)^{(t+1)} = \phi_q(q)^{(t)} - \eta \frac{\nabla_q L}{\|\nabla_q L\|_2}, \quad (12)$$

where η is the step size, and L is the difference between the top-1 similarity and the highest-ranked competitor. The final perturbation norm $\|\delta^*\|_2$ serves as a proxy for the model's confidence:

$$C_{\text{adv}}(q) = \tanh(\|\delta^*\|_2), \quad (13)$$

where δ^* is the minimal perturbation required to flip the top-1 retrieval, and the $\tanh$ function maps the perturbation norm to a bounded confidence score in $[0, 1]$. The L2 norm offers a practical and interpretable proxy for retrieval robustness, as it directly quantifies how far the query embedding must be displaced to alter the retrieval outcome.

Notice that we solve the optimization in Eq. 11 using PGD in CLIP's embedding space. Specifically, we apply small, normalized gradient steps (with a fixed step size) until the top-1 retrieval result changes or a maximum number of iterations is reached. In this setting, the minimal query embedding perturbation required to alter the top-1 retrieval corresponds to the distance to the nearest decision boundary in embedding space – that is, the set of points where another candidate becomes more similar than the current top-1 item. Since the cosine similarity function is 1-Lipschitz continuous on the unit sphere, the magnitude of the required perturbation provides a meaningful proxy for retrieval robustness: larger perturbations imply greater distance to the decision boundary, and thus higher model confidence; smaller perturbations indicate proximity to ambiguity regions where the ranking is unstable. This perspective aligns with traditional margin-based uncertainty estimation in classification tasks (e.g., SVMs), where distance to the decision boundary serves as an uncertainty measure.

In our experiments, we also consider a linear approximation of δ^* that directly estimates the minimal perturbation required to flip the ranking:

$$\delta^* \approx \frac{\text{sim}(q, I_1) - \text{sim}(q, I_2)}{\|\nabla_q(\text{sim}(q, I_1) - \text{sim}(q, I_2))\|_2}. \quad (14)$$

where I_1 and I_2 are the top-1 and top-2 retrieved items.

In our experiments we apply the adversarial perturbation methods to both text-to-image and image-to-text retrieval tasks. In all cases, the perturbation is applied only to the query embedding (either text or image), while the gallery embeddings remain fixed.

4 EXPERIMENTS

In this section, we present a comprehensive evaluation of our proposed uncertainty estimation framework for cross-modal retrieval. We begin by describing the datasets and evaluation metrics used in our experiments. This is followed by an in-depth analysis of top-1-based uncertainty estimation techniques, including a comparison of our approach with state-of-the-art probabilistic embeddings to highlight its effectiveness in terms of calibration and efficiency.

4.1 DATASETS AND METRICS

We evaluate our methods on two standard benchmarks for cross-modal retrieval, enabling reproducibility and comparability with prior work: MSCOCO Lin et al. [2014] and Flickr30K Young et al. [2014].

Flickr30K contains 31,783 images, each paired with five descriptive captions. We follow the standard splits commonly used in cross-modal retrieval benchmarks, such as the CLIP benchmark, with 29,000 images for training, 1,000 for validation, and 1,000 for testing.

MSCOCO-Captions comprises over 123,000 images, each associated with five captions. We adopt the standard 2014 version with 82,783 images for training and 40,504 images for validation/testing. For fair comparison, we follow the established 5K test split protocol, which is widely used in standard becnchmarks.

To assess the quality of our uncertainty estimates, we employ a combination of calibration plots, correlation measures, and rejection curves. Calibration Plots (Reliability Diagrams) visualize the relationship between predicted uncertainty scores and actual retrieval performance (measured by Recall@k). Ideally, well-calibrated models should have points lying close to the diagonal, indicating that the predicted confidence aligns with empirical performance.

Following Upadhyay et al. [2023], we define uncertainty levels by partitioning the dataset based on predicted uncertainty scores. We then compute the Spearman rank correlation (S) to measure the monotonic relationship between uncertainty

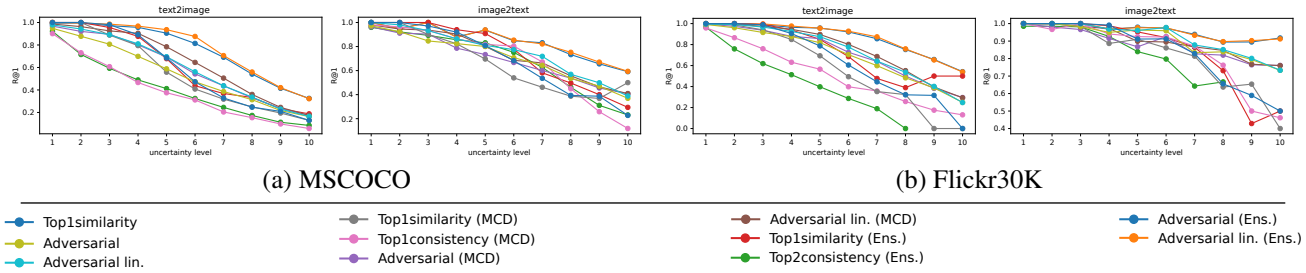


Figure 1: Calibration plots for all considered uncertainty estimation methods on MSCOCO (a) and Flickr30K (b).

levels and Recall@1. A perfectly calibrated model would exhibit a correlation of -1, indicating that performance decreases monotonically with increasing uncertainty.

On the other hand, the R^2 score evaluates how well a linear regression model fits the relationship between uncertainty levels and Recall@1. A higher R^2 indicates a stronger linear trend. We also provide a unified metric ($-SR^2$) which combines both scores to provide a single calibration measure. An ideal model would achieve a score of 1.0, reflecting perfect monotonicity and linearity in the relationship between uncertainty and retrieval performance.

4.2 IMPLEMENTATION DETAILS

For all experiments, we use the ViT-L/14 architecture with original pretrained weights from Radford et al. [2021]. Appendix B provides additional experiments comparing ViT-L/14 and the smaller ViT-B/32 variant.

For the experiments that involve Monte Carlo Dropout (MCD), we perform 50 stochastic forward passes with 0.2 dropout rate during inference to approximate the predictive posterior distribution, which is common practice in the MCD literature to achieve stable uncertainty estimates.

For Deep Ensembles, we construct an ensemble of 12 independently trained ViT-L/14 models sourced from the OpenCLIP repository (Ilharco et al. [2021]). These models are trained on diverse datasets, including OpenAI, LAION, DataComp, MetaCLIP, and DFN. This ensemble configuration enables the capture of diverse model behaviors, contributing to more robust uncertainty estimates through the aggregation of outputs from models with varying inductive biases.

For the Adversarial Perturbation-based uncertainty estimation, we set the perturbation hyperparameters after empirical tuning to a step size of 0.025 and a maximum of 50 iterations, striking a balance between computational efficiency and the effectiveness of the perturbations in revealing model uncertainty. We empirically determined approximately optimal values for these parameters using a hold-out dataset (MSCOCO validation).

4.3 UNCERTAINTY CALIBRATION

Figure 1 presents calibration plots for all considered methods, while Table 1 provides a quantitative assessment of their calibration in terms of the Spearman rank correlation (S) and R^2 scores.

To compute the uncertainty levels used in our analysis, we first define bins based on the range of values produced by each uncertainty measure. Specifically, for each method, we identify the minimum and maximum uncertainty scores and divide this range into 10 equally spaced bins, representing different levels of uncertainty.

Each query is then assigned to one of these bins based on its corresponding uncertainty score. Within each bin, we compute the retrieval performance in terms of Recall@1 (R@1), which reflects the proportion of queries where the correct item is retrieved at the top rank. This binning process allows us to evaluate how well the model's predicted uncertainty aligns with its actual retrieval accuracy, providing insights into the calibration of the uncertainty estimates.

In a well-calibrated model, we expect a monotonic decrease in R@1 performance as the uncertainty level increases—indicating that higher uncertainty corresponds to lower retrieval accuracy. This trend is clearly observed in Figure 1, where R@1 consistently declines across increasing uncertainty levels for most methods, demonstrating effective calibration of the uncertainty estimates.

The results in Table 1 demonstrate all proposed top1-based methods exhibit exceptional calibration performance, as seen in their consistently low Spearman Rank Correlation (S) and high R^2 and -SR^2 scores. These methods directly address uncertainty in retrieval rankings, making them particularly effective for the task at hand.

The *Top-1similarity* baseline achieves near-perfect calibration across both image-to-text and text-to-image retrieval tasks. Its simplicity – using the cosine similarity between the query and the top-1 retrieved item as a confidence score – proves highly effective, yielding a $-SR^2 = 0.95$ for image-to-text and text-to-image retrieval in the MSCOCO dataset. The method based on Adversarial Perturbations on top of the ranking provided by the deterministic model ("*Adversarial*"

	MSCOCO						Flickr30K					
	image2text			text2image			image2text			text2image		
	S	R^2	-SR^2	S	R^2	-SR^2	S	R^2	-SR^2	S	R^2	-SR^2
Upadhyay et al. [2023]	-0.99	0.93	0.93	-0.30	0.35	0.10	-0.70	0.71	0.49	0.70	0.50	0.35
Top1similarity	-1.00	0.95	0.95	-1.00	0.95	0.95	-0.98	0.86	0.84	-1.00	0.94	0.94
Adversarial	-1.00	0.97	0.97	-1.00	0.99	0.99	-0.95	0.87	0.83	-1.00	0.96	0.96
Adversarial lin.	-1.00	0.96	0.96	-1.00	0.98	0.98	-0.95	0.85	0.81	-1.00	0.92	0.92
Top1similarity (MCD)	-0.92	0.88	0.82	-1.00	0.96	0.96	-0.97	0.85	0.83	-1.00	0.96	0.96
Top1consistency (MCD)	-1.00	0.88	0.88	-1.00	0.96	0.96	-0.98	0.77	0.75	-1.00	0.99	0.99
Adversarial (MCD)	-1.00	0.99	0.99	-1.00	0.98	0.98	-0.96	0.93	0.89	-1.00	0.95	0.95
Adversarial lin. (MCD)	-1.00	0.97	0.97	-1.00	0.95	0.95	-1.00	0.95	0.95	-1.00	0.91	0.91
Top1similarity (Ens.)	-0.99	0.92	0.91	-1.00	0.95	0.95	-0.98	0.77	0.75	-0.90	0.85	0.77
Top1consistency (Ens.)	-1.00	0.91	0.91	-1.00	0.96	0.96	-0.98	0.90	0.88	-1.00	0.99	0.99
Adversarial (Ens.)	-0.97	0.89	0.86	-0.99	0.90	0.89	-0.90	0.83	0.75	-1.00	0.84	0.84

Table 1: Uncertainty calibration metrics for all considered methods. The calibration results of ProbVLM (Upadhyay et al. [2023]) are included for reference, though they are not directly comparable to the other methods (see main text for detailed analysis). Note that the ProbVLM results on Flickr30K are based on models trained on MSCOCO in a cross-dataset scenario.

in the table) slightly outperforms the baseline method.

The best uncertainty estimation in terms of average -SR² is the method based on Adversarial Perturbations on top of the MCD ranking – "*Adversarial (MCD)*" in the table. We appreciate that using Monte Carlo Dropout (MCD) or Deep Ensemble (Ens.) improve in some tasks/datasets. Although there is no clear winner overall in terms of calibration, the analysis in section 4.4 offers a distinct analysis that reveals clear differences among methods.

Comparison with ProbVLM

ProbVLM (Upadhyay et al. [2023]), while more sophisticated and capable of converting deterministic embeddings into probabilistic ones, demonstrates weaker calibration performance. It is important to highlight that ProbVLM and the rest of considered methods tackle different problems, and thus, their performance metrics are not directly comparable in every aspect.

ProbVLM introduces a probabilistic adapter over pre-trained Vision-Language Models (VLMs) like CLIP, converting their deterministic outputs into probability distributions. However, the calibration results indicate that its uncertainty estimates do not align as closely with retrieval performance as those of the proposed top1-based methods.

The probabilistic approach in ProbVLM is more flexible, enabling the model to capture uncertainties in multi-modal data and supporting advanced downstream tasks like model selection and active learning, which simpler methods cannot do. However, this increased complexity comes at the cost of calibration in retrieval tasks, as shown by its lower -SR² scores compared to the simpler methods proposed.

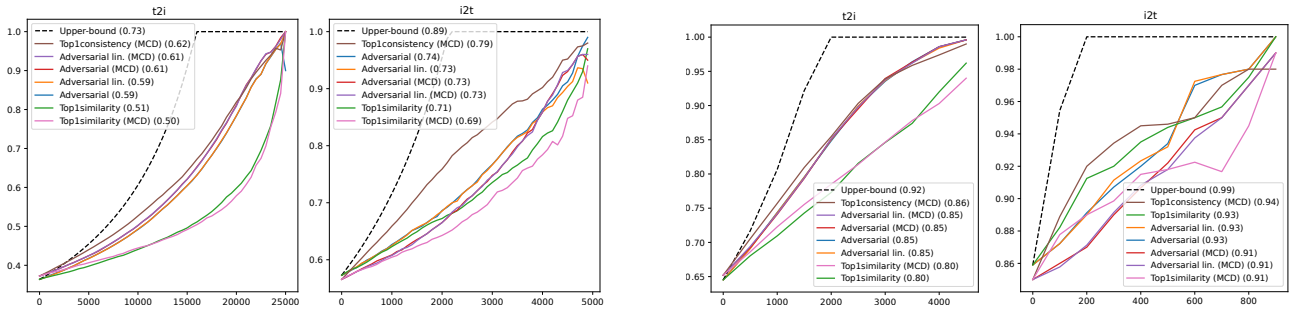
Moreover, ProbVLM relies on training data for cross-modal alignment, making it more computationally expensive and data-dependent. As an example, notice the lower results on Flickr30K in Table 1 for ProbVLM trained on MSCOCO – i.e. in a cross-dataset scenario. In contrast, the proposed methods show that variability/similarity in top-1 retrieval results provides an excellent indicator of retrieval uncertainty, leading to high-quality uncertainty calibration, in a data-agnostic manner.

4.4 REJECTION PLOTS

We complement calibration analysis with rejection plots which show how retrieval performance improves as increasingly uncertain samples are rejected. This helps visualize the utility of uncertainty estimates in practical scenarios, where unreliable predictions may be filtered out to enhance system robustness. Figure 2 presents rejection plots for all considered uncertainty estimation methods.

To implement these plots, we first sort all queries in descending order of uncertainty, starting from the most uncertain to the least uncertain. For each uncertainty estimation method, we progressively remove the most uncertain queries in batches and recalculate the retrieval performance after each removal. This process allows us to observe how performance metrics evolve as the most uncertain samples are systematically excluded.

For text-to-image (t2i) retrieval, where we have a total of 25,000 and 5,000 queries in MSCOCO and Flickr30K respectively, we remove 500 text queries at each step. In the



(a) MSCOCO (b) Flickr30K

Figure 2: Rejection plots for all considered uncertainty estimation methods on MSCOCO (a) and Flickr30K (b). The x-axis represents the number of rejected queries, while the y-axis shows Recall@1. The Area Under the Curve (AUC) for each method is indicated in brackets next to the method names in the figure legends, facilitating direct comparison across methods.

case of image-to-text (i2t) retrieval, we remove 100 image queries per step due to the smaller query set (5,000 and 1,000 respectively). After each batch removal, we compute Recall@1 (R@1) for the remaining queries to track performance changes as increasingly uncertain samples are filtered out.

A well-calibrated uncertainty estimation method should show a monotonic improvement in R@1 as the most uncertain queries are removed. This is because the retained queries are those for which the model is more confident, leading to higher retrieval accuracy. The upper-bound curve in the plots represents the theoretical maximum performance achievable if the most challenging queries were perfectly identified and removed.

In addition to visualizing the rejection curves, we quantify performance by computing the area under the curve (AUC) for each method. The AUC is calculated using the trapezoidal rule, which approximates the region under the curve as a series of trapezoids. The area of each trapezoid is computed based on the retrieval performance at consecutive points along the rejection curve. Mathematically, this is expressed as:

$$\int_a^b f(x)\,dx \approx (b-a)\cdot\tfrac{1}{2}(f(a)+f(b)). \qquad (15)$$

where $f(x)$ represents the retrieval performance (R@1), and $[a, b]$ are the boundaries of each interval corresponding to the rejection steps. We normalize the number of rejected samples such that the maximum possible area under the curve equals 1. This method provides an efficient and accurate approximation of the overall performance across the entire range of rejected samples.

As shown in Figure 2, most methods exhibit a clear upward trend, confirming that their uncertainty estimates effectively identify low-confidence predictions. The computed AUC values – shown in brackets after the methods' names in the figure legends – reflect the overall performance improvement, with higher AUC indicating better utilization of uncertainty estimates. This trend is particularly evident in both MSCOCO and Flickr30K, where performance approaches the upper bound as a large fraction of uncertain queries is rejected, highlighting the effectiveness of the uncertainty estimates in improving retrieval robustness.

Interestingly, unlike the results observed in the calibration analysis, where all proposed methods performed equally well and showed similar trends, the rejection plots reveal a clear distinction in performance across the different methods. Specifically, methods based on *Top-1consistency* (across Monte Carlo Dropout samples) and adversarial perturbations consistently outperform the top-1 similarity baselines. This indicates that while simple similarity-based measures can provide good overall calibration, more sophisticated approaches like MCD-based consistency and adversarial robustness capture deeper aspects of model uncertainty that translate into better real-world performance when uncertain samples are filtered out.

This divergence in findings between calibration and rejection analyses can be attributed to the different aspects of uncertainty each evaluation metric emphasizes. Calibration analysis primarily assesses how well the model's predicted uncertainty scores align with actual performance, focusing on the global relationship between uncertainty and accuracy across all samples. In contrast, rejection analysis places greater emphasis on the relative ranking of uncertainty estimates – it evaluates how effectively the model can prioritize uncertain samples for rejection to maximize performance gains.

While top-1 similarity may provide well-calibrated scores on average, it may lack the fine-grained sensitivity needed to distinguish between subtle differences in uncertainty among hard queries. On the other hand, top-1 consistency (MCD) and adversarial perturbation methods are designed to capture model stability and robustness under perturbations, which are more directly linked to the model's uncertainty in specific decisions. These methods excel in identifying truly

uncertain queries, leading to superior performance in rejection scenarios.

5 CONCLUSION

In this work, we have presented a comprehensive framework for uncertainty estimation in cross-modal retrieval models, exploring different techniques to quantify retrieval confidence. We introduced a range of methods, starting from straightforward top-1 similarity-based measures, progressing through probabilistic approaches like Monte Carlo Dropout (MCD) and Deep Ensembles, and culminating in an adversarial perturbation-based method that assesses uncertainty through retrieval robustness.

Our calibration analysis demonstrated that all proposed methods achieve exceptional calibration performance, with top-1 similarity-based approaches providing strong baseline results. Notably, methods incorporating MCD and adversarial perturbations slightly outperformed the baseline in certain settings, although the differences were not pronounced. This suggests that simple confidence measures, such as cosine similarity to the top-1 retrieved item, can be surprisingly effective for aligning predicted confidence with actual retrieval accuracy.

However, rejection analysis uncovered clear distinctions between the methods. Specifically, techniques based on top-1 consistency across MCD samples and adversarial perturbations consistently outperformed top-1 similarity baselines. These methods excelled at identifying truly uncertain queries, leading to superior performance when filtering out unreliable retrieval rankings. This divergence highlights an important insight: while calibration metrics evaluate global alignment between confidence and performance, rejection analysis is more sensitive to a method's ability to rank uncertainty effectively – a critical factor in real-world applications where decisions are made based on the most confident predictions.

Our comparison with ProbVLM (Upadhyay et al. [2023]) reveals that while ProbVLM offers advanced capabilities through probabilistic modeling – enabling applications like active learning and model selection – it demonstrated weaker calibration compared to the proposed methods. This performance gap is specially notable in cross-dataset scenarios, due to dataset-specific training dependencies. This highlights an inherent strength of our approach – dataset agnosticism and superior generalization.

In conclusion, our findings suggest that top1-based, retrieval-focused predictive uncertainty estimation methods, such as MCD-based rank consistency and adversarial perturbation approaches, are not only computationally efficient but also highly effective in both calibration and robustness evaluations. These methods offer strong, data-agnostic performance without the overhead of complex probabilistic modeling, making them well-suited for real-world cross-modal retrieval applications.

While our MCD and Ensemble-based methods do not require additional training, they do incur extra inference-time computation. This overhead scales linearly with the number of MCD samples or the number of models in the Ensemble; however, these computations are trivially parallelizable in practice, leading to minimal time overhead. Moreover, the computational cost can be further mitigated through selective application – for example, using simpler cosine similarity-based uncertainty (or fewer MCD samples) for routine queries, while reserving more expensive uncertainty estimation for critical or high-risk decisions.

To support reproducibility and further research, the code for all proposed uncertainty estimation methods, along with the evaluation framework used in this work, are made publicly available at `http://github.com/lluisgomez/uCLIP`.

Acknowledgements

This work is funded by the Ramon y Cajal research fellowship RYC2020-030777-I / AEI / 10.13039/501100011033.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736, 2022.

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.

Sanghyuk Chun. Improved probabilistic image-text representations. In *The Twelfth International Conference on Learning Representations*.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. *arXiv preprint arXiv:2101.05068*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, 2016.

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

Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 318–319, 2020.

Lara Hoffmann and Clemens Elster. Deep ensembles from a bayesian perspective. *arXiv preprint arXiv:2105.13283*, 2021.

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

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.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 2017.

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. *Advances in Neural Information Processing Systems*, 35:11934–11946, 2022.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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*. PMLR, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Andrei Neculai, Yanbei Chen, and Zeynep Akata. Probabilistic compositional embeddings for multimodal image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4547–4557, 2022.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

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.

Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7322–7329, 2019.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

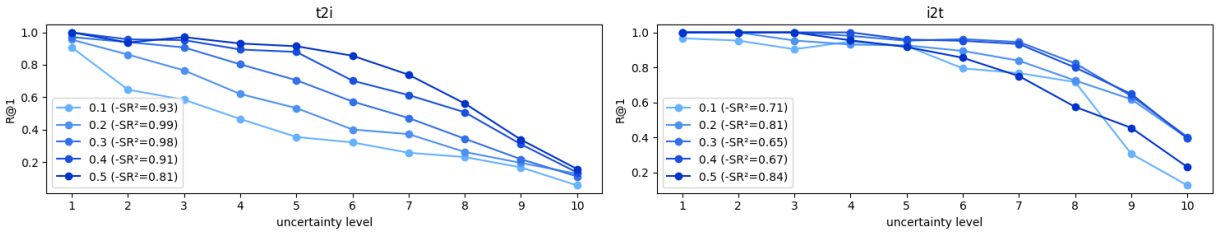
Over the Top-1: Uncertainty-Aware Cross-Modal Retrieval with CLIP (Supplementary Material)

Lluis Gomez[1]

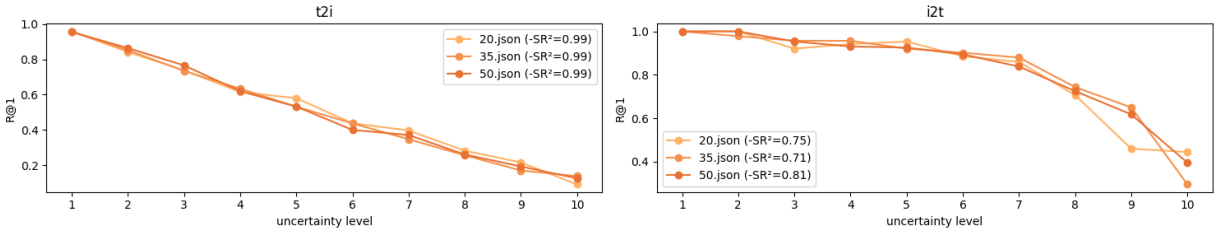
[1] Computer Vision Center, Universitat Autònoma de Barcelona.

A SENSITIVITY ANALYSIS OF MCD HYPERPARAMETERS

In this section, we provide additional results analyzing the sensitivity of the Monte Carlo Dropout (MCD) uncertainty estimation to its key hyperparameters: the dropout rate and the number of stochastic forward passes (samples). In the main paper, we use 50 stochastic forward passes with a dropout rate of 0.2 during inference, here we evaluate the robustness of our top-1 consistency uncertainty estimates on Flickr30K and MSCOCO retrieval tasks under different settings of these hyperparameters. Calibration plots are shown in Figures 3 and 4 respectively.



(a) Calibration when varying the dropout rate with a fixed number of MCD samples ($num_samples = 50$).

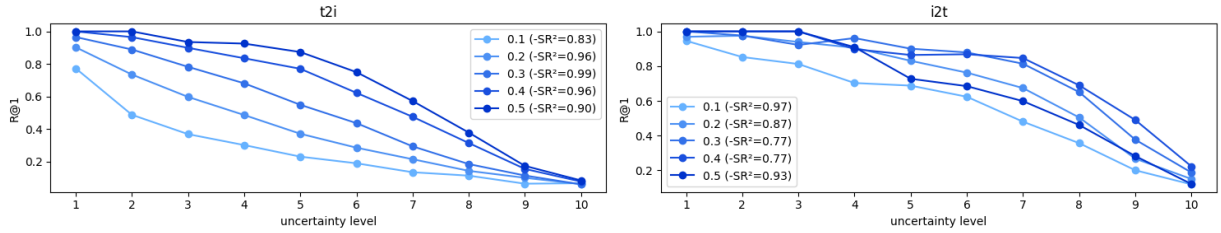


(b) Calibration when varying the number of MCD samples with a fixed dropout rate ($drop_rate = 0.2$).

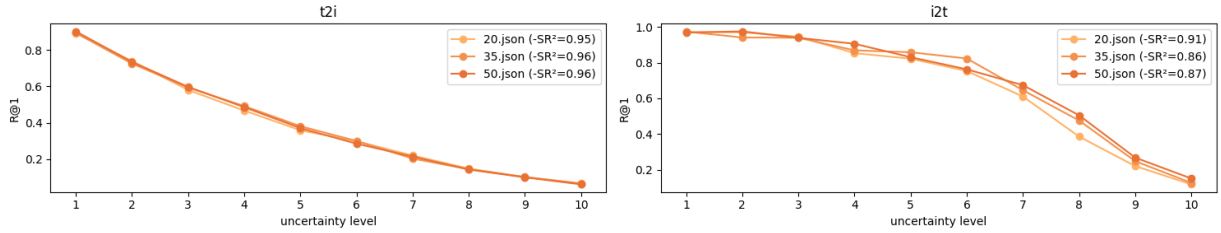
Figure 3: Calibration plots of MCD top1-consistency uncertainty estimation on Flickr30K text-to-image (t2i) and image-to-text (i2t) retrieval tasks for different hyperparameter settings. Each curve corresponds to a different hyperparameter configuration, with the $-SR^2$ calibration score shown in brackets in the respective legend entry.

Effect of Dropout Rate. Figures 3(a) and 4(a) show calibration plots for varying dropout rates while fixing the number of samples to 50. We observe that the uncertainty estimates are relatively robust across typical dropout values, with a dropout rate of 0.2 providing slightly better calibration performance overall.

Effect of Number of Samples. Figure 3(b) and 4(b) show calibration plots for varying the number of samples while fixing the dropout rate to 0.2. As expected, increasing the number of samples leads to more stable uncertainty estimates. Nevertheless, we find that 20 samples already provide a good trade-off between performance and computational cost, while our default choice of 50 samples ensures more stable estimates.



(a) Calibration when varying the dropout rate with a fixed number of MCD samples ($num_samples = 50$).



(b) Calibration when varying the number of MCD samples with a fixed dropout rate ($drop_rate = 0.2$).

Figure 4: Calibration plots of MCD top1-consistency uncertainty estimation on MSCOCO text-to-image (t2i) and image-to-text (i2t) retrieval tasks for different hyperparameter settings. Each curve corresponds to a different hyperparameter configuration, with the $-SR^2$ calibration score shown in brackets in the respective legend entry.

B IMPACT OF MODEL SCALE

To assess the impact of model scale on uncertainty estimation, we conducted additional experiments comparing ViT-L/14 and the smaller ViT-B/32 variant. The results are presented in Table 2.

	MSCOCO						Flickr30K					
	image2text			text2image			image2text			text2image		
	S	R^2	$-SR^2$	S	R^2	$-SR^2$	S	R^2	$-SR^2$	S	R^2	$-SR^2$
Top1similarity (ViT-L/14)	-1.00	0.95	0.95	-1.00	0.95	0.95	-0.98	0.86	0.84	-1.00	0.94	0.94
Top1similarity (ViT-B/32)	-1.00	0.95	0.95	-1.00	0.95	0.95	-1.00	0.83	0.83	-0.98	0.96	0.94
MCD Top1similarity (ViT-L/14)	-0.92	0.88	0.82	-1.00	0.96	0.96	-0.97	0.85	0.83	-1.00	0.96	0.96
MCD Top1similarity (ViT-B/32)	-0.95	0.86	0.82	-1.00	0.96	0.96	-0.98	0.92	0.90	-1.00	0.95	0.95
MCD Top1consistency (ViT-L/14)	-1.00	0.88	0.88	-1.00	0.96	0.96	-0.98	0.77	0.75	-1.00	0.99	0.99
MCD Top1consistency (ViT-B/32)	-0.98	0.84	0.82	-1.00	0.99	0.99	-0.96	0.72	0.70	-1.00	0.98	0.98
MCD Adversarial (ViT-L/14)	-1.00	0.99	0.99	-1.00	0.98	0.98	-0.96	0.93	0.89	-1.00	0.95	0.95
MCD Adversarial (ViT-B/32)	-0.99	0.92	0.91	-0.98	0.87	0.85	-0.99	0.90	0.89	-0.97	0.78	0.76
MCD Adversarial lin. (ViT-L/14)	-1.00	0.97	0.97	-1.00	0.95	0.95	-1.00	0.95	0.95	-1.00	0.91	0.91
MCD Adversarial lin. (ViT-B/32)	-0.95	0.89	0.85	-1.00	0.86	0.86	-0.99	0.91	0.90	-0.99	0.78	0.77

Table 2: Uncertainty calibration metrics for all considered methods using CLIP ViT-L/14 and ViT-B/32.

Our analysis shows that while the larger ViT-L/14 model achieves better calibration metrics overall, the uncertainty estimation performance of ViT-B/32 remains competitive and consistent across tasks. Specifically, we observe that for Top1-similarity and Top1-consistency methods, the relative differences between the two models are moderate, suggesting that these uncertainty estimates are robust to model scale. In contrast, for the adversarial methods, the differences between ViT-L/14 and ViT-B/32 are more pronounced, indicating a stronger dependence on model size.