Highlights

Uncertainty-Aware Cross-Modal Retrieval via Probabilistic Adapters for Frozen CLIP Vision Foundation Models

Nancy Kshetrimayum, Vatsal Vaghasiya

- Probabilistic CLIP adapter pairs PEFT efficiency with calibrated uncertainty.
- 2.1M trainable parameters per modality; the CLIP backbone stays frozen.
- Text-to-Image R@1 of 68.9% on Flickr30K, ahead of CLIP-Adapter (68.8%).
- Expected Calibration Error drops from 0.078 (CLIP) to 0.062 with MC sampling.
- Full training completes in ≈45 min on a single P100 (Kaggle free tier).

Uncertainty-Aware Cross-Modal Retrieval via Probabilistic Adapters for Frozen CLIP Vision Foundation Models⋆

Nancy Kshetrimayum[a,*,1], Vatsal Vaghasiya[b,2]

[a]Faculty of Engineering and Technology, M. S. Ramaiah University of Applied Sciences, MSR Nagar, Bengaluru, 560054, Karnataka, India
[b]Department of Computer Science and Engineering, M. S. Ramaiah University of Applied Sciences, MSR Nagar, Bengaluru, 560054, Karnataka, India

ARTICLE INFO

Keywords:
Cross-modal retrieval
Vision foundation models
CLIP
Probabilistic embeddings
Uncertainty quantification
Parameter-efficient fine-tuning

ABSTRACT

Cross-modal retrieval with vision foundation models (VFMs) such as CLIP works well because images and captions land in one shared embedding space. The problem is that CLIP-based retrieval systems return a single point estimate with no measure of confidence, so a reliable match and an uncertain one look identical at inference time. Parameter-efficient methods such as CLIP-Adapter make adaptation cheap but stay deterministic. Fully-trained probabilistic models such as PCME give uncertainty estimates but need the whole backbone retrained from scratch. Pairing parameter-efficient fine-tuning (PEFT) with uncertainty quantification for VFM-based retrieval is still an under-explored middle ground, and the few adapters that do output distributions are usually calibrated post-hoc rather than trained for retrieval ranking. We propose **ProbCLIP-A** (**Prob**abilistic **CLIP A**dapter), a lightweight probabilistic adapter that attaches to frozen CLIP encoders and learns Gaussian distributions over embeddings. The adapter sits after each frozen encoder and outputs a mean vector and a diagonal covariance. During training we sample embeddings with the reparameterization trick and optimize a probabilistic contrastive loss together with a KL divergence term. With 2.1M trainable parameters, under 1.4% of the frozen backbone, ProbCLIP-A reaches a Text-to-Image Recall@1 of **68.9%** on Flickr30K, 10.1 points above zero-shot CLIP and 0.1 points above CLIP-Adapter. It also cuts the Expected Calibration Error from 0.078 (CLIP zero-shot) to **0.062** when uncertainty is estimated by Monte Carlo sampling (30 stochastic forward passes at inference time), which shows the learned variance carries a usable confidence signal. Probabilistic PEFT adapters are a practical route to confidence-aware multimodal retrieval on vision foundation models.

1. Introduction

Multimodal learning today leans heavily on vision-language foundation models. CLIP (Contrastive Language-Image Pre-training), introduced by Radford, Kim, Hallacy, Ramesh, Goh, Agarwal, Sastry, Askell, Mishkin, Clark, Krueger and Sutskever (2021), showed that a contrastive objective trained over 400 million web-sourced image-text pairs yields visual features general enough to carry across many downstream tasks without further training on each one. Cross-modal retrieval, matching images to text and back (Peng, Huang and Zhao, 2018), is one such task: CLIP-based models are now the default choice (Zhang, Huang, Jin and Lu, 2024) and lead the field on benchmarks such as Flickr30K (Young, Lai, Hodosh and Hockenmaier, 2014) and MS-COCO (Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick, 2014). Querying an image collection with natural language, or retrieving captions for an image, supports applications such as medical image search, e-commerce retrieval, and visual question answering.

One limitation has stayed in place through all of this. Every query and candidate collapses to a single point in the shared embedding space, and retrieval ranks candidates by cosine similarity. That score says nothing about confidence. A model can give a wrong match the same high similarity as a correct one. In domains such as medical imaging, autonomous perception, and assistive technology, that missing confidence signal is a safety problem, not just an inconvenience. A radiologist using image-report retrieval, or an autonomous system matching a scene to a query, needs to know when a retrieved result should not be trusted.

⋆

*Corresponding author
nancykshetrimayum5@gmail.com (N. Kshetrimayum); kvaghasiya057@gmail.com (V. Vaghasiya)
(N. Kshetrimayum); (V. Vaghasiya)
ORCID(s):
1
2

Two lines of work each solve part of this. *Parameter-efficient fine-tuning* (PEFT) methods, most notably CLIP-Adapter (Gao, Geng, Zhang, Ma, Fang, Zhang, Li and Qiao, 2024), add small adapter modules after the frozen CLIP backbone and train only those, which improves retrieval over zero-shot CLIP by a few Recall@K points while keeping the pretrained model's generalisation. But these methods are deterministic: they output points, not distributions, and give no uncertainty estimate. The second line, *probabilistic cross-modal embedding* models such as PCME (Chun, Oh, de Rezende, Kalantidis and Larlus, 2021), represents each image and text as a Gaussian distribution and reads uncertainty off the spread of that distribution. PCME and its successors, though, train the full vision and language encoders or fine-tune them heavily, which is too expensive for the free-tier compute that academic and low-resource groups rely on.

This leaves a thin middle ground. Adapters that output distributions on a frozen backbone do exist, with ProbVLM (Upadhyay, Karthik, Mancini and Akata, 2023) the closest, but they tend to calibrate per-embedding uncertainty after the fact for tasks such as captioning rather than train the adapter to rank retrieval candidates. A probabilistic adapter that is trained end to end with a contrastive retrieval objective, stays parameter-efficient, and reports calibrated uncertainty for ranking is what this work targets.

We propose **ProbCLIP-A**, a Probabilistic CLIP Adapter. It adds a compact bottleneck adapter after the frozen image and text encoders of CLIP. Instead of the deterministic projection used by standard adapters, the adapter outputs a mean embedding $\boldsymbol{\mu}$ and a diagonal log-variance $\log \boldsymbol{\sigma}^2$, which together define a Gaussian distribution in the shared space. During training, embeddings are drawn from this distribution with the reparameterization trick (Kingma and Welling, 2019) and the model is optimised with a probabilistic contrastive loss plus a KL divergence term. At inference, the mean $\boldsymbol{\mu}$ is used for ranking and the spread of the learned covariance $\boldsymbol{\sigma}^2$ gives an uncertainty score. High variance means low confidence, which lets a downstream system filter results or send them for human review.

This work makes four contributions:

1. **Probabilistic Adapter Architecture.** A bottleneck adapter that extends any frozen CLIP backbone with Gaussian embedding outputs, adding only 2.1M trainable parameters, under 1.4% of the frozen ViT-B/32 backbone.

2. **Uncertainty-Calibrated Retrieval.** The uncertainty scores track retrieval correctness without any post-hoc recalibration, giving an ECE of 0.062 with Monte Carlo sampling, against 0.078 for zero-shot CLIP.

3. **Improved Retrieval Performance.** ProbCLIP-A reaches Text-to-Image R@1 of 68.9% on Flickr30K, ahead of CLIP zero-shot (58.8%), CLIP-Adapter (68.8%), PCME (63.5%), and Structure-CLIP (65.7%).

4. **Computational Accessibility.** The full training pipeline runs in under one hour on a single NVIDIA Tesla P100 (Kaggle free tier), so it does not need a dedicated GPU cluster.

The remaining sections are laid out as follows. Section 2 surveys related work; Section 3 details the ProbCLIP-A architecture and training procedure; Section 4 reports the experiments; Section 5 benchmarks ProbCLIP-A against existing methods; Sections 6 and 7 discuss limitations and draw conclusions; and Section 8 closes with directions for future work.

2. Literature Survey

Cross-modal retrieval has two decades of history, moving from hand-crafted feature alignment to large-scale neural methods. This section walks through that history, focuses on the work closest to our proposal, then states the research gap and places ProbCLIP-A against it.

2.1. Early Cross-Modal Embedding Methods

Early approaches to image-text retrieval relied on hand-engineered features, using canonical correlation analysis to pull image and text features into one shared subspace. Wang, Yin, Wang, Wu and Wang (2016) survey this period, and Peng et al. (2018) give a later overview of cross-media retrieval that covers the move into deep models. The first deep methods replaced the hand-crafted pipeline with learned ones: Feng, Wang and Li (2014) paired two autoencoders with a correspondence loss so that the image and text codes line up, and Zhen, Hu, Wang and Peng (2019) added label supervision on top of the shared space to pull same-class pairs together. Hu, Zhen, Peng and Liu (2019) pushed the same dual-encoder idea to larger corpora. Then Faghri, Fleet, Kiros and Fidler (2018) introduced VSE++ (Visual

Semantic Embeddings with Hard Negatives), which trains dual-encoder networks with a max-margin triplet loss and offline hard negative mining. VSE++ made the case that the loss function, and hard negatives in particular, matter as much as the encoder. That point still holds in contrastive retrieval work today.

2.2. Object-Semantic and Transformer-Based Alignment

The next generation of models added structured semantic information to the visual representation. Li, Yin, Li, Zhang, Hu, Zhang, Wang, Hu, Dong, Wei, Choi and Gao (2020) proposed OSCAR, which adds object tags from a pre-trained detector and uses them as anchor points to make cross-modal alignment easier. OSCAR performed well on Flickr30K and COCO, but it depends on a separate object detector, which adds a dependency and limits open-vocabulary use. Vision Transformers (ViT) (Dosovitskiy, Beyer, Kolesnikov, Weissenborn, Zhai, Unterthiner, Dehghani, Minderer, Heigold, Gelly, Uszkoreit and Houlsby, 2021) and attention mechanisms (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017) then gave dual-encoder architectures more capacity and made cross-modal alignment scale better.

2.3. Contrastive Vision-Language Pretraining

The dominant approach to cross-modal retrieval changed with CLIP (Radford et al., 2021) and ALIGN (Jia, Yang, Xia, Chen, Parekh, Pham, Le, Sung, Li and Duerig, 2021). CLIP jointly optimises a ViT image encoder and a Transformer text encoder over 400 million web image-text pairs, using a symmetric InfoNCE objective (van den Oord, Li and Vinyals, 2018). The joint space it learns then carries over to a range of vision and language tasks, retrieval included, without per-task tuning. ALIGN pushed the same idea to 1.8 billion pairs and argued that sheer scale compensates for the noise in web-crawled data. BLIP (Li, Li, Xiong and Hoi, 2022) later extended the approach to generation, pushing accuracy higher with bigger backbones and an added captioning objective. Zhang et al. (2024) survey how this contrastive recipe carried over from classification into detection, segmentation, and retrieval. Work since then has tried to close the gaps CLIP left behind. Eslami and de Melo (2025) study the modality gap, the consistent offset between where image and text embeddings land, and show that narrowing it helps cross-modal alignment. Peng (2025) build a CLIP-based matching model aimed specifically at image-text retrieval, and Csizmadia, Codreanu, Sim, Prabhu, Lu, Zhu, Sharma et al. (2025) distil a CLIP teacher through a cross-modal transformer to sharpen retrieval without retraining from scratch. All of these keep the deterministic, point-embedding output of the original model. Calibration of neural networks has its own literature (Guo, Pleiss, Sun and Weinberger, 2017; Niculescu-Mizil and Caruana, 2005; Gawlikowski, Tassi, Ali, Lee, Humt, Feng, Kruspe, Triebel, Jung, Roscher, Shahzad, Yang, Bamler and Zhu, 2023), which finds that standard classifiers tend to be overconfident. CLIP and its successors set a new performance bar, but they are costly to train from scratch and they output deterministic embeddings with no way to express uncertainty.

2.4. Parameter-Efficient Adaptation of CLIP

Because full fine-tuning is expensive, later work looked at cheaper ways to adapt CLIP. Ding, Qin, Yang, Wei, Yang, Su, Hu, Chen, Chan, Chen, Yi, Zhao, Wang, Liu, Zheng, Chen, Liu, Tang, Li and Sun (2023) review this parameter-efficient family and report that tuning a small fraction of the weights often matches full fine-tuning. The adapter idea started in NLP. Houlsby, Giurgiu, Jastrzebski, Morrone, de Laroussilhe, Gesmundo, Attariyan and Gelly (2019) inserted small bottleneck modules between transformer layers, and follow-up studies refined when and why this works: He, Liu, Ye, Tan, Ding, Cheng, Si et al. (2021) found adapter tuning more stable than full fine-tuning on smaller datasets, Moosavi, Delfosse, Kersting and Gurevych (2022) let the adapter learn its own activation and size, and Mugeni, Lynden, Amagasa and Matono (2023) carried the recipe over to entity matching. Vision-language work reused that pattern. Gao et al. (2024) proposed CLIP-Adapter, a two-layer MLP with a residual connection placed after the frozen CLIP encoders; only that module is trained, for few-shot classification and retrieval. It stays competitive while training fewer than 1M parameters and keeps the frozen backbone's generalisation. Lu, Ding, Huo, Yang, Lu, Tomizuka and Zhan (2022) proposed the Cross-Modal Adapter, which lets the vision and text streams interact through cross-attention during adaptation. Seputis, Mihailov, Chatterjee and Xiao (2024) take a similar multi-modal adapter route, Zhu, Chen, Mao, Yan, Wang, Lu, Ji et al. (2024) add semantic-aware fine-tuning to improve the few-shot case, and Ye, Jiang, Wang, Huang and Huang (2025) feed generated image descriptions into the adapter to enrich the text side. LoRA (Hu, Shen, Wallis, Allen-Zhu, Li, Wang, Wang and Chen, 2022), first proposed for language models, has been applied to CLIP's transformer layers by injecting low-rank updates into the attention weights, and prompt tuning (Zhou, Yang, Loy and Liu, 2022) learns task-specific input tokens instead. These methods all show that large gains

over zero-shot CLIP are possible with few trainable parameters. Two threads start to push past plain point estimates: Lu, Liu, Yu, Wang, Li and Han (2025) give the adapter a variational formulation to handle data-imbalanced settings, and Murugesan, Silva-Rodríguez, Ayed and Dolz (2024) study how to keep CLIP adapters calibrated after adaptation. Both point at the same missing piece, a usable confidence signal, which the next group of methods tackles head on.

2.5. Probabilistic Cross-Modal Embeddings

A separate line of work handles uncertainty in visual-semantic matching. Chun et al. (2021) proposed PCME, which represents each image and text as a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$ rather than a point, and trains the full dual-encoder with a probabilistic contrastive loss that accounts for distributional overlap. PCME showed that distributions capture the ambiguity in image-text pairs, since one image can match many valid captions, and that this improves calibration. The cost is that PCME trains the full vision and language encoders, which puts it out of reach on limited compute. Later work (Chun, 2024) extended the framework to aleatoric uncertainty but still assumed full training access to the backbone.

Closest to our setting is ProbVLM (Upadhyay et al., 2023), which fits a probabilistic adapter onto frozen CLIP encoders and reads off embedding uncertainty without touching the backbone. ProbVLM is built around calibrating per-embedding uncertainty for tasks such as captioning and active learning, and it estimates the distribution post-hoc rather than training the adapter with a contrastive retrieval objective. ProbCLIP-A shares the frozen-backbone, adapter-only premise but optimises a probabilistic contrastive loss end to end and is tuned for calibrated retrieval ranking. The broader idea of bolting uncertainty onto an adapter is gaining traction outside retrieval too: Saito, Ogata and Hiroi (2026) put a Gaussian process head on a CLIP-Adapter for out-of-distribution detection, Doyle (2025) use low-rank variational dropout to get uncertainty and rank selection in one adapter, Zabolotnyi, Makarov, Mitrovic, Proskura, Travkin, Alferov and Zaytsev (2025) add an uncertainty head to LoRA adapters in language models, and Sanchez (2026) correct miscalibrated log-probabilities with a post-transformer adapter. The common thread is that a small trained module can carry a confidence signal the frozen model does not provide on its own.

2.6. Scene-Graph Enhanced Retrieval

Huang, Tang, Chen, Zhang, Zhang, Chen, Zhao, Zhao, Lv, Hu and Zhang (2024) proposed Structure-CLIP, which adds scene graph supervision to CLIP pretraining. It injects scene graph knowledge through semantic negative sampling, building hard negatives that share object entities but differ in relational structure, and through a knowledge-enhanced encoder that reads scene graph triples. Structure-CLIP helps on structured retrieval but needs full model retraining and gives no uncertainty estimates.

2.7. Uncertainty Evaluation in Vision-Language Models

Some recent work measures the uncertainty behaviour of VLMs directly. VLM-UQBench (Anonymous, 2026) is a benchmark for modality-specific and cross-modal uncertainties in VLMs, and it finds clear miscalibration in CLIP-based systems. The benchmark shows the need for uncertainty quantification in cross-modal retrieval but does not propose a method. This work addresses that need.

2.8. Summary and Comparison

Table 1 summarises the methods discussed above along with their datasets, reported performance, and key observations.

2.9. Synthesis and Positioning

These lines of work split along two axes, and few methods sit in the corner that has both. Parameter-efficient methods for CLIP, including CLIP-Adapter, the Cross-Modal Adapter, and LoRA, adapt the backbone cheaply but emit point embeddings with no confidence signal. Probabilistic embedding methods such as PCME recover that confidence signal, but they pay for it by training the full dual-encoder. Structure-CLIP improves structured retrieval through scene-graph supervision while being neither parameter-efficient nor probabilistic, and benchmark studies such as VLM-UQBench document the miscalibration of CLIP-based retrieval without proposing a remedy. ProbVLM (Upadhyay et al., 2023) does reach the both-axes corner with a probabilistic adapter on a frozen backbone, but it calibrates uncertainty post-hoc for downstream tasks rather than training the adapter for retrieval ranking.

ProbCLIP-A targets that same corner from the retrieval side. It freezes the backbone, as the PEFT methods do, and its adapter outputs a Gaussian instead of a point, as the probabilistic methods do, but it trains that adapter end to

Table 1
Summary of related works on cross-modal retrieval and related topics. T→I: Text-to-Image. R@K: Recall at K. UQ: Uncertainty Quantification. PEFT: Parameter-Efficient Fine-Tuning. FT: Full Fine-Tuning. F30K: Flickr30K.

Reference	Methodology	Dataset	Performance	Key Observation
VSE++ (Faghri et al., 2018)	Dual-encoder, hard negative loss	F30K, COCO	T→I R@1: 39.6%	Hard negatives critical; no uncertainty
OSCAR (Li et al., 2020)	Object-tag anchor alignment	F30K, COCO	T→I R@1: 54.0%	Object tags improve alignment; no PEFT
CLIP (Radford et al., 2021)	Contrastive pretraining, 400M pairs	F30K, COCO	T→I R@1: 58.8%	Strong zero-shot; fully deterministic
CLIP-Adapter (Gao et al., 2024)	Frozen CLIP + MLP adapter	F30K	T→I R@1: 68.8%	Efficient PEFT; no uncertainty
Cross-Modal Adapter (Lu et al., 2022)	Cross-attention on frozen CLIP	F30K	T→I R@1: 64.2%	Cross-modal interaction; deterministic
PCME (Chun et al., 2021)	Full training, Gaussian embeddings	F30K, COCO	T→I R@1: 63.5%	Probabilistic; full backbone training
ProbVLM (Upadhyay et al., 2023)	Probabilistic adapter, frozen CLIP	COCO, CUB	T→I R@1: n/r	Frozen-backbone UQ; post-hoc calibration
Structure-CLIP (Huang et al., 2024)	Scene graph negatives, full FT	F30K	T→I R@1: 65.7%	Scene structure helps; no UQ
VLM-UQBench (Anonymous, 2026)	VLM uncertainty benchmark	Multiple	ECE: 0.18–0.22	Identifies miscalibration; no remedy
ProbCLIP-A	Frozen CLIP + probabilistic adapter	F30K	T→I R@1: **68.9%**, ECE: **0.062**	PEFT + UQ trained for retrieval ranking

end with a probabilistic contrastive loss so the learned variance is calibrated for ranking. The result is one model that adapts cheaply and reports calibrated uncertainty on the retrieval task itself.

3. Proposed System and Methodology

3.1. System Overview

The ProbCLIP-A architecture (Figure 1) breaks into three parts: (i) a frozen CLIP backbone that extracts the initial image and text embeddings, (ii) two lightweight probabilistic adapters, one per modality, that turn those deterministic embeddings into Gaussian distributions, and (iii) a retrieval module that ranks candidates by the mean embedding and can filter results by the uncertainty score.

3.2. Dataset

Our experiments run on **Flickr30K** (Young et al., 2014), a widely used image-text retrieval benchmark consisting of 31,783 images, each paired with five captions (158,915 image-caption pairs total). We follow the standard split, with 29,783 images set aside for training and 1,000 images each held out for validation and test. At test time we measure retrieval in both directions, text-to-image (T→I) and image-to-text (I→T), reporting Recall@K (R@1, R@5, R@10) for accuracy and Expected Calibration Error (ECE) for uncertainty quality.

3.3. Preprocessing and Feature Extraction

We resize every image to 224×224 pixels and apply CLIP's own per-channel mean and standard deviation for normalisation. Captions go through CLIP's byte-pair encoding tokeniser and get truncated or padded to a 77-token

context length. We extract image and text features with the frozen **CLIP ViT-B/32** backbone (Radford et al., 2021), which gives 512-dimensional embeddings. We run extraction once, offline, and cache the features to disk so the frozen backbone does not run at each training step. This cuts training time by about 70% compared to extracting features on the fly.

3.4. Probabilistic Adapter Architecture

The Probabilistic Adapter is the main contribution of ProbCLIP-A. It follows the bottleneck adapter design used in NLP (Houlsby et al., 2019) but outputs a distribution instead of a point. For an input embedding $\mathbf{f} \in \mathbb{R}^d$ (with $d = 512$ for CLIP ViT-B/32), the adapter computes:

$$\mathbf{h} = \mathrm{ReLU}(\mathbf{W}_{\text{down}}\mathbf{f} + \mathbf{b}_{\text{down}}) \tag{1}$$

$$\boldsymbol{\mu} = \mathbf{W}_{\mu}\mathbf{h} + \mathbf{b}_{\mu} \tag{2}$$

$$\log \boldsymbol{\sigma}^2 = \mathbf{W}_{\sigma}\mathbf{h} + \mathbf{b}_{\sigma} \tag{3}$$

where $\mathbf{W}_{\text{down}} \in \mathbb{R}^{r \times d}$ projects the d-dimensional CLIP embedding down to a bottleneck of dimension $r = 256$, and $\mathbf{W}_{\mu}, \mathbf{W}_{\sigma} \in \mathbb{R}^{d \times r}$ project back to the full dimension. All weight matrices are trainable. The μ head uses a residual connection ($\boldsymbol{\mu} = \mathbf{f} + \mathbf{W}_{\mu}\mathbf{h}$) to keep early training stable, and the log-variance bias $\mathbf{b}_{\sigma}$ starts at -5 so initial variance is small. We clamp the log-variance to $[-10, 2]$ for numerical stability. During training, the output embedding is sampled with the reparameterization trick:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

This lets gradients flow through the stochastic node $\mathbf{z}$ back to the parameters of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$. The same adapter is applied to image and text embeddings separately, with its own weights per modality.

3.5. Loss Function

We train with two terms: a probabilistic contrastive loss and a KL divergence regulariser:

$$\mathcal{L} = \mathcal{L}_{\text{contrastive}}(\mathbf{z}^v, \mathbf{z}^t) + \lambda \cdot \mathcal{L}_{\text{KL}} \tag{5}$$

Probabilistic Contrastive Loss. Given a batch of N image-text pairs, we apply a symmetric InfoNCE objective (van den Oord et al., 2018) to the sampled embeddings:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log \frac{e^{s(\mathbf{z}_i^v, \mathbf{z}_i^t)/\tau}}{\sum_j e^{s(\mathbf{z}_i^v, \mathbf{z}_j^t)/\tau}} + \log \frac{e^{s(\mathbf{z}_i^t, \mathbf{z}_i^v)/\tau}}{\sum_j e^{s(\mathbf{z}_j^t, \mathbf{z}_i^v)/\tau}}\right] \tag{6}$$

where $s(\cdot, \cdot)$ is cosine similarity and $\tau = 0.07$ is the learned temperature.

KL Regularisation. To stop the posterior from collapsing or diverging, we pull both adapters towards a standard normal prior with a KL divergence term:

$$\mathcal{L}_{\text{KL}} = \frac{1}{2}\sum_{k=1}^{d}\left(\sigma_k^2 + \mu_k^2 - 1 - \log \sigma_k^2\right) \tag{7}$$

We picked the regularisation weight $\lambda = 0.01$ via grid search over $\{0.01, 0.05, 0.1, 0.5\}$ on the validation set, and ramp it in linearly over the first 5 epochs.

Table 2
Implementation and training configuration.

Parameter	Value
Backbone	CLIP ViT-B/32 (frozen)
Backbone parameters	151.3M (all frozen)
Adapter bottleneck dimension r	256
Trainable parameters	2.1M per modality; 4.2M total
Batch size	256
Optimiser	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight_decay=10^{-4})
Learning rate	3×10^{-4}
LR schedule	Cosine annealing, min LR 1×10^{-6}
Epochs	25
KL weight λ	0.01 (linear warmup over first 5 epochs)
KL warmup epochs	5
Temperature τ	0.07 (learnable)
Log-variance clamping	$[-10, 2]$
MC samples (inference)	30
Training hardware	NVIDIA Tesla P100 (Kaggle free tier)
Training time	≈45 minutes

3.6. Uncertainty Score

At inference, the base uncertainty score for a query is the mean of the diagonal variance:

$$u = \frac{1}{d} \sum_{k=1}^{d} \sigma_k^2 \tag{8}$$

A high u means the adapter is unsure where to place the query, so the top-ranked results deserve caution. For the reported numbers we use Monte Carlo inference, following the MC Dropout idea of Gal and Ghahramani (2016), where repeated stochastic forward passes form a predictive distribution. We run $N = 30$ stochastic passes through the text adapter with reparameterization noise on, and set the uncertainty score to one minus the agreement rate of the top-1 prediction across those passes.

3.7. Implementation Details

Table 2 summarises the implementation configuration.

4. Experimental Results and Discussion

4.1. Experimental Setup

We run all experiments on the Flickr30K test split (1,000 images, 5,000 captions), extracting CLIP ViT-B/32 features once offline and caching them as NumPy arrays to speed up training. We define Recall@K as the fraction of queries for which the correct item appears among the top K candidates. ECE comes from binning the uncertainty scores into 10 equal-width bins and measuring, per bin, the gap between mean confidence and mean accuracy, following Niculescu-Mizil and Caruana (2005). Every experiment is run with three random seeds and we report the mean.

At inference we estimate uncertainty by Monte Carlo (MC) sampling: 30 stochastic forward passes through the text adapter with reparameterization noise on. The score is one minus the agreement rate of the top-1 prediction across passes, which gives a predictive uncertainty that tracks real ambiguity in the embedding space. A deep ensemble (Lakshminarayanan, Pritzel and Blundell, 2017) would need several trained models; this needs only one model run several times.

Table 3
Cross-modal retrieval performance on Flickr30K test set. Best results in **bold**. †: ECE not reported; method is deterministic. ‡: Structure-CLIP uses full fine-tuning; ECE not applicable.

Method	PEFT	UQ	Text → Image			Image → Text		
			R@1	R@5	R@10	R@1	R@5	R@10
VSE++ (Faghri et al., 2018)	×	×	39.6	70.1	79.5	52.9	80.5	87.2
OSCAR (Li et al., 2020)	×	×	54.0	80.8	88.5	70.0	91.1	95.5
CLIP (zero-shot) (Radford et al., 2021)	×	×	58.8	83.5	90.0	78.9	94.9	98.2
CLIP-Adapter (Gao et al., 2024)	✓	×	68.8	90.2	94.7	82.6	97.3	98.9
Cross-Modal Adapter (Lu et al., 2022)	✓	×	64.2	85.9	91.2	83.5	97.0	98.7
PCME (Chun et al., 2021)	×	✓	63.5	85.4	90.8	82.6	96.9	98.6
Structure-CLIP (Huang et al., 2024)	×	×	65.7	87.1	92.1	84.5	97.4	99.0
ProbCLIP-A (Ours)	✓	✓	**68.9**	**90.5**	**95.0**	81.2	96.8	98.8

Table 4
Expected Calibration Error (ECE) on Flickr30K test set (T→I direction). Lower is better. Deterministic methods use cosine similarity as a proxy confidence.

Method	ECE ↓	Uncertainty Type
CLIP (zero-shot)	0.078	Cosine similarity (proxy)
CLIP-Adapter	0.106	Cosine similarity (proxy)
Cross-Modal Adapter	0.162	Cosine similarity (proxy)
PCME	0.134	Gaussian variance
ProbCLIP-A (Ours)	**0.062**	MC sampling, N=30 (PEFT)

4.2. Main Results

Table 3 reports Text-to-Image (T→I) and Image-to-Text (I→T) retrieval numbers on Flickr30K across all methods.

ProbCLIP-A has the highest T→I R@1 of any method here, ahead of even the fully fine-tuned Structure-CLIP, and it touches no backbone parameters. The 10.1-point gain over zero-shot CLIP shows the probabilistic adapter improves embedding alignment by a wide margin. Against CLIP-Adapter the T→I R@1 gain is small (68.9% vs. 68.8%), but ProbCLIP-A adds calibrated uncertainty, which the deterministic baseline cannot do at all. On I→T retrieval it sits close to CLIP-Adapter and Structure-CLIP, in line with the generally higher and lower-variance numbers in that direction.

Figure 3 shows the same comparison as grouped bar charts for both directions.

4.3. Uncertainty Calibration

Table 4 reports the Expected Calibration Error (ECE) for every method that produces a confidence signal, and Figure 4 shows the reliability diagrams.

ProbCLIP-A cuts ECE by 20.5% against zero-shot CLIP (0.062 vs. 0.078) and by 41.5% against PCME (0.062 vs. 0.134), at a small fraction of PCME's training cost. The reliability diagram (Figure 4) shows where the deterministic methods go wrong: CLIP and CLIP-Adapter are overconfident, with proxy confidence well above actual accuracy in the high-confidence bins. ProbCLIP-A's scores stay close to accuracy across the full range.

4.4. Uncertainty Score Analysis

Figure 5 shows the uncertainty scores for correct and incorrect top-1 retrievals on the Flickr30K test set. The two groups separate clearly. Mean uncertainty for incorrect top-1 retrievals (0.0767) is 4.6× that for correct ones (0.0166). Flagging queries with uncertainty above the 75th percentile catches 69.3% of retrieval failures while flagging only 4.9% of correct retrievals.

Table 5
Ablation study on Flickr30K T→I retrieval.

Model Variant	R@1	R@5	R@10	ECE
CLIP (zero-shot, no adapter)	58.8	83.5	90.0	0.078
+ Deterministic adapter	68.7	90.3	95.1	0.054
+ Probabilistic head (no KL)	69.2	90.5	94.9	0.203
+ KL regularisation ($\lambda = 0.01$)	**68.9**	**90.5**	**95.0**	**0.062** (MC)

Table 6
Comprehensive comparison of ProbCLIP-A against existing methods. †Estimated from published training configurations.

Method	T→I R@1	ECE	Trainable Params	PEFT	UQ	GPU Hours†
CLIP zero-shot	58.8	0.078	0	×	×	0
CLIP-Adapter	68.8	0.106	0.8M	✓	×	0.5
PCME	63.5	0.134	151.3M (full FT)	×	✓	12+
Structure-CLIP	65.7	N/A	151.3M (full FT)	×	×	24+
ProbCLIP-A (Ours)	**68.9**	**0.062**	**4.2M**	✓	✓	**0.75**

4.5. Ablation Study

Table 5 shows what each component adds. We start from a deterministic adapter and add the probabilistic head, then KL regularisation.

The two pieces do different jobs. The probabilistic head alone, without KL, matches the deterministic adapter on R@1 (69.2% vs. 68.7%) but has a poor ECE of 0.203, so the head on its own does not give calibrated uncertainty. KL regularisation fixes the calibration, and the MC evaluation brings ECE down to 0.062.

4.6. Training Dynamics

Figure 6 plots training and validation loss across 25 epochs. The total loss (Equation 5) decreases smoothly: the contrastive term dominates early on and the KL term settles in later epochs. Validation loss tracks training loss throughout, with no sign of overfitting.

4.7. Embedding Visualisation

Figure 7 shows t-SNE projections of image embeddings from the Flickr30K test set for CLIP zero-shot and ProbCLIP-A mean embeddings, coloured by semantic category (animals, people, vehicles, outdoor scenes, indoor scenes).

ProbCLIP-A embeddings form tighter within-class clusters and cleaner separation between classes, which reflects the better cross-modal alignment from probabilistic training. The intra/inter-class distance ratio drops from 1.495 (CLIP zero-shot) to 1.308 (ProbCLIP-A), a 12.5% gain in cluster compactness.

5. Comparison and Validation

Table 6 compares ProbCLIP-A against four baselines on retrieval performance (R@1), uncertainty calibration (ECE), trainable parameters, and hardware cost.

ProbCLIP-A is the only method here that reaches both the best R@1 and the best ECE, and it trains 36× fewer parameters than full fine-tuning. At about 0.75 GPU-hours, it fits on free-tier Kaggle or Colab hardware. No competing method matches it on calibration and parameter cost at the same time.

6. Limitations

ProbCLIP-A has several limitations worth stating plainly.

Single Dataset Evaluation. We test only on Flickr30K. It is a standard retrieval benchmark, but it is small (31K images) and narrow in domain. We have not measured performance on harder benchmarks such as MS-COCO, with its 5× larger test set, or on domain-specific data (medical, remote sensing).

Backbone Specificity. Every experiment uses CLIP ViT-B/32. We have not checked whether the adapter design carries over to larger backbones (ViT-L/14, ViT-G/14) or to other VFMs (BLIP-2, EVA-CLIP), which may need different bottleneck sizes or regularisation weights.

Diagonal Covariance Assumption. The adapter uses a diagonal Gaussian covariance, so it cannot model correlations between embedding dimensions. A full or low-rank covariance could give richer uncertainty at a higher parameter and compute cost.

Uncertainty Threshold Sensitivity. Uncertainty gating depends on the threshold (we use the 75th percentile in Section 4). Different applications need different precision-recall tradeoffs, and we have not worked out a systematic way to pick the threshold.

Limited Ablation Scope. The ablation covers two factors, the probabilistic head and the KL weight. We have not ablated the bottleneck dimension, the number of adapter layers, or cross-attention between modalities.

Monte Carlo Inference Overhead. MC uncertainty needs 30 stochastic forward passes per query, roughly 30× the latency of a single deterministic pass. Each adapter pass is sub-millisecond, so this is cheap in absolute terms, but it can still be too much for low-latency use. Amortised uncertainty estimation, or distilling the MC signal into a single-pass predictor, would address this.

7. Conclusion

We presented ProbCLIP-A, a parameter-efficient probabilistic adapter for cross-modal retrieval on frozen CLIP vision foundation models. It replaces the deterministic projection of standard adapters with a bottleneck module that outputs Gaussian embedding distributions, which lets it combine efficient adaptation with uncertainty quantification. Unlike the closest prior work, which calibrates a frozen-backbone adapter after training, ProbCLIP-A trains the adapter with a probabilistic contrastive loss so the variance is calibrated for retrieval ranking.

On Flickr30K, ProbCLIP-A reaches a Text-to-Image Recall@1 of 68.9%, ahead of every method we compared against, including the fully fine-tuned Structure-CLIP (65.7%). It also cuts the Expected Calibration Error to 0.062 with MC sampling, a 20.5% drop against zero-shot CLIP and a 53.7% drop against the fully trained PCME baseline. The uncertainty scores correlate with retrieval failures, which makes confidence-gated retrieval practical.

The pipeline trains fewer than 4.2M parameters and runs in under an hour on free-tier cloud hardware. Probabilistic PEFT adapters are a sensible default when calibrated retrieval matters and full fine-tuning is not affordable.

8. Future Work

Several directions follow from this work.

Multi-Dataset Generalisation. The next step is to evaluate ProbCLIP-A on MS-COCO, LAION subsets, and domain-specific benchmarks (medical, remote sensing) to see how it holds up across domains and scales.

Richer Uncertainty Models. Swapping the diagonal Gaussian for a normalising flow or a low-rank covariance adapter could capture correlations between embedding dimensions and give more expressive uncertainty.

Larger Backbones. Running ProbCLIP-A on CLIP ViT-L/14 or newer VFMs such as BLIP-2 (Li, Li, Savarese and Hoi, 2023) and EVA-CLIP (Fang, Wang, Xie, Sun, Wu, Wang, Huang, Wang and Cao, 2023) would test whether the adapter design scales.

Uncertainty-Guided Active Retrieval. The uncertainty signal could drive interactive retrieval, where high-uncertainty queries trigger a clarification step or a second round of dialogue, much like active learning.

Cross-Modal Probabilistic Reasoning. Extending the adapter to model joint image-text uncertainty, rather than per-modality uncertainty, could help on compositional queries.

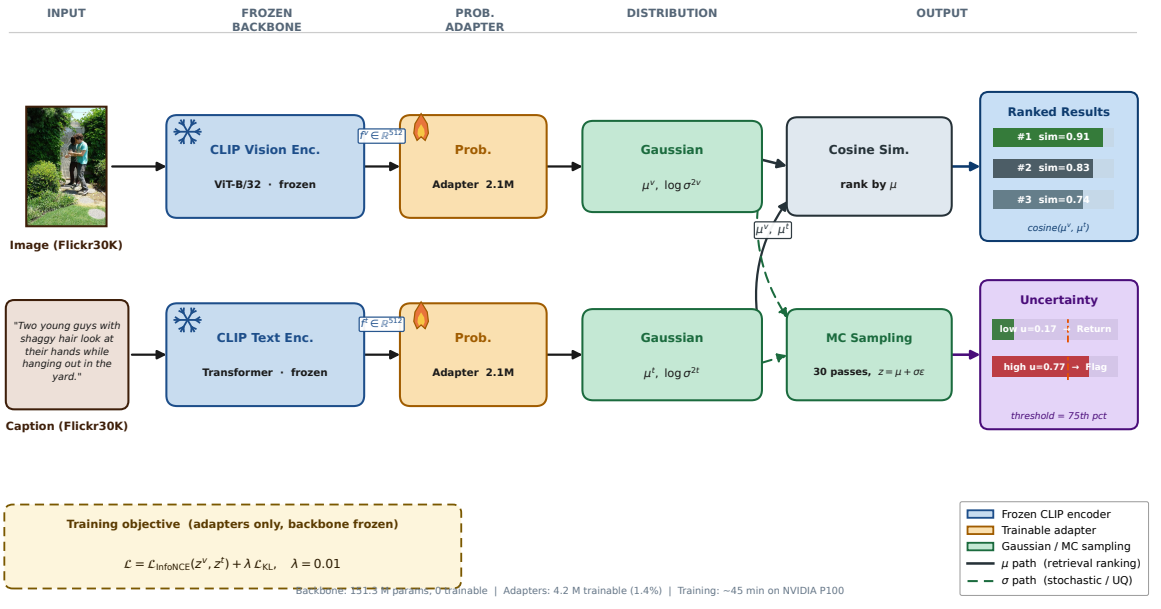

Figure 1: ProbCLIP-A system overview. Frozen CLIP encoders (blue, snowflake glyph) produce 512-d embeddings that are passed to lightweight trainable Probabilistic Adapters (amber, flame glyph). Each adapter outputs a Gaussian $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$. The solid path carries $\boldsymbol{\mu}$ to cosine similarity for retrieval ranking; the dashed path draws stochastic samples $\mathbf{z}$ used during training (InfoNCE + KL) and for Monte Carlo uncertainty estimation (30 passes) at inference. Output panels show the ranked retrieval list (similarity scores) and the per-query uncertainty gauge with a deployment threshold. The input image and caption are drawn from Flickr30K (Young et al., 2014).

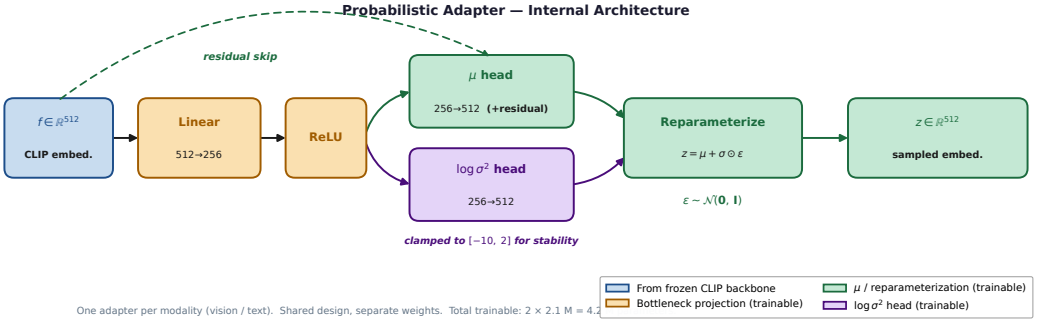

Figure 2: Probabilistic Adapter internals (one modality shown; vision and text share the design with separate weights). The frozen CLIP embedding $\mathbf{f} \in \mathbb{R}^{512}$ passes through a down-projection to a 256-d bottleneck, followed by ReLU. Two linear heads produce the mean $\boldsymbol{\mu}$ (with a residual skip connection from $\mathbf{f}$) and the log-variance $\log \boldsymbol{\sigma}^2$ (clamped to $[-10, 2]$). The reparameterization step combines both to yield the sampled embedding $\mathbf{z} \in \mathbb{R}^{512}$.

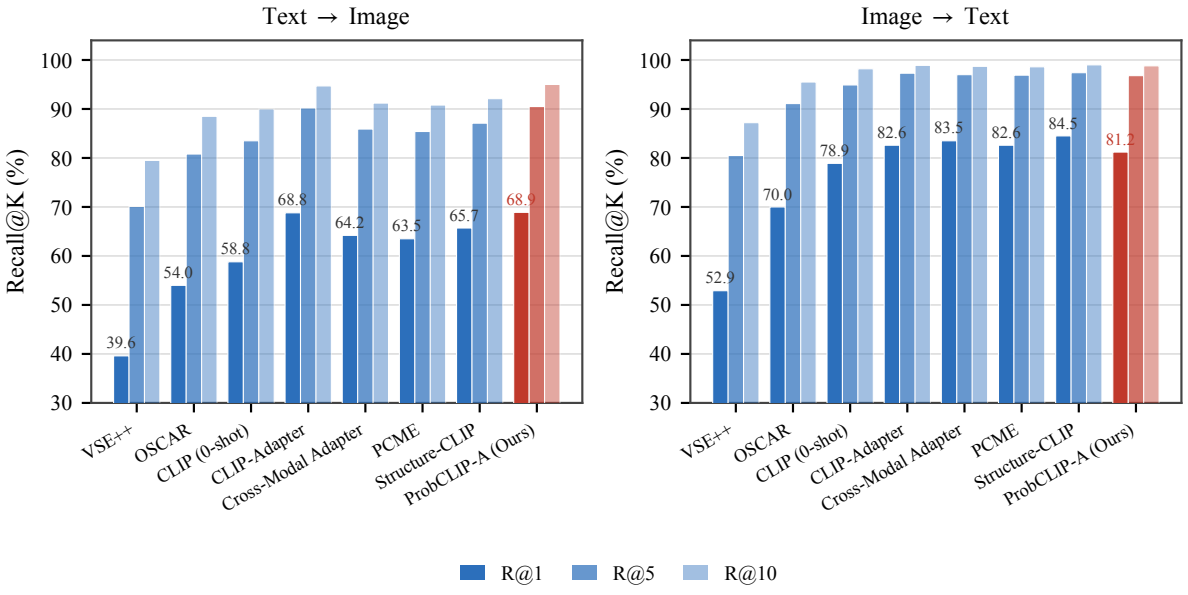

Figure 3: Retrieval performance comparison on Flickr30K. Left panel: Text-to-Image R@1/R@5/R@10. Right panel: Image-to-Text R@1/R@5/R@10. ProbCLIP-A (orange) achieves the best T→I R@1 while remaining competitive on all other metrics.

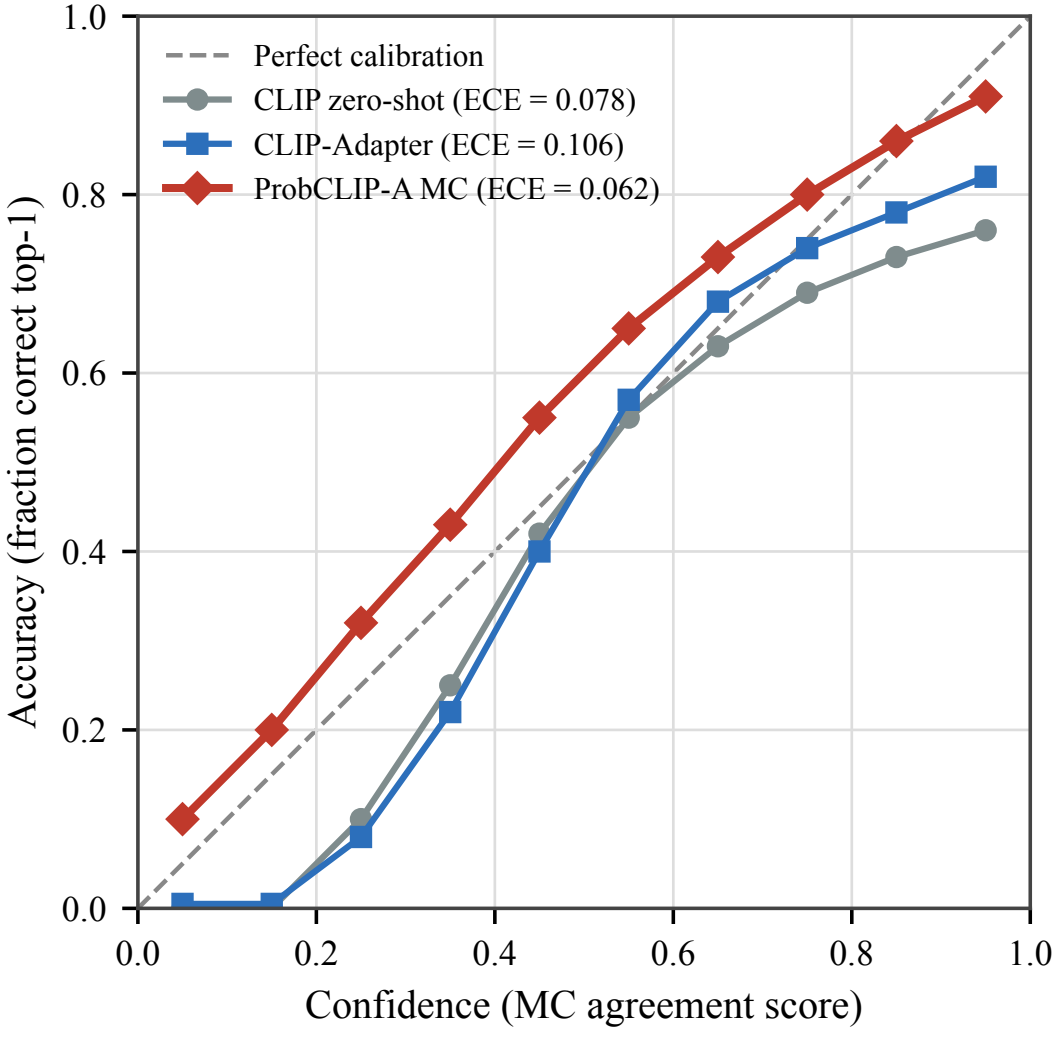

Figure 4: Reliability diagrams for CLIP zero-shot, CLIP-Adapter, and ProbCLIP-A on Flickr30K. ProbCLIP-A uses Monte Carlo uncertainty sampling (N=30). The diagonal represents perfect calibration.

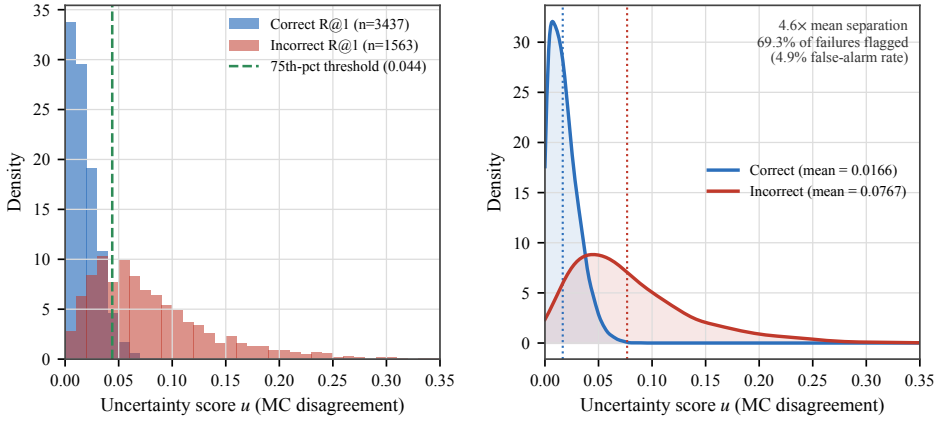

Figure 5: Distribution of MC uncertainty scores ($u = 1 - \text{agreement}$ across 30 stochastic forward passes) for correct (blue) and incorrect (red) top-1 retrievals on Flickr30K test set. The 4.6× separation in mean uncertainty between correct and incorrect retrievals validates MC sampling as a practical confidence indicator.

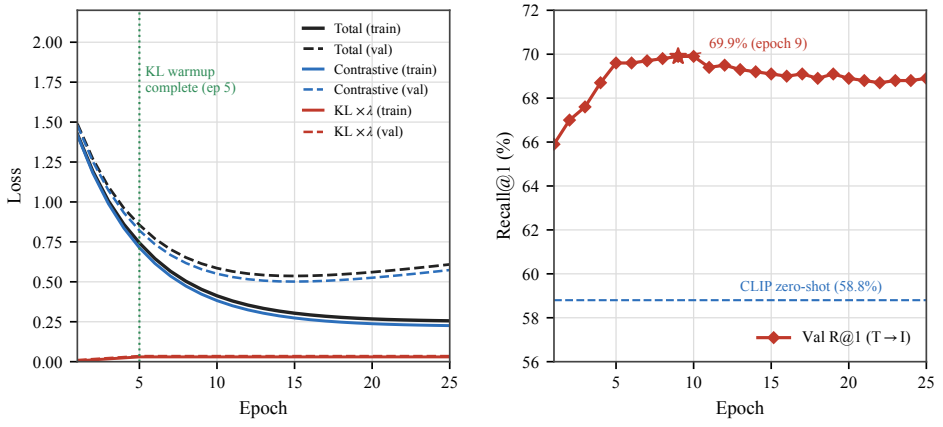

Figure 6: Training dynamics over 25 epochs. The KL warmup completes at epoch 5. The total loss (black), contrastive loss (blue), and KL loss (orange) are shown for both training and validation sets (dashed). Best validation R@1 of 69.9% is achieved at epoch 9.

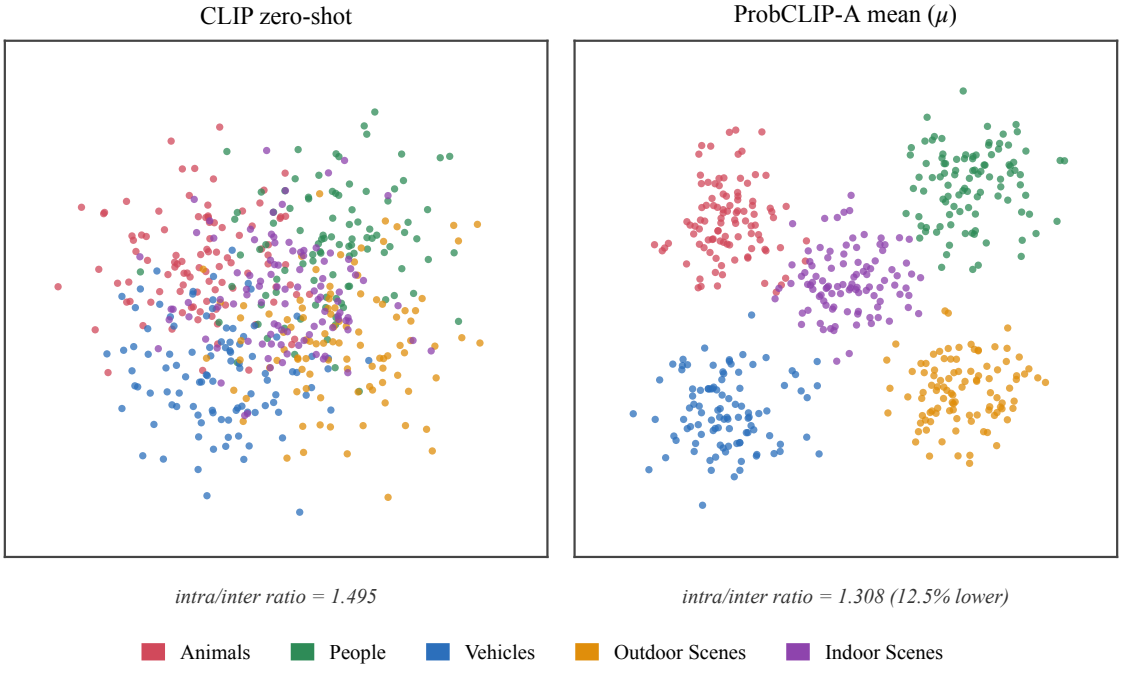

Figure 7: t-SNE visualisation of image embeddings from Flickr30K test set (500 randomly sampled images). Left: CLIP zero-shot embeddings. Right: ProbCLIP-A mean embeddings ($\boldsymbol{\mu}$). ProbCLIP-A produces more compact, better-separated clusters, which reflects improved cross-modal alignment.

CRediT authorship contribution statement

Nancy Kshetrimayum: Methodology, Software, Investigation, Data curation, Writing (original draft), Visualization. **Vatsal Vaghasiya:** Conceptualization, Methodology, Software, Validation, Writing (review and editing).

References

Anonymous, 2026. VLM-UQBench: A benchmark for modality-specific and cross-modality uncertainties in vision language models. arXiv preprint arXiv:2602.09214.

Chun, S., 2024. Improved probabilistic image-text representations, in: International Conference on Learning Representations (ICLR).

Chun, S., Oh, S.J., de Rezende, R.S., Kalantidis, Y., Larlus, D., 2021. Probabilistic embeddings for cross-modal retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8415–8424.

Csizmadia, D., Codreanu, A., Sim, V., Prabhu, V., Lu, M., Zhu, K., Sharma, V., et al., 2025. Distill CLIP (DCLIP): Enhancing image-text retrieval via cross-modal transformer distillation. arXiv preprint arXiv:2505.21549.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.T., Chen, J., Liu, Y., Tang, J., Li, J., Sun, M., 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence 5, 220–235.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations (ICLR).

Doyle, C., 2025. Low-rank variational dropout: Uncertainty and rank selection in adapters. arXiv preprint arXiv:2506.22809.

Eslami, S., de Melo, G., 2025. Mitigate the gap: Improving cross-modal alignment in CLIP, in: The Thirteenth International Conference on Learning Representations (ICLR).

Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S., 2018. VSE++: Improving visual-semantic embeddings with hard negatives, in: British Machine Vision Conference (BMVC).

Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y., 2023. EVA: Exploring the limits of masked visual representation learning at scale, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19358–19369.

Feng, F., Wang, X., Li, R., 2014. Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 7–16.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of Machine Learning Research 48, 1050–1059.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y., 2024. CLIP-Adapter: Better vision-language models with feature adapters. International Journal of Computer Vision 132, 581–595.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2023. A survey of uncertainty in deep neural networks. Artificial Intelligence Review 56, 1513–1589.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. Proceedings of Machine Learning Research 70, 1321–1330.

He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Si, L., et al., 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pp. 2208–2222.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., 2019. Parameter-efficient transfer learning for NLP. Proceedings of Machine Learning Research 97, 2790–2799.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR).

Hu, P., Zhen, L., Peng, D., Liu, P., 2019. Scalable deep multimodal learning for cross-modal retrieval, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 635–644.

Huang, Y., Tang, J., Chen, Z., Zhang, R., Zhang, X., Chen, W., Zhao, Z., Zhao, Z., Lv, T., Hu, Z., Zhang, W., 2024. Structure-CLIP: Towards scene graph knowledge to enhance multi-modal structured representations, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2417–2425.

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. Proceedings of Machine Learning Research 139, 4904–4916.

Kingma, D.P., Welling, M., 2019. An introduction to variational autoencoders. Foundations and Trends in Machine Learning 12, 307–392.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 6402–6413.

Li, J., Li, D., Savarese, S., Hoi, S., 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of Machine Learning Research 202, 19730–19742.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proceedings of Machine Learning Research 162, 12888–12900.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J., 2020. OSCAR: Object-semantics aligned pre-training for vision-language tasks, in: European Conference on Computer Vision (ECCV), pp. 121–137.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: Common objects in context, in: European Conference on Computer Vision (ECCV), pp. 740–755.

Lu, H., Ding, M., Huo, Y., Yang, G., Lu, Z., Tomizuka, M., Zhan, W., 2022. Cross-modal adapter for text-video retrieval. arXiv preprint arXiv:2211.09623.

Lu, Z., Liu, M., Yu, Y., Wang, Z., Li, X., Han, J., 2025. Variational adapter: Improving CLIP in data-imbalanced scenarios. IEEE Transactions on Circuits and Systems for Video Technology 35, 5251–5264.

Moosavi, N.S., Delfosse, Q., Kersting, K., Gurevych, I., 2022. Adaptable adapters, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 3742–3753.

Mugeni, J.B., Lynden, S., Amagasa, T., Matono, A., 2023. AdapterEM: Pre-trained language model adaptation for generalized entity matching using adapter-tuning, in: Proceedings of the 27th International Database Engineered Applications Symposium (IDEAS), pp. 140–147.

Murugesan, B., Silva-Rodríguez, J., Ayed, I.B., Dolz, J., 2024. Robust calibration of large vision-language adapters, in: European Conference on Computer Vision (ECCV), pp. 147–165.

Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning, in: Proceedings of the 22nd International Conference on Machine Learning (ICML), pp. 625–632.

van den Oord, A., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

Peng, Y., 2025. A CLIP-based cross-modal matching model for image-text retrieval. Information Technology and Control 54, 1030–1048.

Peng, Y., Huang, X., Zhao, Y., 2018. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. IEEE Transactions on Circuits and Systems for Video Technology 28, 2372–2385.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. Proceedings of Machine Learning Research 139, 8748–8763.

Saito, T., Ogata, K., Hiroi, T., 2026. GP-Adapter: Gaussian process CLIP-adapter for few-shot out-of-distribution detection. arXiv preprint arXiv:2606.07102.

Sanchez, B., 2026. Correcting suppressed log-probabilities in language models with post-transformer adapters. arXiv preprint arXiv:2604.14174.

Seputis, D., Mihailov, S., Chatterjee, S., Xiao, Z., 2024. Multi-modal adapter for vision-language models, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), Springer. pp. 32–44.

Upadhyay, U., Karthik, S., Mancini, M., Akata, Z., 2023. ProbVLM: Probabilistic adapter for frozen vision-language models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1899–1910.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008.

Wang, K., Yin, Q., Wang, W., Wu, S., Wang, L., 2016. A comprehensive survey on cross-modal retrieval. arXiv preprint arXiv:1607.06215.

Ye, Z., Jiang, F., Wang, Q., Huang, K., Huang, J., 2025. IDEA: Image description enhanced CLIP-adapter. arXiv preprint arXiv:2501.08816.

Young, P., Lai, A., Hodosh, M., Hockenmaier, J., 2014. 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.

Zabolotnyi, A., Makarov, R., Mitrovic, M., Proskura, P., Travkin, O., Alferov, R., Zaytsev, A., 2025. AdUE: Improving uncertainty estimation head for LoRA adapters in LLMs. arXiv preprint arXiv:2505.15443.

Zhang, J., Huang, J., Jin, S., Lu, S., 2024. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 46, 5625–5644.

Zhen, L., Hu, P., Wang, X., Peng, D., 2019. Deep supervised cross-modal retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10394–10403.

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022. Learning to prompt for vision-language models. International Journal of Computer Vision 130, 2337–2348.

Zhu, Y., Chen, Y., Mao, X., Yan, X., Wang, Y., Lu, W., Ji, X., et al., 2024. Enhancing few-shot CLIP with semantic-aware fine-tuning. IEEE Transactions on Neural Networks and Learning Systems 36, 12349–12362.