Post-Hoc Probabilistic Vision-Language Models

ICLR 2026

Dongyoon Kim

26.06.05

Motivation

Vision-Language Models (VLMs) such as CLIP and SigLIP:

- Map images & text to a *shared latent space*
- Evaluate similarity via **cosine similarity**
- Trained on billion-scale datasets

Key problem: deterministic mappings cannot capture uncertainty arising from domain shifts, ambiguous inputs, or distribution change.

Consequences:

- Overconfident predictions on incorrect classes
- Unreliable deployment in safety-critical settings
- Poor active-learning data selection

Related Work & Gaps

Existing Approaches

- **Calibration**[1]: no epistemic UQ
- **Prob. VLMs from scratch**[2]: expensive
- **Adapter fine-tuning**[3]: needs retraining
- **TTA**[4]: $\sim 80\times$ inference cost

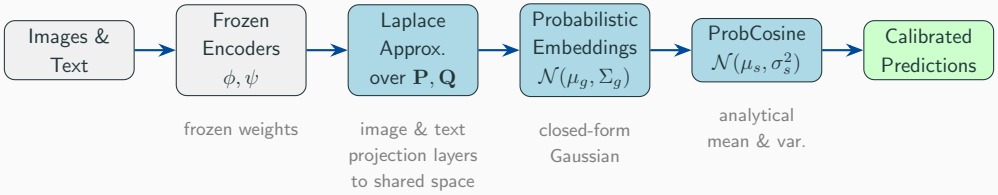
BayesVLM (Ours)

- Post-hoc **Laplace approximation**
- Analytical **ProbCosine** distribution
- No retraining, $< 5\%$ overhead
- Zero-shot & active learning ready

First direct Bayesian formulation of VLM cosine-similarity prediction.

[1] Guo et al., *On Calibration of Modern Neural Networks*, ICML 2017
[2] Chun et al., *Probabilistic Language-Image Pre-Training*, ICLR 2025
[3] Upadhyay et al., *ProbVLM: Probabilistic Adapter for Frozen VLMs*, ICCV 2023
[4] Farina et al., *Frustratingly Easy Test-Time Adaptation of VLMs*, NeurIPS 2024

Method Overview: From VLM to BayesVLM



(1) Frozen encoders
Laplace Approximation applied only to linear projection layers $\mathbf{P}, \mathbf{Q}$ — encoders untouched.

(2) Independent modalities
Separate posteriors per modality; cross-modal interactions preserved at MAP.

(3) No retraining
Post-hoc, training-free, $< 5\%$ runtime overhead vs. vanilla CLIP.

Laplace Approximation for VLM Projections (1/2)

Posterior Approximation

$$p(\theta \mid \mathcal{D}) \approx \mathcal{N}(\theta_{\mathrm{MAP}}, \boldsymbol{\Sigma})$$

where θ_{MAP} is the pre-trained model weights and $\boldsymbol{\Sigma}$ is the posterior covariance:

$$\boldsymbol{\Sigma} = \left(-\nabla^2_\theta \log p(\mathcal{D} \mid \theta) \big|_{\theta_{\mathrm{MAP}}} + \lambda \mathbf{I} \right)^{-1}$$

where λ is the prior precision learned by marginal likelihood maximisation.

Prior

$$p(\theta) = \mathcal{N}(0, \lambda^{-1}\mathbf{I})$$

Implicitly defined by the L2 regularisation (weight decay) used during CLIP pre-training.

Laplace Approximation for VLM Projections (2/2)

KFAC GGN Hessian over image projection $\mathbf{P}$

$$\mathbf{A}_{\text{IMG}} \otimes \mathbf{B}_{\text{IMG}}$$

$$\mathbf{A}_{\text{IMG}} = \tfrac{1}{\sqrt{n}} \sum_i \phi(x_i)\phi(x_i)^\top \qquad \mathbf{B}_{\text{IMG}} = \tfrac{1}{\sqrt{n}} \sum_i \mathbf{J}_{\text{IMG}}^\top \Lambda_{\text{IMG}} \mathbf{J}_{\text{IMG}}$$

where $\phi(x_i)$ are image encoder features, $\mathbf{J}_{\text{IMG}}$ is the Jacobian of the normalised embedding, and Λ_{IMG} is the loss Hessian with respect to the model output.

Resulting Probabilistic Image Embedding

$$p(\mathbf{g} \mid \mathcal{D}) = \mathcal{N}\Big(\mathbf{P}_{\text{MAP}}\phi(x),\ \Big[\phi(x)^\top \tilde{\mathbf{A}}^{-1}\phi(x)\Big]\,\tilde{\mathbf{B}}^{-1}\Big)$$

where $\tilde{\mathbf{A}}, \tilde{\mathbf{B}}$ are the regularised Kronecker factors. The same applies to the text projection $\mathbf{Q}$ and text embedding $\mathbf{h}$.

ProbCosine: Analytical Uncertainty (1/2)

Given $\mathbf{g} \sim \mathcal{N}(\mu_g, \Sigma_g)$ and $\mathbf{h} \sim \mathcal{N}(\mu_h, \Sigma_h)$ with diagonal covariances:

Expected Cosine Similarity

$$\mathbb{E}[S_{\text{COS}}(\mathbf{g}, \mathbf{h})] \approx \frac{\sum_i \mu_{g,i}\, \mu_{h,i}}{\sqrt{\sum_i (\mu_{g,i}^2 + \sigma_{g,i}^2)}\sqrt{\sum_i (\mu_{h,i}^2 + \sigma_{h,i}^2)}}$$

Uses $\mathbb{E}[x^2] = \mu_x^2 + \sigma_x^2$ and the triangle inequality.

Variance of Cosine Similarity

$$\text{Var}[S_{\text{COS}}(\mathbf{g}, \mathbf{h})] = \frac{\sum_i \sigma_{g,i}^2(\sigma_{h,i}^2 + \mu_{h,i}^2) + \sigma_{h,i}^2\mu_{g,i}^2}{\left(\sum_i \mu_{g,i}^2 + \sigma_{g,i}^2\right)\left(\sum_i \mu_{h,i}^2 + \sigma_{h,i}^2\right)}$$

ProbCosine: Analytical Uncertainty (2/2)

Gaussian Approximation of Cosine Similarity

$$p(S_{\mathrm{COS}}(\mathbf{g}, \mathbf{h})) \approx \mathcal{N}(\mathbb{E}[S_{\mathrm{COS}}],\, \mathrm{Var}[S_{\mathrm{COS}}])$$

Final Prediction via Probit Approximation

$$p(y \mid x) = \mathrm{softmax}\left(\frac{\sqrt{t}\, \mathbb{E}[S_{\mathrm{COS}}]}{\sqrt{1 + \frac{\pi}{8} t^2\, \mathrm{Var}[S_{\mathrm{COS}}]}} \right)$$

where t is the learnable temperature parameter learned during CLIP pre-training.

Experiment I: Zero-Shot Uncertainty Quantification

Setup: OpenCLIP ViT-B/32 on 7 benchmarks. Metrics: ACC (%, ↑), NLPD (↓), ECE (%, ↓).

Metric	Method	Flowers	Food	CIFAR-10	CIFAR-100	ImgNet-R	UCF101	SUN397
ACC ↑	CLIP	68.99	80.21	93.61	73.76	74.52	59.82	67.18
	Temp. scaling	68.99	80.21	93.61	73.76	74.52	59.82	67.18
	TTA	68.87	81.68	88.54	65.64	78.29	63.07	68.58
	BayesVLM	68.87	80.43	**93.62**	73.63	74.45	61.43	66.96
NLPD ↓	CLIP	1.90	0.70	0.21	0.97	1.07	1.59	1.16
	Temp. scaling	1.67	0.69	0.21	0.94	1.04	1.46	1.11
	TTA	1.86	0.67	0.35	1.26	0.90	1.50	1.14
	BayesVLM	**1.73**	**0.68**	**0.20**	**0.95**	**1.03**	**1.44**	**1.12**
ECE ↓	CLIP	6.59	3.91	1.45	6.31	5.20	11.52	8.71
	Temp. scaling	5.51	4.74	1.88	3.07	4.80	3.61	2.67
	TTA	9.63	4.18	2.02	5.27	2.88	11.75	9.92
	BayesVLM	**4.22**	**1.69**	**0.72**	**1.92**	**1.78**	**3.57**	**2.06**

Experiment II: Active Learning

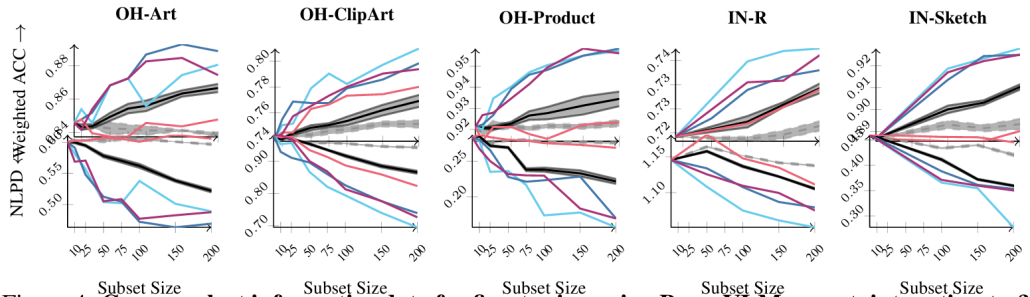


Figure 4: **Can we select informative data for fine-tuning using BayesVLM uncertainty estimates? Yes.** On the OfficeHome data set (OH) and ImageNet variants (IN), when using uncertainty-based scores (EPIG (——) and BALD (——)) to select the fine-tuning data, we achieve better performance compared with Entropy (targeted) (——), Entropy (——), Random selection (targeted) (——), and Random selection (- - -). Thus, highlighting the benefits of using uncertainties from BayesVLM.