Probabilistic Embeddings for Frozen Vision-Language Models: Uncertainty Quantification with Gaussian Process Latent Variable Models

Aishwarya Venkataramanan[1] **Paul Bodesheim**[1] **Joachim Denzler**[1]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany

Abstract

Vision-Language Models (VLMs) learn joint representations by mapping images and text into a shared latent space. However, recent research highlights that deterministic embeddings from standard VLMs often struggle to capture the uncertainties arising from the ambiguities in visual and textual descriptions and the multiple possible correspondences between images and texts. Existing approaches tackle this by learning probabilistic embeddings during VLM training, which demands large datasets and does not leverage the powerful representations already learned by large-scale VLMs like CLIP. In this paper, we propose GroVE, a post-hoc approach to obtaining probabilistic embeddings from frozen VLMs. GroVE builds on Gaussian Process Latent Variable Model (GPLVM) to learn a shared low-dimensional latent space where image and text inputs are mapped to a unified representation, optimized through single-modal embedding reconstruction and cross-modal alignment objectives. Once trained, the Gaussian Process model generates uncertainty-aware probabilistic embeddings. Evaluation shows that GroVE achieves state-of-the-art uncertainty calibration across multiple downstream tasks, including cross-modal retrieval, visual question answering, and active learning.

1 INTRODUCTION

Deep learning has seen remarkable success over the last decade, yet its practical applicability, especially in safety-critical areas is limited by unreliable, overconfident predictions [Abdar et al., 2021].

This has motivated the development of methods to quantify uncertainty in model predictions, including stochastic [Blundell et al., 2015, Gal and Ghahramani, 2016], deterministic [Van Amersfoort et al., 2020, Mukhoti et al., 2023, Venkataramanan et al., 2023a], evidential [Sensoy et al., 2018], and post-hoc approaches [Corbiere et al., 2021], with the aim to produce calibrated confidence values that better reflect the model's actual performance [Guo et al., 2017]. While these methods have shown strong performance in tasks involving data from a single modality, they often struggle in multi-modal settings, such as vision language models (VLMs), where inputs come from different domains, such as images and text [Jung et al., 2022]. The challenge arises because these single-modal approaches fail to capture the uncertainties that emerge from interactions between the different modalities.

VLMs typically encode images and their corresponding text descriptions into vector representations within a joint embedding space. While combining modalities enriches semantics and boosts performance on various tasks [Zhang et al., 2024, Venkataramanan et al., 2023b], it also introduces additional uncertainties. Beyond the inherent uncertainty of each modality, there is an uncertainty due to the ambiguous relationships between images and text. This is illustrated in Figure 1, where each image can correspond to multiple text descriptions, and each text description can be associated with multiple images. Deterministic embeddings from VLMs often fail to capture these uncertainties, motivating the development of probabilistic embeddings [Ji et al., 2023, Chun et al., 2021, Chun, 2023]. Probabilistic embeddings represent a distribution, thereby capturing a range of possible representations for ambiguous or uncertain data. Typically, the embeddings are modeled as Gaussian distributions, and deep neural networks are trained to maximize their likelihood, learning the distribution parameters. However, these methods require training the VLMs from scratch, which requires large-scale datasets, and does not effectively leverage the strong multi-modal representations already provided by the pre-trained large-scale VLMs [Radford et al., 2021, Li et al., 2022b, Singh et al., 2022].

In this work, we introduce GroVE, a method to generate

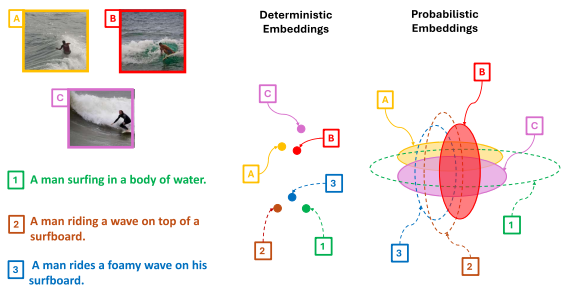

Figure 1: Illustration of uncertainty arising from multiple correspondences between image and text descriptions. Deterministic embeddings represent the instances as fixed points. In contrast, probabilistic embeddings capture uncertainty by modeling text and images as distributions, allowing for multiple reasonable matches.

probabilistic embeddings for VLMs in a post-hoc manner that builds on Gaussian Process Latent Variable Model (GPLVM) [Lawrence, 2003]. GroVE stands for Gaussian Process for Probabilistic VLM Embeddings. A GPLVM models the relationship between a low-dimensional latent space and a high-dimensional observational space using Gaussian Processes (GPs). Traditionally, the latent space is used for dimensionality reduction [Lawrence, 2003, Lalchand et al., 2022], and less commonly for more task-specific applications, such as classification [Eleftheriadis et al., 2014] and cross-modal retrieval [Song et al., 2017]. In our approach, we adopt an extension of the GPLVM framework to a multi-modal context, and show that it provides a principled approach for obtaining probabilistic image and text embeddings from the deterministic embeddings of the large-scale frozen VLMs. To achieve this, we learn a joint low-dimensional latent space, where each pair of image and text embeddings derived from a VLM is represented as a single unified point. The mapping between the latent space and the observed VLM embeddings is established through two GPs: one for image embeddings and one for text embeddings. Our training objective consists of an embedding reconstruction loss to learn this mapping, and a cross-modal alignment that regularizes the latent space to preserve the semantic structure of the data. Once the latent space is learned, the trained GP models are used to obtain probabilistic embeddings for images and texts.

We evaluate GroVE for uncertainty calibration in cross-modal retrieval using CLIP [Radford et al., 2021] and BLIP [Li et al., 2022b] on the following standard benchmarks: common objects datasets MS-COCO [Chen et al., 2015] and Flickr30k[Young et al., 2014], as well as fine-grained datasets CUB-200-2011 [Wah et al., 2011] and Oxford Flowers 102 [Nilsback and Zisserman, 2008]. We further demonstrate the applicability of our approach in an active learning setting. We also evaluate its ability to provide calibrated uncertainty estimates in visual question answering (VQA) using the VQA 2.0 dataset [Goyal et al., 2017]. Our results show that GroVE effectively learns probabilistic embeddings that provide calibrated uncertainty estimates.

The contributions are summarized as follows: i) We propose GroVE, which extends GPLVM and provides a principled approach to obtain probabilistic VLM embeddings for both images and text. ii) We show that GroVE produces calibrated uncertainty estimates for cross-modal retrieval and VQA, and demonstrate its practical utility in active learning. iii) We design GroVE to work in a post-hoc manner on frozen VLMs, avoiding the need for retraining large-scale models from scratch. Code is available: `https://github.com/cvjena/GroVE-Probabilistic_VLM_embeddings.git`

2 RELATED WORK

Uncertainty Quantification in VLMs. Input data ambiguities in VLMs are often addressed by replacing traditional deterministic embeddings with probabilistic embeddings Li et al. [2022a]. PCME [Chun et al., 2021] models image and text embeddings as Gaussians with learned means and variances, optimizing the joint embedding space with a soft cross-modal contrastive loss. PCME++ [Chun, 2023] introduces Closed-Form Sampled Distance (CSD) to compute Gaussian embeddings of images and text for faster uncertainty estimation compared to PCME. MAP [Ji et al., 2023] introduces a Probability Distribution Encoder to model multi-modal representations as probabilistic distributions. However, all these methods require training from scratch, and do not effectively leverage the strong multi-modal representations already learned by the pre-trained large-scale VLMs. ProbVLM [Upadhyay et al., 2023] is a post-hoc approach that trains neural networks to estimate the parameters of Generalized Gaussian distribution for image and text embeddings. Although being straightforward, the prediction of distribution parameters lacks proper probabilistic modeling of statistical processes underlying the sampling of data. Furthermore, neural networks are prone to uncalibrated predictions when presented with out-of-distribution (OOD) data or limited training samples [Guo et al., 2017]. In contrast, our approach leverages GPs, a Bayesian method that inherently incorporates probabilistic reasoning with reliable and theoretically sound uncertainty quantification as well as distance-awareness through the covariance function, which has proven effective in calibrated uncertainty estimation [Liu et al., 2020, Jung et al., 2022].

Post-hoc approaches for uncertainty quantification. Some of the widely used post-hoc calibration techniques for data from a single modality are temperature scaling [Guo

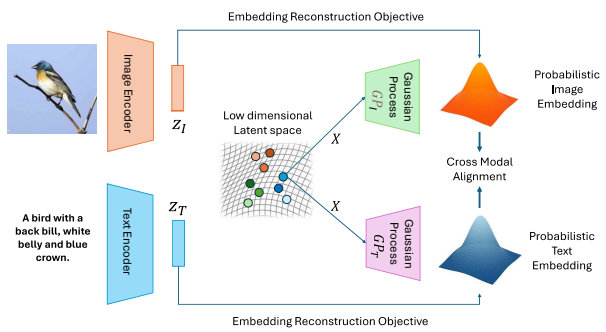

Figure 2: **Method overview of GroVE.** Given deterministic image and text embeddings from a frozen VLM, GroVE learns a joint low dimensional latent space, where each image-text pair is represented by a single point. Two GP models learn to reconstruct the image and text embeddings from the latent space points through single-modal reconstruction and cross-modal alignment objectives. The GP models act as probabilistic mappings that model the uncertainty in both the image and text modalities.

et al., 2017] and Platt scaling [Platt et al., 1999], which adjust the model's predicted probabilities after training to better align predicted confidence scores with actual performance. Test-Time Data Augmentation (TTDA) [Ayhan and Berens, 2018, Wang et al., 2019] quantifies uncertainty by applying various transformations to input data during inference, generating multiple predictions, and measuring the variability among them to assess the uncertainty. A line of work [Corbiere et al., 2021, Yu et al., 2021, Hornauer et al., 2023, Shi and Jain, 2019] focuses on training auxiliary models to quantify uncertainty in the primary model, allowing for uncertainty estimation without impacting the performance of the primary model. Unlike these single-modal approaches, our method captures uncertainty from the relationship between visual and textual modalities, which is crucial for obtaining accurate uncertainty estimates in VLMs [Jung et al., 2022].

3 METHOD

GroVE builds on the GPLVM framework to learn a shared latent space for image and text inputs using GPs. It optimizes this space through single-modal reconstruction and cross-modal alignment loss, generating probabilistic embeddings from deterministic VLM embeddings to capture uncertainty. Figure 2 illustrates the overall pipeline of GroVE.

3.1 PROBLEM DESCRIPTION

Let $\mathcal{D} = \{(I_n, T_n)\}_{n=1}^{N} \subset \mathcal{I} \times \mathcal{T}$ represent a dataset of N paired samples, where $I_n \in \mathcal{I}$ is an image sampled from the image space $\mathcal{I}$, and $T_n \in \mathcal{T}$ is the corresponding text description sampled from the text space $\mathcal{T}$. The VLM maps an image I and a text T into a shared embedding space $\mathcal{Z} \subseteq \mathbb{R}^D$. To achieve this, the VLM consists of an image encoder $f_{\mathcal{I}}^{\theta_{\mathcal{I}}} : \mathcal{I} \to \mathcal{Z}$ with parameters $\theta_{\mathcal{I}}$, and a text encoder $f_{\mathcal{T}}^{\theta_{\mathcal{T}}} : \mathcal{T} \to \mathcal{Z}$ with parameters $\theta_{\mathcal{T}}$. We assume that the VLM has already been trained on a large-scale dataset, and the parameters of $f_{\mathcal{I}}^{\theta_{\mathcal{I}}}$ and $f_{\mathcal{T}}^{\theta_{\mathcal{T}}}$ are fixed as $\theta_{\mathcal{I}}^*$ and $\theta_{\mathcal{T}}^*$, respectively. The encoders have been trained such that, for a given image-text pair (I, T), the resulting embeddings $\mathbf{z}_I = f_{\mathcal{I}}^{\theta_{\mathcal{I}}^*}(I)$ and $\mathbf{z}_T = f_{\mathcal{T}}^{\theta_{\mathcal{T}}^*}(T)$ are positioned close to one another in $\mathcal{Z}$, so that semantically related visual and textual information is aligned. While deterministic VLMs provide fixed embeddings, they lack the ability to represent the uncertainty associated with these embeddings. To address this, we propose GroVE, a method that leverages GPLVM to obtain probabilistic embeddings in a post-hoc manner to model the uncertainties.

3.2 GROVE MODEL

We obtain the image and text embeddings from the frozen VLM on $\mathcal{D}$:

$$\left\{\left(\mathbf{z}_{I_n}, \mathbf{z}_{T_n}\right) = \left(f_{\mathcal{I}}^{\theta_{\mathcal{I}}^*}(I_n), f_{\mathcal{T}}^{\theta_{\mathcal{T}}^*}(T_n)\right)\right\}_{n=1}^{N}, \quad (1)$$

where $\mathbf{z}_{I_n}, \mathbf{z}_{T_n} \in \mathbb{R}^D$ are D-dimensional image and text embeddings, respectively.

To derive probabilistic embeddings using GPLVM, we assume that $\mathbf{z}_{I_n}$ and $\mathbf{z}_{T_n}$ are generated from a shared low-dimensional latent space $\mathcal{X} \subseteq \mathbb{R}^Q$ with $Q \ll D$, where each image-text pair $(\mathbf{z}_{I_n}, \mathbf{z}_{T_n})$ is associated with a common latent point $\mathbf{x}_n \in \mathbb{R}^Q$. We define two GPLVM models $\mathcal{GP}_{\mathcal{I}}$ and $\mathcal{GP}_{\mathcal{T}}$, one for each modality (images from $\mathcal{I}$ and text from $\mathcal{T}$), to learn the mappings $G_{\mathcal{I}} : \mathcal{X} \to \mathcal{Z}_{\mathcal{I}}$ and $G_{\mathcal{T}} : \mathcal{X} \to \mathcal{Z}_{\mathcal{T}}$ from $\mathbf{x}_n$ to the high-dimensional embeddings $\mathbf{z}_{I_n}$ and $\mathbf{z}_{T_n}$, respectively. During GP model training, the latent points $\mathbf{x}_n$ are optimized to maximize the likelihood of the observed embeddings $\mathbf{z}_{I_n}$ and $\mathbf{z}_{T_n}$

GP model definitions. For describing the GPLVM models, we define the matrix $\mathbf{X} \in \mathbb{R}^{N \times Q}$ as the collection of the N latent inputs $\mathbf{x}_n$. Image embeddings $\mathbf{z}_{I_n}$ and text embeddings $\mathbf{z}_{T_n}$ are supposed to be computed from latent functions $G_{\mathcal{I}}$ and $G_{\mathcal{T}}$:

$$\mathbf{z}_{I_n} = G_{\mathcal{I}}(\mathbf{x}_n) + \boldsymbol{\epsilon}_{\mathcal{I}}; \quad \mathbf{z}_{T_n} = G_{\mathcal{T}}(\mathbf{x}_n) + \boldsymbol{\epsilon}_{\mathcal{T}}, \quad (2)$$

with noise terms $\boldsymbol{\epsilon}_{\mathcal{I}}$ and $\boldsymbol{\epsilon}_{\mathcal{T}}$ and a GP prior such that for each dimension d of the embeddings $\mathbf{z}_{I_n}$ and $\mathbf{z}_{T_n}$, the latent function values $\mathbf{g}_I^d, \mathbf{g}_T^d \in \mathbb{R}^N$ of the N samples follow a

multivariate Gaussian distribution:

$$\begin{aligned} \mathbf{g}_{\mathcal{I}}^d &\sim \mathcal{N}(m_{\mathcal{I}}(\mathbf{X}), k_{\mathcal{I}}(\mathbf{X}, \mathbf{X})), \\ \mathbf{g}_{\mathcal{T}}^d &\sim \mathcal{N}(m_{\mathcal{T}}(\mathbf{X}), k_{\mathcal{T}}(\mathbf{X}, \mathbf{X})). \end{aligned} \tag{3}$$

These distributions are parameterized by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$, which defines the covariance matrix between pairs of points in $\mathbf{X}$. For both GPLVM models, we use a constant mean function $m(\mathbf{X}) = \mathbf{m}$ and a radial basis function (RBF) kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\ell^2}\right)$ with length-scale hyperparameter ℓ. However, optimal values for $\mathbf{m}$ and ℓ are learnt separately for each modality $\mathcal{I}$ and $\mathcal{T}$. The likelihood functions are defined as:

$$\begin{aligned} p\big(\mathbf{z}_{\mathcal{I}}^d | \mathbf{g}_{\mathcal{I}}^d\big) &= \prod_{n=1}^{N} p(\mathbf{z}_{I_n}^d | \mathbf{g}_{I_n}^d) = \mathcal{N}\big(\mathbf{g}_{\mathcal{I}}^d, \sigma_{\mathcal{I}}^2 \mathbf{I}\big), \\ p\big(\mathbf{z}_{\mathcal{T}}^d | \mathbf{g}_{\mathcal{T}}^d\big) &= \prod_{n=1}^{N} p(\mathbf{z}_{T_n}^d | \mathbf{g}_{T_n}^d) = \mathcal{N}\big(\mathbf{g}_{\mathcal{T}}^d, \sigma_{\mathcal{T}}^2 \mathbf{I}\big), \end{aligned} \tag{4}$$

where $\sigma_{\mathcal{I}}^2, \sigma_{\mathcal{T}}^2$ are the parameters of the Gaussian noise model, which are learned along with the model parameters during the training.

Embedding Reconstruction Objective. Given the prior and the likelihood, our goal is to estimate the posterior distribution. While the exact inference is possible, it is computationally expensive, with cost $\mathcal{O}(N^3)$. In this work, we adopt a sparse GP with inducing points and variational inference Titsias [2009]. We introduce M inducing points in $\mathcal{X}$ for each modality $\mathcal{I}$ and $\mathcal{T}$, where $M \ll N$. Each inducing point corresponds to an inducing variable, represented as the latent function values $\mathbf{u}_{\mathcal{I}}^d \in \mathbb{R}^M$ and $\mathbf{u}_{\mathcal{T}}^d \in \mathbb{R}^M$, which capture the latent function values at these locations. The key idea is to approximate the true posterior distribution over the latent function values at the observed data points by conditioning on the inducing variables. This reduces the computational complexity of the model to $\mathcal{O}(NM^2)$.

To achieve this, we introduce a variational distribution over the inducing variables as:

$$q(\mathbf{u}_{\mathcal{I}}^d) = \mathcal{N}(\mathbf{u}_{\mathcal{I}}^d | \boldsymbol{\mu}_{\mathcal{I}}^d, \mathbf{S}_{\mathcal{I}}^d); \; q(\mathbf{u}_{\mathcal{T}}^d) = \mathcal{N}(\mathbf{u}_{\mathcal{T}}^d | \boldsymbol{\mu}_{\mathcal{T}}^d, \mathbf{S}_{\mathcal{T}}^d), \tag{5}$$

where $\boldsymbol{\mu}_{\mathcal{I}}^d$ and $\boldsymbol{\mu}_{\mathcal{T}}^d$, $\mathbf{S}_{\mathcal{I}}^d$ and $\mathbf{S}_{\mathcal{T}}^d$ are variational parameters that are optimized during training. These variational parameters, the inducing points, along with the model parameters $\mathbf{m}_{\mathcal{I}}, \mathbf{m}_{\mathcal{T}}, l_{\mathcal{I}}, l_{\mathcal{T}}, \sigma_{\mathcal{I}}^2$ and $\sigma_{\mathcal{T}}^2$ are learned by maximizing the lower bound on the marginal likelihood of the data i.e. the evidence lower bound (ELBO), given by

$$\begin{aligned} \mathcal{L}_{ELBO}^d = \mathbb{E}_{q(\mathbf{g}_{\mathcal{I}}^d)}[\log p(\mathbf{z}_{\mathcal{I}}^d | \mathbf{g}_{\mathcal{I}}^d)] - D_{KL}(q(\mathbf{u}_{\mathcal{I}}^d) || p(\mathbf{u}_{\mathcal{I}}^d)) \\ + \mathbb{E}_{q(\mathbf{g}_{\mathcal{T}}^d)}[\log p(\mathbf{z}_{\mathcal{T}}^d | \mathbf{g}_{\mathcal{T}}^d)] - D_{KL}(q(\mathbf{u}_{\mathcal{T}}^d) || p(\mathbf{u}_{\mathcal{T}}^d)), \end{aligned} \tag{6}$$

where D_{KL} is the Kullback-Leibler (KL) divergence, and is measured between the variational distributions and their corresponding priors obtained by the GP prior evaluated at the inducing points. The embedding reconstruction objective is given by:

$$\mathcal{L}_{emb} = -\sum_{d=1}^{D} \mathcal{L}_{ELBO}^d \tag{7}$$

Cross-modal Alignment Objective. In addition to this reconstruction objective, we introduce a regularization term, so that the predicted distributions of the corresponding image and text embeddings from the GPs match. Aligning these distributions encourages the latent space to learn a shared underlying structure between the modalities, so that semantically related data points are represented by similar latent variables. To enforce this, we define a KL divergence loss function between the distributions of the image and text embeddings from the GP models, which take the forms $\mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{I}}, \hat{\boldsymbol{\Sigma}}_{\mathcal{I}})$ and $\mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{T}}, \hat{\boldsymbol{\Sigma}}_{\mathcal{T}})$ respectively (refer Sec. 3.3 for inference using GP).

The resulting objective $\mathcal{L}_{KL}$ is the mean of the KL divergence in both directions (image-to-text and text-to-image):

$$\begin{aligned} \mathcal{L}_{KL} = \frac{1}{2}[D_{KL}(\mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{I}}, \hat{\boldsymbol{\Sigma}}_{\mathcal{I}}) \| \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{T}}, \hat{\boldsymbol{\Sigma}}_{\mathcal{T}})) + \\ D_{KL}(\mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{T}}, \hat{\boldsymbol{\Sigma}}_{\mathcal{T}}) \| \mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{I}}, \hat{\boldsymbol{\Sigma}}_{\mathcal{I}}))]. \end{aligned} \tag{8}$$

Final Objective. The overall objective function is the weighted sum of the embedding reconstruction loss and the cross-modal alignment loss:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{emb} + \lambda_2 \mathcal{L}_{KL} \tag{9}$$

where λ_1 and λ_2 are trade-off parameters.

3.3 PROBABILISTIC EMBEDDINGS

Once the latent space representation $\mathbf{X}$ is learned, we use $\mathcal{GP}_{\mathcal{I}}$ and $\mathcal{GP}_{\mathcal{T}}$ to predict the probabilistic image and text embeddings. Given a new embedding $\mathbf{z}_*$ (image or text) obtained from the VLM, we first infer its latent representation $\mathbf{x}_*$ by randomly initializing $\mathbf{x}_*$ and iteratively optimizing it with the ELBO. This approximates the posterior distribution $p(\mathbf{x}_* | \mathbf{z}_*, \mathbf{z}_{\mathcal{M}})$, where $\mathcal{M}$ denotes the modality (either $\mathcal{I}$ or $\mathcal{T}$).

From $\mathbf{x}_*$, the probabilistic embedding can be inferred using the respective GP.

Inference using GP. The predictive distribution, which defines the predicted probabilistic embedding is given by:

$$p(\mathbf{g}_*^d) = \int p(\mathbf{g}_*^d | \mathbf{u}_{\mathcal{M}}^d) q(\mathbf{u}_{\mathcal{M}}^d) d\mathbf{u}_{\mathcal{M}}^d \tag{10}$$

Evaluating the integral results in a Gaussian distribution Hensman et al. [2015]:

$$p(\mathbf{g}_*^d) = \mathcal{N}(\mathbf{g}_*^d | \hat{\boldsymbol{\mu}}_*^d, \tilde{\boldsymbol{\Sigma}}_*^d) \tag{11}$$

where the mean $\hat{\boldsymbol{\mu}}_*^d$ and covariance $\hat{\boldsymbol{\Sigma}}_*^d$ of the embedding is:

$$\hat{\boldsymbol{\mu}}_*^d = \mathbf{m}_{\mathcal{M}} + \mathbf{A}(\boldsymbol{\mu}_{\mathcal{M}}^d - \mathbf{m}_{\mathbf{v}_{\mathcal{M}}}) \tag{12}$$

$$\hat{\boldsymbol{\Sigma}}_*^d = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{A}(\mathbf{S}_{\mathcal{M}}^d - k(\mathbf{v}_{\mathcal{M}}, \mathbf{v}_{\mathcal{M}}))\mathbf{A}^T, \tag{13}$$

where $\mathbf{v}_{\mathcal{M}}$ refers to the inducing points of the respective modality, $\mathbf{A} = k(\mathbf{x}_*, \mathbf{v}_{\mathcal{M}})k(\mathbf{v}_{\mathcal{M}}, \mathbf{v}_{\mathcal{M}})^{-1}$, with dimensions $k(\mathbf{v}_{\mathcal{M}}, \mathbf{v}_{\mathcal{M}}) \in \mathbb{R}^{M \times M}$ and $k(\mathbf{x}_*, \mathbf{v}_{\mathcal{M}}) \in \mathbb{R}^M$ and $\mathbf{m}_{\mathbf{v}_{\mathcal{M}}}$ is the prior mean evaluated at $\mathbf{v}_{\mathcal{M}}$.

Uncertainty Quantification. When an embedding $\mathbf{z}_*$ belongs to an ambiguous input, the uncertainty associated with the posterior distribution $p(\mathbf{x}_*|\mathbf{z}_*, \mathbf{z}_{\mathcal{M}})$ increases. This uncertainty is propagated to the predictive distribution, which can be written as: $p(\mathbf{g}_*) = \int p(\mathbf{g}_*|\mathbf{x}_*)p(\mathbf{x}_*|\mathbf{z}_*, \mathbf{z}_{\mathcal{M}})d\mathbf{x}_*$. Here, $p(\mathbf{g}_*|\mathbf{x}_*)$ is a multivariate Gaussian distribution that describes the function values at the fixed point $\mathbf{x}_*$. Thus, a large uncertainty in $\mathbf{x}_*$ increases the variance of the predictive distribution $p(\mathbf{g}_*|\mathbf{z}_*, \mathbf{z}_{\mathcal{M}})$. The uncertainty is captured by $\hat{\boldsymbol{\Sigma}}_*$, which accounts for variance contributions from both the latent space uncertainty and inherent noise in the model's predictions. The final uncertainty is obtained by averaging the uncertainty values across all dimensions in $\hat{\boldsymbol{\Sigma}}_*$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Baselines and Datasets. We evaluate GroVE against six baseline methods: Deterministic, TTDA [Ayhan and Berens, 2018], PFE [Shi and Jain, 2019], PCME [Chun et al., 2021], PCME++ [Chun, 2023], and ProbVLM [Upadhyay et al., 2023], using two VLMs—CLIP [Radford et al., 2021] and BLIP [Li et al., 2022b]—with a focus on uncertainty calibration for downstream tasks. In the deterministic approach, uncertainty is quantified by the cosine distance between the image and text embeddings derived from the VLM. While PFE, PCME and PCME++ are methods to learn probabilistic embeddings for pre-training VLMs, we follow Upadhyay et al. [2023], and adapt them to work in a post-hoc manner. The similarity ranking between probabilistic image and text embeddings is determined by the Wasserstein distance, with embeddings ranked based on the increasing distance, while for the deterministic embeddings, the cosine similarity is used. The implementation details for these methods are provided in Appendix A.2. The methods are evaluated on MS-COCO [Chen et al., 2015], Flickr30k [Young et al., 2014], CUB-200-2011 [Wah et al., 2011] and Oxford Flowers 102 [Nilsback and Zisserman, 2008] for cross-modal retrieval, and VQA2.0 [Goyal et al., 2017] for visual question answering. The captions for the CUB and Flowers datasets were obtained from Reed et al. [2016].

Evaluation Metrics. The cross-modal retrieval is evaluated using the Recall@1 metric. For evaluating uncertainty calibration, we adopt the metrics used in Upadhyay et al. [2023], which computes the Spearman rank correlation (S) between different uncertainty levels and Recall@1, R^2 value for the regression between the uncertainty levels and the Recall@1 performance, and their product $-SR^2$. For an ideal model, the Recall@1 score should decrease with increasing uncertainty, resulting in a S value of -1. A higher R^2 score indicates that with increasing uncertainty levels, the model's performance declines linearly. A higher $-SR^2$ score implies better uncertainty calibration, reflecting both a strong negative correlation and a monotonic decrease in performance with increasing uncertainty. VQA is evaluated using the soft voting accuracy of 10 human-annotated answers [Goyal et al., 2017]. Calibration is evaluated by the Expected Calibration Error (ECE) score between the model's confidence and the soft voting accuracy. Model confidence is computed by first predicting an uncertainty score $u(a)$ for each candidate answer a, and then applying a softmax function over these uncertainty scores. The model confidence is given by $\text{conf}(a) = 1 - \text{softmax}(u(a))$.

Implementation details. The experiments on CLIP were conducted using the ViT-B/32 model as the image encoder, with $D = 512$. For BLIP, we adopt the ViT-B architecture as the image encoder. We trained the GPs with a latent space dimension of $Q = 5$ for MS-COCO, Flickr30k and VQA2.0, and $Q = 10$ for CUB and Flowers alongside trade-off parameters $\lambda_1 = 0.01$ and $\lambda_2 = 400$ and 250 inducing points, determined through grid search. The models were implemented with GPyTorch Gardner et al. [2018], and trained for 200 epochs using the Adam optimizer with a learning rate of $1e^{-5}$ and a batch size of 64. The detailed implementation, including data processing and hyper-parameter tuning is provided in Appendix A.1 and A.3 respectively.

4.2 UNCERTAINTY CALIBRATION IN CROSS-MODAL RETRIEVAL

Quantitative Results. The uncertainty calibration results for CLIP and BLIP is provided in Table 1 and Table 2 respectively. We observe that GroVE demonstrates superior performance across all four datasets in both image-to-text and text-to-image retrieval tasks, outperforming other methods in most cases. A high $-SR^2$ value for GroVE indicates that the model maintains strong performance when uncertainty is low, and the decline in performance is well-aligned with increasing uncertainty scores, indicating effective uncertainty calibration. Interestingly, the Deterministic baseline also performs competitively on the Flickr30k and MS-COCO datasets. This is because the VLMs were trained on datasets with common real-world objects, well-represented in these datasets, allowing the deterministic approach to benefit from familiar image-text pair contexts. However, on fine-grained datasets like CUB and Flowers, which are less represented in the training data, it exhibits a noticeable drop in performance. In these cases, the probabilistic methods outperform

	Method	Flickr			COCO			CUB			Flowers		
		$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$	$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$	$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$	$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$
Image to Text	Deterministic	-0.80±0.00	0.66±0.00	0.52±0.00	-0.80±0.00	0.64±0.00	0.51±0.00	-0.10±0.00	0.05±0.00	0.00±0.00	-0.10±0.00	0.00±0.00	0.00±0.00
	TTDA	0.12±0.03	0.32±0.07	-0.03±0.01	-0.36±0.05	0.38±0.08	0.17±0.05	-0.60±0.00	0.36±0.07	0.21 ±0.04	-0.78±0.04	0.37±0.07	0.28±0.06
	PFE	-0.34±0.06	0.45±0.04	0.13±0.03	0.63±0.05	0.72±0.07	-0.46±0.05	-0.13±0.04	0.28±0.03	0.02±0.01	-0.11±0.05	0.29±0.04	0.04±0.01
	PCME	0.61±0.06	0.18±0.02	-0.11±0.02	-0.63±0.00	0.50±0.03	0.31±0.02	-0.19±0.05	0.13±0.03	0.03±0.01	0.12±0.07	0.04±0.03	0.00±0.01
	PCME++	-0.08±0.04	0.33±0.04	0.04±0.02	-0.30±0.07	0.37±0.04	0.10±0.03	**-0.62±0.05**	0.67±0.05	0.38±0.05	-0.61±0.11	0.55±0.04	0.32±0.08
	ProbVLM	-0.79±0.05	0.52±0.04	0.38±0.04	-0.72±0.04	0.21±0.02	0.14±0.02	-0.33±0.05	0.46±0.04	0.15±0.02	-0.78±0.03	0.47±0.03	0.36±0.03
	GroVE	**-0.87±0.06**	**0.85±0.04**	**0.77±0.05**	**-0.90±0.03**	**0.88±0.04**	**0.79±0.02**	-0.61±0.07	**0.75±0.04**	**0.46±0.06**	**-0.88±0..04**	**0.81±0.01**	**0.70±0.03**
Text to Image	Deterministic	-0.90±0.00	**0.80±0.00**	**0.73±0.00**	-0.80±0.00	0.76±0.00	0.61±0.00	0.60±0.00	0.12±0.00	-0.06±0.00	-0.30±0.00	0.17±0.00	0.05±0.00
	TTDA	0.08±0.06	0.02±0.06	0.01±0.01	-0.61±0.06	0.20±0.05	0.12±0.04	-0.53±0.05	**0.64±0.03**	0.32±0.02	0.04±0.01	0.04±0.00	0.00±0.02
	PFE	-0.68±0.07	0.56±0.05	0.38±0.06	0.33±0.04	0.52±0.02	-0.16±0.02	-0.32±0.07	0.34±0.02	0.09±0.02	0.21±0.04	0.43±0.02	-0.10±0.02
	PCME	0.18±0.08	0.42±0.02	-0.07±0.04	0.86±0.04	**0.84±0.03**	-0.74±0.05	0.57±0.04	0.05±0.00	-0.03±0.00	0.72±0.05	0.45±0.03	-0.29±0.03
	PCME++	-0.13±0.04	0.06±0.03	0.01±0.00	0.02±0.07	0.38±0.02	0.01±0.03	-0.28±0.05	0.02±0.01	0.01±0.00	0.12±0.07	0.47±0.02	-0.06±0.04
	ProbVLM	-0.54±0.03	0.68±0.07	0.34±0.03	0.09±0.02	0.11±0.04	-0.01±0.00	**-0.92±0.03**	0.52±0.05	0.48±0.04	-0.60±0.03	0.16±0.06	0.10±0.03
	GroVE	**-0.92±0.04**	0.74±0.04	0.66±0.04	**-0.81±0.02**	0.81±0.01	**0.65±0.02**	-0.78±0.07	0.60±0.02	**0.49±0.05**	**-0.62±0.08**	**0.66±0.04**	**0.43±0.06**

Table 1: **Uncertainty calibration for cross-modal retrieval using CLIP.** GroVE demonstrates superior performance in uncertainty calibration in majority cases compared to baseline models. The best scores are highlighted in bold and the second-best scores are underlined.

	Method	Flickr			COCO			CUB			Flowers		
		$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$	$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$	$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$	$S\downarrow$	$R^2\uparrow$	$-SR^2\uparrow$
Image to Text	Deterministic	-0.70±0.00	**0.78±0.00**	**0.55±0.00**	-0.80±0.00	**0.84±0.00**	0.67±0.00	0.50±0.00	0.13±0.00	-0.07±0.00	-0.20±0.00	0.05±0.00	0.01±0.00
	TTDA	-0.68±0.04	0.27±0.03	0.19±0.02	-0.72±0.05	0.48±0.04	0.32±0.04	-0.70±0.05	0.50±0.02	0.33±0.03	-0.63±0.04	0.24±0.03	0.13±0.02
	PFE	0.12±0.06	0.37±0.02	-0.04±0.02	0.04±0.00	0.32±0.05	0.00±0.00	0.56±0.06	0.53±0.04	0.32±0.03	0.13±0.04	0.02±0.03	-0.01±0.03
	PCME	-0.31±0.06	0.17±0.04	0.05±0.03	-0.62±0.03	0.24±0.02	0.14±0.02	-0.64±0.03	**0.63**±0.03	0.38±0.03	0.08±0.03	0.25±0.04	-0.03±0.02
	PCME++	-0.68±0.03	0.26±0.03	0.18±0.03	-0.69±0.04	0.50±0.04	0.34±0.03	-0.71±0.04	0.57±0.03	0.40±0.03	-0.69±0.06	0.53±0.02	0.37±0.02
	ProbVLM	0.03±0.07	0.48±0.02	0.02±0.02	-0.61±0.03	0.50±0.04	0.30±0.03	-0.68±0.06	0.60±0.03	0.42±0.04	-0.67±0.00	0.65±0.02	0.46±0.01
	GroVE	**-0.72±0.03**	0.74±0.02	0.51±0.03	**-0.93±0.05**	0.76±0.03	**0.68±0.03**	**-0.89±0.04**	0.60±0.04	**0.54±0.02**	**-0.72±0.07**	**0.72±0.06**	**0.50±0.05**
Text to Image	Deterministic	-0.90±0.00	0.88±0.00	0.79±0.00	**-0.90±0.00**	**0.88±0.00**	**0.80±0.00**	0.40±0.00	0.06±0.00	0.02±0.00	-0.10±0.00	0.00±0.00	0.00±0.00
	TTDA	-0.37±0.03	0.35±0.04	0.14±0.03	0.41±0.06	0.00±0.01	0.00±0.03	-0.68±0.05	0.48±0.06	0.34±0.05	0.09±0.03	0.43±0.02	-0.04±0.04
	PFE	-0.58±0.04	0.50±0.03	0.30±0.04	0.11±0.05	0.15±0.04	-0.02±0.02	-0.78±0.03	0.58±0.02	0.47±0.02	-0.23±0.06	0.01±0.03	0.00±0.03
	PCME	-0.12±0.04	0.50±0.02	0.05±0.01	0.62±0.03	0.42±0.06	-0.25±0.03	-0.68±0.04	0.58±0.03	0.41±0.02	-0.31±0.03	0.26±0.03	0.08±0.04
	PCME++	-0.72±0.06	0.30±0.04	0.21±0.04	-0.48±0.03	0.31±0.02	0.15±0.03	-0.12±0.08	0.00±0.04	0.00±0.02	-0.20±0.07	0.06±0.06	0.01±0.03
	ProbVLM	-0.56±0.04	0.50±0.03	0.31±0.03	-0.12±0.05	0.48±0.04	0.05±0.04	-0.43±0.03	0.50±0.02	0.18±0.03	0.38±0.03	0.02±0.04	-0.01±0.04
	GroVE	**-0.92±0.04**	**0.90±0.03**	**0.81±0.04**	-0.62±0.03	0.36±0.06	0.22±0.04	**-0.89±0.02**	**0.75±0.03**	**0.74±0.03**	**-0.73±0.03**	**0.62±0.04**	**0.44±0.02**

Table 2: **Uncertainty calibration for cross-modal retrieval using BLIP.** GroVE demonstrates superior performance in uncertainty calibration in majority cases compared to baseline models. The best scores are highlighted in bold and the second-best scores are underlined.

the deterministic approach, with GroVE consistently leading across both common object and fine-grained datasets.

Qualitative Results. Given a query image from MS-COCO, we obtain its probabilistic embedding using GroVE. Using the distribution of this embedding, we compute the likelihood of each image in the Flickr30k dataset. Figure 3 shows a t-SNE plot of the mean embeddings on Flickr30k, colored by likelihood scores. The query image depicts children playing on a field. We observe that the images with the highest likelihood scores, share similar semantic content, such as scenes of people playing in fields. In contrast, images with lower likelihood values (close to 0.0) show little to no semantic or visual similarity to the query. Additional results containing the retrieval performance, zero-shot performance, calibration plots and qualitative analysis is provided in Appendix. C.1 and C.2.

4.3 ACTIVE LEARNING

The objective of this experiment is to fine-tune the CLIP model on the CUB dataset with limited labeled data. We estimate the uncertainty of image and text embeddings to identify the most uncertain samples from the unlabeled CUB dataset, which are labeled for fine-tuning. For methods using auxiliary models, we derive uncertainty estimates from models trained on COCO. We sample the top 500 uncertain samples at each step for fine-tuning with contrastive loss. A random sampling baseline is also included. Figure 4 provides the Recall@1 scores achieved in relation to the number of samples used for fine-tuning the CLIP model. GroVE achieves consistently better performance compared to others, demonstrating that its uncertainty estimates effectively identify the informative samples for active learning.

4.4 UNCERTAINTY IN FEW-SHOT SETTING

In this experiment, we explore a practical scenario where labeled training data is scarce. To simulate this, we create a few-shot dataset by randomly selecting three images and their corresponding text descriptions from 150 classes of the CUB dataset as done by Verma et al. [2021]. The probabilistic adapters were trained on this dataset using embeddings obtained from CLIP, and the uncertainty calibration was evaluated for cross-modal retrieval. Table 3 shows the

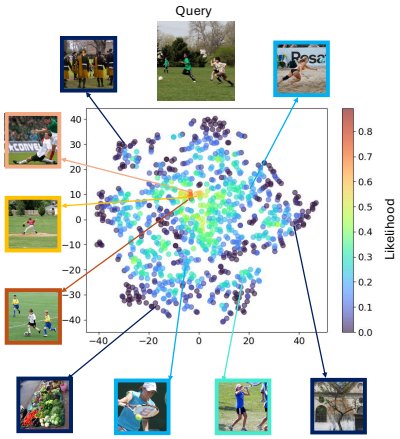

Figure 3: Given a probabilistic query image embedding from COCO, the plot shows a t-SNE visualization of Flickr30k embeddings, colored by their likelihood of belonging to the query distribution. Sample images are shown in colored boxes, where images with high likelihoods share similar semantic and visual content to the query.

Method	Image to Text	Text to Image	Latency (ms)
TTDA	0.03±0.06	0.01±0.04	288.51
PFE	-0.29±0.02	-0.22±0.03	31.59
PCME	0.04±0.03	0.14±0.02	31.60
PCME++	0.27±0.03	0.01±0.04	**31.55**
ProbVLM	-0.12±0.04	-0.43±0.04	32.80
GroVE (M=50)	0.24±0.03	0.22±0.03	47.62
GroVE (M=150)	0.36±0.04	0.31±0.02	142.85
GroVE (M=250)	0.35±0.03	0.31±0.03	392.16
GroVE (exact GP)	**0.39±0.03**	**0.36±0.02**	1130.09

Table 3: **Few-shot uncertainty calibration.** GroVE outperforms other baselines, achieving superior uncertainty calibration in few-shot settings in terms of -SR^2 (↑).

$-SR^2$ scores obtained for the baselines and GroVE with different numbers of inducing points as well as exact GP models, where training and inference is performed without approximations [Williams and Rasmussen, 2006]. The results show that while the calibration performance improves as the number of inducing points increases, GroVE consistently outperforms the baselines in terms of calibration quality. The best performance was achieved with exact GP models and no approximation. However, GroVE is computationally expensive compared to the neural network based approaches, with longer inference time as the number of inducing points increases.

4.5 UNCERTAINTY CALIBRATION FOR VQA

Table 4 shows the accuracy and ECE scores obtained for VQA2.0 using BLIP as the VLM. All the baselines achieve similar accuracy scores, with deterministic achieving the best accuracy. When evaluated for confidence calibration, GroVE achieves the lowest ECE score.

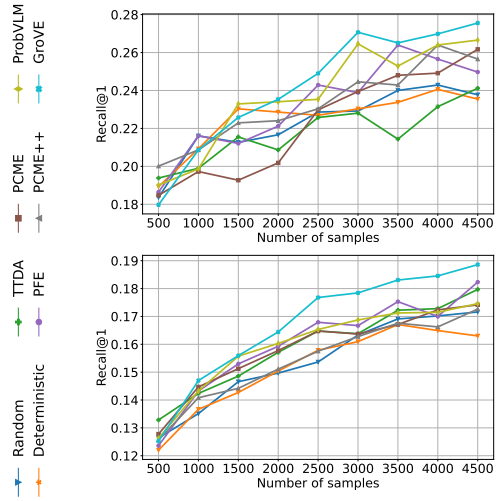

Figure 4: **Active Learning.** The results highlight GroVE's ability to effectively leverage uncertainty estimates to guide sample selection, outperforming the baselines on both image-to-text (left) and text-to-image (right) retrieval.

Method	Accuracy ↑	ECE ↓
Determinsitic	**78.20**	0.56
TTDA	77.67±2.23	0.48±0.06
PFE	76.34±1.98	0.65±0.02
PCME	77.25±1.76	0.64±0.01
PCME++	77.53±1.71	0.64±0.02
ProbVLM	76.66±1.13	0.69±0.01
GroVE	77.48±2.15	**0.24±0.04**

Table 4: **Results for VQA.** While all methods achieve similar accuracy (with the deterministic model performing best), GroVE reaches the best calibration performance in terms of ECE (↓).

4.6 ABLATION ANALYSIS

GP Kernel. We evaluate the performance of the RBF, Matérn ($\nu = 1.5$ and 2.5, where ν is the smoothness parameter) and the cosine similarity kernel on GroVE's performance on the MS-COCO and CUB data. From Table 5, the RBF kernel achieves superior performance compared to the other kernels across both datasets, with improvements up to 53%. The kernels are defined in Appendix B.1.

Latent Space Dimension. We investigate the influence of the latent space dimension Q on GroVE's performance using the MS-COCO dataset. Figure 5 (left) presents the $-SR^2$ scores for various values of Q. Low values of Q lead to information loss, which compromises the model's ability to capture complex patterns in the data. Conversely, high values of Q result in overfitting and make the model more challenging to optimize, resulting in a performance decline.

Kernel	Image to Text		Text to Image	
	COCO	CUB	COCO	CUB
RBF	**0.79±0.02**	0.46±0.06	**0.65±0.02**	**0.49±0.05**
Matérn ($\nu = 1.5$)	0.27±0.03	**0.47±0.05**	0.41±0.04	0.22±0.04
Matérn ($\nu = 2.5$)	0.52±0.05	0.38±0.04	0.43±0.04	0.12±0.05
Cosine Similarity	0.46±0.04	0.39±0.03	0.35±0.03	0.30±0.02

Table 5: **Ablation on choice of GP kernel.** GroVE achieves the best performance on MS-COCO and CUB-200-2011 with the RBF kernel.

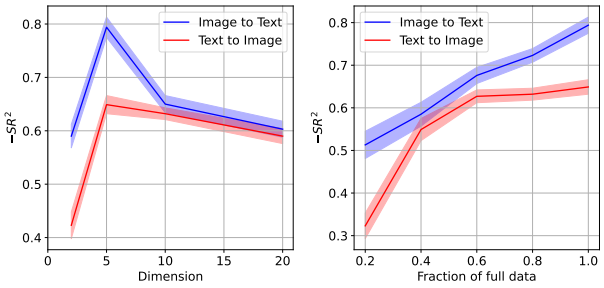

Figure 5: Ablation using MS-COCO: (i) **latent space dimension** (left). Low latent space dimensions results in loss of information, while higher dimensions results in performance degradation due to over-fitting. (ii) **dataset size for training** (right). GroVE achieves good performance with just 60% of the total training dataset.

The optimal performance was observed when $Q = 5$.

Dataset Size. We study the impact of the dataset size on GroVE's performance by training it on various fractions of the COCO training dataset. As shown in Figure 5 (right), the model achieves good performance when trained on 60% of the full dataset. While the uncertainty calibration performance for text-to-image retrieval plateaus beyond this point, the performance for image-to-text retrieval continues to improve almost linearly as more data is utilized.

Cross-modal Alignment. We compare GroVE's KL-divergence-based alignment loss with the MSE loss-based regularization used in Song et al. [2017]. The authors use GPLVM for cross-modal retrieval, regularizing the latent space with the loss function $||k_I - S||^2 + ||k_T - S||^2$, where k_I and k_T are the GP covariance matrices, and S is the latent space similarity matrix. For comparison, we replace $\mathcal{L}_{KL}$ in our method with the MSE loss and experiment with different values of the trade-off parameter λ_2, maintaining $\lambda_1 = 0.01$. As shown in Figure 6 (left), the KL-divergence alignment loss improves uncertainty calibration performance by up to 23%, with the best performance at $\lambda_2 = 400$. Additionally, we evaluate the cross-modal alignment KL-divergence loss against other widely-used probabilistic distance metrics: Jensen-Shannon (JS) divergence and Wasserstein-2 distance. The results in Table 6 indicate that while all metrics perform similarly, KL-divergence offers a slight edge.

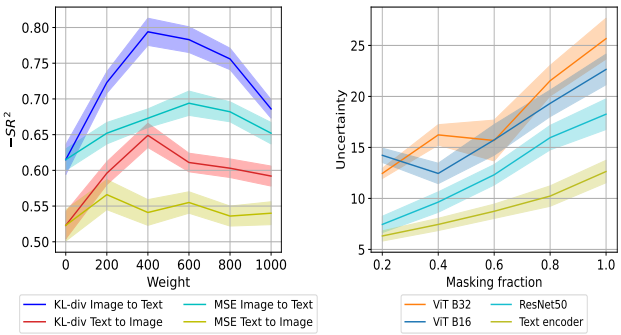

Figure 6: Ablation using MS-COCO: (i) **trade-off parameter** (left). KL-divergence improves uncertainty calibration with optimal performance at $\lambda_2 = 400$, with $\lambda_1 = 0.01$. (ii) **noisy data** (right). With increasing amount of noise in the input data, GroVE predicts higher uncertainty.

Kernel	Image to Text		Text to Image	
	COCO	CUB	COCO	CUB
KL-Divergence	**0.79±0.02**	0.46±0.06	**0.65±0.02**	**0.49±0.05**
JS-Divergence	0.70±0.04	**0.48±0.02**	0.59±0.03	0.44±0.05
Wasserstein-2	0.59±0.04	0.39±0.04	0.60±0.02	0.43±0.04

Table 6: **Ablation on probabilistic distance metric.** GroVE performs better for cross-modal alignment using KL-Divergence compared to other metrics.

Noisy Data. To evaluate the performance of GroVE against noisy inputs, we systematically introduce increasing levels of masking to both the input images and texts. This analysis employs several CLIP image encoder backbones, including ViT-B/32, ViT-B/16, and ResNet50, along with CLIP's text encoder. The results, presented in Figure 6 (right), indicate that as the noise level increases, the uncertainty predicted by GroVE rises steadily as desired.

5 CONCLUSION

This paper introduces GroVE, a post-hoc approach for generating probabilistic embeddings from frozen, pre-trained VLMs to model input data ambiguities. GroVE leverages the GPLVM framework, utilizing GP models to learn a shared, low-dimensional latent space that aligns visual and textual representations. By mapping into this latent space, the GP models generate probabilistic embeddings that provide a measure of uncertainty in the predictions. GroVE demonstrates state-of-the-art performance in uncertainty calibration for cross-modal retrieval, active learning and VQA. One limitation of GroVE is the it is computationally expensive compared to the neural network based methods. In latency-sensitive scenarios, neural network-based stochastic models like Neural Processes [Garnelo et al., 2018] offer a viable alternative to GPs. Future work will focus on assessing their uncertainty calibration performance for VLMs.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021.

Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.

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.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Sanghyuk Chun. Improved probabilistic image-text representations. *arXiv preprint arXiv:2305.18171*, 2023.

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

Charles Corbiere, Nicolas Thome, Antoine Saporta, Tuan-Hung Vu, Matthieu Cord, and Patrick Perez. Confidence estimation via auxiliary models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6043–6055, 2021.

John Duchi. Derivations for linear algebra and optimization. *Berkeley, California*, 3(1):2325–5870, 2007.

Stefanos Eleftheriadis, Ognjen Rudovic, and Maja Pantic. Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. *IEEE transactions on image processing*, 24(1):189–204, 2014.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Jacob Gardner, Geoff Pleiss, Kilian Q Weinberger, David Bindel, and Andrew G Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. *Advances in neural information processing systems*, 31, 2018.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

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.

James Hensman, Alexander Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360. PMLR, 2015.

Julia Hornauer, Adrian Holzbock, and Vasileios Belagiannis. Out-of-distribution detection for monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1911–1921, 2023.

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.

Myong Chol Jung, He Zhao, Joanna Dipnall, Belinda Gabbe, and Lan Du. Uncertainty estimation for multi-view data: The power of seeing the whole picture. *Advances in Neural Information Processing Systems*, 35:6517–6530, 2022.

Vidhi Lalchand, Aditya Ravuri, and Neil D Lawrence. Generalised gaussian process latent variable models (gplvm) with stochastic variational inference. *arXiv preprint arXiv:2202.12979*, 2022.

Neil Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems*, 16, 2003.

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, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022b.

Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512, 2020.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

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.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58, 2016.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6902–6911, 2019.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. Multimodal gaussian process latent variable models with harmonization. In *Proceedings of the IEEE international conference on computer vision*, pages 5029–5037, 2017.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

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.

Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

Aishwarya Venkataramanan, Assia Benbihi, Martin Laviale, and Cédric Pradalier. Gaussian latent representations for uncertainty estimation using mahalanobis distance in deep classifiers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4488–4497, 2023a.

Aishwarya Venkataramanan, Martin Laviale, and Cédric Pradalier. Integrating visual and semantic similarity using hierarchies for image retrieval. In *International Conference on Computer Vision Systems*, pages 422–431. Springer, 2023b.

Vinay Kumar Verma, Ashish Mishra, Anubha Pandey, Hema A Murthy, and Piyush Rai. Towards zero-shot learning with fewer seen class examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2241–2251, 2021.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

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.

Xuanlong Yu, Gianni Franchi, and Emanuel Aldea. Slurp: Side learning uncertainty for regression problems. *arXiv preprint arXiv:2110.11182*, 2021.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Probabilistic Embeddings for Frozen Vision-Language Models: Uncertainty Quantification with Gaussian Process Latent Variable Models (Supplementary Material)

Aishwarya Venkataramanan[1] **Paul Bodesheim**[1] **Joachim Denzler**[1]

[1]Computer Vision Group, Friedrich Schiller University Jena, Germany

A ADDITIONAL IMPLEMENTATION DETAILS

This section provides details on the data processing steps for training both the baseline models and GroVE, implementation details for each baseline, and the hyper-parameter tuning procedure applied for GroVE.

A.1 DATASETS

For the experiments, we use MS-COCO, Flickr30k, CUB-200-2011, and Oxford Flowers 102 dataset.

MS-COCO Chen et al. [2015] is a widely used cross-modal retrieval dataset includes 123,287 images, each image annotated with 5 captions describing common objects. The training set comprises 113,287 images, while both the validation and test sets contain 5,000 images each. Different papers apply varying evaluation protocols on the 5,000 test images in the COCO dataset. Some cross-modal retrieval papers report results on the full 5,000 test set, while others use 1,000 unique test images, averaging results over 5 random splits. In our study, we follow the former approach, presenting results based on the entire 5,000 test set.

Flickr30k Young et al. [2014] is a widely used cross-modal retrieval dataset comprising 31,783 images, each image annotated with 5 captions describing common objects. The dataset is split into 29,783 training images, with 1,000 images each in the validation and test sets.

CUB-200-2011 Wah et al. [2011] is a fine-grained bird species dataset comprising 11,788 images across 200 categories, with each image paired with 10 captions sourced from Reed et al. [2016]. Following the split protocol in Chun et al. [2021], the dataset includes 7,067 training images, 1,754 validation images, and 2,967 test images.

Oxford Flowers 102 Nilsback and Zisserman [2008] is a fine-grained flowers dataset comprising 8,189 images across 102 categories, with each image paired with 10 captions sourced from Reed et al. [2016]. Following the split protocol in Upadhyay et al. [2023], the dataset includes 7,034 training images, 750 validation and 805 test images.

All images are resized to 224×224, suitable for input to VLMs. All methods are trained and evaluated using the same dataset splits for the comparison.

A.2 BASELINE METHODS

In this section, we provide the implementation and training details for the baseline methods.

TTDA Ayhan and Berens [2018]. During inference, data augmentations are applied to the input, generating multiple variations to estimate prediction uncertainty. For each augmented version, a prediction is generated, and the variance across these predictions reflects the model's uncertainty. Image augmentations include random resized cropping and horizontal flipping (applied with a probability of 0.3), while text data undergoes random word masking with a 0.3 probability. The model is run for 10 passes on these augmentations, to obtain the image and text uncertainty.

PFE Shi and Jain [2019], PCME Chun et al. [2021] and PCME++ Chun [2023]. During training, we adapt these methods to process image and text embeddings derived from a frozen VLM. Following Upadhyay et al. [2023], the model architecture consists of two Multi-Layer Perceptrons (MLPs)—one for images and one for text. Each MLP has an input layer that reduces the embedding dimension to 256, a hidden layer of 256 units, and an output layer that maps from 256 back to the original embedding dimension. We apply the respective loss functions to learn covariances for a Gaussian distribution, with mean values matching the VLM's deterministic embeddings. Training is conducted for 200 epochs using the Adam optimizer with a learning rate of 10^{-8} and batch size of 64. The learning rate was fixed using a grid search over values $\{1e^{-4}, 1e^{-5}, 1e^{-6}, 1e^{-7}, 1e^{-8}\}$

ProbVLM Upadhyay et al. [2023]. We follow the training procedure outlined in the original paper. The model architecture consists of two MLPs—one for image embeddings and one for text embeddings—similar to previous methods. Training is conducted with the Adam optimizer for 100 epochs, using a learning rate of 10^{-4} and a batch size of 64.

A.3 HYPER-PARAMETER TUNING

GroVE introduces the following hyper-parameters which were obtained using grid-search: latent space dimension Q, and the trade-off parameters λ_1 and λ_2. For Q, we evaluated values $Q \in \{2, 5, 10, 20, 50, 128, 256\}$. For the trade-off parameters, we used $\lambda_1 \in \{1, 0.1, 0.01, 0.001\}$, and $\lambda_2 \in \{0, 200, 400, 600, 800, 1000\}$. based on the grid-search results, the optimal setting for Q was $Q = 5$ for MS-COCO and Flickr30k, and $Q = 10$ for CUB-200-2011 and Oxford Flowers 102. The trade-off parameters that achieved the best performance were $\lambda_1 = 0.01$ and $\lambda_2 = 400$. The number of inducing points was selected from {100, 150, 200, 250, 300, 350}, with model performance plateauing beyond 250 points. Finally, the learning rate of $1e^{-5}$ was selected based on a grid-search over values $\{1e^{-1}, 1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}, 1e^{-6}\}$.

B DEFINITIONS

This section presents the definitions of the GP kernels and the probabilistic distance metrics used in the ablation study.

B.1 GP KERNELS

In Table. 5, we presented the results for ablation study of various kernels for GP. In this section, we provide the definitions and formulas for the kernels used.

Radial Basis Function (RBF). The RBF kernel, also known as the Gaussian kernel, is a popular choice in GPs. It assumes that closer data points in input space have higher similarity. The RBF kernel between two points $\mathbf{x_i}$ and $\mathbf{x_j}$ is defined as:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{\|\mathbf{x_i} - \mathbf{x_j}\|}{l^2}\right) \tag{14}$$

where l is the length scale parameter, controlling how quickly the similarity decreases with distance in input space.

Matérn. The Matérn kernel generalizes the RBF kernel, defined as

$$k(\mathbf{x_i}, \mathbf{x_j}) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{x_i} - \mathbf{x_j}\|\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{x_i} - \mathbf{x_j}\|\right) \tag{15}$$

where ν controls the smoothness of the resulting function, l is the length scale parameter, Γ is the Gamma function and K_ν is a modified Bessel function.

Cosine Similarity. This is a linear kernel with normalized inputs, and is defined as:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \frac{\mathbf{x_i}^T \mathbf{x_j}}{\|\mathbf{x_i}\|\|\mathbf{x_j}\|} \tag{16}$$

B.2 PROBABILISTIC DISTANCES

Sec. 4.6 presented an ablation study on the choice of the probability distance metric for cross-modal alignment (refer Table. 6). The definitions of the probabilistic distance metrics for two multivariate Gaussians $p = \mathcal{N}(\hat{\boldsymbol{\mu}}_\mathcal{I}, \hat{\boldsymbol{\Sigma}}_\mathcal{I})$ and $q = \mathcal{N}(\hat{\boldsymbol{\mu}}_\mathcal{T}, \hat{\boldsymbol{\Sigma}}_\mathcal{T})$, are as follows:

Kullback-Liebler Divergence. The KL Divergence quantifies the difference between two probability distributions, and is defined as Duchi [2007]:

$$D_{KL}(p\|q) = \frac{1}{2}\left[\text{tr}(\hat{\boldsymbol{\Sigma}}_{\mathcal{T}}^{-1}\hat{\boldsymbol{\Sigma}}_{\mathcal{I}}) + (\hat{\boldsymbol{\mu}}_{\mathcal{T}} - \hat{\boldsymbol{\mu}}_{\mathcal{I}})^T\hat{\boldsymbol{\Sigma}}_{\mathcal{T}}^{-1}(\hat{\boldsymbol{\mu}}_{\mathcal{T}} - \hat{\boldsymbol{\mu}}_{\mathcal{I}}) - D + \log\left(\frac{\det(\hat{\boldsymbol{\Sigma}}_{\mathcal{T}})}{\det(\hat{\boldsymbol{\Sigma}}_{\mathcal{I}})}\right)\right], \tag{17}$$

where $\text{tr}(.)$ is the trace and $\det(.)$ is the determinant of a matrix. Note that the KL-Divergence is asymmetric; thus, we calculate the total cross-alignment loss as the mean of the KL divergences in both directions (refer Eq. 8): $\frac{1}{2}[D_{KL}(p\|q) + D_{KL}(q\|p)]$.

Jensen-Shannon (JS) Divergence. The JS Divergence is obtained by averaging the KL divergences between each distribution and the average distribution. The JS divergence is defined as:

$$D_{JS}(p\|q) = \frac{1}{2}\left(D_{KL}(p\|m) + D_{KL}(q\|m)\right) \tag{18}$$

where $m = \frac{1}{2}(p+q)$ is the mean distribution of p and q.

Wasserstein-2 distance. The Wasserstein-2 distance quantifies the cost of transforming one distribution into another. This is defined as:

$$W_2^2(p,q) = \|\hat{\boldsymbol{\mu}}_{\mathcal{I}} - \hat{\boldsymbol{\mu}}_{\mathcal{T}}\|^2 + \text{tr}\left(\hat{\boldsymbol{\Sigma}}_{\mathcal{I}} + \hat{\boldsymbol{\Sigma}}_{\mathcal{T}} - 2\left(\hat{\boldsymbol{\Sigma}}_{\mathcal{I}}^{1/2}\hat{\boldsymbol{\Sigma}}_{\mathcal{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{I}}^{1/2}\right)^{1/2}\right) \tag{19}$$

C ADDITIONAL RESULTS

This section presents additional quantitative and qualitative results.

C.1 CROSS-MODAL RETRIEVAL

C.1.1 Calibration plots

Figure 7 and Figure 8 show the calibration plots for the CLIP and BLIP models, respectively. Calibration plots are obtained by binning uncertainty values, referred to as uncertainty levels and computing Recall@1 for each bin. From the plots, GroVE maintains a more consistent alignment between decreasing uncertainty and increasing Recall@1.

C.1.2 Retrieval performance

Table 7 presents the Recall@1 scores for various baselines using CLIP. The score for the Deterministic baseline was computed by retrieving the nearest image/text embedding based on cosine similarity to the query text/image from the deterministic embeddings generated by the CLIP model. For the other baselines, retrieval was performed by selecting the image/text embedding with the minimum Wasserstein distance to the query, using the probabilistic image/text embeddings. Results show that GroVE achieves a good performance on the fine-grained CUB and Flowers dataset, whereas deterministic achieves the best scores in MS-COCO and Flickr30k dataset.

C.1.3 Qualitative Analysis

A t-SNE visualization of the probabilistic embeddings from GroVE on a subset of the CUB dataset is provided in Figure 9, where the uncertainty corresponds to the area of the embedding. The plot shows that images and texts with similar semantic content are clustered together, and the probabilistic embeddings are able to capture the uncertainty arising from the data ambiguities.

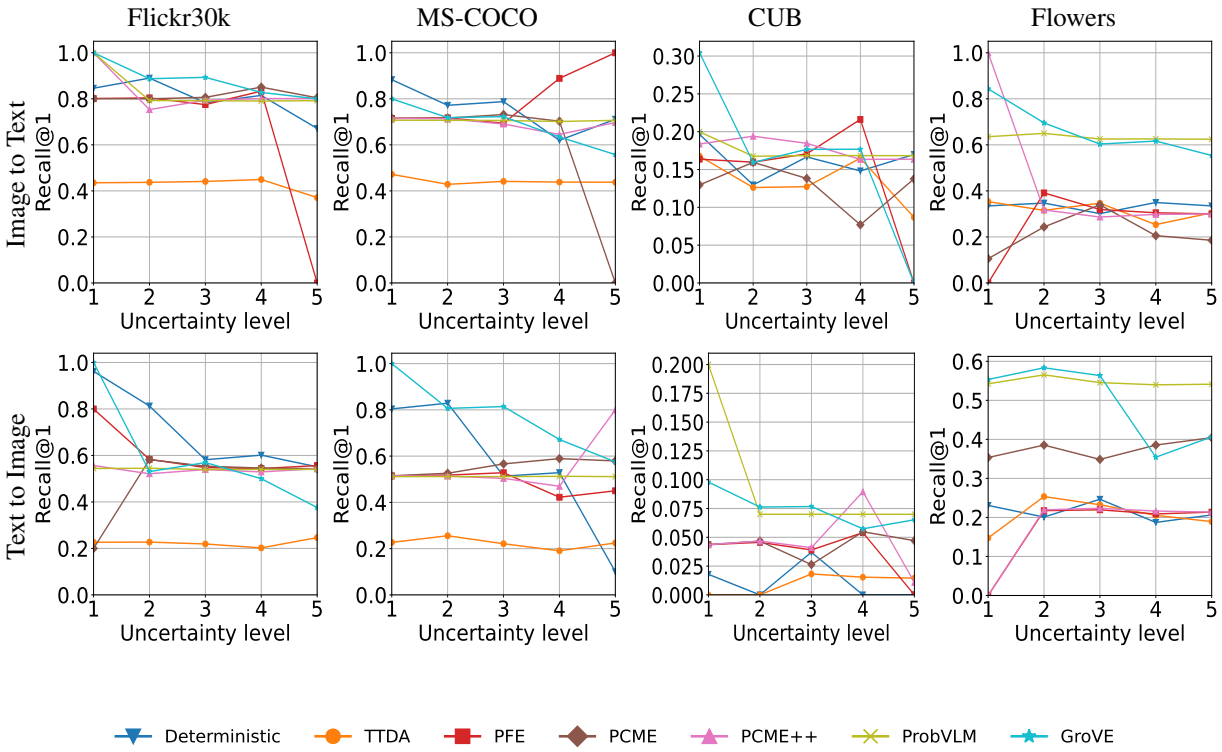

Figure 7: **Evaluation of uncertainty calibration** for embeddings obtained from CLIP on Image-to-Text retrieval. For perfect calibration, the Recall@1 should show a monotonic decrease as uncertainty levels increase. GroVE exhibits a more consistent relationship between increasing uncertainty and performance degradation compared to the baseline methods.

	Method	Flickr	COCO	CUB	Flowers
Image to Text	Deterministic	**0.801**	**0.715**	**0.532**	0.357
	TTDA	0.423	0.326	0.133	0.289
	PFE	0.238	0.213	0.101	0.102
	PCME	0.392	0.246	0.129	0.134
	PCME++	0.423	0.397	0.124	0.111
	ProbVLM	0.491	0.303	0.136	0.245
	GroVE	0.569	0.512	0.307	**0.402**
Text to Image	Deterministic	**0.543**	**0.515**	0.141	0.109
	TTDA	0.202	0.139	0.046	0.057
	PFE	0.298	0.219	0.023	0.024
	PCME	0.092	0.102	0.099	0.029
	PCME++	0.133	0.125	0.087	0.058
	ProbVLM	0.104	0.156	0.005	0.102
	GroVE	0.241	0.288	**0.343**	**0.379**

Table 7: **Retrieval performance using CLIP.** Table shows the Recall@1 scores obtained using the different baselines. GroVE achieves the best scores for the fine-grained datasets.

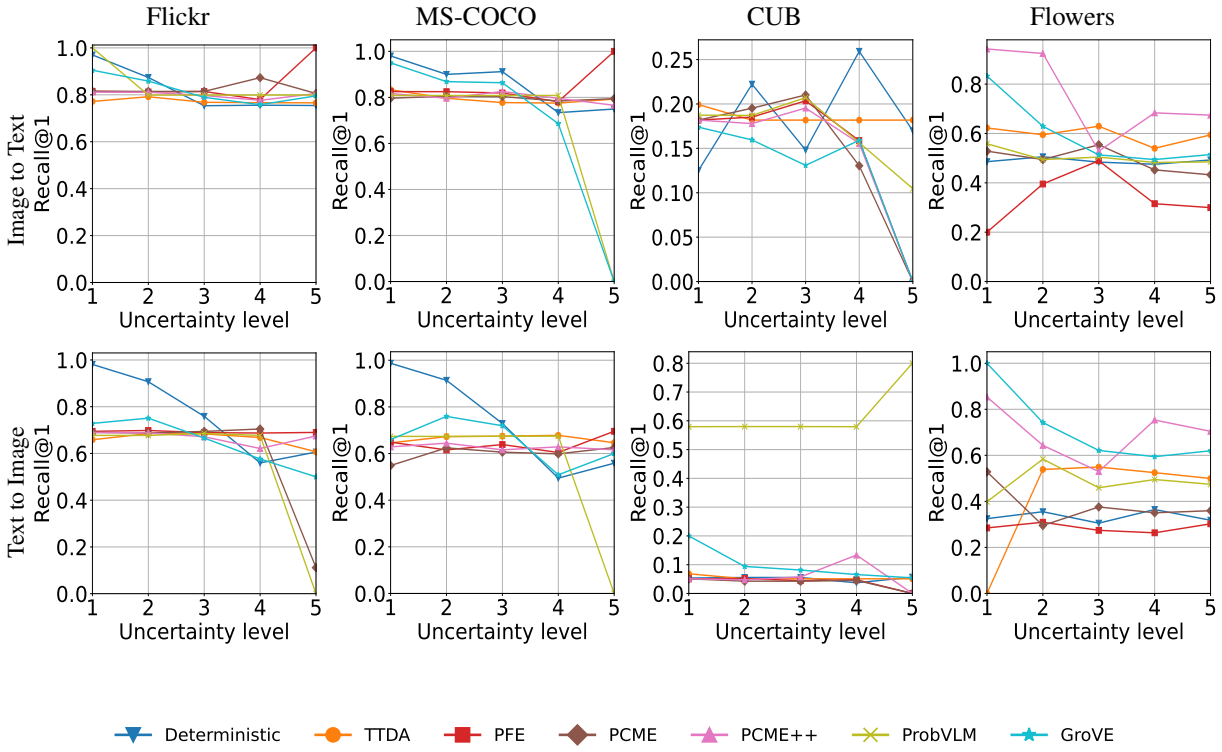

Figure 8: **Evaluation of uncertainty calibration** for embeddings obtained from BLIP on Image-to-Text (top) and Text-to-Image (bottom) retrieval tasks. For perfect calibration, the Recall@1 should show a monotonic decrease as uncertainty levels increase. GroVE exhibits a more consistent relationship between increasing uncertainty and performance degradation compared to the baseline methods.

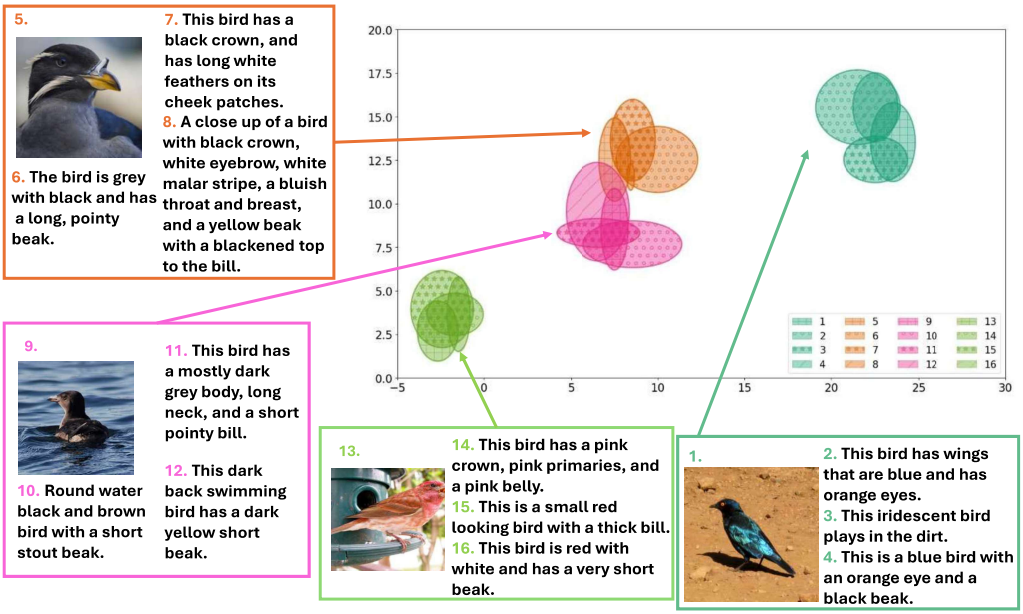

Figure 9: **t-SNE visualization of the probabilistic representations** generated by GroVE on a subset of the CUB-200-2011 dataset. Starting from deterministic embeddings provided by frozen VLMs, GroVE produces corresponding probabilistic representations that capture input ambiguities.

	Method	Flickr	Flowers	CUB
Image to Text	PFE	0.01±0.03	0.38±0.04	0.02±0.02
	PCME	0.04±0.02	0.13±0.04	0.09±0.06
	PCME++	0.01±0.02	0.48±0.03	0.03±0.02
	ProbVLM	0.55±0.03	0.19±0.04	0.15±0.04
	GroVE	**0.74±0.03**	**0.69±0.02**	**0.41±0.03**
Text to Image	PFE	0.41±0.03	0.02±0.03	0.04±0.01
	PCME	0.24±0.03	-0.01±0.02	0.02±0.03
	PCME++	-0.43±0.03	0.05±0.03	0.03±0.02
	ProbVLM	0.14±0.05	0.01±0.03	0.00±0.01
	GroVE	**0.42±0.02**	**0.09±0.03**	**0.04±0.02**

Table 8: **Zero-shot uncertainty calibration - MS-COCO.** GroVE outperforms other baselines in most cases, achieving superior uncertainty calibration in zero-shot settings. The best scores are highlighted in bold and the second-best scores are underlined.

	Method	Flickr	COCO	Flowers
Image to Text	PFE	0.00±0.04	0.02±0.03	-0.13±0.03
	PCME	0.46±0.02	0.01±0.05	0.02±0.04
	PCME++	0.40±0.03	0.10±0.02	0.44±0.03
	ProbVLM	0.15±0.02	0.38±0.03	0.18±0.03
	GroVE	**0.59±0.03**	**0.45±0.03**	**0.50±0.04**
Text to Image	PFE	-0.01±0.02	0.31±0.04	-0.12±0.03
	PCME	0.14±0.02	-0.19±0.03	0.15±0.04
	PCME++	0.13±0.02	**0.52±0.03**	0.36±0.03
	ProbVLM	0.01±0.03	0.01±0.02	0.02±0.03
	GroVE	**0.76±0.03**	0.42±0.02	**0.37±0.03**

Table 9: **Zero-shot uncertainty calibration - CUB-200-2011.** GroVE outperforms other baselines in most cases, achieving superior uncertainty calibration in zero-shot settings. The best scores are highlighted in bold and the second-best scores are underlined.

C.2 ZERO-SHOT UNCERTAINTY CALIBRATION

We evaluate the generalization of uncertainty calibration across methods that use auxiliary models for probabilistic embeddings on out-of-distribution datasets. Two CLIP experiments are conducted by training on MS-COCO and CUB, then evaluating on their respective unseen datasets. Table 8 and 9 presents the $-SR^2$ scores with models trained on MS-COCO and CUB respectively. The models trained on MS-COCO show a strong performance on Flickr30k due to its similarity to MS-COCO, thereby exhibiting better generalization. There is a drop in performance on the more fine-grained Flowers and CUB datasets, particularly for text-to-image retrieval. GroVE, however, demonstrates better generalization than the baseline methods for both the experiments.

The calibration results for the experiments are presented in Figure 10 and Figure 11, respectively, where GroVE maintains a more consistent alignment between decreasing uncertainty and increasing Recall@1.

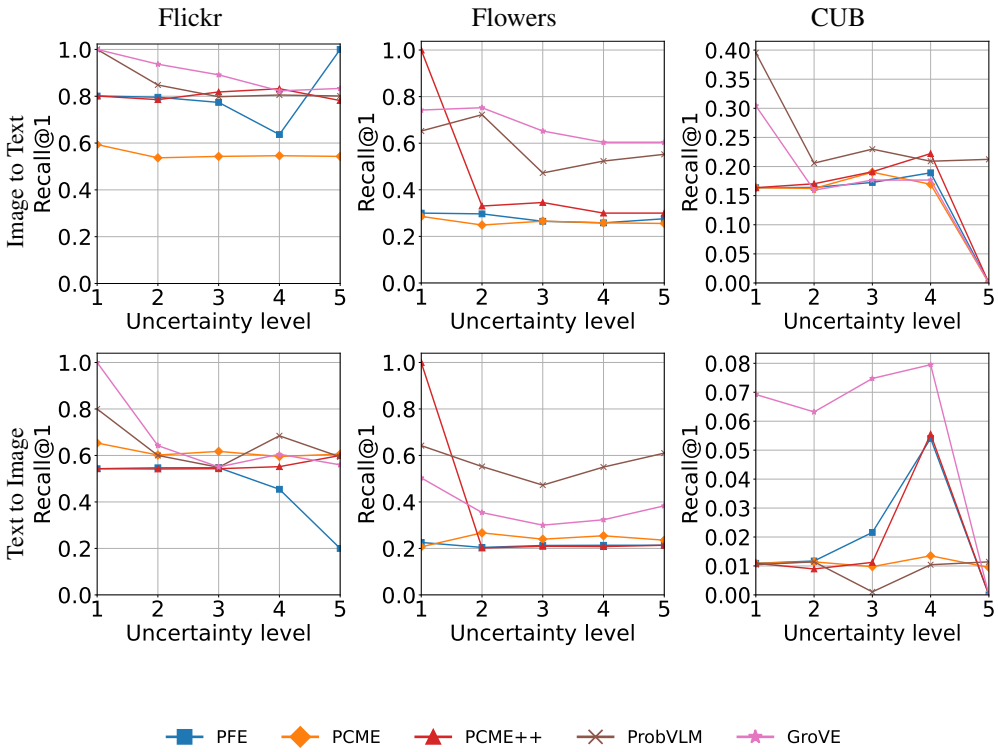

Figure 10: **Evaluation of zero-shot uncertainty calibration using MS-COCO** for embeddings obtained from CLIP on Image-to-Text (top) and Text-to-Image (bottom) retrieval tasks. For perfect calibration, the Recall@1 should show a monotonic decrease as uncertainty levels increase. GroVE exhibits a more consistent relationship between increasing uncertainty and performance degradation compared to the baseline methods.

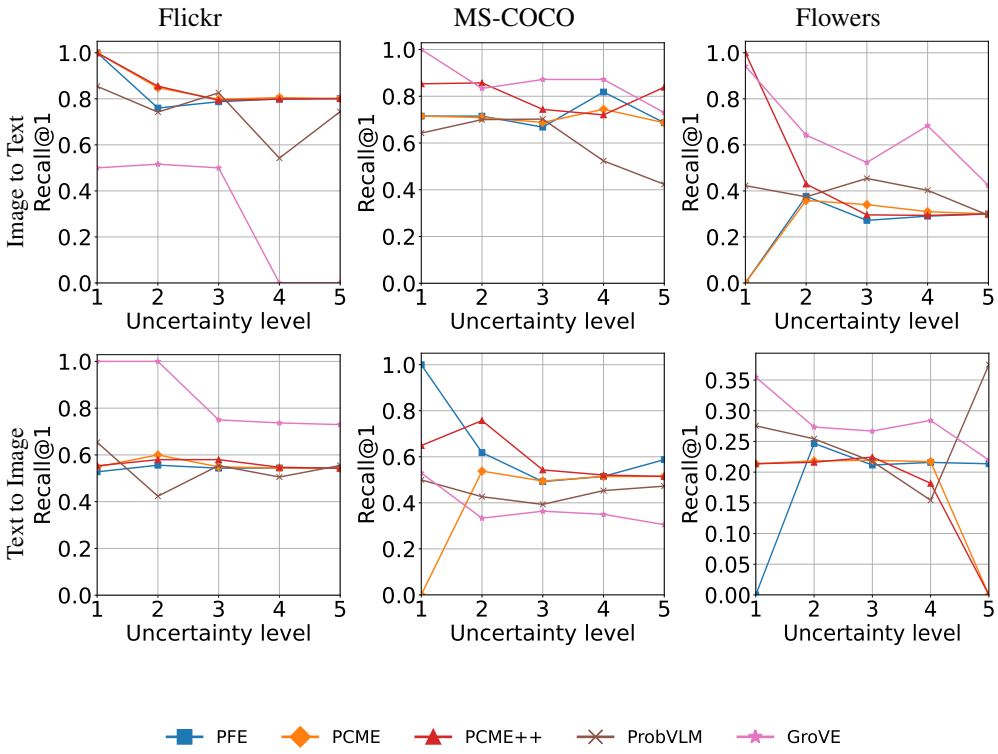

Figure 11: **Evaluation of zero-shot uncertainty calibration using CUB-200-2011** for embeddings obtained from CLIP on Image-to-Text (top) and Text-to-Image (bottom) retrieval tasks. For perfect calibration, the Recall@1 should show a monotonic decrease as uncertainty levels increase. GroVE exhibits a more consistent relationship between increasing uncertainty and performance degradation compared to the baseline methods.

C.3 FEW-SHOT UNCERTAINTY CALIBRATION

Table 10 shows the Recall@1 scores for the cross-modal retrieval task for the auxiliary models trained using limited data from the synthetic CUB dataset. The performance of the neural network based methods drop, which is expected given the insufficient number of data points for the model to generalize. Note that deterministic and TTDA are agnostic to the few shot setting since they work directly on the VLM embeddings for the prediction. Among the methods using auxiliary models, GroVE achieves a higher score, leveraging the ability of GPs to generalize well even with limited data because of their distance awareness property by capturing structure through kernel functions. Moreover, as the number of inducing points increases, GroVE's performance improves, with the best results achieved when performing exact GP.

Method	Image to Text	Text to Image
Determinsitic	**0.532**	<u>0.141</u>
TTDA (10 passes)	<u>0.133±0.003</u>	0.046±0.011
PFE	0.062±0.001	0.026±0.010
PCME	0.074±0.002	0.031±0.005
PCME++	0.063±0.003	0.031±0.003
ProbVLM	0.081±0.001	0.034±0.005
GroVE (M=50)	0.062±0.002	0.035±0.009
GroVE (M=150)	0.084±0.004	0.049±0.004
GroVE (M=250)	0.086±0.003	0.056±0.004
GroVE (exact GP)	0.103±0.002	**0.182±0.002**

Table 10: Retrieval performance using Recall@1 scores and inference speed per instance for few-shot experiment using CUB-200-2011. The best results are highlighted in bold and the second best are underlined.

Table 11 shows the results for uncertainty calibration for few-shot classification using the CUB-200-2011 dataset, with the same settings as described in Section 4.4, and the key difference that we perform multi-class classification instead of cross-modal retrieval. We evaluate using accuracy values and $-SR^2$ scores for uncertainty calibration. For computing the $-SR^2$ scores, we adopt a procedure similar to that used for cross-modal retrieval. Specifically, we divide the uncertainty values into bins, and instead of using Recall@1, we use accuracy scores to compute the $-SR^2$ values. The results show that GroVE achieves a better accuracy and uncertainty calibration in comparison to the other baselines.

Method	Accuracy	$-SR^2$
TTDA	15.08±0.04	0.23±0.08
PFE	13.26±0.03	0.03±0.04
PCME	14.89±0.05	0.16±0.07
PCME++	13.65±0.03	0.22±0.04
ProbVLM	14.62±0.05	0.38±0.03
GroVE	**15.32±0.06**	**0.62±0.02**

Table 11: Results for few-shot classification using CUB-200-2011.