GP-Adapter: Gaussian Process CLIP-Adapter for Few-Shot Out-of-Distribution Detection

1st Taisei Saito
Ricoh Company, Ltd.
Ebina, Japan
Taisei.Saito1@jp.ricoh.com

2nd Koretaka Ogata
Ricoh Company, Ltd.
Ebina, Japan
koretaka.ogata@jp.ricoh.com

3rd Takafumi Hiroi
Ricoh Company, Ltd.
Ebina, Japan
takafumi.hiroi@jp.ricoh.com

***Abstract*—We propose GP-Adapter, a training-free framework that augments CLIP (Contrastive Language-Image Pre-training) with Gaussian Process (GP) uncertainty modeling for few-shot classification and out-of-distribution (OOD) detection. While CLIP achieves strong zero-shot recognition, it yields deterministic similarity scores and offers limited uncertainty information, which is critical under distribution shift and data scarcity. GP-Adapter constructs modality-specific, class-wise one-class GPs on top of frozen CLIP embeddings using an RBF kernel for image features and a linear kernel for text prompts and fuses their predictive statistics to produce a variance-aware confidence score for OOD detection. The method requires no fine-tuning of the CLIP backbone and relies only on a small K-shot cache and lightweight hyperparameter selection, with memory cost scaling as $O(CK^2)$ for C classes and K shots. Experiments on ImageNet and multiple OOD benchmarks show that GP-Adapter provides competitive few-shot performance and consistently improves OOD detection when combined with prompt-learning baselines, highlighting the complementarity between GP-based uncertainty modeling and prompt learning. Overall, our results suggest that integrating probabilistic inference with large pre-trained vision-language models can improve reliability in low-data and distribution-shifted settings. Code is available at https://github.com/tms-byte/GP-Adapter**

***Index Terms*—out-of-distribution detection, few-shot adaptation, gaussian process, reliable AI, uncertainty estimation**

I. Introduction

Modern deep learning systems are increasingly deployed in open-world environments, where distribution shifts are inevitable and reliable operation requires the ability to detect out-of-distribution (OOD) inputs that differ from the training data. In many practical applications such as medical imaging, industrial inspection, and autonomous systems, models frequently encounter OOD samples, and failing to detect such samples may lead to unreliable or unsafe behavior. These considerations make OOD detection a fundamental requirement for trustworthy AI systems [1]. In addition, labeled training data is often severely limited in practice. This combination of limited data and distribution shift poses a significant challenge: with few labeled examples, decision boundaries become highly uncertain, and models are prone to overconfident predictions on both in-distribution (ID) and OOD inputs. As a result, principled uncertainty estimation is essential for robust OOD detection in such settings [2].

Vision-language models (VLMs) such as CLIP (Contrastive Language–Image Pre-training) [3] have demonstrated strong generalization capabilities by aligning images and text in a shared embedding space learned from large-scale web data. However, CLIP relies on deterministic similarity scores and lacks explicit mechanisms for uncertainty modeling. Recent CLIP-based approaches for OOD detection and lightweight adaptation often depend on prompt tuning or gradient-based optimization [4]–[6]. These methods introduce additional learnable parameters, which can be unstable or prone to overfitting under extremely limited data, and are not always suitable for deployment scenarios where gradient-based fine-tuning is costly or undesirable.

Bayesian models, such as Gaussian Processes (GPs) [7], offer a principled framework for uncertainty estimation, particularly well-suited to low-data regimes. GPs provide closed-form predictive distributions that explicitly capture predictive uncertainty, making them attractive for OOD detection [8]. However, integrating GP-based inference with large-scale multimodal models like CLIP in an efficient and training-free manner remains an open challenge.

To address these issues, we propose **GP-Adapter**, a training-free framework that augments frozen CLIP representations with GP-based uncertainty modeling for few-shot classification and OOD detection. GP-Adapter constructs modality-specific one-class GPs on CLIP-derived image and text embeddings using a small K-shot labeled cache, and combines their predictive statistics to compute a variance-aware confidence score for OOD detection. Importantly, all GP parameters are computed analytically or optimized via lightweight grid search, enabling rapid adaptation in real-world environments where gradient-based fine-tuning is undesirable. Because each class-specific one-class GP is built from only K support samples for each of the C classes, the memory footprint scales as $O(CK^2)$, making GP-Adapter scalable to large-scale datasets such as ImageNet-1k [9]. Moreover, our results show that GP-based uncertainty modeling is complementary to prompt-learning approaches: combining GP-Adapter with prompt-tuning baselines consistently strengthens OOD detection while preserving ID accuracy.

The contributions of our work are summarized as follows:

- We propose GP-Adapter, a training-free framework that augments frozen CLIP with GP-based uncertainty modeling for OOD detection under severe data scarcity.

- We introduce a modality-specific, class-wise one-class GP formulation that models image and text embeddings independently and combines their predictive statistics to produce a variance-aware confidence score.
- Extensive experiments demonstrate that GP-Adapter improves few-shot OOD detection performance across multiple benchmarks without requiring additional fine-tuning or OOD datasets, and further gains are obtained when integrating it with prompt-learning methods.

II. Related Work

Early OOD detection methods are primarily based on confidence scores derived from the output of trained discriminative neural networks. Hendrycks and Gimpel [10] proposed the Maximum Softmax Probability (MSP), while Liu *et al.* [11] introduced energy-based scores to improve OOD detection performance. Other approaches exploit feature-space information. ViM [12] models the in-distribution feature subspace, and deep nearest neighbor methods [13] detect OOD samples based on distances in the embedding space. Uncertainty-based approaches leverage Bayesian neural networks (BNNs) [2] or Gaussian Processes (GPs) [7] to estimate predictive uncertainty for OOD detection. Prior work has shown that Bayesian inference enables neural networks to assign higher uncertainty to out-of-distribution samples, providing a useful signal for OOD detection [8], [14], [15].

Recent works have explored OOD detection in the context of CLIP without retraining the backbone. MCM [16] utilizes the maximum class probability as an OOD score, demonstrating that CLIP's zero-shot predictions already contain useful signals for OOD detection. GL-MCM [17] further extends this idea by incorporating global feature statistics into the MCM score. Tip-Adapter [18] proposes a training-free, cache-based few-shot adaptation mechanism for improving in-distribution (ID) classification, and an extension for OOD detection has also been proposed [19]. Prompt-learning approaches have also been adapted for OOD detection. CoOp [4] introduces learnable prompt vectors to adapt CLIP in few-shot settings via gradient-based optimization. Building upon CoOp, LoCoOp [5] incorporates background-aware regularization to explicitly improve OOD detection performance. Other methods enhance OOD detection through alternative mechanisms. CLIPN [6] introduces explicit negative textual concepts for zero-shot OOD detection by adding a negative text encoder trained with "No" prompts, NPOS [20] leverages synthetic outliers generated in the feature space during training, and SeTAR [21] employs selective low-rank approximation to suppress non-discriminative components in CLIP representations. In contrast, our work introduces a Gaussian Process-based uncertainty modeling layer on top of CLIP embeddings, enabling principled OOD detection and few-shot classification without gradient-based training or additional OOD data.

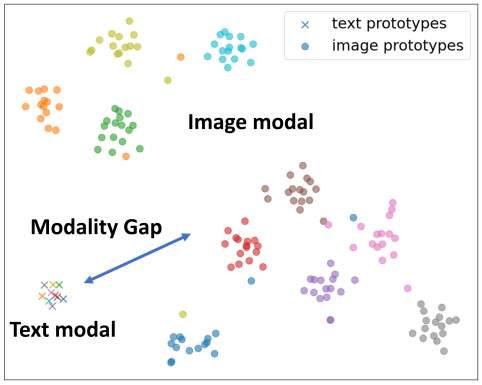


Fig. 1: t-SNE visualization of CLIP embeddings. Image and text embeddings exhibit distinct clustering patterns with a systematic offset, suggesting differences in their geometric structure within the shared embedding space.

III. Preliminaries

A. Gaussian Process Regression

A Gaussian Process is a non-parametric Bayesian model that defines a distribution over functions. It is characterized by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$, also known as kernel, which defines the similarity between input points. The kernel function plays a central role in controlling the smoothness and generalization behavior of the function by encoding prior assumptions about the function's structure. Formally, a GP is written as:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')).$$

Given training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ are input features and $y_i \in \mathbb{R}$ are noisy observations of the latent function $f(\mathbf{x}_i)$, we assume the following observation model:

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where σ^2 is the variance of the independent Gaussian noise added to each observation.

Under this model, the joint distribution of the training outputs and the function value at a test point $\mathbf{x}_*$ is given by:

$$\begin{bmatrix} \mathbf{y} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K + \sigma^2 I & k_* \\ k_*^\top & k_{**} \end{bmatrix}\right).$$

Here, $K \in \mathbb{R}^{n \times n}$ denotes the kernel matrix computed over the training inputs, where each entry $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. The vector $k_* \in \mathbb{R}^n$ represents the kernel evaluations between the training inputs and the test input $\mathbf{x}_*$, such that $(k_*)_i = k(\mathbf{x}_i, \mathbf{x}_*)$. The scalar $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$ is the kernel value at the test input itself.

The predictive distribution at test point $\mathbf{x}_*$ is Gaussian with mean and variance:

$$\mu_* = k_*^\top (K + \sigma^2 I)^{-1} \mathbf{y}, \tag{1}$$

$$\sigma_*^2 = k_{**} - k_*^\top (K + \sigma^2 I)^{-1} k_*. \tag{2}$$

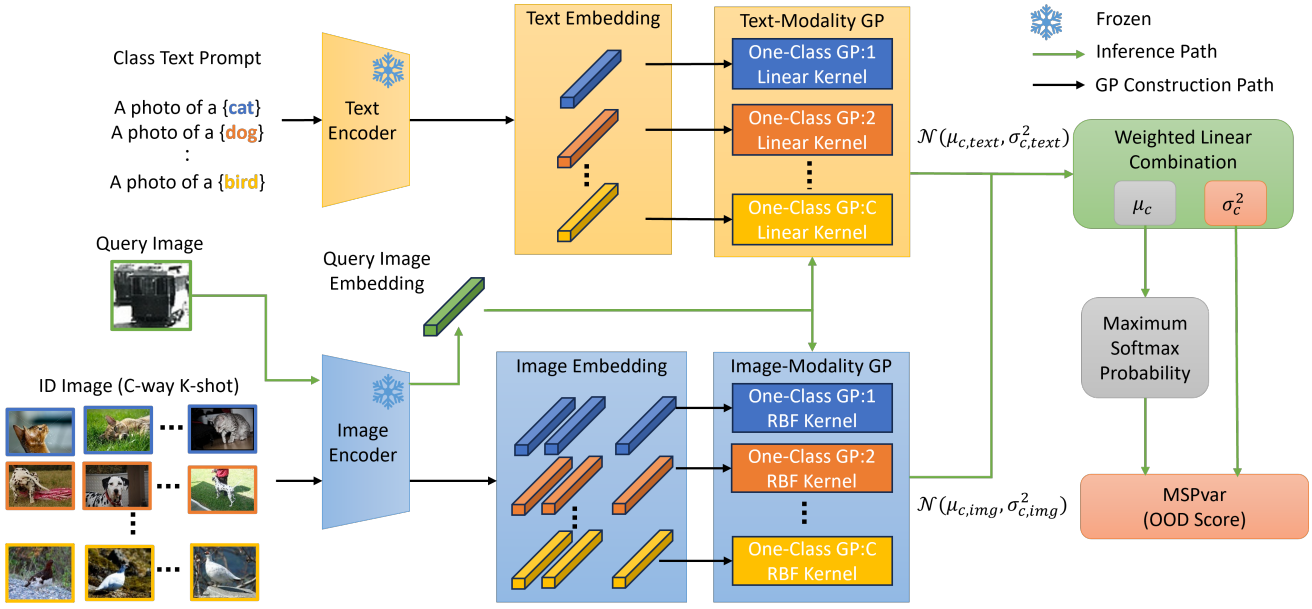


Fig. 2: **Overview of GP-Adapter with one-class Gaussian Processes.** Frozen CLIP encoders produce image embeddings from a small K-shot support set and a query image, and text embeddings from class-specific text prompts. For each class, modality-specific one-class Gaussian Processes output predictive means and variances. These are fused via a weighted combination and converted into OOD scores using the proposed variance-aware MSP.

This formulation allows GPs to provide not only point predictions, computed as the predictive mean, but also uncertainty estimates, computed by the predictive variance, which are crucial for tasks such as few-shot learning and out-of-distribution (OOD) detection. A larger predictive variance indicates lower confidence in the prediction, whereas a smaller variance indicates higher confidence.

B. CLIP Representations

CLIP is a multimodal representation learning framework that embeds images and text into a shared semantic space using separate modality-specific encoders.

Given an image $\mathbf{x}_{\text{img}}$ and a set of candidate class descriptions $\{\mathbf{t}_1, \ldots, \mathbf{t}_C\}$, CLIP produces ℓ_2-normalized feature vectors $\mathbf{f}_{\text{img}} \in \mathbb{R}^d$ and $\mathbf{f}_{\text{text},c} \in \mathbb{R}^d$ for each class c. The prediction score for class c is computed using cosine similarity:

$$s_c = \mathbf{f}_{\text{img}}^\top \mathbf{f}_{\text{text},c}. \tag{3}$$

The predicted class is then given by:

$$\hat{c} = \arg\max_c \; s_c. \tag{4}$$

These scores can be interpreted as similarity-based logits and are commonly used for zero-shot classification.

Although CLIP aligns image and text embeddings in a shared space, Fig. 1 suggests a residual cross-modal misalignment, where the image and text embeddings form distinct clusters and exhibit a systematic offset. This observation highlights that, despite semantic alignment, the two modalities exhibit different geometric characteristics in the embedding space.

IV. METHODOLOGY

A. Problem Setup

We consider a few-shot classification setting with C classes and a small labeled cache of K examples per class, denoted as $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{CK}$, where x_i is an image and $y_i \in \{1, \ldots, C\}$ is its class label. Given a query image x_q, our objective is to predict its class label for ID classification and leverage predictive uncertainty as a signal for OOD detection.

B. Overview of our GP-Adapter

Fig. 2 illustrates the architecture of **GP-Adapter**, which augments CLIP with GP-based Bayesian inference for few-shot OOD detection. The pipeline combines CLIP's multimodal representations with GP-based uncertainty estimation to enable uncertainty-aware predictions under data scarcity and distribution shift.

We begin by constructing a K-shot support set for each class c. For the image modality, we encode these C-way K-shot labeled images using CLIP's image encoder to obtain image embeddings. For the text modality, we use class-specific text prompts (e.g., a photo of a {class}) and encode them using CLIP's text encoder to obtain text embeddings. Following [3], we adopt a prompt-ensemble strategy and use the standard set of seven ImageNet templates. These embeddings are then used to construct modality-specific kernel matrices: an RBF kernel for image embeddings and a linear kernel for text embeddings, which are used to construct independent one-class GPs for each modality.

At inference time, the frozen CLIP image encoder encodes the query image x_q into an ℓ_2-normalized embedding $z_q =$

$f_{\text{img}}(x_q)$. For each class c, we construct two independent one-class GP regressors, one for the image modality and one for the text modality. Given the original multi-class labels $y_i \in \{1, \ldots, C\}$, we assign a positive regression target $\tilde{y}_i^{(c)} = 1$ to all support samples belonging to class c when constructing the one-class GPs for each modality.

Using these targets, the two GPs produce predictive distributions over a latent class score $s_c(z_q)$ for the query embedding:

$$p(s_c(z_q) \mid z_q)_{\text{img}} = \mathcal{N}\big(\mu_{c,\text{img}}(z_q), \sigma^2_{c,\text{img}}(z_q)\big)\,,$$
$$p(s_c(z_q) \mid z_q)_{\text{text}} = \mathcal{N}\big(\mu_{c,\text{text}}(z_q), \sigma^2_{c,\text{text}}(z_q)\big)\,.$$

The predictive mean and variance are computed using the standard GP regression formulas given in (1) and (2). For the image modality, the GP prediction is obtained by computing kernel values between the query image embedding and the support image embeddings. For the text modality, the query image embedding is compared to the fixed text embeddings via the linear kernel, leveraging the alignment property of CLIP's embedding space, and the GP prediction is obtained using the corresponding kernel values and the precomputed Gram matrix among text embeddings.

Assuming independence between the two modalities, we fuse their predictive distributions via a weighted linear combination. The resulting predictive mean and variance for class c are given by:

$$\mu_c(z_q) = \alpha\mu_{c,\text{img}}(z_q) + (1-\alpha)\mu_{c,\text{text}}(z_q),$$
$$\sigma^2_c(z_q) = \alpha^2\sigma^2_{c,\text{img}}(z_q) + (1-\alpha)^2\sigma^2_{c,\text{text}}(z_q),$$

where $\alpha \in [0, 1]$ controls the contribution of each modality. In our experiments, we set $\alpha = 0.15$ unless otherwise specified.

We convert the fused predictive means $\{\mu_c(z_q)\}_{c=1}^C$ into class probabilities using the softmax function:

$$p(y = c \mid z_q) = \frac{\exp(\mu_c(z_q))}{\sum_{c'=1}^C \exp(\mu_{c'}(z_q))}. \tag{5}$$

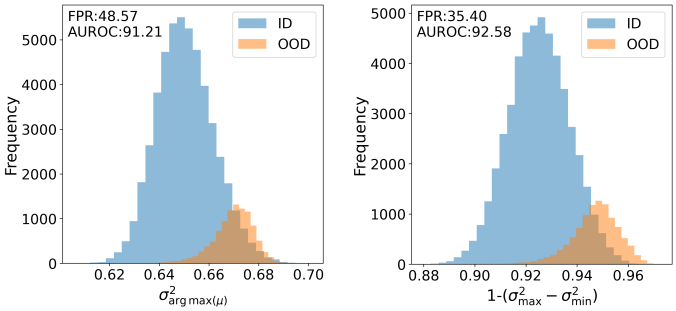
Following the Maximum Softmax Probability (MSP) criterion, the confidence score for a query sample is defined as:

$$\text{MSP}(z_q) = \max_c\, p(y = c \mid z_q). \tag{6}$$

To incorporate predictive uncertainty, we further define a normalized variance-aware confidence score that accounts for the range of predictive variances across classes:

$$\text{MSP}_{\text{var}}(z_q) = \text{MSP}(z_q)\left(1 + \frac{\sigma^2_{\max}(z_q) - \sigma^2_{\min}(z_q)}{\sigma^2_{\max}(z_q) + \sigma^2_{\min}(z_q)}\right), \tag{7}$$

where $\sigma^2_{\max}(z_q)$ and $\sigma^2_{\min}(z_q)$ represent the maximum and minimum predictive variances across classes, respectively. Lower MSP_{var} values indicate higher uncertainty and are used to identify OOD inputs. For OOD detection, we apply a threshold λ: if $\text{MSP}_{\text{var}}(z_q) \geq \lambda$, the sample is detected as ID; otherwise, detected as OOD. This design is motivated by empirical observations that the variance range across classes provides a stronger signal for separating ID and OOD samples than the variance of the most confident class alone. Since GP predictive variance is typically lower near the support set and higher far from it, ID queries yield a wide max–min variance range, whereas OOD queries tend to produce more uniformly high variances and thus a narrower range. As illustrated in Fig. 3, incorporating class-wise variance range leads to improved separation between ID and OOD distributions.



(a) Predicted-class variance (b) Max-min variance range

Fig. 3: Range of GP predictive variances across classes improves separation between ID and OOD samples.

C. Kernel Selection

A key design choice in GP-Adapter is the use of modality-specific kernels, which is motivated by the geometric properties of CLIP embeddings. Although CLIP aligns image and text representations in a shared embedding space, the two modalities exhibit different statistical characteristics and residual misalignment, as illustrated in Fig. 1. This observation suggests that applying a single distance-sensitive kernel uniformly across modalities may lead to distorted similarity estimates.

For the image modality, we use the Radial Basis Function (RBF) kernel, which assumes local smoothness and captures non-linear relationships in the feature space. In CLIP, image embeddings exhibit a geometry where Euclidean distance provides a meaningful measure of visual similarity, making the RBF kernel a natural choice. The kernel is defined as:

$$k_{\text{img}}(\mathbf{z}, \mathbf{z}') = \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}'\|^2}{2\theta^2}\right), \tag{8}$$

where θ is the length-scale parameter.

For the text modality, we apply a linear kernel on ℓ_2-normalized prompt embeddings. Since CLIP computes image-text similarity using cosine similarity, the inner product between ℓ_2-normalized embeddings is directly aligned with CLIP's standard scoring mechanism. This choice avoids introducing an additional distance metric that may amplify residual cross-modal misalignment. Formally, the text kernel is given by:

$$k_{\text{text}}(\mathbf{t}, \mathbf{t}') = \mathbf{t}^\top \mathbf{t}', \tag{9}$$

and the cross-covariance between a query image embedding $\mathbf{z}_q$ and a text embedding $\mathbf{t}$ is computed as $k_{\text{text}}(\mathbf{t}, \mathbf{z}_q) = \mathbf{t}^\top \mathbf{z}_q$.

By restricting distance-sensitive kernels to the image modality and using an inner-product-based kernel for text prompts, GP-Adapter respects the modality-specific geometry of CLIP

embeddings. This design reduces sensitivity to cross-modal misalignment and leads to more stable similarity estimation and uncertainty modeling in OOD scenarios.

D. Learning Kernel Hyperparameters

To effectively model uncertainty in few-shot settings while maintaining scalability, we optimize kernel hyperparameters independently for each one-class Gaussian Process. Specifically, for each class c, a separate GP is constructed using only its K-shot support samples.

Accordingly, the training inputs for class c are given by $\mathbf{X}^{(c)} \in \mathbb{R}^{K\times d}$, consisting of K support embeddings, and the corresponding positive regression targets $\tilde{\mathbf{y}}^{(c)} \in \mathbb{R}^K$ are defined as in Sec. IV-B. For each class-specific GP, kernel hyperparameters are selected by maximizing the log marginal likelihood:

$$\log p(\tilde{\mathbf{y}}^{(c)} \mid \mathbf{X}^{(c)}) = -\frac{1}{2}\tilde{\mathbf{y}}^{(c)\top}\big(K^{(c)} + \sigma^2 I\big)^{-1}\tilde{\mathbf{y}}^{(c)} - \frac{1}{2}\log\big|K^{(c)} + \sigma^2 I\big| - \frac{K}{2}\log 2\pi,$$

where $K^{(c)}$ denotes the kernel matrix computed from $\mathbf{X}^{(c)}$.

In practice, we assume a nearly noise-free observation model and fix the noise variance to $\sigma^2 = 10^{-6}$ for numerical stability. We parameterize the kernel matrix as:

$$K^{(c)}(\mathbf{X}^{(c)}, \mathbf{X}^{(c)}) = \sigma_f^{2(c)} K_0(\mathbf{X}^{(c)}, \mathbf{X}^{(c)} \mid \theta^{(c)}), \tag{10}$$

where $\sigma_f^{2(c)}$ is the signal variance and $\theta^{(c)}$ denotes the kernel length-scale. For a fixed $\theta^{(c)}$, the optimal signal variance admits a closed-form solution:

$$\hat{\sigma}_f^{2(c)} = \frac{1}{K}\tilde{\mathbf{y}}^{(c)\top} K_0(\mathbf{X}^{(c)}, \mathbf{X}^{(c)} \mid \theta^{(c)})^{-1}\tilde{\mathbf{y}}^{(c)}. \tag{11}$$

For the image modality, the length-scale $\theta^{(c)}$ of the RBF kernel is selected via grid search, while the signal variance $\sigma_f^{2(c)}$ of the RBF kernel is updated in closed form for each candidate length-scale using (11), resulting in a profile likelihood optimization. For the text modality, the signal variance $\sigma_f^{2(c)}$ of the linear kernel is updated in closed form without iterative search.

To mitigate overfitting under extremely limited data, we impose an upper bound τ on the log marginal likelihood during length-scale selection: candidate length-scales whose log marginal likelihood exceeds τ are excluded, and we select the best length-scale among the remaining candidates. That is, we maximize the log marginal likelihood subject to $\log p(\tilde{\mathbf{y}}^{(c)} \mid \mathbf{X}^{(c)}) \le \tau$. Unless otherwise specified, we set $\tau = -5$ in all experiments.

Importantly, since each class-specific GP is built from only K support samples, the memory footprint for class-wise kernel matrices scales as $O(CK^2)$ rather than $O(N^2)$, where C denotes the number of classes and N the total number of training samples. This design enables scalable GP-based uncertainty modeling even for large-scale datasets such as ImageNet-1k under few-shot settings.

V. Experiments

A. Experimental Setting

Datasets. We evaluate our method on ImageNet-1k [9] as the ID classification benchmark. For OOD detection, we use iNaturalist [24], SUN [25], Places [26], and Textures [27] which are commonly used OOD test datasets. These OOD datasets are disjoint from ImageNet-1k classes.

Few-shot protocol. We sample $K \in \{1, 2, 4, 8, 16\}$ labeled images per class from the ID training set to construct the few-shot cache. No additional fine-tuning of CLIP is performed. For fair comparison, all few-shot methods are optimized using the same class splits and K-shot support samples.

Implementation Details. All experiments are implemented in PyTorch, and the Gaussian Process modules are implemented using GPyTorch. All training and evaluation are conducted on a single NVIDIA Tesla V100 GPU with 32GB memory. All reported results are averaged over three random seeds to ensure robustness. CLIP ViT-B/16 [28] is used as the default backbone. The RBF kernel length-scale θ is selected via grid search over a range from 0.10 to 2.0 with a step size of 0.05. Unless otherwise specified, all hyperparameters are fixed to their default values across all experiments.

Baselines. We compare our method with several existing methods for zero-shot OOD detection, fine-tuned OOD detection, and few-shot OOD detection. For zero-shot OOD detection, we compare with MCM [16], GL-MCM [17], and SeTAR [21], which are simple yet strong OOD detection methods. For few-shot OOD detection, we compare with CoOp [4] and LoCoOp [5]. For fine-tuned OOD detection, we compare with unimodal methods, MSP [22], ODIN [23], ViM [12], KNN [13], as well as multimodal methods, NPOS [20] and CLIPN [6].

Evaluation Metrics. For ID classification on the ImageNet-1k dataset, we compute Top-1 accuracy on the ID test set. For OOD detection, we treat ID samples as negative and OOD samples as positive. We report threshold-free AUROC (higher is better) and FPR95 (false positive rate at 95% TPR; lower is better).

B. Experimental Results

Table I shows OOD detection performance across multiple datasets. Among few-shot methods, GP-Adapter achieves competitive performance compared to few-shot prompt-learning baselines such as CoOp and LoCoOp, despite not relying on gradient-based optimization.

Notably, combining GP-Adapter with prompt-learning methods consistently improves OOD detection performance across datasets. In particular, GP-Adapter+LoCoOp achieves the best average AUROC and the lowest average FPR95, demonstrating that GP-based uncertainty modeling is complementary to prompt learning. These results indicate that GP-Adapter can effectively enhance existing few-shot CLIP adaptation methods by modeling predictive uncertainty. Furthermore, GP-Adapter-based methods outperform zero-shot baselines and are competitive with fine-tuned OOD detection

TABLE I: Comparison of OOD detection performance with ImageNet-1k as ID data. ↓ indicates lower is better, ↑ indicates higher is better. † indicates results reported in [20], ‡ indicates results reported in [6], and †‡ indicates results reported in [21].

Method	iNaturalist		SUN		Places		Texture		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Zero-shot										
MCM [16]	30.94	94.61	37.67	92.56	44.76	89.76	57.91	86.10	42.82	90.76
GL-MCM [17]	15.18	96.71	30.42	93.09	38.85	89.90	57.93	83.63	35.47	90.83
SeTAR [21]†‡	26.92	94.67	35.57	92.79	42.64	90.16	55.83	86.58	40.24	91.05
Fine-tuned										
MSP [22]†	54.05	87.43	73.37	78.03	72.98	78.03	68.85	79.06	67.31	80.64
ODIN [23]†	30.22	94.65	54.04	87.17	55.06	85.54	51.67	87.85	47.75	88.80
ViM [12]†	32.19	94.65	54.01	87.19	60.67	83.75	53.94	87.17	50.45	88.19
KNN [13]†	29.17	91.52	52.62	86.12	57.02	84.43	54.93	87.12	48.44	87.30
NPOS [20]†	16.58	96.19	43.77	90.44	49.62	87.54	43.65	89.10	38.91	90.82
CLIPN [6]‡	23.94	95.27	**26.17**	**93.93**	**33.45**	90.93	40.83	**92.28**	31.10	93.10
Few-shot (16-shot)										
CoOp [4]	29.63	93.59	35.33	92.42	42.31	89.89	41.58	90.59	37.21	91.62
LoCoOp [5]	23.78	95.20	30.75	93.89	37.27	91.24	41.70	91.05	33.38	92.85
GP-Adapter (ours)	15.78	96.78	37.54	92.36	42.32	90.37	48.52	89.37	36.04	92.12
GP-Adapter+CoOp (ours)	16.24	96.53	32.78	93.02	37.27	**91.58**	**39.27**	91.74	31.39	93.22
GP-Adapter+LoCoOp (ours)	**12.66**	**97.23**	31.52	93.60	37.28	91.56	39.47	91.94	**30.23**	**93.58**

TABLE II: Comparison in ID accuracy on ImageNet-1k validation data.

Method	Top-1 Accuracy
MCM	68.53
CoOp	71.93
LoCoOp	71.63
GP-Adapter(ours)	70.25
GP-Adapter+CoOp(ours)	**72.10**
GP-Adapter+LoCoOp(ours)	71.51

TABLE III: Effect of upper bound τ on OOD detection performance (evaluated under GP-Adapter+LoCoOp).

τ	FPR95↓	AUROC↑
∞	33.12	92.93
5	33.29	93.02
0	**29.99**	93.55
-5	30.23	**93.58**
-10	33.43	93.08

approaches, highlighting their effectiveness under data-scarce settings.

In addition to OOD detection, we evaluate ID classification accuracy. As shown in Table II, GP-Adapter preserves competitive ID accuracy compared to existing baselines. Moreover, integrating GP-Adapter with CoOp improves Top-1 accuracy over the original CoOp method, while its combination with LoCoOp maintains comparable performance. These results confirm that the proposed method enhances robustness to distribution shifts without sacrificing in-distribution classification accuracy.

C. Ablation Study

Ablation on the Upper Bound τ. Table III shows the effect of the upper bound τ that prevents overfitting in few-shot settings, on OOD detection performance. In this ablation, we use the LoCoOp-learned prompts (GP-Adapter+LoCoOp) and fix the modality weight α to 0.15, while evaluating $\tau \in \{-10, -5, 0, 5, \infty\}$ on OOD detection performance. Here, $\tau = \infty$ indicates disabling the upper bound. $\tau = 0$ or $\tau = -5$ yield improved OOD detection performance, achieving lower FPR95 and higher AUROC.

TABLE IV: Effect of modality ratio α on OOD detection performance (evaluated under GP-Adapter+LoCoOp).

α	FPR95↓	AUROC↑
0.05	31.94	93.16
0.1	30.46	93.47
0.15	**30.23**	**93.58**
0.2	32.09	93.29

Ablation on the ratio of modality α. Table IV shows the effect of the modality weight α that controls the contribution of image and text modalities in the combined predictive distribution. In this ablation, we use the LoCoOp-learned prompts (GP-Adapter+LoCoOp) and fix the upper bound τ to -5, while varying $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$. We observe that $\alpha = 0.15$ yields improved OOD detection performance, achieving lower FPR95 and higher AUROC.

Ablation on Variance-Aware MSP Calibration. We investigate the effect of incorporating GP predictive variance into the OOD score using the proposed MSP_{var} defined in (7). Fig. 4 compares OOD detection performance with and without variance-aware MSP calibration under different prompt settings. Across all settings, incorporating GP predictive variance consistently improves AUROC, demonstrating the effectiveness of variance-aware MSP calibration. While the improvement for GP-Adapter alone is modest, substantial performance gains are observed when combined with

TABLE V: Comparison of OOD detection performance with ResNet-50 backbone. ↓ indicates lower is better, ↑ indicates higher is better.

Method	iNaturalist		SUN		Places		Texture		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM	31.98	93.86	46.09	**90.75**	60.56	85.67	60.00	85.72	49.66	89.00
CoOp	34.35	93.17	**44.97**	90.60	55.63	86.35	**42.82**	**89.94**	44.40	90.02
LoCoOp	39.86	92.62	51.55	90.23	58.66	86.34	49.07	89.18	49.79	89.60
GP-Adapter (ours)	29.37	94.22	46.97	90.39	56.08	86.66	55.32	87.26	46.93	89.63
GP-Adapter+CoOp (ours)	**23.07**	**95.36**	47.37	89.88	**54.26**	**86.92**	44.39	89.87	**42.27**	**90.51**
GP-Adapter+LoCoOp (ours)	24.25	95.25	51.40	89.72	55.96	86.76	48.53	89.56	45.03	90.32

TABLE VI: Comparison of OOD detection performance with ImageNet-100 as ID data. ↓ indicates lower is better, ↑ indicates higher is better.

Method	iNaturalist		SUN		Places		Texture		Average	
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MCM	18.13	96.77	36.45	94.54	34.52	94.36	41.22	92.25	32.58	94.48
CoOp	19.77	96.55	19.69	96.27	23.78	95.22	15.95	96.91	19.80	96.24
LoCoOp	7.37	98.16	18.43	**96.80**	22.88	95.71	**15.07**	**97.20**	15.94	**96.97**
GP-Adapter (ours)	**2.46**	**99.19**	19.09	96.62	**21.60**	**96.13**	23.97	95.86	16.78	96.95
GP-Adapter+CoOp (ours)	12.89	97.44	**17.83**	96.51	21.61	95.57	17.37	96.71	17.42	96.56
GP-Adapter+LoCoOp (ours)	3.91	98.89	20.50	96.31	23.41	95.46	15.53	96.98	**15.84**	96.91

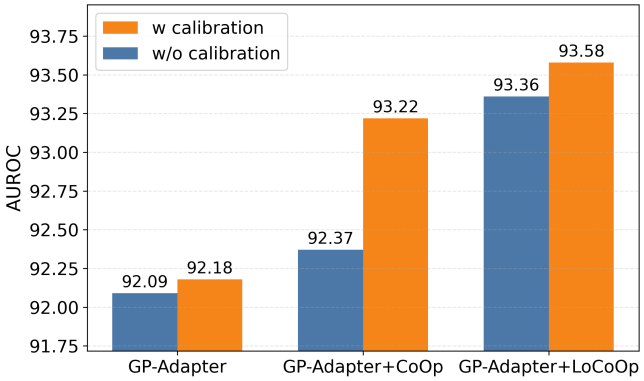
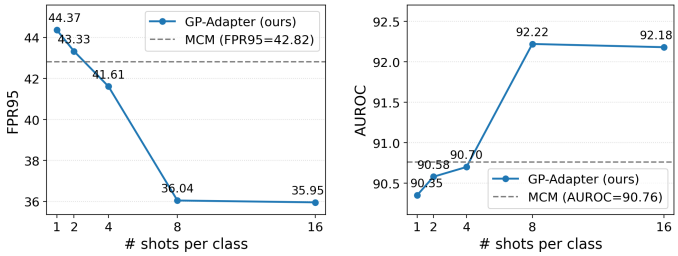


Fig. 4: Ablation of variance-aware MSP calibration (MSP_{var}) for GP-Adapter predictions under different prompt settings.



(a) FPR95 vs. shots (b) AUROC vs. shots

Fig. 5: Effect of the number of shots per class on OOD detection performance, compared to MCM baseline.

prompt-learning methods. In particular, GP-Adapter+CoOp shows a significant increase in AUROC when MSP_{var} is applied. This suggests that variance-aware MSP calibration is especially beneficial when prompt learning sharpens class confidence, helping mitigate overconfident predictions. For GP-Adapter+LoCoOp, performance gains are smaller but still consistent, indicating that variance-aware calibration provides complementary benefits even when the base method is already strong.

Ablation on the Number of Shots. Fig. 5 shows the effect of varying the number of shots per class on OOD detection performance. In this experiment, we fix the modality weight α to 0.15. For numerical stability, we set the upper bound $\tau = 0$ for the 1-shot and 2-shot settings, and use $\tau = -5$ for all other shot settings. GP-Adapter exhibits comparable performance to the MCM baseline in extremely low-shot regimes (1,2,4 shots), and begins to consistently outperform it at 8 shots, with the largest improvement observed between 4 and 8 shots settings, where both FPR95 decreases substantially and AUROC increases.

Different Backbone. Table V shows OOD detection performance using CLIP with a ResNet-50 (RN50) backbone [29]. GP-Adapter achieves competitive OOD detection performance compared to MCM, CoOp, and LoCoOp. Moreover, combining GP-Adapter with CoOp or LoCoOp continues to yield performance gains under the RN50 backbone.

Small Scale ID Dataset (ImageNet-100). We further evaluate GP-Adapter under a reduced ID dataset setting using ImageNet-100, following the same set of 100 ImageNet classes as prior work [20]. Table VI shows OOD detection performance when the number of ID classes is limited. Under this setting, GP-Adapter achieves strong OOD detection performance compared to existing baselines. In particular, GP-Adapter attains the best FPR95 and AUROC on iNaturalist, and maintains competitive average performance across all OOD datasets. Moreover, combining GP-Adapter with prompt-learning methods continues to provide stable performance.

These results suggest that GP-Adapter remains effective even when the number of ID classes is substantially reduced.

VI. Conclusion

In this paper, we proposed GP-Adapter, a lightweight adaptation method that augments frozen CLIP representations with class-wise one-class Gaussian Processes for few-shot classification and OOD detection. GP-Adapter avoids gradient-based optimization and relies on closed-form GP inference with simple kernel hyperparameter selection, allowing rapid adaptation with memory footprint scaling as $O(CK^2)$ in few-shot settings. Extensive experiments on ImageNet-1k and multiple OOD benchmarks demonstrate that GP-Adapter achieves strong and competitive OOD detection performance, and remains effective when the number of ID classes is substantially reduced or when different backbone architectures are used. Moreover, GP-Adapter is complementary to prompt-learning methods such as CoOp and LoCoOp, and their combination yields stable performance gains. Our ablation analysis further shows that incorporating GP predictive variance into OOD scoring provides additional improvements. Future work includes extending GP-Adapter to other few-shot recognition settings, as well as exploring variational sparse Gaussian Process approximations to broaden its applicability beyond extremely low-shot regimes.

References

[1] S. Lu, Y. Wang, L. Sheng, L. He, A. Zheng, and J. Liang, "Out-of-distribution detection: A task-oriented survey of recent advances," *ACM Comput. Surv.*, 2025.

[2] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48, 2016, pp. 1050–1059.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 139, 2021, pp. 8748–8763.

[4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[5] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Locoop: Few-shot out-of-distribution detection via prompt learning," in *Advances in Neural Information Processing Systems*, 2023, pp. 76 298–76 310.

[6] H. Wang, Y. Li, H. Yao, and X. Li, "Clipn for zero-shot ood detection: Teaching clip to say no," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1802–1812.

[7] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[8] Y. Chen, C.-L. Sung, A. Kusari, X. Song, and W. Sun, "Uncertainty-aware out-of-distribution detection with gaussian processes," *arXiv preprint arXiv:2412.20918*, 2024.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[10] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017, poster session.

[11] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[12] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4911–4920.

[13] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 20 827–20 840.

[14] T. Saito, K. Ando, and T. Asai, "Extending binary neural networks to bayesian neural networks with probabilistic interpretation of binary weights," *IEICE TRANSACTIONS on Information*, vol. E107-D, no. 8, pp. 949–957, 2024.

[15] K. Minagawa, T. Saito, S. Kojima, K. Ando, and T. Asai, "Out-of-distribution data detection using bayesian convolutional neural network with variational inference," in *2024 International Joint Conference on Neural Networks (IJCNN)*, 2024, pp. 1–8.

[16] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

[17] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Gl-mcm: Global and local maximum concept matching for zero-shot out-of-distribution detection," *International Journal of Computer Vision*, vol. 133, no. 12, pp. 3586–3596, 2025.

[18] R. Zhang, Z. Wei, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *Proceedings of the 17th European Conference on Computer Vision (ECCV)*, 2022, pp. 493–510.

[19] X. Chen, Y. Li, and H. Chen, "Dual-adapter: Training-free dual adaptation for few-shot out-of-distribution detection," *arXiv preprint arXiv:2405.16146*, 2024.

[20] L. Tao, X. Du, X. Zhu, and Y. Li, "Non-parametric outlier synthesis," in *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[21] Y. Li, B. Xiong, G. Chen, and Y. Chen, "Setar: Out-of-distribution detection with selective low-rank approximation," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[22] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, pp. 7068–7081.

[23] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[24] G. Van Horn, O. Mac Aodha, Y. Song, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8769–8778.

[25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492, 2010.

[26] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[27] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 3606–3613.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.