ReBaPL: Repulsive Bayesian Prompt Learning

Yassir Bendou[1,⋆,†], Omar Ezzahir[1,⋆], Eduardo Montesuma[1,⋆]
Gabriel Mahuas[1], Victoria Schevchenko[1], Mike Gartrell[2,†]
[1] Sigma Nova, Paris, France [2]Rhizome Labs, Paris, France
first-name.second-name@sigmanova.ai
first-name.second-name@rhizomelabs.ai

Abstract

Prompt learning has emerged as an effective technique for fine-tuning large-scale foundation models for downstream tasks. However, conventional prompt learning methods are prone to overfitting and can struggle with out-of-distribution generalization. To address these limitations, Bayesian prompt learning has been proposed, which frames prompt optimization as a Bayesian inference problem to enhance robustness. This paper introduces Repulsive Bayesian Prompt Learning (ReBaPL), a novel method for Bayesian prompt learning, designed to efficiently explore the complex and often multimodal posterior landscape of prompts. Our method integrates a cyclical step-size schedule with a stochastic gradient Hamiltonian Monte Carlo (SGHMC) algorithm, enabling alternating phases of exploration to discover new modes, and exploitation to refine existing modes. Furthermore, we introduce a repulsive force derived from a potential function over probability metrics (including Maximum Mean Discrepancy and Wasserstein distance) computed on the distributions of representations produced by different prompts. This representation-space repulsion diversifies exploration and prevents premature collapse to a single mode. Our approach allows for a more comprehensive characterization of the prompt posterior distribution, leading to improved generalization. In contrast to prior Bayesian prompt learning methods, our method provides a modular plug-and-play Bayesian extension of any existing prompt learning method based on maximum likelihood estimation. We demonstrate the efficacy of ReBaPL on several benchmark datasets, showing superior performance over state-of-the-art prompt learning methods.

1. Introduction

Large-scale vision-language models (VLMs), such as CLIP [34], have demonstrated remarkable zero-shot generalization capabilities across a wide array of visual tasks. Many works have shown better performance on downstream task by adapting VLMs for few-shot image classification [1, 12, 45–47]. Among these methods, prompt learning has emerged as an efficient model adaptation technique, enabling VLMs and other foundation models to be tailored for specific downstream tasks by learning continuous prompt vectors, instead of fine-tuning the entire model. However, this approach is not without its drawbacks. Standard prompt learning, which typically optimizes prompts through maximum likelihood estimation (MLE), is prone to overfitting on training data, leading to diminished generalization performance on out-of-distribution (OOD) samples and unseen classes.

The initial MLE-based prompt learning method, known as CoOp [47], is especially prone to overfitting [27]. In an attempt to mitigate this issue, a subsequent MLE-based prompt learning method, known as CoCoOp [46], proposed learning an instance-specific continuous prompt that is conditioned on the input image. While CoCoOp performs better than CoOp, it can still suffer from generalization issues.

In contrast to prompt learning for the text modality only, other recent approaches involve multi-modal prompt learning for both image and text representations [4, 16, 21, 43]. These approaches introduce learnable prompt tokens at varying depths in the transformer layers in the image and text encoders, and then generally use a vision-language coupling function to induce learning of prompts in a shared embedding space.

Other MLE prompt learning methods [22, 26] generally focus on regularization-based approaches to mitigate overfitting [46, 48]. One such approach, PromptSRC [22], uses self-regulating constraints to prevent prompts from losing the generalized knowledge of the pretrained model. The ProDA method [26] is a probabilistic approach, where the

⋆ Equal contribution.
† Corresponding author.
Our code is available at https://github.com/SigmaNova/ReBaPL

prompt distribution is fit to a Gaussian distribution using MLE, with a regularization term to improve prompt diversity. While effective at mitigating overfitting, these methods guide the learning process towards a single optimal solution manifold and may not fully capture the complex, potentially multi-modal posterior distribution of effective prompts.

Rather than seeking a single point estimate with regularization, an alternative paradigm is to characterize the full distribution over prompts through Bayesian inference. Bayesian prompt learning methods [4, 5, 9, 23, 33] frame prompt learning as a Bayesian inference problem. These methods aim to enhance robustness and improve generalization by inferring a posterior distribution over the prompt space. This probabilistic perspective naturally introduces regularization, which helps prevent the model from learning spurious features and overfitting to the training set. Most Bayesian prompt learning approaches estimate the posterior using a variational unimodal Gaussian approximation, which limits diversity. To address this limitation, the Adaptive Particle-based Prompt Learning (APP) approach [5] uses a Wasserstein gradient flow and Stein Variational Gradient Descent (SVGD) to approximate multiple modes in the posterior using a variational distribution represented by a collection of interacting particles. VaMP [4] extends multi-modal (text-image) prompt learning by introducing variational inference to model prompts as probabilistic latent variables rather than deterministic parameters. This enables instance-specific, uncertainty-aware prompt generation, where text prompts are dynamically conditioned on input image features and sampled from learned posterior distributions, with a class-aware prior providing semantic regularization. While VaMP incorporates uncertainty modeling into multi-modal prompting, it relies on a variational approximation with a unimodal Gaussian posterior, which may limit its ability to capture the full complexity of multi-modal prompt distributions.

We propose a novel Repulsive Bayesian Prompt Learning (ReBaPL) approach based on cyclical stochastic gradient Hamiltonian Monte Carlo (rcSGHMC). In contrast to the deterministic SVGD method used in APP, our approach leverages an efficient MCMC algorithm where the posterior is represented as a collection of samples, coupled to a repulsive force based on the interaction between prompt representations. This allows our method to provide a richer representation of the shape of the high-density regions around multiple modes in the posterior, which results in better generalization to novel classes without overfitting on the base classes. In contrast to prior Bayesian prompt learning methods, our method provides a modular plug-and-play Bayesian extension of any existing MLE-based prompt learning method. We implement and perform experiments with ReBaPL running as a plug-and-play learning algorithm on top of two existing prompt learning methods: MaPLe [21] and MMRL [16].

Our contributions include:

- Repulsive cyclical SGHMC: We propose the rcSGHMC algorithm for approximating complex multimodal posteriors. Through the use of Hamiltonian dynamics, cyclical learning rates, and repulsion between prompt representations, our method is effective at exploring the complex posteriors with multiple modes that are found in prompt learning.
- Representation-based repulsion: Rather than comparing parameters directly in weight space, our method introduces a repulsive potential based on probability metrics, including Maximum Mean Discrepancy (MMD) and Wasserstein distance, between the distributions of representations. This allows us to capture the functional similarity between prompts through their induced representations, encouraging exploration of functionally diverse modes in the posterior.
- We show experimentally that our approach provides rich characterization of the space of diverse prompts, which improves generalization performance on base-to-novel tasks, cross-dataset transfer, and domain generalization.

2. Background

In this section we introduce the core principles of our ReBaPL method, namely: prompt learning (Section 2.1), Bayesian learning (Section 2.2), and probability metrics (Section 2.3). Figure 1 provides an overview of our approach.

2.1. Prompt Learning

Contrastive Language-Image Pretraining (CLIP) [34] is a pre-training method that learns a joint latent space for texts and images. Its main principle is *encoding* an image through a Convolutional Neural Net (CNN), e.g., a ResNet [17]) or a Transformer, e.g., a ViT [10], and a text snippet through a Transformer [38]. From a pre-trained CLIP, it is possible to perform downstream tasks, such as classification. For example, given an image x, one can measure the alignment with a textual prompt T in the latent space. A good candidate for the textual prompt is *"A photo of [CLS]"*, where *"[CLS]"* corresponds to the class with which one is measuring the alignment. In this sense, CLIP already demonstrates its own *zero-shot* generalization ability, mainly due to pre-training. Prompt learning takes this idea further, where the prompt is treated as an *optimization variable during learning* [47].

Early prompt learning methods for VLMs, such as CoOp [47] and CoCoOp [46], focus exclusively on learning prompts in the *language branch* of CLIP. While these approaches demonstrate improved performance on downstream tasks, they adopt a *unimodal* prompting strategy that only partially adapts the pre-trained model. This limitation

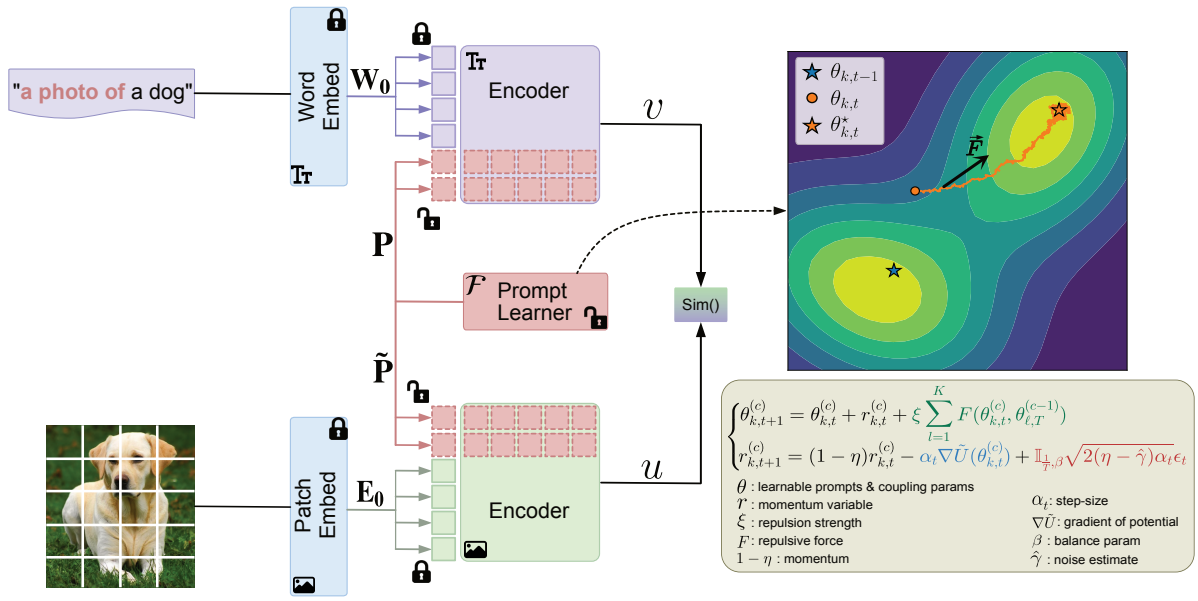

Figure 1. Overview of our proposed ReBaPL approach, in a multi-modal prompt learning setting. Text and image encoders receive text ($\mathbf{W_0}$) and image ($\mathbf{E_0}$) embeddings as input, combined with learnable tokens of context prompts ($\mathbf{P}$). The terms pertaining to the exploration and sampling stages and the repulsion force are colored in blue, red and green respectively.

motivated the development of *multi-modal prompt learning*, which recognizes that both the vision and language encoders should be adapted simultaneously to achieve optimal alignment between modalities [21]. While our ReBaPL approach can be run on top of any MLE-based prompt learning method, we focus primarily on multi-modal prompt learning for the remainder of this paper, since these methods tend to significantly outperform unimodal methods in terms of predictive performance [16, 21].

Multi-modal Prompt Learning Given CLIP's dual-encoder architecture with text encoder $\mathcal{L}$ and image encoder $\mathcal{V}$, multi-modal prompt learning approaches [4, 16, 21] introduce learnable context tokens in *both* branches at multiple depths. Let $\mathcal{V} = \{V_i\}_{i=1}^{K}$ and $\mathcal{L} = \{L_i\}_{i=1}^{K}$ denote the K transformer layers in the vision and language branches, respectively. In the case of MaPLe[21], MMRL [16], and VaMP [4], prompt tuning is further implemented in deeper layers in the encoders.

For the *language branch*, learnable tokens $\{P_i \in \mathbb{R}^{d_{\text{text}}}\}_{i=1}^{b}$ are introduced alongside the input word embeddings $W_0 = [w_0^1, w_0^2, \ldots, w_0^N] \in \mathbb{R}^{N \times d_{\text{text}}}$, forming the input $[P^1, P^2, \ldots, P^b, W_0]$ to the first transformer layer. Additional learnable tokens are introduced in the transformer blocks of the language encoder $\mathcal{L}_i$, in deeper layers up to depth $J < K$:

$$[\_, W_i] = \mathcal{L}_i([P_{i-1}, W_{i-1}]) \quad i = 1, 2, \ldots, J, \tag{1}$$

where the learned prompts P_i are processed at each layer, and $[\cdot, \cdot]$ is the concatenation operation. After depth J, subsequent layers process the prompts from layer J, and the final text representation v is computed by projecting the text embeddings to a common vision-language embedding space:

$$[P_j, W_j] = \mathcal{L}_j([P_{j-1}, W_{j-1}]) \quad j = J+1, \ldots, K, \tag{2}$$

$$v = \text{TextProj}(w_K^N) \tag{3}$$

Similarly, for the *vision branch*, learnable tokens $\{\tilde{P}_i \in \mathbb{R}^{d_{\text{vis}}}\}_{i=1}^{b}$ are introduced alongside the patch embeddings E_0 and class token c_0. The vision prompts are processed through transformer layers of the vision encoder $\mathcal{V}_i$ analogously to the language prompts, with learnable tokens introduced up to depth J, where u is the final image representation is obtained by projecting to a common vision-language embedding space:

$$[c_i, E_i, \_] = \mathcal{V}_i([c_{i-1}, E_{i-1}, \tilde{P}_{i-1}]) \quad i = 1, 2, \ldots, J, \tag{4}$$

$$[c_j, E_j, \tilde{P}_j] = \mathcal{V}_j([c_{j-1}, E_{j-1}, \tilde{P}_{j-1}]) \quad j = J+1, \ldots, K, \tag{5}$$

$$u = \text{ImageProj}(c_K) \tag{6}$$

Vision-Language Coupling. A key insight of recent multi-modal prompt learning is that the vision and language prompts should not be learned independently [16, 21]. To ensure mutual synergy between modalities, a *coupling function* $\mathcal{F}$ explicitly conditions the vision prompts on their language counterparts:

$$\tilde{P}_i = \mathcal{F}_i(P_i), \tag{7}$$

where the implementation of $\mathcal{F}_i : \mathbb{R}^{d_{\text{text}}} \rightarrow \mathbb{R}^{d_{\text{vis}}}$ depends on the particular multi-prompt learning method. For MaPLe [21], $\mathcal{F}$ is implemented as a learnable linear projection. In MMRL [16], the coupling function uses visual tokens embedded in a shared latent space, which are initialized by sampling from a Gaussian, and then uses a separate linear projection to generate modality-specific prompts. VaMP [4] employs a different coupling strategy, where text prompts are generated by layer-specific MLPs that map frozen CLIP image features to prompt tokens, creating sample-specific text prompts, while vision prompts remain shared and learnable across samples. For all of these approaches the coupling function acts as a bridge between the two modalities, allowing mutual gradient propagation and promoting synergistic adaptation. This explicit conditioning discourages independent unimodal solutions and encourages learning of prompts in a shared embedding.

Given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ of images x_i and labels $y_i \in \{1, \cdots, C\}$, multi-modal prompt learning optimizes the language prompts $\{P_i\}$, vision prompts $\{\tilde{P}_i\}$, and coupling functions $\{\mathcal{F}_i\}$ via:

$$\theta^\star = \arg\min_\theta -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|u_i, \theta), \tag{8}$$

$$p(y_i|u_i, \theta) = \frac{\exp(\text{cossim}(u_i, v_{y_i})/\tau)}{\sum_{c=1}^{C} \exp(\text{cossim}(u_i, v_c)/\tau)}, \tag{9}$$

where θ encompasses all learnable prompts and coupling parameters, u_i is the image embedding from the vision encoder (now influenced by vision prompts $\tilde{P}_i$), v_c is the text embedding for class c (influenced by language prompts P_i), and τ is a temperature parameter.

2.2. Bayesian Learning

Bayesian learning is a framework for reasoning under uncertainty, which is particularly relevant to data-scarce applications [31]. Starting from Bayes theorem [36], for a dataset $\mathcal{D}$ and a model parameterized by θ, the posterior probability over parameters $p(\theta|\mathcal{D})$ can be modeled as,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}, \text{ then } p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta). \tag{10}$$

In practice, we want to estimate $\log p(\theta|\mathcal{D})$, or, equivalently (modulo the normalization constant $p(\mathcal{D})$) $\log p(\mathcal{D}|\theta) + \log p(\theta)$. The first term is the log-likelihood, and the second term is the prior over parameters. Henceforth, we call $U(\theta) = -\log p(\mathcal{D}|\theta)$ the *potential*.

One way to estimate the posterior in Bayesian learning is *sampling* from $p(\theta|\mathcal{D})$, i.e., acquiring samples with high probability under this distribution. This can be done by adding the proper amount of noise to a gradient-based optimization algorithm. Indeed, assuming $\theta_1, \cdots, \theta_K \sim p(\theta)$, we can flow these *samples* with Langevin dynamics,

$$\theta_{k,t+1} = \theta_{k,t} - \alpha_t \nabla U(\theta_{k,t}) + \sqrt{2\alpha_t}\epsilon_t, \tag{11}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathrm{I}_d)$. This equation corresponds to the discretization in time of the Langevin Stochastic Differential Equation (SDE), $\dot{\theta}_k(t) = -\nabla U(\theta_k(t))dt + \sqrt{2}dW_t$, where dW_t is a standard $d-$dimensional Wiener process. Solving this SDE is potentially intractable, since computing $\nabla U(\theta_k) = -\nabla \log p(\mathcal{D}|\theta)$ involves evaluating the likelihood term over the complete dataset. This motivated [41] to propose the Stochastic Gradient Langevin Dynamics (SGLD) algorithm, which estimates the potential's gradient, ∇U, over mini-batches from $\mathcal{D}$, that is,

$$\tilde{U}(\theta) = -\frac{m}{n}\sum_{i=1}^{m} \log p(x_i|\theta) + \log p(\theta),$$

where $m \ll n$ is the mini-batch size. However, this algorithm suffers from slow convergence in high dimensions.

To solve the limitations of SGLD, [3] proposes momentum variables r with a friction term, thus introducing Stochastic Gradient Hamiltonian Monte Carlo (SGHMC),

$$\begin{cases} \theta_{k,t+1} = \theta_{k,t} + r_{k,t} \\ r_{k,t+1} = r_{k,t} - \alpha_t \nabla \tilde{U}(\theta_{k,t}) - \eta r_{k,t} + \sqrt{2(\eta - \hat{\gamma})\alpha_t}\epsilon_t. \end{cases} \tag{12}$$

Here, $\hat{\gamma}$ estimates the stochastic gradient noise. The friction term η counteracts this noise to ensure convergence to the correct stationary distribution. This allows SGHMC to combine the computational efficiency of mini-batching with the rapid exploration of momentum-based dynamics.

In a further development, [44] proposed alternating between exploration and sampling stages for enhancing the diversity of samples in Stochastic Gradient Markov Chain Monte Carlo (SGMCMC) methods. First, they update the learning rate using a cosine scheduler. Second, they alternate between iterations of exploration and sampling cyclically. This approach is known as Cyclical SGMCMC.

Remark. *(Nomenclature of methods) As discussed in [44], both SGLD and SGHMC can be unified under the common framework of SGMCMC methods. This nomenclature hints at the fact that these algorithms are stochastic gradient discretizations of continuous-time Markov processes designed to sample from the posterior distribution, differing only in how they incorporate momentum, friction, and injected noise. For the remainder of the paper, we adopt this convention for Bayesian methods.*

2.3. Probability Metrics

In this section, we present metrics between probability distributions that will be used in our method. Recall that, given

a subset $\Omega \in \mathbb{R}^d$, $\mathcal{P}(\Omega)$ denotes the set of probability distributions on Ω. A probability metric is a metric on elements of $\mathcal{P}(\Omega)$. In the following, we discuss how to estimate these metrics based on finite samples from p and q, denoted $\{z_i^{(p)}\}_{i=1}^n$ and $\{z_j^{(q)}\}_{j=1}^m$, respectively.

The Maximum Mean Discrepancy (MMD) metric, proposed by [15], computes a distance between p and q based on a kernel $\kappa : \Omega \times \Omega \to \mathbb{R}$. More specifically,

$$d(p,q)^2 = \frac{1}{n^2}\sum_{i,j=1}^{n}\kappa(z_i^{(p)}, z_j^{(p)}) + \frac{1}{m^2}\sum_{i,j=1}^{m}\kappa(z_i^{(q)}, z_j^{(q)}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\kappa(z_i^{(p)}, z_j^{(q)}). \tag{13}$$

Meanwhile, the Wasserstein distance comes from Optimal Transport (OT) [29]. This distance computes the least amount of effort or energy to transport one distribution into another. In mathematical terms,

$$d(p,q)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\gamma_{ij}^{\star}\|z_i^{(p)} - z_j^{(q)}\|_2^2, \tag{14}$$

where $\gamma^\star = \text{argmin}_{\gamma\in\Gamma}\sum_{i=1}^n\sum_{j=1}^m\gamma_{ij}\|z_i^{(p)} - z_j^{(q)}\|_2^2$ is called the *optimal transport plan*. Here, $\Gamma = \{\gamma \in \mathbb{R}_+^{n\times n} : \sum_{i=1}^n\gamma_{ij} = m^{-1}, \sum_{j=1}^m\gamma_{ij} = n^{-1}\}$ is the set of feasible transportation plans.

In either of these cases, one can see $d(p,q)$ as a function of its samples, i.e., $\{\{z_i^{(p)}\}_{i=1}^n, \{z_j^{(q)}\}_{j=1}^m\} \mapsto d(p,q)$. In this sense, Equations 13 and 14 compute a distance between groups of points. This will be useful in the next section (c.f. Equation 18), as probability metrics will serve as a way of capturing the geometry of the weight space. More details are available in Appendix C.

3. Bayesian Prompt Learning with Repulsive Cyclical SGHMC

Our goal in this section is to present a Bayesian view of prompt learning. We recall that $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ is a dataset of i.i.d. samples. Each text prompt and image in $\mathcal{D}$ is encoded, i.e., $v_{y_i} = \text{TextProj}(w_K^N)$ and $u_i = \text{ImageProj}(c_K)$. Starting from Equations 8 and 9, and under the i.i.d. assumption, the log-likelihood is written as $\log p(\mathcal{D}|\theta) = \sum_{i=1}^n p(y_i|u_i,\theta)$. Now, using Equation 9, multi-modal prompt learning corresponds to Maximum A Posteriori (MAP) estimation:

$$\theta^\star_{\text{MAP}} = \arg\max_\theta \log p(\theta) + \sum_{i=1}^{n} p(y_i|u_i,\theta), \tag{15}$$

where θ includes all learnable prompts and coupling parameters. As a consequence of Equation 15, MaPLe and MMRL inherit the intrinsic limitations of MAP estimation under data scarcity. In particular, they are subject to overfitting, and do not account for predictive uncertainty.

Our goal is to use a Bayesian setting, which allows us to sample high quality prompts from the posterior $p(\theta|\mathcal{D})$. The main insight is that the underlying prompt landscape has many equally good (in terms of training loss) prompts that have different generalization capabilities (test loss or accuracy). On the one hand, we want to generate samples that are likely under the posterior $p(\theta|\mathcal{D})$. On the other hand, we want to enhance sample diversity to uncover multiple modes in the landscape. To achieve **both** of these goals, we propose the Repulsive Cyclical SGMCMC (rcSGMCMC) algorithm. Our insight is that, in addition to using the cyclical SGMCMC schedule of [44], we can further encourage exploration with a repulsion term between prompts.

Remark. *(Notation) In the following, we describe our method mathematically. We use k to denote different samples in MCMC, t to denote the iteration, and c to denote cycles. In this sense, each cycle $c = 1, \cdots, C$ is composed of $t = 1, \cdots, T$ iterations across the $k = 1, \cdots, K$ samples.*

Given a balance parameter β, we define the exploration stage where $\frac{t}{T} \leq \beta$, and the sampling stage where $\frac{t}{T} > \beta$. We sample using Equation 16 :

$$\begin{cases} \theta_{k,t+1}^{(c)} = \theta_{k,t}^{(c)} + r_{k,t}^{(c)} + \xi\sum_{\ell=1}^K F(\theta_{k,t}^{(c)}, \theta_{\ell,T}^{(c-1)}), \\ r_{k,t+1}^{(c)} = (1-\eta)r_{k,t}^{(c)} - \alpha_t\tilde{\nabla U}(\theta_{k,t}^{(c)}) \\ \qquad\qquad + \mathbb{I}_{\frac{t}{T},\beta}\sqrt{2(\eta-\hat{\gamma})\alpha_t}\epsilon_t, \end{cases} \tag{16}$$

for $\epsilon_t \sim \mathcal{N}(0, \mathrm{I}_d)$, and $\mathbb{I}_{\frac{t}{T},\beta} = 1$ if $\frac{t}{T} > \beta$ and 0 otherwise. These two stages are scheduled within cycles, as in Zhang et al. [44]. We provide further mathematical motivation for these equations in our supplementary materials.

The main feature that distinguishes Equation 16 is the repulsion of samples from the current cycle c, $\{\theta_{k,t}^{(c)}\}_{k=1}^K$, away from those of the previous cycle, $\{\theta_{\ell,T}^{(c-1)}\}_{\ell=1}^K$. We refer readers to Figure 1 for a conceptual illustration of the effect of this force on the SGHMC algorithm, i.e., mode exploration in the posterior.

We model the repulsive force through another potential, $V(\theta,\theta')$, so that, $F(\theta,\theta') = -\nabla_\theta V(\theta,\theta')$. Intuitively, this potential should be large for similar parameters, and small for different ones. A natural potential is,

$$V(\theta,\theta') = \frac{1}{d_\Theta(\theta,\theta')^2 + \epsilon}, \tag{17}$$

where $d_\Theta : \Theta^2 \to \mathbb{R}$ is a distance in the space of parameters Θ. We propose comparing the representations extracted by

Algorithm 1 Repulsive cSGHMC Training with Cycle Restarts and Inter-Cycle Repulsion

Require: Initial step-size α_0, number of iterations T, number of cycles C, proportion of exploration β, momentum $1-\eta$, repulsion strength ξ, noise estimate $\hat{\gamma}$.

for Cycle $c = 1, \cdots, C$ **do**
  for Iteration $t = 1, \cdots, T$ **do**
    Set α_t with cosine scheduling
    $F_{k,t}^{(c)} = \begin{cases} \sum_{\ell=1}^{K} F(\theta_{k,t}^{(c)}, \theta_{\ell,T}^{(c-1)}) & c > 1 \\ 0 & \text{otherwise} \end{cases}$
    $n_t = \begin{cases} \sqrt{2(\eta-\hat{\gamma})\alpha_t}\epsilon_t & \frac{\text{mod}(t-1, \lceil T/C \rceil)}{\lceil T/C \rceil} > \beta \\ 0 & \text{otherwise} \end{cases}$
    $\theta_{k,t+1}^{(c)} = \theta_{k,t}^{(c)} + r_{k,t}^{(c)} + F_{k,t}^{(c)}$,
    $r_{k,t+1}^{(c)} = (1-\eta) r_{k,t}^{(c)} - \alpha_t \nabla \tilde{U}(\theta_{k,t}^{(c)}) + n_t$
  end for
end for

the networks, i.e., $u_{\theta,i}$ for a representation of image x_i; Algorithm 1 gives our complete repulsive cSGHMC approach.

Modeling the geometry of the weight space is challenging [25] due to invariance to permutations (e.g. [13]) and data scarcity in this space. Therefore, we propose comparing parameters θ and θ' in terms of the distribution of their activations, i.e.,

$$d_\Theta(\theta, \theta') = d_{\mathcal{P}(\mathcal{U})}(U_\theta, U_{\theta'}), \tag{18}$$

where $U_\theta = \{u_{\theta,i}\}_{i=1}^n$ and $d_{\mathcal{P}(\mathcal{U})}$ is a distance over the set of probability distributions over $\mathcal{U}$. The MMD [15] and Wasserstein distance [29] are of main interest to our applications, which are described in Section 2.3.

While computing the MMD or the Wasserstein distance incurs an additional computational overhead, Equation 18 is computed on the level of mini-batches (e.g., 32 samples). Since the complexity of these distances scales with the number of samples (i.e., $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ respectively), the overall computational cost of $d_{\mathcal{P}(\mathcal{U})}$ is negligible (c.f. Appendix A.4).

After the execution of Algorithm 1, we produce a set of network parameter samples $\{\{\theta_{k,T}^{(c)}\}_{k=1}^K\}_{c=1}^C$, which includes prompts and parameters of the coupling function. As a result, we compute predictions based on the ensembling of these samples, namely,

$$p(y|x) = \sum_{c=1}^{C} \sum_{k=1}^{K} \omega_{c,k} p(y|x, \theta_{k,T}^{(c)}),$$

where $\omega_{c,k}$ is the importance of each $\theta_{k,T}^{(c)}$. For simplicity, we use uniform weighting $\omega_{c,k} = (C \cdot K)^{-1}$. Running inference on the $C \cdot K$ models is embarrassingly parallelizable (c.f. Appendix A.5) and adds a small computational overhead.

4. Experiments

Overview. We assess the effectiveness of our proposed approach across three distinct evaluation protocols: base-to-novel class generalization, domain generalization, and cross-dataset transfer. Following established protocols in recent work on prompt learning [16, 21, 46, 47], all experiments are conducted under a 16-shot learning scenario, where we are provided with only 16 labeled training samples per category. We run experiments on ReBaPL-based extensions of the MaPLe and MMRL methods for multi-modal prompt learning. Full details on our experiments are available in the supplementary material (Appendix A).

4.1. Base to Novel Generalization

In this protocol, the available classes within each dataset are partitioned into two disjoint subsets: base categories and novel categories. During training, the model has access exclusively to labeled examples from the base categories. Evaluation is then performed on both base and novel subsets, enabling us to measure both the model's adaptation performance on seen classes and its capacity to preserve zero-shot generalization on previously unseen classes. This evaluation is carried out across 11 benchmark classification datasets: ImageNet [8], Caltech101 [11], OxfordPets [32], StanfordCars [24], Flowers102 [30], Food101 [2], FGVCAircraft [28], SUN397 [42], UCF101 [37], DTD [6], and EuroSAT [18].

In Table 1 we present the performance of our method in the base-to-novel generalization task. We compare our method to the CLIP [34] baseline, and other methods including CoOp [47], and CoCoOp [46], APP [5], PromptSRC [22], MaPLe [21] and MMRL [16]. We re-run some of the leading methods (PromptSRC, MaPLe, and MMRL) to ensure a fair comparison and avoid inconsistent results due to different hardware. While VaMP [4] is a method related to ours, its code is not publicly available, and thus we do not include it our experiments. We demonstrate the effectiveness of our proposed ReBaPL approach by running it as extensions of MaPLe and MMRL.

In comparison with MaPLe, our method consistently improves performance on base classes, and in many cases also novel classes. For some datasets (e.g., OxfordPets, StanfordCars, and FGVCAircraft), MaPLe outperforms our method by a small margin. Overall, we consistently improve over MaPLe in terms of the harmonic mean of base and novel accuracy, striking a balance in generalization.

Concerning MMRL, while we improve on both base and novel classes, our gains are mostly on the novel classes. This finding highlights our claim that our repulsive cyclical SGHMC algorithm, and hence ReBaPL, improves generalization by exploring the posterior landscape more thoroughly. On average, our method improves MMRL on both base and novel classes, with a larger margin on the novel

Table 1. **Base to novel generalization results**. Overall, our ReBaPL method improves the base performance of both MaPLe and MMRL, with a considerable increase on the FGVCAircraft and EuroSAT datasets. We used MMD for the MaPLe + ReBaPL method, and Wasserstein distance for the MMRL + ReBaPL method, which respectively performed best. An asterisk (*) denotes methods we re-ran using the same random seeds and hardware. Delta values (Δ) show improvements from adding ReBaPL (green for positive, red for negative).

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [34]	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp [47]	82.69	63.22	70.83	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp [47]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
APP [5]	83.0	65.8	72.61	69.9	63.2	66.4	95.2	91.0	93.0	96.8	88.3	92.4
PromptSRC* [22]	84.93	74.49	78.61	76.77	67.8	72.01	98.07	94.03	96.01	95.27	97.23	96.24
MaPLe* [21]	82.03	75.03	78.37	74.96	66.97	70.74	97.83	94.87	96.33	95.20	98.13	96.64
MaPLe* + ReBaPL	83.28	76.08	79.52	76.06	68.80	72.25	98.35	94.87	96.58	**96.17**	**97.77**	**96.96**
Δ	+1.25	+1.05	+1.15	+1.10	+1.83	+1.51	+0.52	0.0	+0.25	+0.97	-0.36	+0.32
MMRL* [16]	85.54	76.52	80.59	77.55	67.43	72.14	98.93	**94.60**	**96.72**	95.27	97.23	96.24
MMRL* + ReBaPL	**85.74**	**77.44**	**81.38**	**77.90**	**68.83**	**73.09**	**99.10**	94.27	96.62	95.93	97.57	96.74
Δ	+0.20	+0.92	+0.79	+0.35	+1.40	+0.95	+0.17	-0.33	-0.10	+0.66	+0.34	+0.50

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [34]	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp [47]	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp [47]	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
APP [5]	85.9	69.5	76.8	96.8	61.0	74.8	84.6	86.1	85.4	44.9	26.0	33.0
PromptSRC* [22]	77.93	75.5	76.7	97.83	77.13	86.26	90.60	91.53	91.06	41.43	23.67	30.13
MaPLe* [21]	72.45	74.90	73.65	96.33	73.33	83.27	90.80	92.10	91.45	36.60	34.90	35.73
MaPLe* + ReBaPL	74.73	74.57	74.65	97.43	74.37	84.35	**90.83**	**92.13**	**91.48**	38.00	34.03	35.91
Δ	+2.28	-0.33	+1.0	+1.10	+1.04	+1.08	+0.03	+0.03	+0.03	+1.4	-0.87	+0.18
MMRL* [16]	**81.23**	75.00	77.99	98.70	76.83	86.40	90.60	91.53	91.06	**45.70**	37.1	40.95
MMRL* + ReBaPL	81.20	**75.37**	**78.17**	**98.80**	**77.23**	**86.70**	90.73	91.60	91.16	45.13	**38.57**	**41.59**
Δ	-0.03	+0.37	+0.18	+0.10	+0.40	+0.30	+0.13	+0.07	+0.10	-0.57	+1.47	+0.64

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [34]	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp [47]	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp [47]	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
APP [5]	80.6	73.3	76.8	78.4	48.9	60.2	93.6	47.6	63.1	86.2	69.2	76.8
PromptSRC* [22]	92.93	74.17	80.53	83.33	61.3	70.64	92.93	74.17	82.5	87.10	78.53	82.6
MaPLe* [21]	81.07	77.90	79.45	81.30	57.50	67.36	92.47	77.03	84.05	83.97	77.70	80.71
MaPLe* + ReBaPL	81.83	**79.87**	80.84	82.50	61.60	70.53	95.30	79.20	86.51	84.97	79.73	82.27
Δ	+0.76	+1.97	+1.39	+1.2	+4.10	+3.17	+2.83	+2.17	+2.46	+1.0	+2.03	+1.56
MMRL* [16]	83.23	79.27	81.20	85.63	**65.13**	73.99	95.80	77.20	85.50	88.30	79.70	83.78
MMRL* + ReBaPL	**83.20**	79.37	**81.24**	**86.10**	65.00	**74.08**	**96.73**	**83.63**	**89.71**	**88.33**	**80.40**	**84.18**
Δ	-0.03	+0.10	+0.04	+0.47	-0.13	+0.09	+0.93	+6.43	+4.21	+0.03	+0.70	+0.40

classes. Overall the harmonic mean is also improved. Our method has substantial gains on the EuroSAT and ImageNet datasets, establishing a new state-of-the-art.

4.2. Cross-Dataset Transfer

To evaluate robustness under dataset shift, we adopt a source-to-target transfer protocol. The model is first trained on the complete set of 1000 ImageNet categories using the 16-shot setting, and evaluated on the remaining 10 datasets without any additional fine-tuning. This setup allows us to measure how well learned prompt adaptations transfer to entirely different data distributions and visual domains.

We see from Table 2 that our ReBaPL-based models provide significant gains over the baseline MaPLe and MMRL methods. Notably, our MMRL + ReBaPL approach provides the highest average accuracy of 67.62%, indicating better generalization performance than competing methods.

4.3. Domain Generalization

To assess robustness against domain shift and out-of-distribution scenarios, we train exclusively on ImageNet and evaluate on four domain-shifted variants: ImageNetV2 [35], ImageNet-Sketch [40], ImageNet-A [20], and ImageNet-R [19].

We show in Table 3 that our ReBaPL methods improve generalization on out-of-domain datasets, compared to the underlying MLE-based prompt learning method. We also see that our MaPLe + ReBaPL method provides the best performance on the ImageNet-A target. Finally, we see that our MMRL + ReBaPL method provides the highest accu-

Table 2. Comparison of our ReBaPL method with previous state-of-the-art multi-modal prompt learning methods on cross-dataset evaluation across 10 datasets. An asterisk (*) denotes methods we re-ran using the same random seeds and hardware. Delta values (Δ) show improvements from adding ReBaPL (green for positive, red for negative).

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers101	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Average
PromptSRC* [22]	68.96	93.47	90.33	65.87	70.40	83.66	24.13	67.17	46.40	45.97	67.67	65.50
MaPLe* [21]	67.96	93.17	90.20	65.97	71.07	86.33	23.23	67.23	47.20	45.70	66.27	65.63
MaPLe + ReBaPL	68.66	93.80	90.73	66.30	72.57	86.40	24.37	67.70	47.17	50.90	67.87	66.77
Δ	+0.70	+0.63	+0.53	+0.33	+1.50	+0.07	+1.14	+0.47	-0.03	+5.20	+1.60	+1.14
MMRL* [16]	70.13	94.30	91.0	66.20	71.53	86.20	26.07	67.50	46.93	49.83	69.13	66.87
MMRL + ReBaPL	**71.0**	**94.60**	**91.97**	**66.63**	**72.57**	**86.43**	**26.33**	**67.90**	**47.27**	**52.97**	**69.50**	**67.62**
Δ	+0.87	+0.30	+0.97	+0.43	+1.04	+0.23	+0.26	+0.40	+0.34	+3.14	+0.37	+0.75

	Source	Target			
Method	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R
PromptSRC* [22]	68.96	62.50	**48.60**	49.63	**75.77**
MaPLe* [21]	67.96	61.57	47.70	48.80	75.33
MaPLe + ReBaPL	68.66	62.30	48.50	**49.73**	75.40
Δ	+0.70	+0.73	+0.80	+0.93	+0.07
MMRL* [16]	70.13	62.20	47.80	48.90	75.03
MMRL + ReBaPL	**71.00**	**62.50**	48.40	49.63	75.63
Δ	+0.87	+0.30	+0.60	+0.73	+0.60

Table 3. Comparison of robustness on out-of-distribution datasets. Delta values (Δ) show improvements from adding ReBaPL (green for positive, red for negative).

Table 4. Ablation on the use of repulsion in ReBaPL, averaged over 11 datasets.

Method	Base	Novel	HM
MaPLe	82.03	75.03	78.37
+ ReBaPL (No Repulsion)	83.39	75.47	78.93
+ ReBaPL (Wasserstein)	83.39	75.86	79.44
+ ReBaPL (MMD)	83.28	**76.08**	**79.52**

racy on the source ImageNet dataset. Taken together, these results show that our ReBaPL approach is effective at improving out-of-domain generalization performance without hindering the performance on the source domain.

4.4. Ablation

We ablate a few design choices in our method. We compare 3 main choices: ReBaPL without repulsion (i.e., $F(\theta, \theta') := 0$) vs. ReBaPL with repulsion based on the Wasserstein distance and the MMD. This ablation seeks to isolate the benefit of using repulsion, and demonstrate that our method is robust to the choice of probability metric. We evaluate our methods on all datasets listed in Table 1.

Overall, as shown in Table 4, our method has stable performance for both the Wasserstein distance and MMD. The gap in harmonic mean between these choices is 0.08%, which is marginal compared to the improvement over MaPLe (around 1%). Furthermore, even without repulsion, our ReBaPL method improves over MaPLe, highlighting the benefits of sampling from the posterior distribution. By adding the repulsion term we improve performance for novel classes. This finding supports our claim that adding repulsion allows us to explore the posterior landscape more thoroughly, and thus improve generalization.

5. Conclusion

In this paper we have introduced Repulsive Bayesian Prompt Learning (ReBaPL), a novel approach that addresses overfitting and out-of-distribution generalization challenges in prompt learning methods. Our key contribution is the repulsive cyclical SGHMC (rcSGHMC) algorithm, which leverages Hamiltonian dynamics with cyclical learning rate schedules to alternate between exploration and exploitation phases when sampling from complex multimodal posterior distributions. By introducing a representation-based repulsive force derived from probability metrics (MMD and Wasserstein distance) computed on representation distributions, our method captures functional similarity between prompts and encourages diverse mode discovery while preventing premature convergence. Unlike prior Bayesian prompt learning approaches that rely on restrictive unimodal approximations or deterministic variational methods, ReBaPL provides a modular, plug-and-play framework that can extend any existing MLE-based prompt learning method with principled Bayesian inference. Through comprehensive experiments we have demonstrated superior generalization by maintaining a richer characterization of the prompt posterior landscape. Future work includes exploring alternative probability metrics beyond MMD and Wasserstein distance, such as Sinkhorn divergence or information-theoretic measures, and developing adaptive cyclical mechanisms that automatically adjust the number of cycles based on convergence diagnostics or posterior diversity to improve efficiency and performance.

References

[1] Yassir Bendou, Amine Ouasfi, Vincent Gripon, and Adnane Boukhayma. Proker: A kernel perspective on few-shot adaptation of large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25092–25102, 2025. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer, 2014. 6

[3] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1683–1691, 2014. 4

[4] Silin Cheng and Kai Han. Vamp: Variational multi-modal prompt learning for vision-language models. In *NeurIPS 2025*. 1, 2, 3, 4, 6

[5] Youngjae Cho, HeeSun Bae, Seungjae Shin, Yeo Dong Youn, Weonyoung Joo, and Il-Chul Moon. Make prompts adaptable: Bayesian modeling for vision-language prompt learning with data-dependent prior. In *AAAI*, pages 11552–11560, 2024. 2, 6, 7

[6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014. 6

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 4

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 6

[9] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Bayesian prompt learning for image-language model generalization. In *IEEE/CVF International Conference on Computer Vision*, pages 15237–15246, 2023. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition - Workshops*, page 178. IEEE, 2004. 6

[12] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 1

[13] Yoav Gelberg, Tycho FA van der Ouderaa, Mark van der Wilk, and Yarin Gal. Variational inference failures under model symmetries: Permutation invariant posteriors for bayesian neural networks. *arXiv preprint arXiv:2408.05496*, 2024. 6

[14] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018. 4

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012. 5, 6, 3

[16] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25015–25025, 2025. 1, 2, 3, 4, 6, 7, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[18] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 6

[19] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE International Conference on Computer Vision*, pages 8320–8329, 2021. 7

[20] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 7

[21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 1, 2, 3, 4, 6, 7, 8

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15144–15154. IEEE, 2023. 1, 6, 7, 8

[23] Mingyu Kim, Jongwoo Ko, and Mijung Park. Bayesian principles improve prompt learning in vision-language models. In *AISTATS*, 2025. 2

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE Conference on Computer Vision and Pattern Recognition - Workshops*, pages 554–561. IEEE, 2013. 6

[25] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 6

[26] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceed-*

ings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5206–5215, 2022. 1

[27] Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, and Changsheng Xu. Understanding and mitigating overfitting in prompt tuning for vision-language models. *arXiv [cs.CV]*, 2022. 1

[28] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[29] Eduardo Fernandes Montesuma, Fred Maurice Ngole Mboula, and Antoine Souloumiac. Recent advances in optimal transport for machine learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 5, 6, 3, 4

[30] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 6

[31] Theodore Papamarkou, Maria Skoularidou, Konstantina Palla, Laurence Aitchison, Julyan Arbel, David Dunson, Maurizio Filippone, Vincent Fortuin, Philipp Hennig, José Miguel Hernández-Lobato, et al. Position: Bayesian deep learning is needed in the age of large-scale ai. *arXiv preprint arXiv:2402.00809*, 2024. 4

[32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012. 6

[33] Zhen Qu, Xian Tao, Xinyi Gong, Shichen Qu, Qiyu Chen, Zhengtao Zhang, Xingang Wang, and Guiguang Ding. Bayesian prompt flow learning for zero-shot anomaly detection. *arXiv [cs.CV]*, 2025. 2

[34] 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. 1, 2, 6, 7

[35] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 7

[36] Paul J Smith. Statistics: A bayesian perspective, 1997. 4

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[39] Cédric Villani et al. *Optimal transport: old and new*. Springer, 2008. 4

[40] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 7

[41] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011. 4

[42] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010. 6

[43] Shijun Yang, Xiang Zhang, Wanqing Zhao, Hangzai Luo, Sheng Zhong, Jinye Peng, and Jianping Fan. Multi-modal mutual-guidance conditional prompt learning for vision-language models. *arXiv [cs.CV]*, 2025. 1

[44] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019. 4, 5

[45] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 1

[46] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1, 2, 6

[47] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 6, 7

[48] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv [cs.CV]*, 2025. 1