Supplementary material for Bayesian Prompt Learning for Image-Language Model Generalization

Table 9: Shared hyperparameters used to generate all results in the main paper.

Hyperparameters	**Values**
Batch Size	1
Input Size	224×224
Input Interpolation Method	"Bicubic"
Input Mean	[0.48145466, 0.4578275, 0.40821073]
Input STD	[0.26862954, 0.26130258, 0.27577711]
Transformation	["random resized crop", "random flip", "normalize"]
Optimizer	SGD
Learning Rate	$2e-3$
LR Scheduler	"cosine"
Warmup Epoch	1
Warmup Type	"Constant"
Warmup LR	$1e-5$
Backbone	ViT-B/16
Prompt Length	4
Prompt Initialization	"a photo of a {class}"
Number of Shots	16

1. Hyperparameters

In this section, we provide the detailed hyperparameter settings in Tables 9 and 10 that are used to generate results in the main paper for each dataset. There are two sets of hyperparameter. In Table 9, we report the shared hyperparameters among unconditional and conditional Bayesian prompt learning. Table 10 contains parameters that are optimized per dataset.

Table 10: **Dataset-specific hyper-parameters** used to generate all results in the main paper. In this table, we provide the number of Monte Carlo samples (MC) and also the number of epochs used to optimize our unconditional and conditional Bayesian prompt learning.

	ImageNet	**Caltech101**	**OxfordPets**	**StanfordCars**	**Flowers102**	**Food101**	**FGVC**	**SUN397**	**DTD**	**EuroSAT**	**UCF101**
MC	10	20	40	20	10	20	10	20	40	20	5
Epochs	10	20	20	40	40	20	10	10	10	60	20

2. More Ablations

Vision encoder alternatives. All previous experiments benefit from ViT-B/16 as the vision encoder's backbone following [54, 55, 32]. For completeness, in Figure 5, we replace this vision encoder with a Resnet50 and Resnet100 and examine its impact on unseen prompt generalization task for one random seed. As reported, the visual transformer outperforms the Resnet alternatives on 10 out of 10 benchmarks due to the fact that a more over-parameterized model is able to extract better generalizable features. Hence, we suggest training and evaluating Bayesian prompt learning on visual transformer for better model performance.

Comparison with Fixed-Prompt Baseline. This section presents a comparison between conditional Bayesian prompt learning and the fixed-prompt baseline, such as CLIP, as shown in Table 11. We assess their performance in terms of unseen prompt generalization (Task I) and cross-domain prompt generalization (Task III). In the fixed-prompt approach, prompts remain non-learnable and are usually hand-engineered. In contrast, our approach involves training prompts and adapting them to downstream tasks. The experiments in Table 11 demonstrate that our proposed method outperforms the CLIP model on 7 out of 11 datasets in Task I and on all 4 datasets in Task III.

Table 11: Comparison between conditional Bayesian prompt learning performance and CLIP model on unseen prompt generalization **(Task I)** and cross-domain prompts generalization **(Task III)**. Our model consistently performs better than CLIP model acroos all tasks.

	CLIP	**Ours**
Task I		
Caltech101	94.00	**94.93**±0.1
DTD	59.90	**60.80**±0.5
EuroSAT	64.05	**75.30**±0.7
FGVCAircraft	**36.29**	35.00±0.5
Flowers102	**77.80**	70.40±1.8
Food101	91.22	**92.13**±0.1
ImageNet	68.14	**70.93**±0.1
OxfordPets	97.26	**98.00**±0.1
StanfordCars	**74.89**	73.23±0.2
SUN397	75.35	**77.87**±0.5
UCF101	**77.50**	75.77±0.1
Average	74.22	**74.94**±0.2
Task III		
ImageNetV2	60.83	**64.23**±0.1
ImageNet-Sketch	46.15	**49.20**±0.0
ImageNet-A	47.77	**51.33**±0.1
ImageNet-R	73.96	**77.00**±0.1
Average	57.18	**60.44**±0.1

These results underscore the effectiveness of our proposed approach.

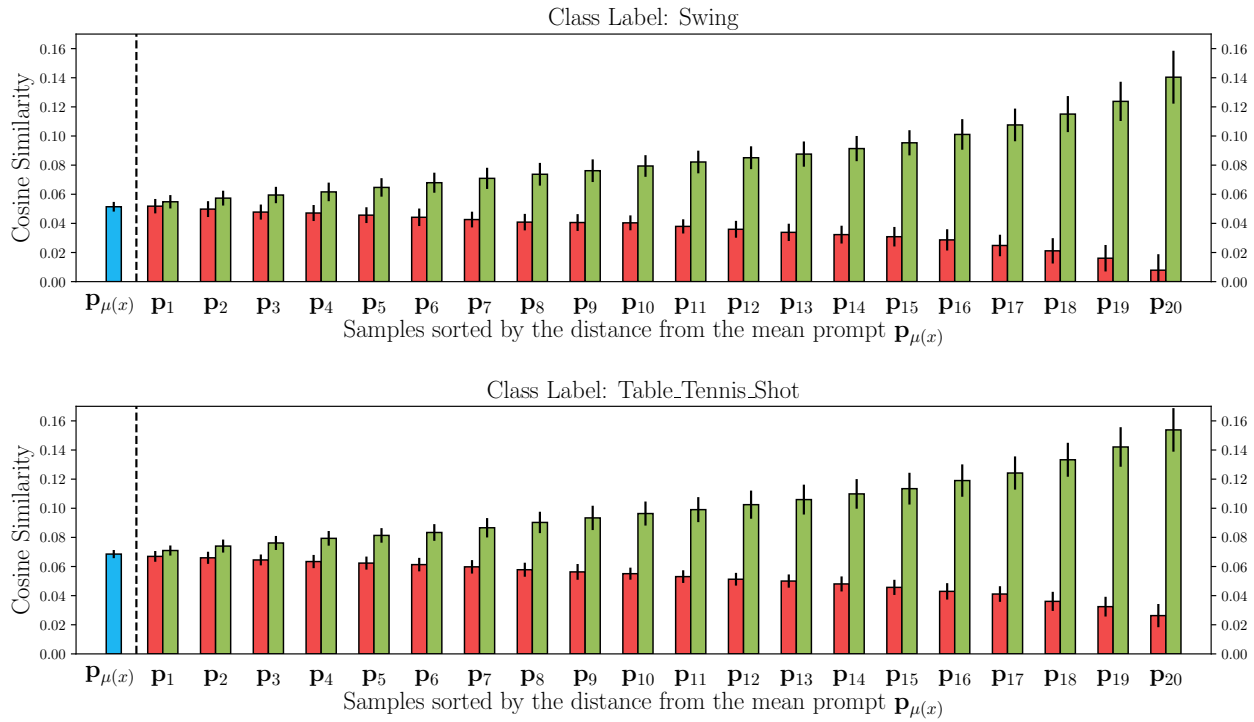

Figure 1: **Variational distribution interpretation** on the UCF101 dataset. The text encoding of the mean prompt $\mathbf{p}_{\mu(\mathbf{x})}$ (■) is the most similar to the image encoding. As we move further away from the mean prompt, the cosine similarity scores between the text encoding and image encoding decrease further (■). When we ensemble the text encoding of different prompts the cosine similarity increases (■), where the maximum similarity is obtained when all text encodings are combined.

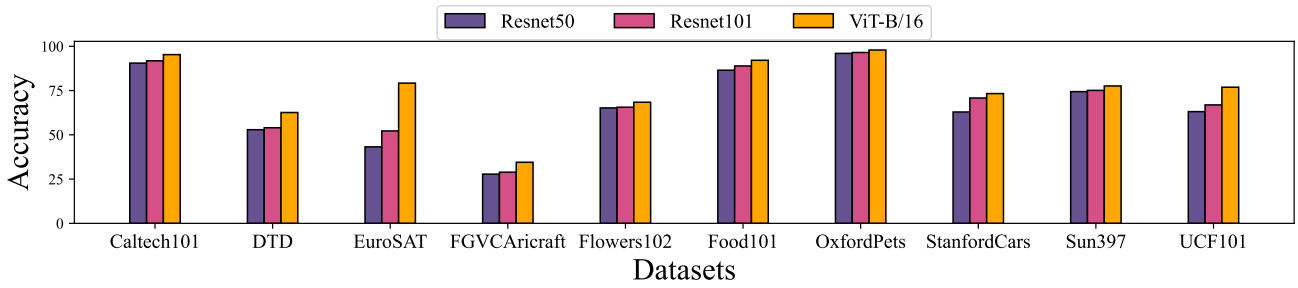

Figure 5: **Ablation of different vision encoder backbones with respect to unseen prompt generalization**. A more over-parameterized model leads to better generalization performance across all datasets.