Bayesian Prompt Learning for Image-Language Model Generalization

Mohammad Mahdi Derakshani et al.

Jonathan Hu
ECSE 626
Dec 3, 2025

Few-Shot Learning in the Real World

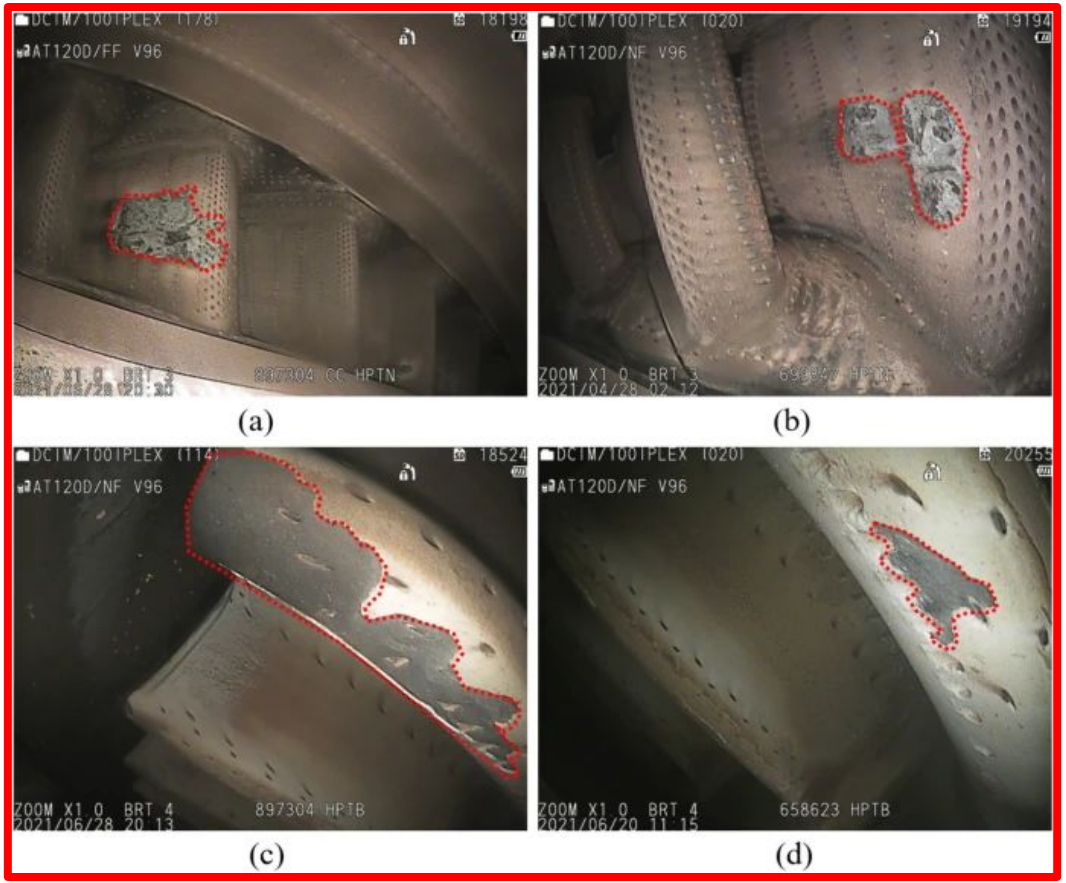
Why do we care about few-shot learning?

- Modern VLMs like CLIP are trained once on large dataset of image-text pairs
- Zero-shot CLIP performance is good, but not excellent
- In the real world, we face new, specialized tasks where:
  - Labeled data is scarce
  - Finetuning entire model is undesirable
- **Objective:** Adapt a pre-trained, frozen model like CLIP to a novel downstream task using very few labeled samples (few-shot regime)
  - Robust to distribution shift and out-of-distribution examples (generalization)

Anomaly Detection

CBC News

Jet Engine Inspection



Deep learning-based defects detection of certain aero-engine blades and vanes with DDSC-YOLOv5s [Li et al., 2022], SP's Airbuz,

Existing Methods: CoOp/CoCoOp/ProDA

What's been tried?

- **Context Optimization (CoOp)**
  - Single global soft prompt
  - Good in-domain, weaker on unseen classes
- **Conditional Context Optimization (CoCoOp)**
  - Prompt residual conditioned on image
  - Better unseen-class & cross-dataset generalization
- **ProDA**
  - Learn an ensemble of prompts
  - Fit a multivariate Gaussian over classifier weights, use mean at test time

Bayesian Prompt Learning

ProDA

- Learn a distribution over the **output embeddings** of the prompts
- Still deterministic at inference (no sampling)
- Can not condition prompts on the image

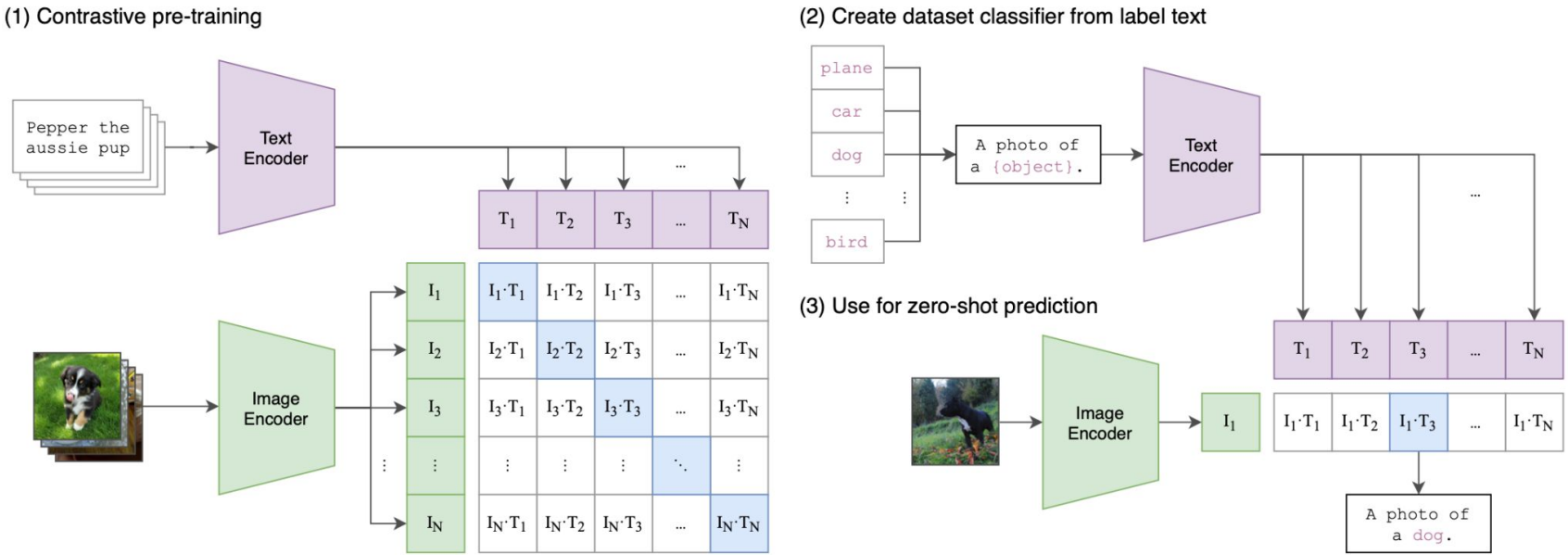
Bayesian Prompt Learning

- Learn a distribution over **input embeddings** of the prompts
- Sample the learned distribution at test time
- Unconditional and conditional

Theory time!

CLIP Recap

Learn joint embedding of image and text via contrastive learning



Learning Transferable Visual Models From Natural Language Supervision [Radford et al., 2021]

VLM Recap

Normalized image embedding $\mathbf{z} = \dfrac{f(\mathbf{x})}{\|f(\mathbf{x}))\|_2}$

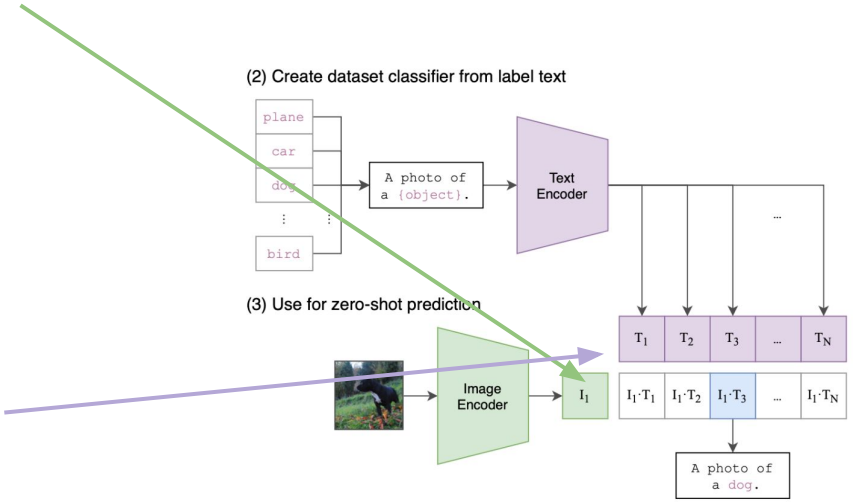
$\mathbf{x}$: image

$f(\cdot)$: image encoder

Normalized text embedding $\mathbf{w} = \dfrac{g(\mathbf{t})}{\|g(\mathbf{t})\|_2}$

$\mathbf{t}$: text

$g(\cdot)$: text encoder



Learning Transferable Visual Models From Natural Language Supervision [Radford et al., 2021]

Zero-shot CLIP for Image Classification

Generate category descriptions $\{t_c\}_{c=1}^{C}$

C : number of classes

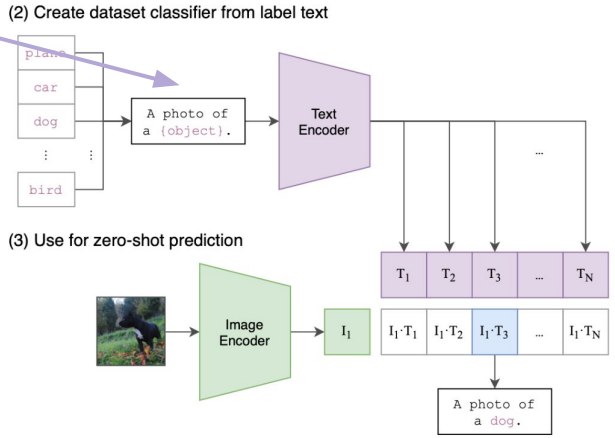
$g(\cdot)$

Generate text embedding w for each category

$$\mathbf{w}_{1:C} = [\mathbf{w}_1^T, \ldots, \mathbf{w}_C^T]^T$$

Prediction probability of class y for test sample x

$$p(y|\mathbf{x}) = \frac{e^{\mathbf{z}^T \mathbf{w}_y / \tau}}{\sum_{c=1}^{C} e^{\mathbf{z}^T \mathbf{w}_c / \tau}}$$



Learning Transferable Visual Models From Natural Language Supervision [Radford et al., 2021]

Context Optimization (CoOp)

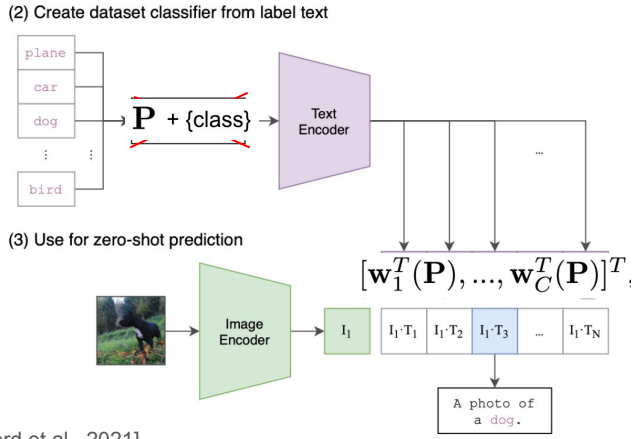
Learn prompt P with a few training samples $\mathcal{D}^{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$

Learnable continuous prompt $\mathbf{P} \in \mathbb{R}^{p \times e}$

$\mathbf{t}_c(\mathbf{P})$ = $\mathbf{P}$ + {class}

Randomly initialized

$$\mathbf{w}_{1:C}(\mathbf{P}) = [\mathbf{w}_1^T(\mathbf{P}), ..., \mathbf{w}_C^T(\mathbf{P})]^T, \text{where } \mathbf{w}_c(\mathbf{P}) = g(\mathbf{t}_c(\mathbf{P}))$$



Learning Transferable Visual Models From Natural Language Supervision [Radford et al., 2021]

Limitation: Overfitting to Seen Classes

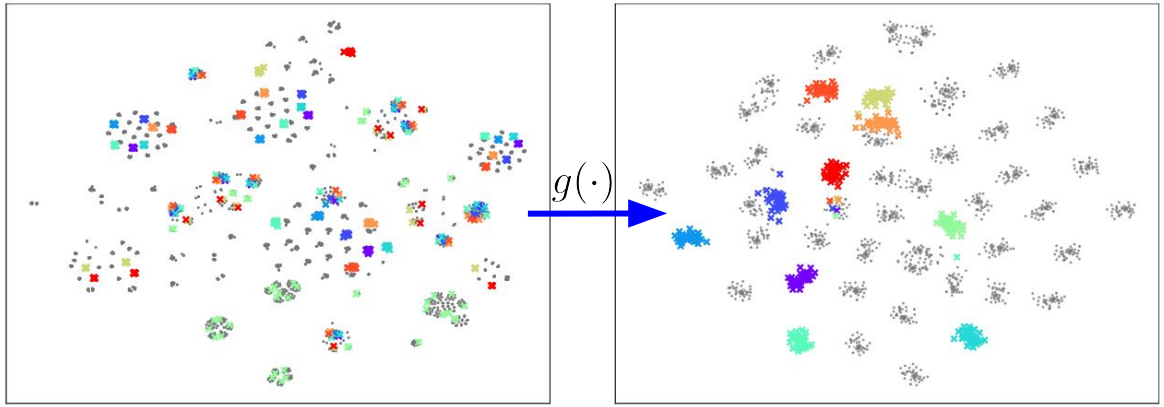
In few-shot settings, CoOp tends to:

- Learn prompts that are highly tuned to the training classes even with data augmentation
- Hurts performance on new / unseen classes or shifted domains

Why?

- We train one fixed prompt to fit the labeled training set as well as possible
- The model has no notion of uncertainty over prompts: it commits to one very specific solution

Input vs Output Embeddings



(a) Input embeddings (b) Output embeddings

Figure 2. **The t-SNE [41] visualization of the descriptions for** 50 **random categories on ImageNet.** The descriptions of each category are generated by 80 *hand-crafted* prompts presented by CLIP [31]. For clarity, we randomly select 10 categories and highlight them with different colors. Other categories are in gray. (**a**) The input embeddings of the text encoder, which are obtained by feeding the raw text into the embedding layer. Various descriptions within a category are scattered in the space, resulting in difficulty representing their distribution. (**b**) The output embeddings of the text encoder about category descriptions. Relying on the capability of the text encoder, the output embeddings of the descriptions within a category are close to each other, allowing them to be modeled with a simple distribution. (Best viewed in color.)

$\mathbf{t}_c(\mathbf{P})$ = $\mathbf{P}$ + {class}

$$\mathbf{w}_c(\mathbf{P}) = g(\mathbf{t}_c(\mathbf{P}))$$

Prompt Distribution Learning [Lu et al., 2022]

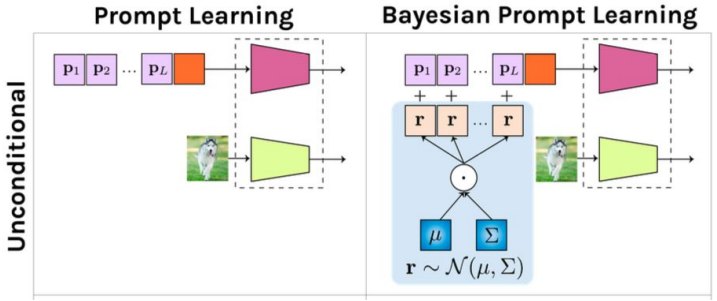
Unconditional Bayesian Prompt Learning

From a single prompt to a distribution

- Represent each prompt as base + residual:

$$\mathbf{p}_\gamma = [\mathbf{p}_1 + \mathbf{r}, \ldots, \mathbf{p}_L + \mathbf{r}]$$

- Introduce a latent residual **r**
  - same **r** applied to all prompt tokens
  - ***r*** is random, drawn from a learned distribution
  - Base tokens $\mathbf{p}_1, \ldots, \mathbf{p}_L$ are still learned



- Unconditional BPL:
  - Distribution over r does not depend on image x

In an ideal world…

- Find the posterior distribution over residual **r** that maximizes the marginal likelihood:

$$p(y = c \mid \mathbf{x}) = \int \frac{\exp\left(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(\mathbf{t}_c(\mathbf{r}))\right)}{\sum_{c'} \exp\left(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(\mathbf{t}_{c'}(\mathbf{r}))\right)} p_\gamma(\mathbf{r})\, d\mathbf{r}.$$

- Intractable to integrate over the entire 512-D prompt space

real posterior distribution over r

Variational Inference & ELBO

Introduce a surrogate distribution over residuals

- Learn a surrogate Gaussian distribution $\boldsymbol{\pi}_\phi(\mathbf{r}) \sim \mathcal{N}(\mu, \Sigma)$ that maximizes the marginal likelihood (optimality)
- How? By maximizing the variational bound (ELBO)

prior defined as N(0,I)

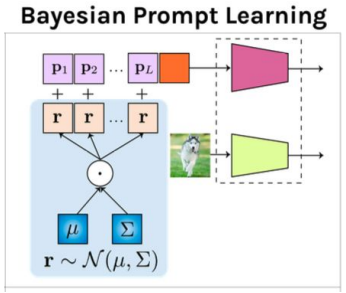
$$\log p(y|\mathbf{x}) \geq \mathbb{E}_{\pi_\phi(\mathbf{r})}\left[\log p(y \mid \mathbf{x},\, \mathbf{r})\right] - D_{\mathrm{KL}}\left[\pi_\phi(\mathbf{r}) \,\|\, p_\gamma(\mathbf{r})\right]$$

data fit term

regularization term

Optimization & Inference (Unconditional)

How do we train and use it?



- CLIP encoders f, g **frozen**
- Learn
  - base prompt tokens $\mathbf{p}_1, \cdots, \mathbf{p}_L$
  - mean and covariance $\boldsymbol{\pi}_\phi(\mathbf{r}) \sim \mathcal{N}(\mu, \Sigma)$
- Training
  - Reparameterization trick
  - Monte Carlo training objective: $\hat{\mathcal{L}}_{\text{BPL}} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{(\mathbf{x}_i, y_i)} \log p(y_i \mid \mathbf{x}_i, \mathbf{r}_k) + D_{\text{KL}}\big(\boldsymbol{\pi}_\phi(\mathbf{r}) \,\|\, p_\gamma(\mathbf{r})\big).$
- Inference (test time)

$$\log p(y|\mathbf{x}) \geq \mathbb{E}_{\pi_\phi(\mathbf{r})}\big[\log p(y \mid \mathbf{x}, \mathbf{r})\big] - D_{\text{KL}}\big[\pi_\phi(\mathbf{r}) \,\|\, p_\gamma(\mathbf{r})\big]$$

$$\mathbf{p}_\gamma = [\mathbf{p}_1 + \mathbf{r}_\gamma, \mathbf{p}_2 + \mathbf{r}_\gamma, \cdots, \mathbf{p}_L + \mathbf{r}_\gamma],$$

$$p(y = c \mid \mathbf{x}) \approx \frac{1}{K}\sum_{k=1}^{K} p(y = c \mid \mathbf{x}, \mathbf{r}_k), \quad \mathbf{r}_k \sim \boldsymbol{\pi}_\phi(\mathbf{r})$$

Results

Experimental Setup: FGVCAircraft

- **Dataset:** FGVCAircraft (fine-grained aircraft recognition)
  - 100 classes, 100 images per class
- **Backbone:** ViT-B/16
- **Few-shot regime:** 16 shots per class for training
- **Class split**
  - Base classes = used for training prompts
  - Novel classes = held out, test generalization
- **Baselines**
  - Zero-shot CLIP
  - CoOp
  - Unconditional BPL

Source: FGVCAircraft

Class: Falcon 900

Class: Cessna 560

Class: Tornado

Overview of paper's results

Table 1: **Task I: unseen prompts generalization** comparison between conditional Bayesian prompt learning and alternatives. Our model provides better generalization on unseen prompts compared to CoOp, CoCoOp and ProDA.

	CoOp [55]	CoCoOp [54]	ProDA [32]	Ours
Caltech101	89.81	93.81	93.23	**94.93**±0.1
DTD	41.18	56.00	56.48	**60.80**±0.5
EuroSAT	54.74	60.04	66.00	**75.30**±0.7
FGVCAircraft	22.30	23.71	34.13	**35.00**±0.5
Flowers102	59.67	**71.75**	68.68	**70.40**±1.8
Food101	82.26	91.29	88.57	**92.13**±0.1
ImageNet	67.88	70.43	70.23	**70.93**±0.1
OxfordPets	95.29	97.69	97.83	**98.00**±0.1
StanfordCars	60.40	**73.59**	71.20	73.23±0.2
SUN397	65.89	76.86	76.93	**77.87**±0.5
UCF101	56.05	73.45	71.97	**75.77**±0.1
Average	63.22	71.69	72.30	**74.94**±0.2

Manufacturer

Boeing or Airbus

Source: Wikipedia, Radio-Canada

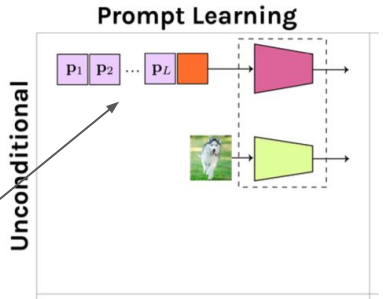
Variant

Example: Bombardier Global 6500 vs Global 7500

Results: Accuracy

Context Optimization (CoOp)

- **Image transforms**
  - Random resized crop
  - Random horizontal flip
- **Epochs:** 200
- **Trained on all 100 classes**
- **3 seeds**



Approach	Dataset	Backbone	Context length	Shots	CoOp	Ours
Context Optimization	FGVC Aircraft	ViT-B/16	**4**	16	40.27%	39.67%
Context Optimization	FGVC Aircraft	ViT-B/16	**8**	16	42.57%	41.90%
Context Optimization	FGVC Aircraft	ViT-B/16	**16**	16	43.53%	43.78%

Learning to prompt for vision-language models [Zhou et al., 2022]

Results: Accuracy

Task I: Unseen Prompts Generalization

- 100 classes
  - Train on half of the classes
  - Test on unseen half
- BPL
  - Reduced context length from 16 to 4
  - 10 MC samples, 10 epochs

Table 4: **Effect of variational formulation.** Formulating prompt learning as variational inference improves model generalization on unseen prompts compared to a non-Bayesian baseline [55], for both the unconditional and conditional setting.

	DTD	EuroSAT	FGVC	Flowers102	UCF101
Baseline	41.18	54.74	22.30	59.67	56.05
Unconditional	58.70	71.63	33.80	75.90	74.63
Conditional	60.80	75.30	35.00	70.40	75.77

Approach	Dataset	Backbone	Context length	MC Samples	Epochs	BPL	Ours
BPL - Unconditional	FGVC Aircraft	ViT-B/16	4	10	10	34.17%	34.45%
BPL - Unconditional	FGVC Aircraft	ViT-B/16	8	10	10	32.53%	31.61%
BPL - Unconditional	FGVC Aircraft	ViT-B/16	16	10	10	31.10%	30.39%

Bayesian Prompt Learning [Derakshani et al., 2023]

Results: Accuracy

Task I: Unseen Prompts Generalization

- 100 classes
  - Train on half of the classes
  - Test on unseen half
- BPL: Reduced context length from 16 to 4

Table 4: **Effect of variational formulation.** Formulating prompt learning as variational inference improves model generalization on unseen prompts compared to a non-Bayesian baseline [55], for both the unconditional and conditional setting.

	DTD	EuroSAT	FGVC	Flowers102	UCF101
Baseline	41.18	54.74	22.30	59.67	56.05
Unconditional	58.70	71.63	33.80	75.90	74.63
Conditional	60.80	75.30	35.00	70.40	75.77

Approach	Dataset	Backbone	Context length	Shots	Original	Ours
Zero-shot CLIP	FGVC Aircraft	ViT-B/16				23.13%
Context Optimization	FGVC Aircraft	ViT-B/16	4	16	**22.30%**	**25.25%**
BPL - Unconditional	FGVC Aircraft	ViT-B/16	4	16	**34.17%**	**34.45%**

Bayesian Prompt Learning [Derakshani et al., 2023]

Results: Accuracy

Task I: Unseen Prompts Generalization

- Weak baseline in the paper
  - 200 epochs + 4 tokens = overfitting
- BPL still better than CoOp, but improvement not as dramatic

Table 4: **Effect of variational formulation.** Formulating prompt learning as variational inference improves model generalization on unseen prompts compared to a non-Bayesian baseline [55], for both the unconditional and conditional setting.

	DTD	EuroSAT	FGVC	Flowers102	UCF101
Baseline	41.18	54.74	22.30	59.67	56.05
Unconditional	58.70	71.63	33.80	75.90	74.63
Conditional	60.80	75.30	35.00	70.40	75.77

Approach	Dataset	Backbone	Context length	Shots	Epochs	Ours
Context Optimization	FGVC Aircraft	ViT-B/16	4	16	200	25.25%
Context Optimization	FGVC Aircraft	ViT-B/16	4	16	100	27.17%
Context Optimization	FGVC Aircraft	ViT-B/16	4	16	75	27.63%
Context Optimization	FGVC Aircraft	ViT-B/16	4	16	50	28.33%
Context Optimization	FGVC Aircraft	ViT-B/16	4	16	25	**30.41%**
BPL - Unconditional	FGVC Aircraft	ViT-B/16	4	16	10	34.45%

Bayesian Prompt Learning [Derakshani et al., 2023]

Discussion

Key modeling assumptions

- **Global residual:** a single vector **r** shifting all prompt tokens captures useful prompt variability
- **Gaussian residual distribution**: residuals live roughly in a single Gaussian blob in the text-embedding space
- **Frozen CLIP encoder**

Overview

- **Pros:**
  - Simple conceptual change: one global latent r
  - Easy to implement on top of CoOp.
  - Acts as a Bayesian regularizer in prompt space, improving generalization to unseen classes.
- **Weaknesses:**
  - Extra cost at inference due to sampling (though small for modest K)
  - Gaussian form of the surrogate residual distribution may not accurately capture prompt distribution
  - Little improvement over past approaches on certain tasks/datasets

Ablation on MC Samples

Where does it break?

- Table 6 is Conditional BPL

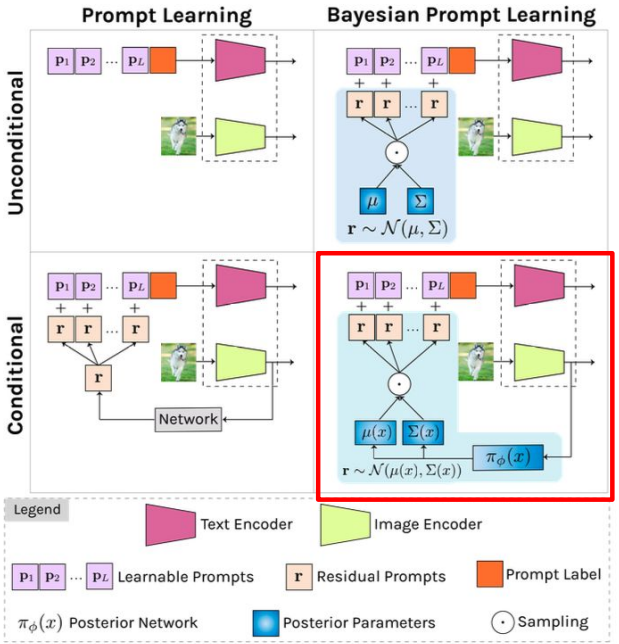


Table 6: **Influence of Monte Carlo sampling** on unseen prompts accuracy. As demonstrated, increasing the number of Monte Carlo samples boosts performance initially but reaches a plateau after 20 samples for all datasets.

	DTD	EuroSAT	FGVC	Flowers102	UCF101
1	56.40	64.50	33.00	72.30	75.60
2	60.00	67.40	33.90	73.90	76.20
5	62.20	71.00	34.20	74.00	76.60
10	61.60	73.60	34.40	73.50	77.00
20	62.90	74.80	35.00	74.00	77.10
40	62.60	76.10	35.50	73.80	77.15
80	63.50	76.20	34.70	74.20	77.20

Bayesian Prompt Learning [Derakshani et al., 2023]

Ablation on MC Samples

Where does it break?

- Performance of unconditional BPL diverges significantly from conditional BPL in low MC sample regime

Table 6: **Influence of Monte Carlo sampling** on unseen prompts accuracy. As demonstrated, increasing the number of Monte Carlo samples boosts performance initially but reaches a plateau after 20 samples for all datasets.

	DTD	EuroSAT	FGVC	Flowers102	UCF101
1	56.40	64.50	33.00	72.30	75.60
2	60.00	67.40	33.90	73.90	76.20
5	62.20	71.00	34.20	74.00	76.60
10	61.60	73.60	34.40	73.50	77.00
20	62.90	74.80	35.00	74.00	77.10
40	62.60	76.10	35.50	73.80	77.15
80	63.50	76.20	34.70	74.20	77.20

			BPL				Ours			
Approach	**MC Samples**	**Epochs**	**Seed 1**	**Seed 2**	**Seed 3**	**Average**	**Seed 1**	**Seed 2**	**Seed 3**	**Average**
BPL - Unconditional	**1**	10	3.30%	4.00%	4.10%	**3.80%**	2.70%	31.91%	1.38%	**12.00%**
BPL - Unconditional	**2**	10	22.70%	4.70%	5.40%	**10.93%**	5.52%	1.44%	27.11%	**11.36%**
BPL - Unconditional	**5**	10	34.20%	34.90%	32.70%	**33.93%**	28.19%	34.13%	36.23%	**32.85%**
BPL - Unconditional	**10**	10	34.80%	33.70%	34.00%	**34.17%**	34.73%	34.07%	34.55%	**34.45%**

Bayesian Prompt Learning [Derakshani et al., 2023]

Task III

	CoOp [55]	CoCoOp [54]	Ours
Task III: cross-domain prompts generalization			
ImageNetV2	**64.20**	64.07	**64.23**±0.1
ImageNet-Sketch	47.99	48.75	**49.20**±0.0
ImageNet-A	49.71	50.63	**51.33**±0.1
ImageNet-R	75.21	76.18	**77.00**±0.1
Average	59.27	59.88	**60.44**±0.1

Impact

1. First to frame prompt learning from Bayesian perspective and formulate it as variational inference problem

2. Models input prompt space in probabilistic manner which is compatible with prompt learning approaches that condition on the image

3. Demonstrated positive impact of Bayesian regularization on generalization to unseen prompts across different datasets and domains

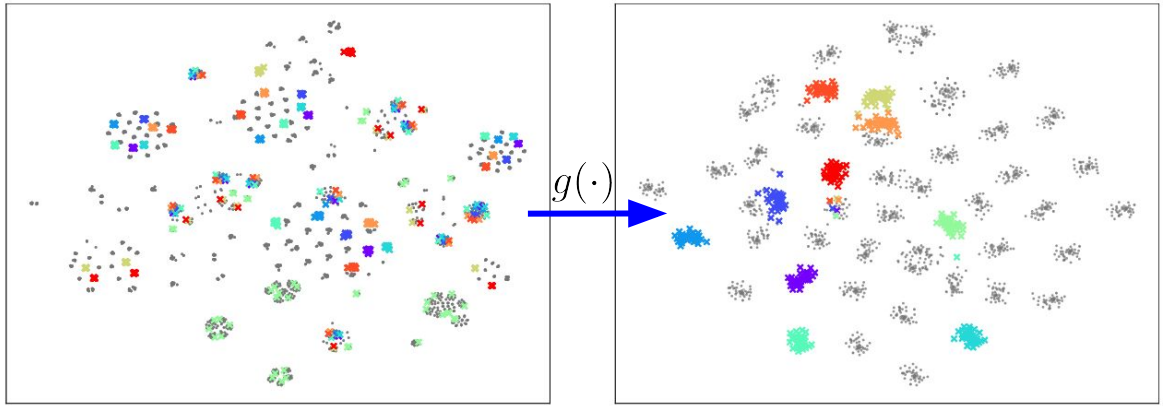
Questions

- Why do you think CLIP-based models still perform poorly on fine-grained classification tasks like FGVCAircraft?
- What are the fundamental limitations of prompt learning in the text embedding space?
- In what situation would the Gaussian assumption over the residual distribution break?
- Why do you think the performance collapsed when the number of MC samples was reduced?

References

1. Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. **Bayesian prompt learning for image-language model generalization**. ICCV, 2023
2. Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. **Prompt distribution learning.** *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5206–5215, June 2022.
3. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. **Conditional prompt learning for vision-language models**. *In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
4. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. **Learning to prompt for vision-language models**. *International Journal of Computer Vision* (IJCV), 2022.
5. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. **Learning transferable visual models from natural language supervision**. *In Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

Input vs Output Embeddings



(a) Input embeddings (b) Output embeddings

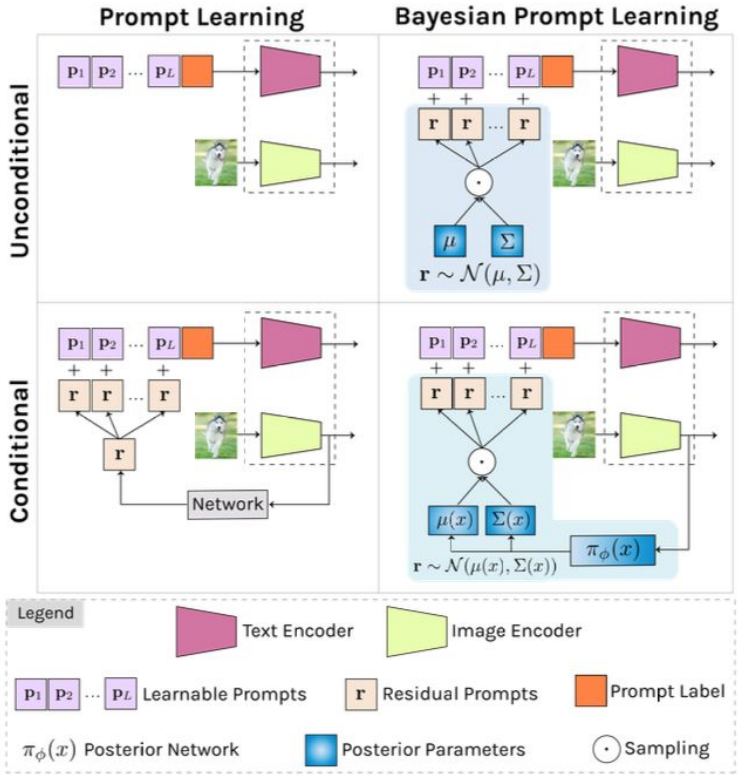
Figure 2. **The t-SNE [41] visualization of the descriptions for** 50 **random categories on ImageNet.** The descriptions of each category are generated by 80 *hand-crafted* prompts presented by CLIP [31]. For clarity, we randomly select 10 categories and highlight them with different colors. Other categories are in gray. (**a**) The input embeddings of the text encoder, which are obtained by feeding the raw text into the embedding layer. Various descriptions within a category are scattered in the space, resulting in difficulty representing their distribution. (**b**) The output embeddings of the text encoder about category descriptions. Relying on the capability of the text encoder, the output embeddings of the descriptions within a category are close to each other, allowing them to be modeled with a simple distribution. (Best viewed in color.)

$\mathbf{t}_c(\mathbf{P})$ = $\mathbf{P}$ + {class}

$$\mathbf{w}_c(\mathbf{P}) = g(\mathbf{t}_c(\mathbf{P}))$$

Prompt Distribution Learning [Lu et al., 2022]

Overview


Prompt Learning
Bayesian Prompt Learning
Unconditional
Conditional
Network
$\mathbf{r} \sim \mathcal{N}(\mu, \Sigma)$
$\mathbf{r} \sim \mathcal{N}(\mu(x), \Sigma(x))$
$\pi_\phi(x)$
Legend
Text Encoder
Image Encoder
Learnable Prompts
Residual Prompts
Prompt Label
Posterior Network
Posterior Parameters
Sampling