# Towards Generative Class Prompt Learning for Fine-grained Visual Recognition


**Soumitri Chattopadhyay[1], Sanket Biswas[2], Emanuele Vivoli[3], Josep Lladós[2]**

1. Department of Computer Science, University of North Carolina at Chapel Hill, USA
2. Computer Vision Center & Computer Science Department, Universitat Autònoma de Barcelona, Spain
3. MICC, University of Florence, Italy



BMVC 2024 — The 35th British Machine Vision Conference, 25th - 28th November 2024, Glasgow, UK


## Motivation, Challenges, and Contributions

### Limitations of CLIP-based representations

- Fine-grained category names are often highly dataset-specific that ***lack in semantic visual cues***
- CLIP's knowledge is about natural visual content, *cannot be adopted directly to* ***unseen domains***
- Visual concepts that are ***hard to describe by language*** (e.g. fractal patterns, abstract imagery) yield spurious representations from CLIP during prompting

Core underlying issue: ***Suboptimality of raw CLIP representations, which often lack fine-grained visual semantic awareness.*** We use *Generative Models to capture fine-grained visual information!*

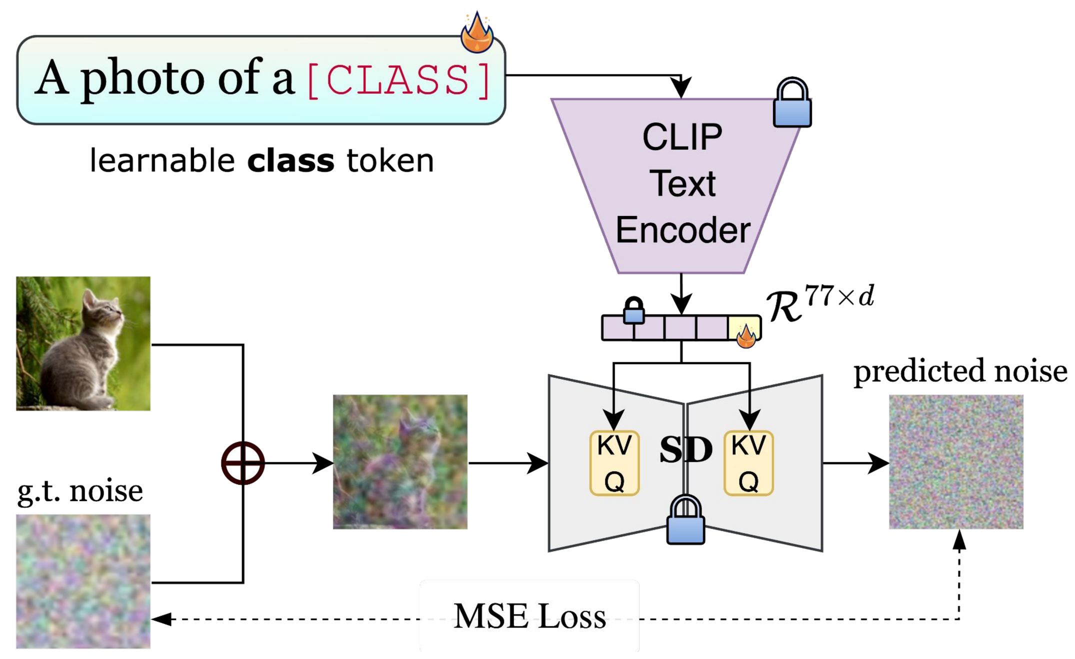
### Contributions

- We propose a *generative class prompt learning (GCPL)* baseline, leveraging pre-trained diffusion models to tackle CLIP's limitations.
- GCPL explicitly conditions CLIP class embeddings with *fine-grained visual semantic knowledge* via **generation-aided learning**.
- We further extend it, advocating for learning stronger vision-induced textual representations with ***inter-class discriminative knowledge.***

*To our best knowledge – one of the first attempts to introduce* ***generation guided prompting*** *for few-shot VLM adaptation!*

## GCPL: Generative Class Prompt Learning

- Inject learnable `[CLASS]` token via handcrafted prompt into CLIP *(only this token is* ***trainable!****)*
- Use it to condition a T2I LDM, optimizing **L2 loss** w.r.t. the few-shot support set.

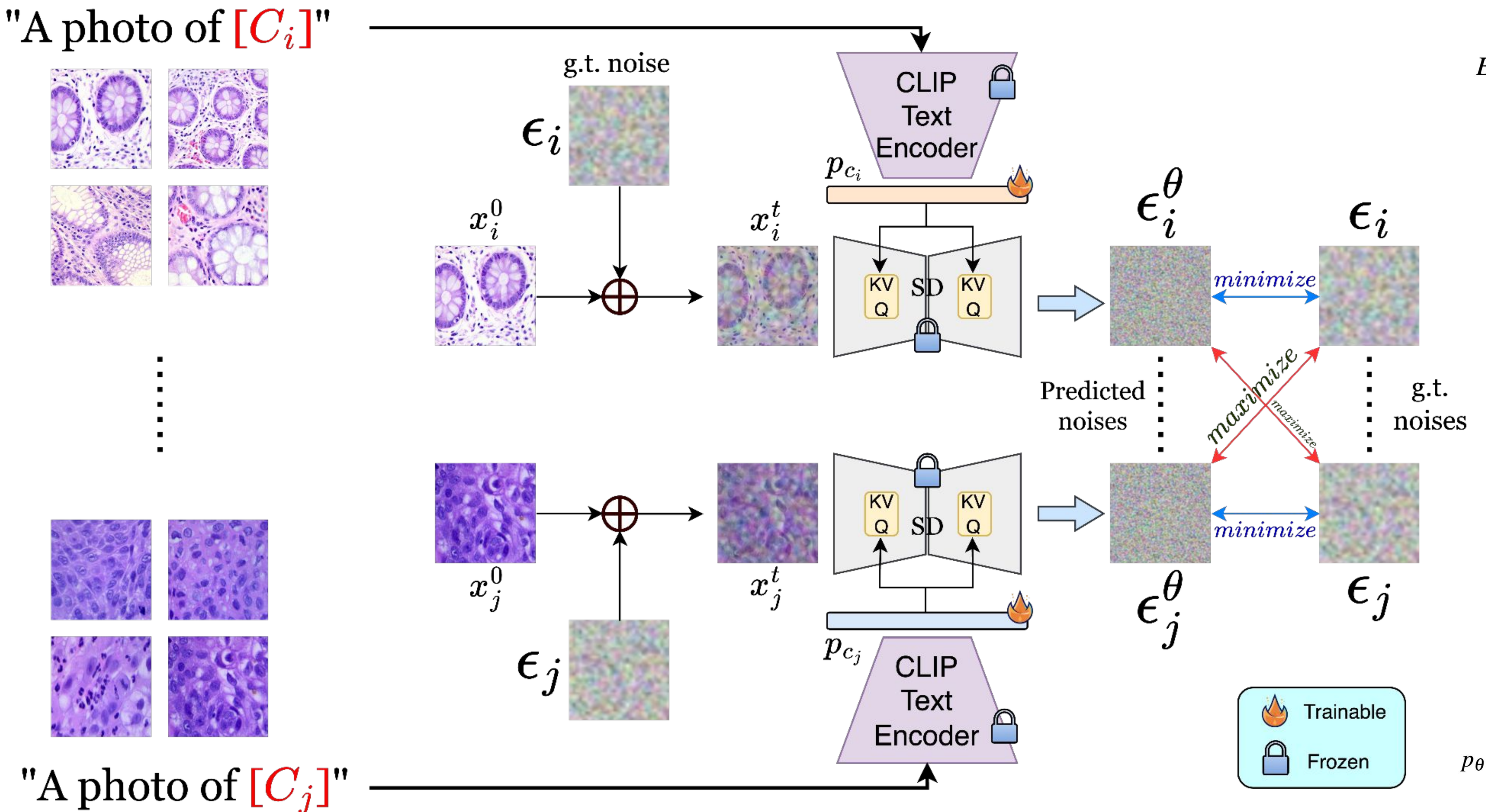



$$\mathcal{L}_{GCPL} = \mathbb{E}_{x\sim\mathcal{E}(x),p_c,\varepsilon\sim\mathcal{N}(O,1),t}\left[\|\varepsilon_c - \varepsilon_c^{\theta}(x_t,t,c_\theta(p_c))\|_2^2\right]$$

$$p_c^* = \arg\min_{p_c} \mathbb{E}_{x\sim\mathcal{E}(x),p_c,\varepsilon\sim\mathcal{N}(O,1),t}\left[\|\varepsilon_c - \varepsilon_c^{\theta}(x_t,t,c_\theta(p_c))\|_2^2\right]$$

## CoMPLe: Contrastive Multi-Class Prompt Learning

**Extends GCPL to multi-class setting – all class prompts are jointly optimized by additionally enforcing divergence of the noise predictions across other classes.**




$$\mathcal{L}_{CoMPLe} = \frac{1}{B}\sum_{i=j}^{B}\mathbb{E}_{x\sim\mathcal{E}(x),p_{c_j},\varepsilon\sim\mathcal{N}(O,1),t}\left[\|\varepsilon_{c_i} - \varepsilon_{c_j}^{\theta}(x_t^j,t,c_\theta(p_{c_j}))\|_2^2\right] - \lambda\cdot\frac{1}{B(B-1)}\sum_{i\neq j}^{B}\mathbb{E}_{x\sim\mathcal{E}(x),p_{c_j},\varepsilon\sim\mathcal{N}(O,1),t}\left[\|\varepsilon_{c_i} - \varepsilon_{c_j}^{\theta}(x_t^j,t,c_\theta(p_{c_j}))\|_2^2\right]$$

## Few-shot Diffusion Classifier

### Inference pipeline after training

- ELBO approximation for LDMs:

$$ELBO = -\mathbb{E}_\varepsilon\left[\sum_{t=2}^{T} w_t\|\varepsilon-\varepsilon_\theta(x_t,c)\|_2^2 - \log p_\theta(x_0\mid x_1,c)\right]+C$$
$$= -\mathbb{E}_{\varepsilon,t}\left[\|\varepsilon-\varepsilon_\theta(x_t,c)\|_2^2\right]+C$$

- Bayes' theorem gives us:

$$p_\theta(c_i\mid x) = \frac{p(c_i)p_\theta(x|c_i)}{\sum_j p(c_j)p_\theta(x|c_j)}$$

- Simplifying using ELBO:

$$p_\theta(c_i\mid x) = \frac{\exp\{-\mathbb{E}_{\varepsilon,t}\left[\|\varepsilon-\varepsilon_\theta(x_t,c_i)\|_2^2\right]\}}{\sum_j\exp\{-\mathbb{E}_{\varepsilon,t}\left[\|\varepsilon-\varepsilon_\theta(x_t,c_j)\|_2^2\right]\}}$$

- Conditioning signal *c* is derived from *few-shot learned [CLASS] prompts* ⇒ **few-shot diffusion classifier!**

$$c_i = c_\theta(p_{c_i})$$

$$p_\theta(c_i\mid x) = \frac{1}{\sum_j exp\{\mathbb{E}_{\varepsilon,t}\left[\|\varepsilon-\varepsilon_\theta(x_t,c_\theta(p_{c_i}))\|_2^2 - \|\varepsilon-\varepsilon_\theta(x_t,c_\theta(p_{c_j}))\|_2^2\right]\}}$$

(please refer to paper for details.)

## Quantitative Results: Few-shot Classification

### Medical imaging datasets

- Zero-shot methods *completely fail* on the unseen domain.
- **GCPL** and **CoMPLe** *significantly boosts* performance over prior SoTA.
- **Prompt learning** is very noisy for unseen domain (i.e. medical datasets) – as seen from **high variances.**
- **GCPL** and **CoMPLe** are lot consistent and robust across unseen domains.

| Method | CRC5k | ISIC2018 | LC25000 |
|---|---|---|---|
| **Zero-Shot** | | | |
| CLIP | 21.49 | 14.43 | 25.40 |
| Diffusion Classifier | 24.16 | 10.41 | 17.29 |
| **Adapter** | | | |
| Tip-Adapter | 59.90±2.18 | 33.88±7.26 | 80.48±1.93 |
| Tip-Adapter-F | 71.44±2.46 | 40.32±5.19 | 86.02±1.59 |
| **Prompt learning** | | | |
| CoCoOp | 60.91±2.98 | 24.67±6.54 | 73.86±4.19 |
| KgCoOp | 59.90±5.17 | 29.16±6.82 | 75.87±3.88 |
| MaPLe | 40.56±16.12 | 30.33±13.67 | 71.96±5.22 |
| PromptSRC | 56.45±18.28 | 44.18±7.02 | 77.54±1.51 |
| **Ours** | | | |
| Ours-GCPL | 74.76±1.94 | 48.84±2.13 | 93.44±0.78 |
| Ours-CoMPLe | **76.36**±1.82 | **49.27**±2.59 | **94.83**±0.28 |

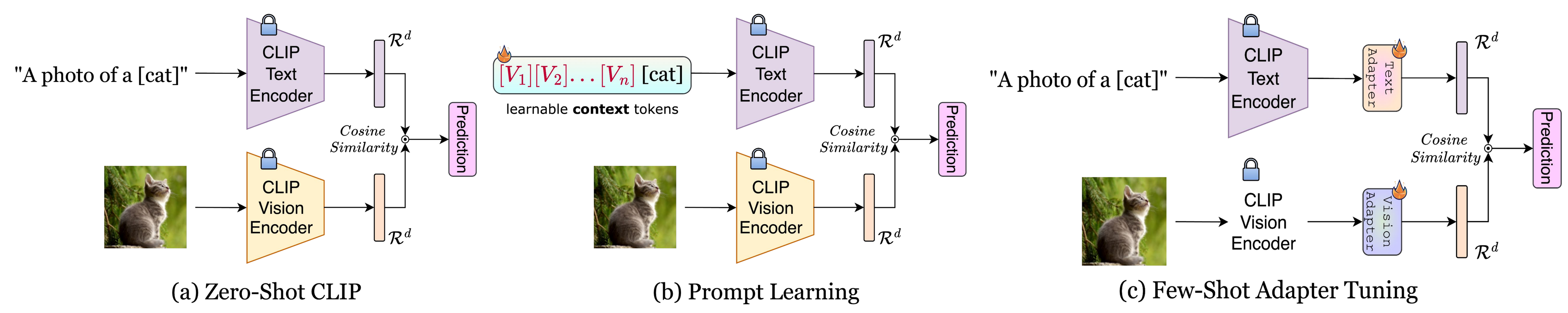
### Fine-grained natural image datasets

- Mostly observe *high gains* over prior few/zero-shot methods.

| Method | StanfordCars | Cornseeds | Flowers102 | Fractals |
|---|---|---|---|---|
| **Zero-Shot** | | | | |
| CLIP | 65.56 | 18.47 | 70.73 | 9.25 |
| Diffusion Classifier | 76.77 | 17.77 | 54.21 | 6.25 |
| **Adapter** | | | | |
| Tip-Adapter | 65.82±0.51 | 34.27±3.97 | 89.28±0.55 | 81.49±1.22 |
| Tip-Adapter-F | 75.14±0.35 | 39.61±2.88 | 94.25±0.43 | 86.16±0.54 |
| **Prompt learning** | | | | |
| CoCoOp | 71.57±0.76 | 36.56±5.42 | 87.84±0.48 | 67.89±1.29 |
| KgCoOp | 78.76±0.61 | 38.45±4.84 | 91.97±0.44 | 72.84±0.93 |
| MaPLe | 74.39±0.43 | 34.37±15.44 | 93.96±0.61 | 76.91±6.55 |
| PromptSRC | 83.33±0.35 | 33.69±4.55 | **97.06**±0.27 | **93.45**±0.52 |
| **Ours** | | | | |
| Ours-GCPL | **88.47**±0.27 | 43.42±2.84 | 93.45±1.39 | 90.76±2.23 |
| Ours-CoMPLe | 87.69±1.47 | **45.79**±2.12 | 90.73±1.05 | 88.83±1.57 |

## Experimental Setup

**Competitors:** existing VLM adaptation paradigms

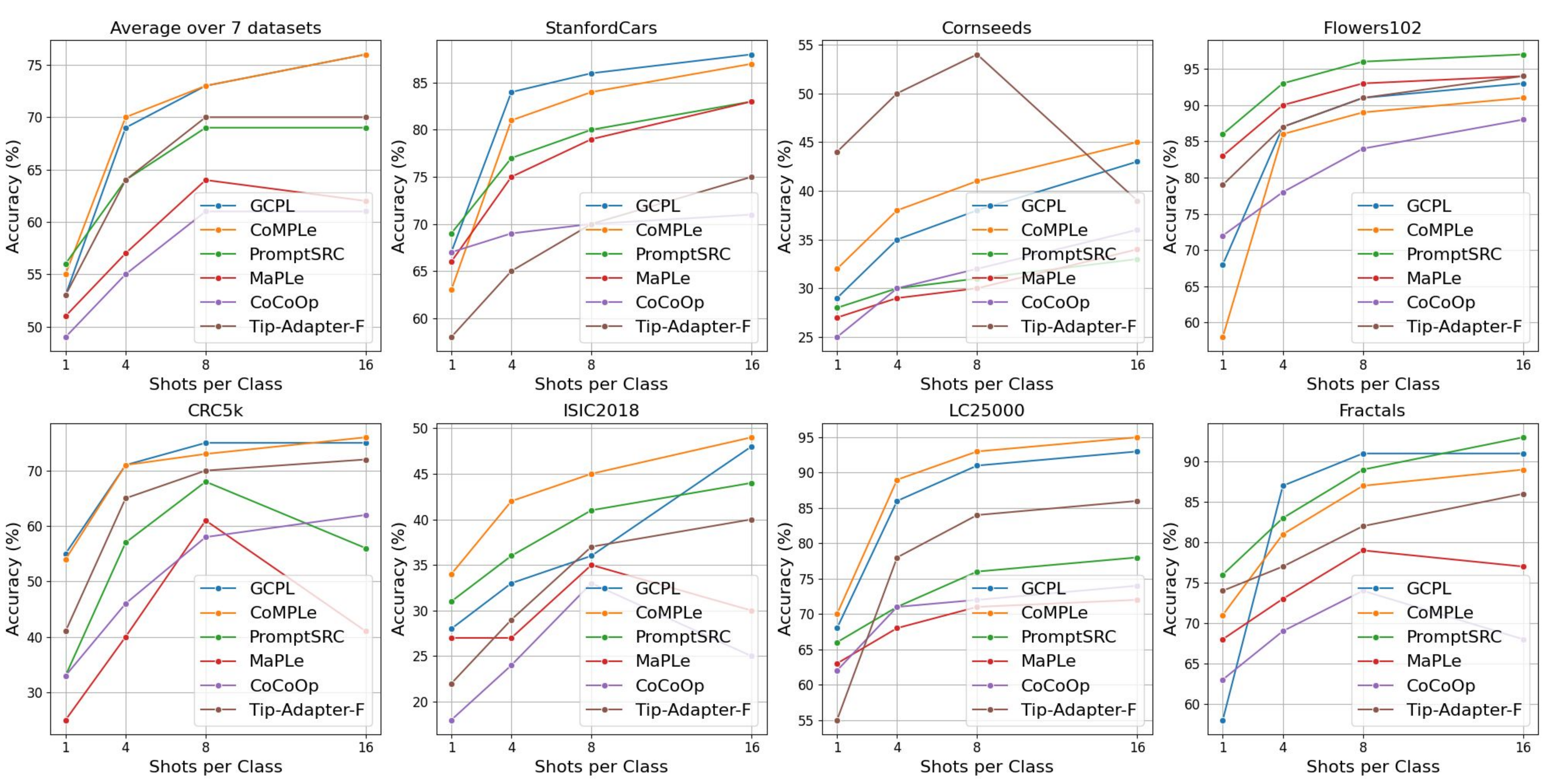



(a) Zero-Shot CLIP (b) Prompt Learning (c) Few-Shot Adapter Tuning

**Datasets:** (a) fine-grained natural images; (b) medical images; (c) abstract patterns

| Dataset | Visual concept | Prompt template | Initializer word |
|---|---|---|---|
| StanfordCars | Vehicular variants | "A photo of `[CLASS]`, a type of car." | car |
| Cornseeds | Natural images, agriculture | "A photo of `[CLASS]` corn seed." | seed |
| CRC5k | Histopathology | "`[CLASS]` tissue." | tissue |
| ISIC2018 | Dermatology | "`[CLASS]` skin lesion." | skin |
| LC25000 | Histopathology | "`[CLASS]` tissue." | tissue |
| Fractals | Abstract imagery | "`[CLASS]` fractal." | fractal |

## Ablation Study: Varying number of shots per class





**Acknowledgements:** This work acknowledges the Spanish projects GRAIL PID2021-126808OB-I00, DocAI 2021- SGR-01559, the CERCA Program / Generalitat de Catalunya, and PhD Scholarship from AGAUR 2023 FI-3- 00223. The authors also acknowledge Gedas Bertasius and Feng Cheng of UNC Chapel Hill for constructive discussions and hardware resources.



For more details, please refer to the arXiv version of our paper at: https://arxiv.org/abs/2409.01835 or email authors at: soumitri@cs.unc.edu | sbiswas@cvc.uab.es | evivoli@cvc.uab.es. Thanks for visiting!