Learning to Prompt with Text Only Supervision for Vision-Language Models

Muhammad Uzair Khattak[1], **Muhammad Ferjad Naeem**[5], **Muzammal Naseer**[2], **Luc Van Gool**[3], **Federico Tombari**[4,5]

[1]MBZ University of AI
[2]Computer Science Department and Center of Secure Cyber-Physical Security Systems, Khalifa University
[3]INSAIT
[4]TU Munich
[5]Google

Abstract

Foundational vision-language models like CLIP are emerging as a promising paradigm in vision due to their excellent generalization. However, adapting these models for downstream tasks while maintaining their generalization remains challenging. In literature, one branch of methods adapts CLIP by learning prompts using images. While effective, these methods often rely on image-label data, which is not always practical, and struggle to generalize to new datasets due to overfitting on few-shot source data. Another approach explores training-free methods by generating class captions from large language models (LLMs) and performing prompt ensembling, but these methods often produce static, class-specific prompts that cannot be transferred to new classes and incur additional costs by generating LLM descriptions for each class separately. In this work, we aim to combine the strengths of both approaches by learning prompts using only text data derived from LLMs. As supervised training of prompts in the image-free setup is non-trivial, we develop a language-only efficient training approach that enables prompts to distill rich contextual knowledge from LLM data. Furthermore, by mapping the LLM contextual text data within the learned prompts, our approach enables zero-shot transfer of prompts to new classes and datasets, potentially reducing the LLM prompt engineering cost. To the best of our knowledge, this is the first work that learns generalized and transferable prompts for image tasks using only text data. We perform evaluations on 4 benchmarks, where ProText improves over ensembling methods while being competitive with those using labeled images.

Code — https://github.com/muzairkhattak/ProText

Introduction

Foundational Vision-Language models (Radford et al. 2021; Jia et al. 2021; Yu et al. 2022; Lai et al. 2023), which are large DNNs pre-trained on web-scale data have started to become increasingly popular in the vision community due to their wide scale applicability and generalization abilities. Among these, Vision-Language models (VLMs) such as CLIP (Radford et al. 2021) stand out as the latest highlights which leverage contrastive pre-training on massive image-text pairs from the internet. During pre-training, CLIP learns to align image-text samples in a shared feature space. This allows CLIP to encode open-vocabulary concepts and generalize to zero-shot tasks such as image recognition (Kim et al. 2022), object detection (Feng et al. 2022), and image segmentation (Lüddecke and Ecker 2022).

	Method	Do not require images	Transfer to unseen datasets
Prompt learning methods	CoOp	✘	✓
	CoCoOp	✘	✓
	MaPLe	✘	✓
	ProGrad	✘	✓
	TCP	✘	✓
Prompt ensembling methods (LLM)	DCLIP	✓	✘
	WaffleCLIP-Concept	✓	✘
	CuPL	✓	✘
	ProText (Ours)	✓	✓

Table 1: Existing methods improve CLIP's generalization by learning prompts with image supervision or using non-transferable prompt ensembling with LLM knowledge. In contrast, our approach aims to learn prompts with text-only data which are transferable to new datasets and classes.

CLIP consists of two encoders to encode image and text inputs respectively. For zero-shot classification, a handcrafted prompt such as '`a photo of a CLS`' is used as the text input. Text features of classes are compared with visual feature and class with highest similarity is assigned as predicted label. Improving the quality of text templates such as adding attributes (An et al. 2023), or class-specific details (Pratt et al. 2023; Jin et al. 2021) has shown to improve CLIP performance. However, designing high-quality prompts (also known as textual-prompt enhancement) that can best describe test image remains a key challenge, as image content is not known in advance.

In literature, numerous techniques have been proposed to adapt CLIP for recognition tasks. One branch of methods (Zhou et al. 2022a,b; Chen et al. 2022; Huang, Chu, and Wei 2022; Shu et al. 2022; Lu et al. 2022) treat text prompts as learnable vectors and optimize them using task-specific objectives such as cross-entropy. As prompts are learned in the embedding space, this allows them to be used with classes and datasets beyond those on which they were trained on. While effective over the baseline CLIP, most of these methods require annotated image labels to optimize the prompts which is often impractical, especially in real-world scenarios such as medical imaging, remote sensing, security, surveil-

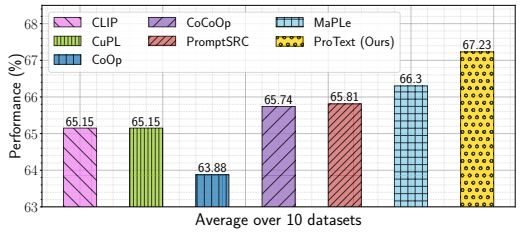

Figure 1: Without using any images for training, ProText using text-only data improves over CLIP, CuPL, and prior 16-shot image-supervised methods in challenging cross-dataset transfer settings. Prompt ensembling-based CuPL (training-free) performs same as CLIP as it cannot transfer class-specific LLM templates to cross-datasets.

lance, etc. Moreover, these methods tend to overfit on few-shot source samples and struggle to retain CLIP's generalization, especially in cross-dataset settings.

Alternatively, several methods (Pratt et al. 2023; Menon and Vondrick 2023; Roth et al. 2023) have proposed training-free variants of prompt ensembling by using Large Language Models (LLMs). Instead of using hand-crafted templates, these methods obtain dataset or class-specific descriptors and captions from LLMs to enhance text features. These enriched features aim to better represent content that could occur in test images, leading to improvements over the baseline CLIP. Although these methods do not require images, the knowledge acquired from LLMs is mostly specific to each class and not directly transferable to unseen classes and datasets as no optimization is performed. Additionally, generating LLM descriptions for each concept separately incurs additional LLM serving and prompt engineering costs.

In this work, we present a new paradigm to improve CLIP's generalization. Our motivation comes from combining the strengths of prompt learning and prompt ensembling approaches while effectively addressing their limitations. To this end, we introduce ProText: **Pro**mpt Learning with **Text**-Only Supervision. In contrast to previous methods, our approach instead proposes to learn prompts using text only data obtained from LLMs. As supervised training of prompts is not trivial due to image-free setting, we develop a novel training framework that allows prompts to learn and extract rich contextual knowledge from LLM data. Moreover, as LLM contextual knowledge is mapped within the learned prompts, it enables zero-shot transfer of prompts to new classes and datasets, potentially leading to a substantial reduction in LLM serving and prompt engineering cost.

As shown in Tab. 1, ProText adopts a text-only approach to learn prompts, and in addition the adapted CLIP transfers well to unseen classes and datasets, therefore addressing the transferability limitations of LLM-based prompt ensembling methods. We demonstrate the effectiveness of ProText by performing extensive evaluations on 4 benchmarks. On challenging cross-dataset transfer setting, ProText without using any visual information achieves an average gain of +2.08% over CLIP while surpassing the performance of previous best image-supervised prompt learning method MaPLe (Khattak et al. 2023a) by +0.93% (Fig. 1). Further, ProText with text-only supervision performs competitively against prior methods in domain generalization, base-to-novel class, and text-only supervised setting. Our main contributions are summarized as follows:

- Our work explores the task of prompt learning in CLIP using text-only data. Our method harmonically combines the strengths of prompt learning and prompt ensembling methods to improve CLIP's generalization.
- To optimize prompts with text-only data, we develop a training approach that allows prompts to learn a mapping function by distilling rich contextual information derived from LLM textual data.
- As LLM contextual knowledge is mapped within the learned prompts, this enables prompts to be directly used with new classes and datasets potentially reducing the additional LLM serving and prompt engineering cost.
- We validate the effectiveness of our method through extensive experiments on four benchmarks. Our ProText approach improves the generalization of CLIP across various settings and fares competitive to approaches that explicitly use labeled image samples during training.

Related Work

Foundational Vision-Language models (VLMs). VLMs (Radford et al. 2021; Jia et al. 2021; Yu et al. 2022; Yuan et al. 2021; Yao et al. 2021; Naeem et al. 2023b) leverage joint image-text pretraining using internet-scale data in a self-supervised fashion. Using the contrastive learning objective, VLMs learn rich multi-modal features by attracting together the features of paired images and texts while repelling un-paired image-text features in a joint feature space. The resulting model learns open-vocabulary concepts interpretable through natural language suitable for downstream discriminative vision tasks such as open-vocabulary image classification (Khattak et al. 2023a; Lu et al. 2022; Chen et al. 2022; Zhou et al. 2022b; Naeem et al. 2022, 2023a), detection (Du et al. 2022; Zhou et al. 2022c; Minderer et al. 2022; Liu et al. 2023; Bangalath et al. 2022), and segmentation (Ghiasi et al. 2022; Li et al. 2022; Liang et al. 2023). Although promising, adapting VLMs effectively while maintaining their original generalization remains a challenge. In this work, we propose a novel method to adapt CLIP with *transferable* prompt learning through *text modality* supervision to improve its performance on *vision modality* tasks.

Prompt Learning for VLMs. Prompt Learning (Chen et al. 2022; Zhou et al. 2022a,b; Lu et al. 2022; Derakhshani et al. 2023; Shu et al. 2022; Yao, Zhang, and Xu 2024; Zhu et al. 2023) has emerged as an effective fine-tuning strategy to adapt large-scale models. This approach adds a small number of learnable embeddings along with model inputs which are optimized during training while the rest of the model is kept frozen. As the pre-trained model remains unchanged, this technique has become particularly effective for VLMs such as CLIP, where maintaining the model's original generalizability is crucial. CoOp (Zhou et al. 2022b) is the pioneering prompt learning method for CLIP which learns text prompts to fine-tune CLIP. CoCoOp (Zhou et al. 2022a) improves CoOp's generalization by conditioning text prompts on visual features. MaPLe (Khattak et al. 2023a)

proposes multi-modal prompts to adapt both vision and language branches of CLIP. UPL (Huang, Chu, and Wei 2022) adopts an unsupervised prompt learning approach to fine-tune CLIP. PromptSRC (Khattak et al. 2023b) improves prompt learning from a regularization perspective by using additional loss functions during training. While these methods improve baseline CLIP performance, most of them require labeled image samples, which is less practical, and generating pseudo-labels is often less effective. In contrast, we present a novel prompt learning approach that improves CLIP generalization without relying on any visual samples during training.

Training-Free Text Prompt Enhancement. With the emergence of LLMs like GPT-3 (Brown et al. 2020), several approaches (Menon and Vondrick 2023; Roth et al. 2023; Pratt et al. 2023) have explored their potential for improving zero-shot generalization of CLIP. Instead of using handcrafted templates for generating class features, these methods use LLMs to generate high-level concepts, class captions, and/or attributes which are used in one form or another to enrich text features. DCLIP (Menon and Vondrick 2023) generates fine-grained per-class language descriptors and ensemble its similarity with image to produce classification scores. WaffleCLIP (Roth et al. 2023) matches DCLIP performance with random descriptors and show further gains by data-specific concepts generated via LLMs. CuPL (Pratt et al. 2023) query LLMs to generate class-specific prompt descriptions for prompt ensembling. While effective, most of these approaches generate class-specific text from LLMs which are not directly transferable to unseen classes and new datasets since no training is performed. Our work aims to leverage the same LLM data via a novel text-only prompt learning technique that allows the transfer of learned prompts to unseen classes and new datasets.

Text-only Supervision for VLMs. Recently, several methods (Nukrai, Mokady, and Globerson 2022; Zhang et al. 2023; Cho et al. 2023; Gu, Clark, and Kembhavi 2023) have been proposed to adapt VLMs with text-only supervision. DrML (Zhang et al. 2023) trains a classifier on text features and transfers it to image features at inference for the error slice discovery problem. PromptStyler (Cho et al. 2023) adopts a similar approach for domain generalization tasks and learns a fixed-head classifier on different style prompts, where the style prompts are learned separately before training the classifier. While these methods perform text-only training similar to ours, they adopt the *cross-modal transfer principle* (e.g., transferring a text classifier to images). Secondly, they struggle to transfer to new classes due to the fixed head classifier network. On the other hand, our approach is motivated from the direction of *text-prompt enhancement* and it transfers well to unseen classes.

Method

Given the language interpretable nature of VLMs such as CLIP (Radford et al. 2021), they are naturally suited for zero-shot recognition tasks. However, to achieve full potential of CLIP's generalization for downstream tasks, adaptation still appears to be necessary. One line of methods adopts prompt learning (Lu et al. 2022; Khattak et al. 2023a; Zhou et al. 2022a,b) to re-purpose CLIP features for downstream data. These methods often require image samples with labels to learn the prompts. Another line of methods adopts training-free prompt ensembling (Pratt et al. 2023; Menon and Vondrick 2023; Roth et al. 2023) using LLMs. Although prompt ensembling do not require image information, these methods mostly generate class-specific LLM prompts that are not directly transferable to new classes and datasets.

In this work, we explore a new paradigm for learning transferable prompts for VLMs using text-only supervision. Our adaptation framework, ProText: **Pro**mpt Learning with **Text** only supervision aims to learn *transferable* prompts *without* utilizing images, aiming to combine the text-only and transferability characteristics from training-free and image-based prompt learning methods respectively. Fig. 2 shows our overall framework. First, we curate text-only LLM caption data using class names of a given dataset and a LLM such as GPT-3 (Brown et al. 2020). As a text-supervised approach, ProText only requires CLIP text encoders during training. Specifically, we employ one frozen encoder with learnable prompts and a second frozen encoder without learnable prompts. Learnable prompts with class-name templates are input to the prompted text encoder to obtain the class-name template feature, and a frozen text encoder generates LLM template feature from its description obtained from LLM data. Next, we employ a contextual mapping training objective which maps class-name template feature to the LLM template feature. Contextual mapping allows the prompts to learn a generalized mapping function that embeds rich contextual knowledge from LLM data within the prompt vectors. As prompts are learned in the embedding space, they are directly compatible with new classes and datasets. At inference, the learned prompts are used with CLIP model for standard zero-shot CLIP inference.

Preliminaries

Contrastive Language-Image Pre-training (CLIP). CLIP consist of an image encoder f and a text encoder g which maps image and text input into visual and textual feature respectively. We denote CLIP parameters as $\theta_{\texttt{CLIP}} = \{\theta_f, \theta_g\}$ where θ_f and θ_g refer to the image and text encoder parameters, respectively. Input image $\boldsymbol{X}$ is divided into M patches which are linearly projected to produce patch tokens and a learnable class token CLS is prepended resulting in the final sequence as $\tilde{\boldsymbol{X}} = \{\text{CLS}, \boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_M\}$. The image encoder f encodes the input patches via multiple transformer blocks to produce a latent visual feature $\tilde{\boldsymbol{f}} = f(\tilde{\boldsymbol{X}}, \theta_f)$, where $\tilde{\boldsymbol{f}} \in \mathbb{R}^d$. Next, the corresponding class label y is embedded in a text template, such as 'a photo of a [CLASS]' which can be formulated as $\tilde{\boldsymbol{Y}} = \{\text{SOS}, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L, \boldsymbol{c}_k, \text{EOS}\}$. Here $\{\boldsymbol{t}_l|_{l=1}^{L}\}$ and $\boldsymbol{c}_k$ are the word embeddings corresponding to the text template and the label y, respectively while SOS and EOS are the learnable start and end token embeddings. The text encoder g encodes $\tilde{\boldsymbol{Y}}$ via multiple transformer blocks to produce the latent text feature as $\tilde{\boldsymbol{g}} = g(\tilde{\boldsymbol{Y}}, \theta_g)$, where $\tilde{\boldsymbol{g}} \in \mathbb{R}^d$. For zero-shot inference, text features of text template with class labels $\{1, 2, \cdots, C\}$ are matched with image feature $\tilde{\boldsymbol{f}}$ as

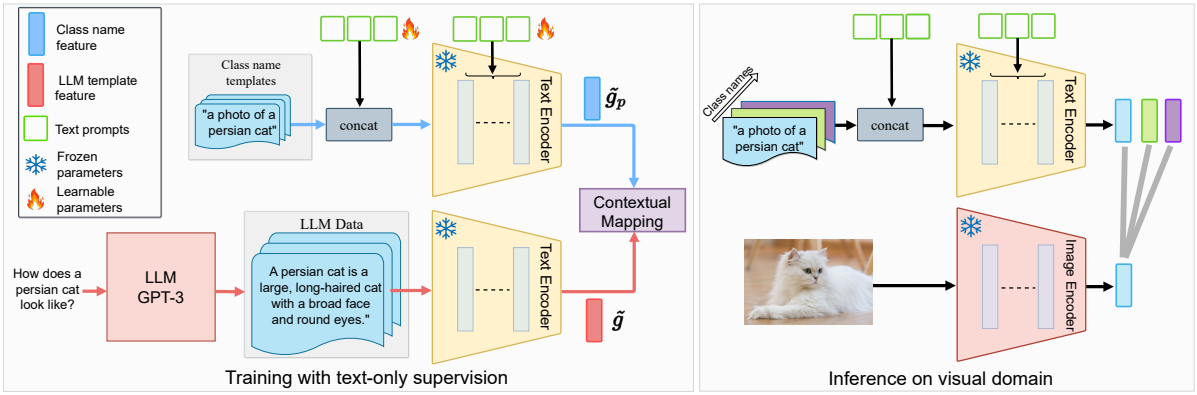

Figure 2: Overview of ProText. **(Left)** First, diverse captions are generated for training classes using GPT-3. During training, CLIP text encoders generate **prompted class-name feature** ($\tilde{g}_p$) from class-name templates with learnable prompts and **frozen LLM template feature** ($\tilde{g}$) from LLM generated templates. Next, we employ contextual mapping loss to guide learnable prompts to learn a mapping from prompted class-name feature to LLM template feature containing more information about the class. This allows the learned prompts to exploit internal knowledge of text encoder complemented by LLM descriptions. **(Right)** At inference, learned prompts are used with class-names and zero-shot CLIP inference protocol is followed. Moreover, contextual information from LLM descriptions mapped to the learned prompts enables its transfer to new classes and datasets.

$\frac{\texttt{exp}(\texttt{sim}(\tilde{\boldsymbol{g}}\cdot\tilde{\boldsymbol{f}})\tau)}{\sum_{i=1}^{C}\texttt{exp}(\texttt{sim}(\tilde{\boldsymbol{g}}_i\cdot\tilde{\boldsymbol{f}})\tau)}$, where $\texttt{sim}()$ denotes the cosine similarity and τ is the temperature.

Prompt Learning with CLIP. Being a parameter efficient tuning method, prompt learning has emerged as a popular technique to adapt vision-language models like CLIP. Since most of the model is kept frozen during adaptation, prompt learning aims to reduce overfitting. Learnable prompts are appended either at the image side (Bahng et al. 2022), text encoder side (Zhou et al. 2022a,b), or both sides. In this work, we learn hierarchical prompts at the text encoder named Deep Language Prompting (DLP) (Khattak et al. 2023a) formulated as follows.

T learnable language prompts $\boldsymbol{P_t} = \{\boldsymbol{p_t^1}, \boldsymbol{p_t^2}, \cdots, \boldsymbol{p_t^T}\}$ are appended with text input tokens, resulting in $\tilde{\boldsymbol{Y}}_{\boldsymbol{p}} = \{\text{SOS}, \boldsymbol{P_t}, \boldsymbol{t}_1, \boldsymbol{t}_2, \cdots, \boldsymbol{t}_L, \boldsymbol{c}_k, \text{EOS}\}$. The text encoder processes $\tilde{\boldsymbol{Y}}_{\boldsymbol{p}}$ and prompted text feature is obtained as $\tilde{\boldsymbol{g}}_{\boldsymbol{p}} = g(\tilde{\boldsymbol{Y}}_{\boldsymbol{p}}, \theta_g)$. We use deep prompting which learns hierarchical prompts at subsequent transformer blocks of text encoder. Visual feature $\tilde{\boldsymbol{f}}$ is obtained without utilizing learnable prompts. To adapt CLIP on image classification task on dataset $\mathcal{D}$, prompts $\boldsymbol{P_t}$ are optimized in a supervised fashion using labeled image samples with cross-entropy loss, $\mathcal{L}_{\text{CE}}$.

$$\mathcal{L}_{\text{CE}} = \arg\min_{\boldsymbol{P_t}} \mathbb{E}_{(\boldsymbol{X},y)\sim\mathcal{D}}\, \mathcal{L}(\text{sim}(\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{g}}_{\boldsymbol{p}}), y). \quad (1)$$

Prompt Ensembling with LLM Captions. Several methods adapts CLIP via training-free prompt ensembling techniques for textual-prompt enhancement. These methods use LLMs to obtain diverse captions, attributes, or high-level concepts of class names. The corresponding text features are either averaged (Pratt et al. 2023) or the similarity score of each attribute with the image is calculated to obtain classification scores (Menon and Vondrick 2023). In this work, we show comparisons with a strong ensembling baseline CuPL (Pratt et al. 2023). Specifically, a LLM $\mathcal{F}$ such as GPT-3 (Brown et al. 2020) is used to generate class-specific descriptions for class labels $\{1, 2, \cdots, C\}$ using queries such as *'How does a* CLASS *look like'*. CLIP text features of the same class description are averaged together, which serves as the ensembled text features. Finally, zero-shot inference is performed with those ensembled text features.

Prompt Learning with Text-Only Supervision

While LLM-based prompt ensembling methods are effective in adapting CLIP, they face notable challenges as outlined:

LLM Prompts transferability limitation. LLM-based prompt ensembling approaches like CuPL (Pratt et al. 2023) generate class-specific LLM descriptions for inference. We note that a caption generated for class A (seen class) cannot be directly utilized for class B (unseen class). While open-source LLMs exhibit lower performance, proprietary ones such as GPT-3 are required for generating data for new classes and datasets leading to additional serving costs.

Our work aims to address the aforementioned limitations within a unified framework. Below we discuss our strategy for curating text-to-text data using LLMs for training, followed by our text-only prompt learning framework.

Text-Only LLM Data for Prompt Learning As discussed earlier, learning prompts for downstream datasets typically require image-labels pairs. Orthogonal to this, we study a text-only approach and leverage LLMs to curate text data for prompt learning which consists of text inputs and text outputs. Given a set of classes $\{c_i\}_{i=1}^{C}$, we prepare text inputs $\{L_{\text{inputs}}^i\}_{i=1}^{C}$ by wrapping each class name in a standard hand-written text template,

$$L_{\text{inputs}}^i = \text{'a photo of a } c_i\text{'}.$$

Next, text outputs corresponding to the L_{inputs} are prepared. Specifically, we query GPT-3 to generate detailed descriptions for each class name c_i. Similar to CuPL (Pratt

et al. 2023), we prompt GPT-3 with different queries Q conditioned on class names such as *'How does a c_i look like?'* and *'How can you identify a c_i?"* to obtain text outputs,

$$L^i_{\text{outputs}} = \mathcal{F}(Q|c_i).$$

Similar to (Pratt et al. 2023), we generate M text outputs per query Q and use N different queries, resulting in $M \times N$ text outputs per class category. We associate all L_{outputs} with the corresponding single L_{inputs} for each class c_i. As LLMs are pre-trained on internet-scale text corpora, they possess the capability of generating very diverse and high-quality descriptions and captions for different class categories which results in high-quality text outputs. Finally we combine L_{inputs} and L_{outputs} to create LLM based text-to-text data for text only prompt learning, $\mathcal{D}_{\text{PROMPT}} = \{L^i_{\text{inputs}}, L^i_{\text{outputs}}\}^{M\times N\times C}_{i=1}$. We refer the readers to the appendix section for additional details on the choice of LLM prompts and examples of $\mathcal{D}_{\text{PROMPT}}$.

Contextual Mapping with Prompt Learning We utilize the LLM text-to-text data $\mathcal{D}_{\text{PROMPT}}$ for learning generalized and transferable prompts using a contextual mapping strategy that effectively learns a mapping function that maps standard class name templates such as '`a photo of a` c_i' to the text feature generated from a LLM description which contains more information about the class c_i. In other words, contextual mapping allows learnable prompts to map L_{inputs} to L_{outputs} in the text feature space of CLIP. The mapping function is realized in the form of learnable prompts. We study alternatives for mapping functions in ablation studies.

For an i_{th} training sample from $\mathcal{D}_{\text{PROMPT}}$ consisting of a text-to-text pair $\{L_{\text{inputs}}, L_{\text{outputs}}\}_i$, we obtain prompted class-name feature $\tilde{\boldsymbol{g}}_p$ for L^i_{inputs} using learnable prompts and frozen LLM feature $\tilde{\boldsymbol{g}}$ for L^i_{outputs} without the prompt vectors within the pre-trained latent space of CLIP text encoder. We then impose a contextual mapping constraint between $\tilde{\boldsymbol{g}}_p$ and $\tilde{\boldsymbol{g}}$ text features as follows,

$$\mathcal{L}_{\text{mapping}} = \frac{1}{d}\sum_{i=1}^{d} ||\tilde{\boldsymbol{g}}_{\boldsymbol{p}} - \tilde{\boldsymbol{g}}||^2_2. \tag{2}$$

We utilize MSE loss objective to enforce contextual mapping from L^i_{inputs} to L^i_{outputs}. We study other choices of consistency objectives in our ablations ().

Motivation for $\mathcal{L}_{\text{mapping}}$: Contextual mapping objective aims to train a generalized mapping function using learnable prompts in the text embedding space of CLIP to enhance text features. These enriched features are aligned with the LLM descriptions (L^i_{outputs}) for a given class. When trained using all classes together, the mapping function generalizes well and becomes adaptable for new concepts, effectively enabling the transferability of class-specific LLM descriptions to unseen classes and datasets. Learned prompts are used with new classes (as a prefix), by applying the same mapping and enables its transfer. This protocol is analogous to pretraining a model on a source dataset for transfer learning. Consequently, this reduces the per-dataset overhead associated with LLM serving and prompt engineering.

Inference. Once text prompt vectors are learned in our framework in the text domain, they are utilized with CLIP for downstream visual domain inference using a standard zero-shot CLIP inference setup. As shown in Fig. 2 (right), the learned prompts $\boldsymbol{P_t}$ are concatenated with each given class name to generate prompted text features $\{\tilde{\boldsymbol{g}}_{\boldsymbol{p}}\}^C_{i=1}$. Subsequently, zero-shot inference is performed using the prompted text features along with the input image feature $\tilde{\boldsymbol{f}}$ to generate classification scores for test images.

Experiments

Evaluation settings

We perform evaluations in 4 benchmark settings. Prompt ensembling methods and ProText utilize text-only LLM data for adapting CLIP while image-supervised prompt learning methods use image-label pairs for training. Supervised text-only setting benchmarks are discussed in the appendix.

Baselines. Our experimental setup is text-only, where our main baselines are LLM-based prompt ensembling methods. We mainly compare with CuPL (best among prompt ensembling methods) which uses LLM descriptions. We additionally compare results with image-supervised methods (indirect competitors) for holistic benchmark comparison.

Base-to-Novel Generalization. We evaluate the generalization of methods within a dataset. Following previous methods (Zhou et al. 2022a), we split each dataset into base and novel classes. Models are trained on base classes and evaluated on the test set of base and novel classes. ProText and CuPL use LLM text data of base classes in training.

Cross-dataset transfer. We evaluate the generalization ability of models trained on ImageNet-1k (Deng et al. 2009) source dataset by directly transferring it on cross-datasets.

Domain Generalization This setting evaluates the robustness of methods on out-of-distribution datasets. We train models on the ImageNet-1k source dataset and evaluate its performance on four ImageNet variants with domain shifts.

Datasets. For the aforementioned benchmarks, we use same datasets as followed by previous works (Zhou et al. 2022b,a; Khattak et al. 2023a). For cross-dataset transfer, domain generalization, and base-to-novel generalization, we use 11 datasets that cover multiple recognition tasks. These includes ImageNet (Deng et al. 2009) and Caltech101 (Fei-Fei, Fergus, and Perona 2004) which contains generic objects; OxfordPets (Parkhi et al. 2012), StanfordCars (Krause et al. 2013), Flowers102 (Nilsback and Zisserman 2008), Food101 (Bossard, Guillaumin, and Gool 2014), and FGVCAircraft (Maji et al. 2013) for fine-grained classification, SUN397 (Xiao et al. 2010) for scene recognition, UCF101 (Soomro, Zamir, and Shah 2012) for action recognition, DTD (Cimpoi et al. 2014) for texture classification, and EuroSAT (Helber et al. 2019) for satellite images. For domain generalization, we train models on ImageNet (Deng et al. 2009) as a source dataset and use ImageNet-A (Hendrycks et al. 2021b), ImageNet-R (Hendrycks et al. 2021a), ImageNet-S (Wang et al. 2019) and ImageNetV2 (Recht et al. 2019) for OOD evaluation.

Implementation details. We use pretrained ViT-B/16 CLIP model from OpenAI (Radford et al. 2021). ProText is with Deep Language Prompting in the first 9 transformer blocks of the CLIP text encoder. For cross-dataset transfer and do-

Method	ImageNet Acc.
1: CLIP (ICML'21)	66.72
2: CLIP-Attribute	67.60
3: CLIP-80	68.32
4: DCLIP (ICLR'23)	68.03
5: Waffle CLIP (ICCV'23)	68.34
6: CuPL (ICCV'23)	69.62
7: ProText-Attribute	68.05
8: ProText-80	68.48
9: ProText-CuPL	**70.22**

Table 2: Using same text data, learning contextual prompts with text-only supervision improves CLIP performance in comparison to the prompt ensembling techniques.

main generalization setting, we train ProText using $T = 4$ and $T = 16$ language prompts with 10 and 200 epochs respectively. Similar to (Wang et al. 2019), ProText and zero-shot CLIP use additional concepts where available with its prompts such as '`a photo of a CLS, a type of flower`' for OxfordFlowers. For base-to-novel and supervised text-only settings, ProText uses optimal prompt length and epoch configuration for each dataset. Optimal training configuration is obtained through hyper-parameter search on validation split of datasets. We use CuPL LLM text-data and generate descriptions for datasets not provided by CuPL using GPT-3 DaVinci-002 model. AdamW optimizer is used with 5 warm-up epochs for training using a 16-GB V100 GPU.

Effectiveness of Text-Only Supervision

We first present an ablation to motivate our approach of learning prompts with text-only supervision. We train ProText with 3 types of text data and evaluate on ImageNet. ProText-Attribute uses 46 templates from (An et al. 2023) containing common image attributes such as rotation, blurriness, etc. ProText-80 is trained on standard 80 templates provided by CLIP (Radford et al. 2021) and ProText-CuPL is trained on class-specific LLM data employed by our main baseline CuPL (Pratt et al. 2023) for its ensembling approach. In Tab. 2, we compare ProText with CLIP and LLM-based ensembling methods. Prompt ensembling with attribute templates and 80 templates improves over CLIP single template result. Among the LLM-based ensembling methods, CuPL provide highest result of 69.62%. In contrast, ProText uses a learning-based approach and shows competitive performance using same text data. ProText-Attribute provides gain of 0.45% over CLIP-Attribute while roughly maintaining its performance against CLIP-80. When equipped with CuPL LLM text-data, ProText surpasses CuPL by 0.60% leading to highest performance against all methods. These results motivate that instead of prompt ensembling, one can achieve competitive results by using the same available text data to learn prompts.

Base to novel class generalization

We now compare results in base-to-novel generalization setting where training data for only base classes are available. For CuPL (Pratt et al. 2023), we use base-class LLM templates for base classes and zero-shot CLIP results for novel classes. ProText uses base-class LLM templates for training and transfer the learned prompts for new classes at inference. Results are shown in Tab. 3. CuPL outperforms zero-shot CLIP on base classes while maintaining its performance on novel classes as LLM prompts for new classes are not available. ProText shows consistent improvements over CuPL on base classes for 11 datasets. Furthermore, with the same LLM base-class data as CuPL, ProText effectively transfers learned prompts towards novel classes and improves CLIP and CuPL novel class performance by 2.76% averaged across 11 datasets. This shows the advantage of ProText prompts to benefit unseen class performance potentially reducing the LLM prompt serving cost by half.

Dataset	CLIP			CuPL			ProText (Ours)		
	B	N	HM	B	N	HM	B	N	HM
ImageNet	72.4	68.1	70.2	74.3	68.1	71.1	**75.0**	**71.4**	**73.1**
Caltech101	96.8	94.0	95.4	97.2	94.0	95.6	**98.1**	**95.6**	**96.8**
OxfordPets	91.2	97.3	94.1	94.4	97.3	95.8	**94.9**	**98.0**	**96.5**
StanfordCars	63.4	74.9	68.7	63.6	74.9	68.8	**64.5**	**76.1**	**69.8**
Flowers102	72.1	77.8	74.8	**74.4**	77.8	76.1	**74.4**	**78.4**	**76.4**
Food101	90.1	91.2	90.7	89.9	91.2	90.6	**90.2**	**91.9**	**91.1**
Aircraft	27.2	**36.3**	31.1	30.6	**36.3**	**33.2**	**30.9**	34.1	32.4
SUN397	69.4	75.4	72.2	76.0	75.4	75.7	**76.2**	**79.1**	**77.6**
DTD	53.2	59.9	56.4	62.9	59.9	61.3	**63.1**	**61.6**	**62.3**
EuroSAT	56.5	64.1	60.0	59.6	64.1	61.8	**59.7**	**80.9**	**68.7**
UCF101	70.5	77.5	73.9	75.3	77.5	76.4	**75.5**	**79.5**	**77.5**
Average	69.3	74.2	71.7	72.6	74.2	73.4	**72.9**	**76.9**	**74.9**

Table 3: **Base-to-novel setting.** ProText enables the transferability of prompts to new classes and improves over CuPL.

Cross-dataset transfer

In cross-dataset transfer setting, we compare ProText with CLIP (Radford et al. 2021), CuPL (Pratt et al. 2023), and image-supervised prompt learning methods. Since class-specific ImageNet LLM prompts limit its transfer to other datasets in CuPL, we assign CLIP results to CuPL for cross-datasets. Image-supervised methods (Zhou et al. 2022a,b) are trained with 16-shot ImageNet data. We show our results in Tab. 4. CuPL improves ImageNet performance of CLIP by ensembling ImageNet LLM captions, while its cross-dataset results remain same as CLIP. In contrast, ProText addresses the transferability challenges of CuPL using prompts trained with the same ImageNet LLM data. The learned prompts are directly used with CLIP for cross-datasets leading to absolute average gains of +2.1% against CLIP and CuPL. With ProText, one can notably reduce proprietary LLM serving and prompt engineering costs as prompts learned on one dataset are effectively transferable to other datasets. For holistic comparison, we compare ProText with 16-shot image-supervised methods. Without using any visual samples, ProText shows effective generalization on cross-datasets with the highest average accuracy of 67.23%.

Domain generalization experiments

We present the results for domain generalization task in Table 5. As the domain shift variants of ImageNet share class names with ImageNet, CuPL employs prompt ensembling for each dataset and provides an average gain of +2.84% over CLIP. In contrast, ProText with learned prompts shows an additional gain of +0.44% against CuPL averaged over 4

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
Methods utilizing labeled visual samples												
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
TCP	71.40	93.97	91.25	64.69	71.21	86.69	23.45	67.15	44.35	51.45	68.73	66.29
Zero-shot & Prompt ensembling methods												
CLIP	66.72	92.98	89.13	65.29	71.30	86.11	**24.90**	62.59	44.56	47.84	66.83	65.15
CuPL	69.62	92.98	89.13	65.29	71.30	86.11	**24.90**	62.59	44.56	47.84	66.83	65.15
Prompt learning with text-only supervision												
ProText (Ours)	**69.80**	**94.81**	**91.01**	**66.00**	**72.35**	**86.66**	24.72	**67.34**	**47.93**	**51.86**	**69.60**	**67.23**

Table 4: **Cross-dataset transfer**. CuPL and CLIP perform same for cross-datasets as CuPL source data cannot transfer to cross-datasets. Image-based models are trained on 16-shot ImageNet samples. ProText uses same ImageNet text data as CuPL.

	Source	Target				
	ImageNet	-V2	-S	-A	-R	Avg.
Methods utilizing labeled visual samples						
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	50.90	76.98	60.27
TCP	71.20	64.60	49.50	51.20	76.73	60.51
Zero-shot & Prompt ensembling methods						
CLIP	66.72	60.83	46.15	47.77	73.96	57.18
CuPL	69.62	63.27	49.02	50.72	77.05	60.01
Prompt learning with text-only supervision						
ProText (Ours)	**70.22**	**63.54**	**49.45**	**51.47**	**77.35**	**60.45**

Table 5: **Domain generalization.** Prompt learning methods are trained on ImageNet and evaluated on OOD datasets.

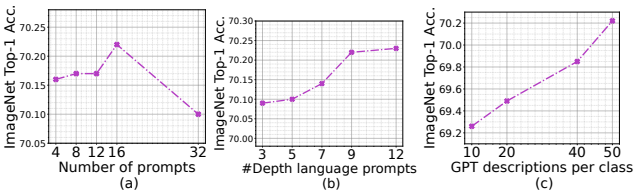

Figure 3: **Ablation studies: (a)** Prompt length. **(b)** Prompt depth. **(c)** LLM data size vs. performance.

Method	ImageNet Top1
1: ProText-80 templates	68.48
2: ProText-Alpaca	67.10
3: ProText-GPT-3	**70.22**

Table 6: Ablation on different text data for training.

Method	ImageNet Top1
1: ProText-contrastive loss	68.12
2: ProText- $L1$ loss	69.96
3: ProText-MSE loss	**70.22**

Table 7: Ablation: loss for contextual mapping.

datasets. Moreover, ProText fairs competitively with image-supervised methods by showing consistent improvements over CoOp, CoCoOp, and MaPLe. ProText aims to serve as an effective alternative to improve the robustness of VLMs when no visual information is available for training.

Ablative analysis

Choice of LLM for generating text data. Consistent with CuPL (Pratt et al. 2023), ProText by default uses GPT-3 (Brown et al. 2020) descriptions. Here we ablate on a recent open-source Mixtral-8x7B LLM (Jiang et al. 2024) in Tab. 6. ProText improves over CuPL-Mixtral by 4.38% and fares competitively to ProText-GPT. Stronger future open LMMs will further narrow the gap with GPT. **Loss metric in contextual mapping.** We ablate on choice of loss used for the contextual mapping module in Tab. 7. Distance-based losses improve over contrastive loss. For contrastive loss, batch-wise similarity is maximized between $\{L_{\text{inputs}}, L_{\text{outputs}}\}$ (positive) while it is minimized with unpaired samples (negatives). This causes multiple instances of the same class to be treated as negatives leading to lower performance. **Contextual mapping network.** While ProText employs prompt learning to learn contextual mapping from LLM descriptions, here we ablate on other choices. MLP adapters (Gao et al. 2023) at the output of CLIP encoders perform relatively lower having IN-Acc of 69.36%. ProText using LoRA (Hu et al. 2021) provides 70.19% IN-Acc., suggesting ProText compatibility with recent parameter efficient finetuning methods. **Prompt length and prompt depth.** Fig. 3 (a) shows the effect of prompt length for training ProText. Setting prompt length to 16 leads to optimal performance. Fig. 3 (b) shows the effect of prompt depth on final performance where prompt depth of 9 shows optimal results. **Training data size for text-supervision.** To assess the effect of LLM template data size on ProText, we ablate on the number of descriptions per class in Fig. 3 (c). Increasing descriptions per each class improves the results suggesting a positive correlation between performance and quantity of training data.

Conclusion

Prompt learning and LLM-based ensembling are effective techniques to improve CLIP's generalization. However, prompt learning often requires labeled images, while LLM ensembling methods are dominantly class-specific and not directly transferable to new classes. In this work, we explore a new direction to adapt CLIP by learning generalized prompts with text-only supervision, without using visual data. We formulate a training strategy for prompts to learn a mapping function that distills contextual knowledge from LLM text data within the prompts. The context learned by these prompts transfers well to unseen classes and datasets, potentially reducing the LLM prompt engineering and serving cost. We perform evaluations on four benchmarks where our text-only approach performs competitive to previous methods, including those utilizing labeled images.

References

An, B.; Zhu, S.; Panaitescu-Liess, M.-A.; Mummadi, C. K.; and Huang, F. 2023. More Context, Less Distraction: Improving Zero-Shot Inference of CLIP by Inferring and Describing Spurious Features. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*.

Bangalath, H.; Maaz, M.; Khattak, M. U.; Khan, S. H.; and Shahbaz Khan, F. 2022. Bridging the gap between object and image-level representations for open-vocabulary detection. *NeurIPS*, 35: 33781–33794.

Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101–mining discriminative components with random forests. In *ECCV*, 446–461. Springer.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *NeurIPS*, 33: 1877–1901.

Chen, G.; Yao, W.; Song, X.; Li, X.; Rao, Y.; and Zhang, K. 2022. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. In *ICLR*.

Cho, J.; Nam, G.; Kim, S.; Yang, H.; and Kwak, S. 2023. Promptstyler: Prompt-driven style generation for source-free domain generalization. In *ICCV*, 15702–15712.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *CVPR*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255. Ieee.

Derakhshani, M. M.; Sanchez, E.; Bulat, A.; da Costa, V. G. T.; Snoek, C. G.; Tzimiropoulos, G.; and Martinez, B. 2023. Bayesian prompt learning for image-language model generalization. In *CVPR*, 15237–15246.

Du, Y.; Wei, F.; Zhang, Z.; Shi, M.; Gao, Y.; and Li, G. 2022. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 14084–14093.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, 178–178. IEEE.

Feng, C.; Zhong, Y.; Jie, Z.; Chu, X.; Ren, H.; Wei, X.; Xie, W.; and Ma, L. 2022. PromptDet: Towards Open-vocabulary Detection using Uncurated Images. In *ECCV*.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2023. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 1–15.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 540–557. Springer.

Gu, S.; Clark, C.; and Kembhavi, A. 2023. I Can't Believe There's No Images! Learning Visual Tasks Using only Language Supervision. In *ICCV*, 2672–2683.

Helber, P.; Bischke, B.; Dengel, A.; and Borth, D. 2019. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *J-STARS*, 12(7): 2217–2226.

Hendrycks, D.; Basart, S.; Mu, N.; Kadavath, S.; Wang, F.; Dorundo, E.; Desai, R.; Zhu, T.; Parajuli, S.; Guo, M.; et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 8340–8349.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021b. Natural adversarial examples. In *CVPR*, 15262–15271.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, T.; Chu, J.; and Wei, F. 2022. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICLR*, 4904–4916. PMLR.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jin, W.; Cheng, Y.; Shen, Y.; Chen, W.; and Ren, X. 2021. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023a. Maple: Multi-modal prompt learning. In *CVPR*, 19113–19122.

Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M.-H.; and Khan, F. S. 2023b. Self-regulating Prompts: Foundational Model Adaptation without Forgetting. In *ICCV*, 15190–15200.

Kim, K.; Laskin, M.; Mordatch, I.; and Pathak, D. 2022. How to Adapt Your Large-Scale Vision-and-Language Model.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *ICCV*, 554–561.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven Semantic Segmentation.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 7061–7070.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt distribution learning. In *CVPR*, 5206–5215.

Lüddecke, T.; and Ecker, A. 2022. Image Segmentation Using Text and Image Prompts. In *CVPR*, 7086–7096.

Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Menon, S.; and Vondrick, C. 2023. Visual Classification via Description from Large Language Models. In *ICLR*.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. 2022. Simple open-vocabulary object detection. In *ECCV*, 728–755. Springer.

Naeem, M. F.; Khan, M. G. Z. A.; Xian, Y.; Afzal, M. Z.; Stricker, D.; Van Gool, L.; and Tombari, F. 2023a. I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Zero-Shot Image Classification. In *CVPR*.

Naeem, M. F.; Xian, Y.; Gool, L. V.; and Tombari, F. 2022. I2dformer: Learning image to document attention for zero-shot image classification. *NeurIPS*.

Naeem, M. F.; Xian, Y.; Zhai, X.; Hoyer, L.; Van Gool, L.; and Tombari, F. 2023b. SILC: Improving Vision Language Pretraining with Self-Distillation. *arXiv preprint arXiv:2310.13355*.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *ICVGIP*, 722–729. IEEE.

Nukrai, D.; Mokady, R.; and Globerson, A. 2022. Text-Only Training for Image Captioning using Noise-Injected CLIP. In *NeurIPS*, 4055–4063.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *CVPR*, 3498–3505. IEEE.

Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 15691–15701.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 5389–5400. PMLR.

Roth, K.; Kim, J. M.; Koepke, A.; Vinyals, O.; Schmid, C.; and Akata, Z. 2023. Waffling around for Performance: Visual Classification with Random Words and Broad Concepts.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35: 14274–14289.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Wang, H.; Ge, S.; Lipton, Z.; and Xing, E. P. 2019. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, volume 32.

Xiao, J.; Hays, J.; Ehinger, K. A.; Oliva, A.; and Torralba, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 3485–3492. IEEE.

Yao, H.; Zhang, R.; and Xu, C. 2024. TCP: Textual-based Class-aware Prompt tuning for Visual-Language Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23438–23448.

Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2021. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *ICLR*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Zhang, Y.; HaoChen, J. Z.; Huang, S.-C.; Wang, K.-C.; Zou, J.; and Yeung, S. 2023. Diagnosing and rectifying vision models using language.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *IJCV*, 130(9): 2337–2348.

Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; and Misra, I. 2022c. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 350–368. Springer.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2023. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15659–15669.