A Retrospect to Multi-prompt Learning across Vision and Language

Ziliang Chen[1,3], Xin Huang[2], Quanlong Guan[1]*, Liang Lin[2] Weiqi Luo[1]

[1]Jinan University [2]Sun Yat-sen University [3]Pazhou Laboratory

c.ziliang@yahoo.com, huangx353@mail2.sysu.edu.cn, {Gql,lwq}@jnu.edu.cn, linliang@ieee.org

Abstract

The vision community is undergoing the unprecedented progress with the emergence of Vision-Language Pretraining Models (VLMs). Prompt learning plays as the holy grail of accessing VLMs since it enables their fast adaptation to downstream tasks with limited resources. Whereas existing researches milling around single-prompt paradigms, rarely investigate the technical potential behind their multi-prompt learning counterparts. This paper aims to provide a principled retrospect for vision-language multi-prompt learning. We extend the recent constant modality gap phenomenon to learnable prompts and then, justify the superiority of vision-language transfer with multi-prompt augmentation, empirically and theoretically. In terms of this observation, we propose an Energy-based Multi-prompt Learning (EMPL) to generate multiple prompt embeddings by drawing instances from an energy-based distribution, which is implicitly defined by VLMs. So our EMPL is not only parameter-efficient but also rigorously lead to the balance between in-domain and out-of-domain open-vocabulary generalization. Comprehensive experiments have been conducted to justify our claims and the excellence of EMPL.

1. Introduction

Recent years have witnessed the rise of multimodal intelligence, in particular, Vision-Language Pre-training models (VLMs), *e.g.*, CLIP [40], ALIGN [24], achieving downstream tasks in low resources by converting the prior knowledge behind large language models (LLMs) [10, 4]. Given a pair of image encoder (*e.g.*, ResNet [19], ViT [11], *etc*) and text encoder (*i.e.*, LLMs), VLMs align visual features with their corresponding textual description embeddings via contrastive learning [12, 18, 56]. So provided a text known as *prompt*, VLMs may rapidly adapt to diverse tasks [12, 32] by matching harmony visual patterns with the textual description. The principle sheds a new light in computer vision for in-domain and out-of-domain generalization.

The impressive cross-modal transferability behind VLM

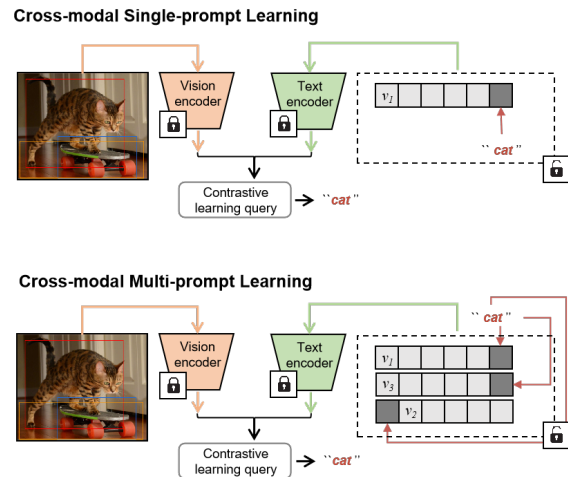*indicate corresponding author.

Figure 1. The overview of cross-modal single-prompt learning and multi-prompt learning (MPL). With more prompt templates, MPL brings new opportunities and challenges as discussed in the community, yet seldom giving a systematic investigation and solution.

typically owes to the problem-customized prompting style, yet demanding a great magnitude of trials and errors for selecting the ideal prompt template from a pool of candidates. Tedious workloads are consumed and do not guarantee the optimal prompt template either. Instead of the prompt engineering, *prompt learning / tuning* [55] sidesteps the obstacle using soft prompting: outside of the words related with what we are interested in, the rest of textual slots in the template are replaced by a sequence of learnable context vectors ahead of the text encoder. In this principle, the optimal prompt template could be achieved by fine-tuning the learnable context vectors along with the given textual semantic while keeping the rest parameters of VLMs frozen for the optimization. The data-driven merit increasingly arouses a flood of interests in the community [54, 42, 46].

Despite the significant progress, existing work of prompt learning focused on a single template whereas multi-prompt learnable context templates remain under-explored (Fig.1). Recent NLP advance argue that instructing LLMs via more prompts may trigger its underlying in-context learning ability to master new skills [4]. In this regard, multi-prompting

is also deemed to be a promising trend for VLMs. On the other hand, existing studies remain confused of how multiple prompts work for VLMs, particularly from two aspects. The first is ***vision-language transferrability*** [53]. Prompt augmentation eliminates the average cross-modal transfer shift while [54] showed that increasing the scale of context vector tokens resulted in the detrimental effect, implying a larger cross-modal disparity. The second is ***open-vocabulary (OV) generalization***, *i.e.*, the model awareness of unseen classes. Different from multple textual descriptions, learning with more prompts suggests more parameters. It casted a doubt of overfitting to training classes and put unseen classes at the risk of model generalization [5].

In this paper, we provided several principled insights to understand multi-prompt learning *empirically* and *theoretically*. We first consider the cross-modal embedding space following the constant modality-gap phenomenon found by [53]. With regards to our empirical observations, we extend the conclusion to learnable prompts to show that more learnable prompts might reduce the constant modality gap more significantly. In terms of constant modality gaps, we further proved the existence of ***cross-modal unidentifiability issue***: a paradox confusing the cross-modal model with a single prompt template in visual recognition. It could be restrained by multi-prompting empirically, thus, interpreting why multi-prompt learning could outperform single prompt for the sake of vision-language transferrability.

In terms of our retrospect, the main challenge of multi-prompt learning refers to its generalizability. Derived from this concern, we propose a new methodology Energy-based Multi-Prompt Learning (EMPL) for striking the balance between in-domain generalization and open-vocabulary generalization abilities. EMPL implicitly defines an energy-based [27] prompt distribution that simultaneously use image and prompt as the variable. With this regard, our method could be rigorously treated as modeling the uncertainty to explore the image-prompt embedding pairs with concepts out of the training domains, whereas also well generalizes to examples belonging to in-domain classes. The prompts are iteratively generated via a stochastic Markov Chain Monte Carlo (MCMC) sampler [50], which is parameter-efficient, sensitive of input knowledge from vision-text encoders, and more importantly, general enough to cooperate with existing prompt learning strategies to upgrade the performances. Experiments are comprehensively conducted to validate our claims and the superiority of our approach.

2. Related Work

Vision-Language Pre-trained (VLM) models. VLM models, which unify the two most commonly used modalities, vision and language, have gained great popularity due to the success of pre-trained models [4] in CV and NLP. Among numerous VLM models [40, 24], CLIP [40] is the most widely used and representative one. It utilizes a pair of image and text encoders to receive information from both modalities and leverages a large amount of paired image and text data collected from the Internet. In contrast to other VLM models that use Masked language modeling [25, 30], Masked region prediction [47, 45], etc., CLIP utilizes feature vectors from both encoders to train with the Contrastive Learning [40] strategy, which successfully aligns the feature space of both modalities and has been widely employed for a variety of downstream tasks [17, 15, 34].

Prompt learning. Prompt tuning [38, 41], a technique derived from the field of natural language processing (NLP), has gained great popularity [55, 54, 31] in the field of VLP in recent years, which has the ability to unleash the potential of pre-trained multimodal models. CoOp [55], a well-known text branching technique, eliminates the need for manual prompt design by transforming input context tokens into learnable vectors. CoCoOp [54], its successor, overcomes its generalization issues by taking visual features into account when creating prompts. Bahng et al.[2] propose a visual prompt approach by adding task-specific, learnable visual signals into images. Additionally, prompt learning has been employed to equip VLP models with the ability to tackle a variety of tasks, including open-vocabulary object detection [12], semantic segmentation [32, 42], and scene graph generation [20].

Multi-prompt learning. More recently, the advantages of building multiple context templates for prompt learning have been empirically verified. For instance, [31, 9] provided a distributional point of view to model the learnable template, in which the diversity across templates were emphasized; [5] trained multiple prompts by decreasing the cross-modality optimal transport [39, 6] across the prompt embeddings and visual embeddings to match different visual aspects by different templates; [14] learns different prompts to specify domain information, *etc*. These work demonstrate the promising outlook of multi-prompt learning for VLMs, whereas their solutions are typically heurstic and specific, limited to inspire the research in this thread.

3. Background

Here we give a brief review of cross-modal prompt engineering and learning, then generalize the notations to multi-prompt strategies prepared for the elaboration of our work.

Contrastive Language-Image Pre-training. CLIP consists of a pair of visual encoder f and text encoder h, which take ResNet / ViT and BERT [10] as their backbones. Given an image $\boldsymbol{x}$ with its label c contained in the description $\boldsymbol{y}(c)$, CLIP extracts a feature $f = f(\boldsymbol{x})$ by the visual encoder, then taking the text encoder $h(\cdot)$ to align f with the text embedding $\boldsymbol{h}_c = h(\boldsymbol{y}(c))$ generated from $\boldsymbol{y}(c)$, *e.g.*, "`a cropped photo of {c}`". f and h are trained with tons of image-caption pairs to bridge the modalities by contrastive repre-

sentation learning based on the prediction probability:

$$P(\texttt{class}=c|\boldsymbol{x}) = \frac{\exp\left(\text{sim}(f, \boldsymbol{h}_c)/\gamma\right)}{\sum_{i=1}^{K}\exp\left(\text{sim}(f, \boldsymbol{h}_{c_i})/\gamma\right)}, \quad (1)$$

where c is supposed to classify into the K classes $\{c_i\}_{i=1}^{K}$; $\text{sim}(\cdot, \cdot)$ denotes a metric function such as cosine similarity, and γ denotes the temperature of Softmax.

Prompt Learning. The template $\boldsymbol{y}(\cdot)$ is engineered before matching f and $\boldsymbol{h}_c$ by CLIP. Instead, CoOp [55] learns a set of vectors $\boldsymbol{v}=\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m\}$ to replace $\boldsymbol{y}(\cdot)$, each of which was understood as a pseudo word embedding and m denotes the length of the learnable word slots. So prompting becomes $\boldsymbol{h_v}(c)=\{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_m, \boldsymbol{v}(c)\}$ where $\boldsymbol{v}(c)$ denotes the embedding of the class name of c. Therefore given a pre-trained CLIP, CoOp tunes the prompt parameter $\boldsymbol{v}$ to achieve a downstream task with a frozen CLIP:

$$P_{\boldsymbol{v}}(\texttt{class}=y|\boldsymbol{x}) = \frac{\exp\left(\text{sim}(f, \boldsymbol{h_v}(c))/\gamma\right)}{\sum_{i=1}^{K}\exp\left(\text{sim}(f, \boldsymbol{h_v}(c_i))/\gamma\right)}, \quad (2)$$

in which $\boldsymbol{h_v}(c)$ can be specified to represent a wider range of prompt learning paradigms [54, 42].

Generic Notations of Multi-prompt Learning. Most existing work of multi-prompt engineering and learning are derived from Eq.1 and Eq.2, therefore we may extend their notations to generally represent the multi-prompt methods. In particular, we employ $\boldsymbol{H}(c; \mathcal{V})$ instead of $\boldsymbol{h}_c$ to denote a set of prompts composed of a vocabulary $\mathcal{V}$ and contains the word c. CLIP-derived prompt engineering approaches can be generally concluded into:

$$P(\boldsymbol{x})[c] = \frac{\exp\left(\text{sim}(f(\boldsymbol{x}), \boldsymbol{H}(c; \mathcal{V}))/\gamma\right)}{\sum_{i=1}^{K}\exp\left(\text{sim}(f(\boldsymbol{x}), \boldsymbol{H}(c_i; \mathcal{V}))/\gamma\right)}, \quad (3)$$

where $\text{sim}(\cdot, \cdot)$ denotes a generic metric to estimate the difference between the feature $f(\boldsymbol{x})$ and the multi-prompt embeddings $\boldsymbol{H}(c; \mathcal{V})$. If $\boldsymbol{H}(c; \mathcal{V})$ can be learned, we use $\phi=\{\boldsymbol{\theta}, \boldsymbol{v}\}$ to denote the context $\boldsymbol{v}$ and other learnable parameters $\boldsymbol{\theta}$, then rewrite Eq.3 by $\boldsymbol{H}_\phi(c; \mathcal{V})$:

$$P_\phi(\boldsymbol{x})[c] = \frac{\exp\left(\text{sim}(f(\boldsymbol{x}), \boldsymbol{H}_\phi(c; \mathcal{V}))/\gamma\right)}{\sum_{i=1}^{K}\exp\left(\text{sim}(f(\boldsymbol{x}), \boldsymbol{H}_\phi(c_i; \mathcal{V}))/\gamma\right)}. \quad (4)$$

In terms of task goals, the vocabulary $\mathcal{V}$ in Eq.4 is only interested in the words related with the task, which has already summed up a set of multi-prompt learning methods [31, 6].

4. Embedding Geometry behind Prompts

The core of CLIP-derived prompt learning hinges on the prompt template's ability to solve vision tasks with text embeddings as inquiry proxies. The cross-modal transferability is achieved for a pair of an image $\boldsymbol{x}$ and its description

$\boldsymbol{y}(c)$ if a vision classifier outputs similar predictions on their embeddings. While given a matched image-text pair, the embeddings extracted by CLIP counter-intuitively persisted a modality gap [53, 28]. Our work demonstrated this geometrical phenomenon also widely exist in prompt learning with virtual description $\boldsymbol{v}(c)$ and results in *cross-modal non-identifiability issues* in single-prompt learning. To this, the technical merit of multi-prompt strategies can be verified from the view of cross-modal transferrability.

Modality Gaps between Images and Prompts. As proposed in [28], the modality gap is caused by contrastive optimization and can be categorized into two types: the *individual*-level modality gap $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$ differentiates the embeddings for a image-text pair $(\boldsymbol{x}, \boldsymbol{y})$; the *class*-level modality gap $\boldsymbol{g}(c)$ differentiates the average between the embeddings for images and text related with the class c, namely,

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) &= f(\boldsymbol{x}) - h(\boldsymbol{y}), \quad \forall(\boldsymbol{x}, \boldsymbol{y}) \sim P_{\mathcal{X} \times \mathcal{Y}}; \\ \boldsymbol{g}(c) &= \mathbb{E}_{\boldsymbol{x} \sim P_{\mathcal{X}|c}} f(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{y} \sim P_{\mathcal{Y}|c}} h(\boldsymbol{y}). \end{aligned} \quad (5)$$

Observed across a range of contrastive multimodal models, *the modality gaps $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{g}_c$ could be approximated by a constant vector* [53]. These embedding geometrical properties provide a new explanation why prompt learning can outperform CLIP: learning to prompt might implicitly reduce the modality gap constant between image-text pairs.

To procure the evidences of our guessing, we investigate the embedding geometry derived from the prompt embeddings extracted from the single-prompt learner CoOp [55] and multi-prompt learner ProDA [31]. The means and variances of the magnitude ($|\boldsymbol{g}|$) and direction ($\cos(\boldsymbol{g}, \mathbb{E}_{\boldsymbol{g}}\boldsymbol{g})$) are estimated to justify whether the individual and class modality gaps can be approximated by a constant vector[1]. In Fig.2, CoOp and three ProDA variant models with different scale of prompts preserve the average gaps with trivial variances in their magnitudes and directions, thus, their modality gaps have been approximated by some constant vectors, respectively. On account of this observation, prompt learner differ in their modality gap magnitudes: CoOp trained from CLIP can further minimize the gap magnitude, however, underperforms ProDA with the prompt augmentation strategy: adding prompts notably leads to closing the modality gaps. Hence *the failure of the overextended scale of prompts more likely results from the overfitted model rather than the incompetence of bridging the modality disparity.*

Cross-modal Non-identifiability Issue. In terms of the modality gap, there is a issue that might happen if we use a single template to prompt a cross-modal contrastive model. Concretely, let's consider two images $\boldsymbol{x}_i, \boldsymbol{x}_j$ with mutually exclusive concepts in visual realism, *e.g.*, $\boldsymbol{x}_i$ belongs to c_1, c_2 and $\boldsymbol{x}_j$ belongs to c_2, c_3. Given a single prompt template $\boldsymbol{v}$ to convey these concepts, feature-embedding pairs

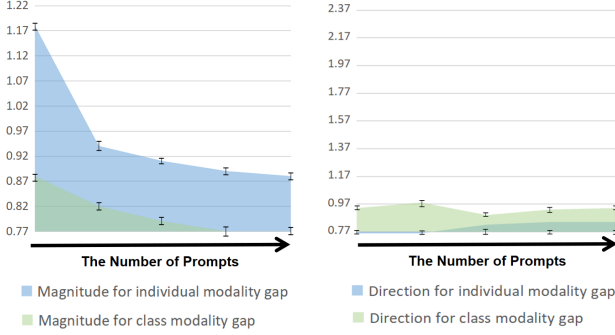---

[1]We follows the same evaluation setup in [53].

Figure 2. Magnitude (**M**) and Direction (**D**) of individual modality gap (IMG) and class modality gap (CMG) on MsCOCO [29]. We gradually increase the number of prompts by switching models as CLIP→CoOp→ProDA→ProDA(x2)→ProDA(x4), to observe the change of IMG and CMG.

$(f(\boldsymbol{x}_i), h_{\boldsymbol{v}}(c_1))$, $(f(\boldsymbol{x}_i), h_{\boldsymbol{v}}(c_2))$, and $(f(\boldsymbol{x}_j), h_{\boldsymbol{v}}(c_2))$ are supposed to be approximated by the individual-level modality gap constant vector $\boldsymbol{c}$. Given this, we can prove $f(\boldsymbol{x}_j)$ far away from $h_{\boldsymbol{v}}(c_1)$ with the same constant vector $\boldsymbol{c}$,

Proposition 1. *Individual-level cross-modal non-identifiability (Informal) Suppose a single-prompt learning model $\big(f(\cdot), h_{\boldsymbol{v}}(\cdot)\big)$ satisfies the constant individual-level modality gap. Given each pair of images $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ with mutually exclusive concepts, it is not able to distinguish them by single-prompting with their exclusive concepts.*

The formal statement and proof refer to our Appendix.A. Derived from the result, the image $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can not be distinguished in terms of the proxy $\boldsymbol{v}(c_1)$, which should have been distinguished since the concept c_1 is exclusive for the image $\boldsymbol{x}_i$ in terms of the image $\boldsymbol{x}_j$.

Here we discussed a simple case for illustrating the issue: suppose we having a pair of *horse* (mutual concept) images, where the first image refers to the scene of *a man* (exclusive concept) *riding a horse* and the second describes that *horses and cows* (exclusive concept) *drink nearby a river*. Prompting the images by the given captions is capable to differentiate the images in terms of their exlusive concepts *man* and *cow*. However, given a single prompt only built with learnable context vectors in $\boldsymbol{v}$, inquiring with *cow* or *man* is hard to classify these images if the context optimization satisfies the constant modality gap pressumption, in which the single template was optimized to overlook the other descriptive information beyond the classes (*e.g.*, *riding* and *nearby river*). The case is general since the images with mutually exclusive concepts may also refer to semantic information incorporated from the visual encoder [54].

What's worse, the class-level modality gap resembles the tragedy across different groups, arousing the chaos to identify embedding sets belonging to different concepts:

Proposition 2. *Population-level cross-modal non-*

identifiability (Informal) Suppose a single-prompt learning model $\big(f(\cdot), h_{\boldsymbol{v}}(\cdot)\big)$ satisfies the constant individual-level modality gap. Given image groups $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ have mutually exclusive concepts, it is not able to distinguish the groups via single-prompting with their group-specific concepts.

The existence of non-identifiability validates *the remarkably more effecting of multi-prompt learning strategy compared with single-prompt learning*: learning to prompt with multiple templates suggests remodeling the diversity of captions that could helpfully alleviate the issues. It is reflected by evaluating different prompt models on their prediction consistencies to images with mutually exclusive classes. Some evidences agreed with our conjecture in our empirical studies on Language-to-Image Retrieval and Multi-label classification (Appendix.C).

5. Energy-based Multi-prompt Learning

In the previous section, we discussed why multi-prompt learning benefits the cross-modal transferrability, though it does not reflect a model's generalization ability to adapt the cases beyond the prompt-tuning stage. The OV generalization is remarkable in CLIP yet it might rapidly deteriorate by prompt-tuning the backbone due to the traded-off performance on in-domain images. Subsequent techniques developed to resist the degeneration [55] were barely motivated by the multi-prompt learning charateristics. It is somehow because multi-prompt learning with more templates is supposed to bring more learnable context tokens, increasing the risk of overfitting. Our concern rises from this regard:

Is there a multi-prompt learning algorithm that simultaneously benefit the cross-modal transferrability (in-domain) and the OV generalization (out-domain)?

To kill the two birds with one stone, we reinterpret multi-prompt learning from a persepective of energy-based models (**EBMs**) [27], where mutiple prompt templates are reproduced by drawing instances from an underlying EBM-based prompt distribution. The paradigm of Energy-based Multi-Prompt Learning (**EMPL**) is briefly shown in Fig.3 and we further elaborate the methodology to demonstrate its potential to address our concern.

Energy-based Models. The formulation of EBMs consists of an energy function $E(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}$ that maps a D-dimensional datapoint into a scalar. It is implemented by a neural network with the parameter θ, then $E_\theta(\cdot)$ is trained to assign the low energy to observed configurations of variables and deliver the high energy to unobserved ones. So given a dataset without the knowledge of its underlying density $p(Z)$, EBMs enable $p_\theta^{(\text{EBM})}(Z) = \frac{\exp(-E_\theta(Z))}{\int_{\boldsymbol{z}\in\mathcal{Z}}\exp(-E_\theta(\boldsymbol{z}))}$ to approximate the density function. The variable Z in EBMs mostly refers to images or image features in previous work [16, 36, 49]. In contrast, our EMPL depends on an energy

function $E_\phi(Z) = E_\phi(X, H)$ with Z constructed by the *image variable* X and the *prompt variable* H. Therefore the multiple prompts in $\boldsymbol{H}_\phi$ (Eq.4) are obtained via drawing prompt instances from a EBM-based conditional prompt distribution $p_\phi^{(\text{EBM})}(H|X)$, *i.e.*,

$$\forall\, \boldsymbol{x} \sim p(X) = P_\mathcal{X},$$
$$\boldsymbol{H}_\phi \sim p_\phi^{(\text{EBM})}(H|\boldsymbol{x}) = \frac{\exp(-E_\phi(\boldsymbol{x}, H))}{\int_{H\sim\mathcal{H}} \exp(-E_\phi(\boldsymbol{x}, H))}, \quad (6)$$

in which $\mathcal{H}$ denotes the space of soft prompts and the energy function is derived from the contrastive score $P_\phi(\cdot)$ in Eq.4.

Energy-based Open-vocabulary Learning. Given an open vocabulary $\mathcal{V}$, we elaborate the meta-learning objective with the energy function $E_\phi(X, H)$ for improving in-domain and out-of-domain generalization. Instead of a K-word vocabulary in Eq.4, the open vocabulary $\mathcal{V}$ demands a *meta-classifier* of predicting arbitrary classes across visual recognition tasks. So EMPL is suggested to incorporate all the words in $\mathcal{V}$ and accordingly, $E_\phi(X, H)$ should support meta-learning to achieve a set of K-class visual recognition tasks. Each task is constructed by K' ($0 < K' < K$) observed classes and the other $K - K'$ classes refer to the unseen class names that have appeared in the open vocabulary $\mathcal{V}$. Hence given each task $\mathcal{T}_i$ with $K - K'$ observed classes $\mathcal{V}_i$ and unseen classes $\mathcal{U}_i$, we define the task-specific energy function $E_\phi(X, H; \mathcal{T}_i)$ to capture the out-of-domain uncertainty:

$$E_\phi(X, H; \mathcal{T}_i) = \log \sum_{c\sim U_i} P_\phi(X, H)[c]$$
$$= \log \sum_{c\sim\mathcal{U}_i} \frac{\exp\left(\frac{\text{sim}(f(X), H(c;\mathcal{V}_i\cup\mathcal{U}_i))}{\gamma}\right)}{\sum_{i=1}^K \exp\left(\frac{\text{sim}(f(X), H(c;\mathcal{V}_i\cup\mathcal{U}_i))}{\gamma}\right)}, \quad (7)$$

in which $\boldsymbol{x}$, $\boldsymbol{H}_\phi$ from Eq.4 is rewritten into X, H, respectively, for representing random variables in the energy function in the range of the training data distribution. $E_\phi(X, H; \mathcal{T}_i)$ derived from $E_\phi(X, H)$ is suggested to meta-learn unseen concepts in U_i across different tasks during training. Then, EMPL objective is defined as

$$\min_\phi \mathbb{E}_{\mathcal{T}_i}\left[\underbrace{\mathbb{E}_{p(X,H|\mathcal{T}_i)}\left[\sum_{c\sim V_i} -\log P_\phi(X, H)[c]\right]}_{\text{Generic prompt learning goal}}\right.$$
$$\left.\underbrace{-\lambda\mathbb{E}_{p_{\overline{\phi}}^{(\text{EBM})}(X,H|\mathcal{T}_i)}\left[E_\phi(X, H; \mathcal{T}_i)\right]}_{\text{EBM uncertainty modeling}}\right], \quad (8)$$

where $p(X, H|\mathcal{T}_i)$ denotes the image-prompt training pairs extracted with their labels for achieving the maximum log-likelihood prompt-tuning goal in terms of the task $\mathcal{T}_i$ (the first term); $p_{\overline{\phi}}^{(\text{EBM})}(X, H|\mathcal{T}_i)$ represents the image-prompt joint distribution derived from the task-specific energy function
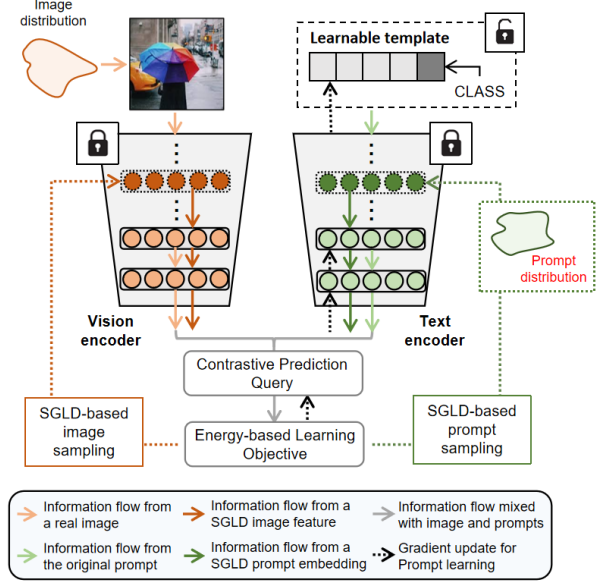

Figure 3. The paradigm of EMPL (best viewed in color). Briefly speaking, EMPL defines a prompt distribution based upon a EBM with the variables lying in the image feature and prompt embedding spaces. It categorizes an image with multiple prompts iteratively drawn from the EBM-based distribution by SGLD samplers.

$E_\phi(X, H; \mathcal{T}_i)$. It is noteworthy that, $\overline{\phi}$ denotes the parameter frozen of the current iteration, hence $p_{\overline{\phi}}^{(\text{EBM})}(X, H|\mathcal{T}_i)$ is only available of generating image-prompt training pairs to compute the expectation for the second term. The hyper-parameter λ balances the strengths between in-domain image recognition and open-vocabulary concept exploration.

The energy function $E_\phi(X, H; \mathcal{T}_i)$ interacts with prompt learning by the second term, encouraging the low energy in terms of the image-prompt pairs extracted from $p(X, H|\mathcal{T}_i)$, in turn, stay high if the $p_{\overline{\phi}}^{(\text{EBM})}(X, H|\mathcal{T}_i)$ deviates from the marginal image-prompt distribution $p(X, H|\mathcal{T}_i)$. The training process can be proved to endow the image-prompt pairs with a profound property:

Proposition 3. *Provided an arbitrary task $\mathcal{T}_i$ constructed to classify an image into either observed classes V_i with avaiable training images or unseen classes $\mathcal{U}_i$ without unavailable training images, if $p(X, H|\mathcal{T}_i)$ denotes the marginal distribution of image-prompt pairs extracted from training data distribution, and $p_\phi(X, H|\mathcal{U}_i) = \sum_{c\sim\mathcal{U}_i} P_\phi(X, H)[c]$ denotes the distribution of image-prompt contrastive prediction marginalized over all unseen classes in $\mathcal{U}_i$, the optimization objective of EMPL (Eq.8) encourages $p(X, H|\mathcal{T}_i)$ and $p_\phi(X, H|\mathcal{U}_i)$ negatively correlate with each other.*

The theoretical result derived from uncertainty modeling [49], enlightens us to understand the superiority of EMPL. In particular, when a well-trained EMPL model is provided with in-domain images, the energy-based prompt distribu-

tion is encouraged to assign a low contrasive score to any prompt without matching the images with correct classes since in-domain image-prompt pairs well match $p(X, H|\mathcal{T}_i)$ so that squeezes the value of $P_\phi(X, H)[c]$ for all c that falls within $\mathcal{U}_i$. With images drawn from the other domains or unseen classes, the image-prompt pairs are far from $p(X, H|\mathcal{T}_i)$, equivalently to increase $p_\phi(X, H|\mathcal{U}_i)$ for exploring the proper matching between the images and the unseen classes in U_i.

SGLD Sampling and Prompting. It is pivotally important of sampling image-prompt pairs from the energy-based distribution $p_\phi^{(\text{EBM})}(\cdot)$ since it does not only provide training instances for the second term in Eq.8 but also generate $\boldsymbol{H}_\phi$ to categorize $\boldsymbol{x}$ by the multi-prompting (Eq.6). We employ Stochasitc Gradient Langevin Dynamics (SGLD) [50] to alternatively execute the sampling process[2]:

$$
\begin{aligned}
\boldsymbol{x}^{t+1} &= \boldsymbol{x}^t - \frac{\alpha}{2}\frac{\partial E_\phi(\boldsymbol{x}^t, \boldsymbol{h}^t)}{\partial \boldsymbol{x}^t} + \sqrt{\alpha}\epsilon_1, \qquad \epsilon_1 \sim \mathcal{N}(0; I), \\
\boldsymbol{h}^{t+1} &= \boldsymbol{h}^t - \frac{\alpha}{2}\frac{\partial E_\phi(\boldsymbol{x}^{t+1}, \boldsymbol{h}^t)}{\partial \boldsymbol{h}^t} + \sqrt{\alpha}\epsilon_2, \quad \epsilon_2 \sim \mathcal{N}(0; I),
\end{aligned}
\tag{9}
$$

where t and α denote the iteration and the step-size in the stochastic process; ϵ_1 and ϵ_2 are random noises drawn from a Gaussian distribution, respectively. The sampling process run in the feature space of $\boldsymbol{x}^{t+1}$ and the embedding space of $\boldsymbol{h}^{t+1}$. It significantly reduces the computational burden.

Comparison with Other Prompt-Distribution Methods. Although previous multi-prompt learning efforts [31, 1] have treated prompts as instances drawn from a distribution, our EMPL are predominant from some aspects. Concretely, the previous methods draw prompts from a specific type of density function or have to maintain a predefined context vector collection. Instead, EMPL defines the prompt distribution via a EBM derived from multi-prompt learning objective, where multiple prompts are dynamically drawn via executing a SGLD process, requiring little extra parameters beyond the base contexts. It prevents multi-prompt learning from the higher risk of overfitting. Besides, the EBM-based prompt distribution typically generates dynamic prompts conditioned on the visual feature. It resembles the spirit of CoCoOp distinct from other works.

6. Experiments

In this section, we conduct comprehensive experiments to evaluate EMPL with diverse prompt learning approaches across three tasks, *e.g.*, *base-to-new generalization*, *cross-domain generalization*, and *cross-dataset transfer learning*. It provides the answer of our previous concern.

Table 1. Comparison of single-prompt and multi-prompt learning baselines in the base-to-new generalization setting. H (Harmonic mean [51]) measures the generalization trade-off. Different background colors indicate the corresponding group of abalating EMPL for CoOp and ProDA.

	Single-prompt			Multi-prompt learning			
	CLIP	CoOp	CoCoOp	ProDA	PLOT*	CoOp (+EMPL)	ProDA (+EMPL)
Base	69.34	82.66	80.47	81.56	75.90	**82.73** (+0.07)	82 (+0.44)
New	74.22	63.22	71.69	72.29	67.6	70.93 (+7.71)	**73.27** (+0.98)
H	71.69	71.65	75.83	76.65	71.8	76.38 (+4.73)	**77.39** (+0.74)

6.1. Experimental Setup

Benchmarks. The three tasks with fifteen datasets evaluate cross-modal prompt learners from different aspects. In terms of the base-to-new generalization and cross-dataset transfer setups, it takes ImageNet [8], Caltech101 [13] for normal object recognition; SUN397 [52] for scene recognition; UCF101 [44] for action recognition; DTD [7] for texture classification; EuroSAT [21] for satellite image classification; and OxfordPets [37], StanfordCars [33], Flowers102 [35], Food101 [3], FGVCAircraft [26] for fine-grained image recognition derived from diverse scenarios. For the domain generalization, it trains models on ImageNet, then report the evaluation on ImageNetV2 [43], ImageNet-Sketch [48], ImageNet-A [23], and ImageNet-R [22]. The evaluation metric refer to the average accuracies and we additionally report the Harmonic mean [51] for base-to-new generalization, which is broadly regarded to judge the traded-off performances between the base and new classes.

Baselines. Beyond our methodology and CLIP [40], we consider CoOp [55], CoCoOp [54], ProDA [31], and PLOT [5] for our comparison. Their backbones are typically derived from the open-source CLIP[3], in which CoOp, CoCoOp are supposed to be the single-prompt learning baselines and ProDA, PLOT denote the multi-prompt learning baselines (We take PLOT* instead of PLOT to indicates its backbone distinct from other baselines). Note that EMPL is orthogonal to most existing prompt-tuning based methods and can be deployed to improve their strategies by energy-based multi-prompting. In this regard, our implementation for EMPL are derived from CoOp and ProDA[4], whereas their original objectives have been considered as the first term of Eq.8 and their prompt generations are substituted by our SGLD-based sampler rather than their primitive strategies. We use EMPL(+CoOp) and EMPL(+ProDA) to represent their marriages, respectively.

Implementation. EMPL(+CoOp) and EMPL(+ProDA)

---

[2]In terms of the meta-learning formulation in Eq.8, $E_\phi(\cdot, \cdot)$ is replaced by $E_\phi(\cdot, \cdot; \mathcal{T}_i)$ to denote the SGLD sampling process executed for the task.

[3]https://github.com/openai/CLIP

[4]EMPL might also suit CoCoOp and PLOT in the spirit whereas it is prohibitively implemented by their open-source versions due to the heavy memory consumption for training.
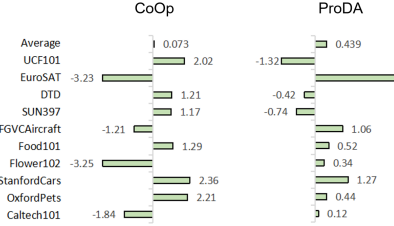
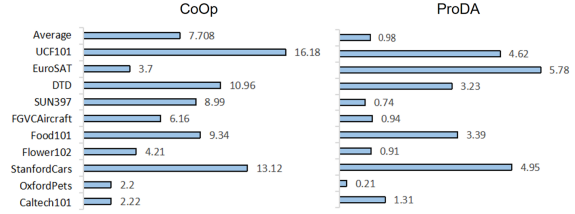Figure 4. The performance change in base classes in 11 datasets.

Figure 5. The performance change in new classes in 11 datasets.

Table 2. Comparison of single-prompt and multi-prompt learning baselines for cross-domain generalization.(best viewed in color)

	Source	Target			
	ImageNet	-V2	-Sketch	-A	-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp	**71.51**	64.20	47.99	49.71	75.21
CoOp(+EMPL)	70.89 ↓	64.91 ↑	48.64 ↑	51.27 ↑	76.01 ↑
CoCoOp	71.02	64.07	**48.75**	50.63	76.18
ProDA	71.41	65.14	46.78	51.62	75.67
ProDA(+EMPL)	71.17 ↓	64.79 ↓	48.42 ↑	**52.35** ↑	76.84 ↑

are implemented with the public codes of CoOp and ProDA, in which we reformulate their learning objectives by introducing an energy-based function derived from the learnable contexts to specify SGLD-based samplers. The details of SGLD-based sampler and open-vocabulary meta-learning strategy refer to our Appendix.

6.2. Base-to-new Generalization

Task setup. It requires the prompt learner trained for few-shot generalization on the 10 datasets with three different random seeds. Each dataset is divided into two disjoint subsets with base classes and new classes, where baseline models are trained with base classes via few-shot learning and evaluated on both base and new classes in the test dataset. For a fair comparison, we follow the dataset split and the number of shots in [54] during training.

Results. Due to the space limitation, we report the average base-class and new-class accuracis along with their Harmonic mean over all datasets in Table.1, then providing the performance ablation for each dataset (Fig.4,5). As demonstrated in Table.1, CoOp is a competitive rival in Base-class generalization with regard to its outperformance compared with other baselines beyond the range of our EBML. But its superiority remains a doubt of overfitting because its accuracy rapidly drops while coming to the unseen classes. The performance discrepancy could be greatly mitigated when CoOp takes the prompts generated by EMPL. The marriage leads to +7.71 performance gain on the new classes, driving the Harmonic mean to 76.38 that sufficiently defeats all single-prompt learning baselines. CoCoOp is famous as a complementary strategy to CoOp, while its pipeline suffers from the low inference efficiency due to its instance-specific prompt scheme demanding an independent forward pass for each prompt. With this regard, CoCoOp hardly becomes, or combines with a multi-prompt strategy and thus, inevitably fall behind all ProDA variants. Notice that, EMPL endows ProDA with visually-encoded information conducted by the SGLD-based prompting scheme (Eq.9). It results in the uppermost trade off in the base-to-new generalization .

We further ablate EMPL in the CoOp and ProDA across all datasets. As shown in Fig.4, EMPL benefits the majority of tasks (7 of 10 in CoOp and 6 of 10 in ProDA) with moderate margins whereas also produces unexpected negative

effects to CoOp and ProDA for the minority. It probably owes to the conservative tendency to the observed classes for open-vocarbulary meta-learning, in which the energy function aims to maintain the exploitation-exploration balance between in-domain classes and out-of-domain classes (Proposition.3). Notwithstanding, the base-class predominances of CoOp and ProDA go on without sacrificing the new-class generalizability in Fig.5.

6.3. Cross-domain Generalization

Task setup. Distinct from the base-to-new setup, cross-domain generalization attempts to examine the baselines in terms of their resiliences against domain shift and adversarial robustness: models trained in ImageNet are evaluated on its four target dataset variants for different purposes.

Results. As reported in Table.2, our EMPL enhances the cross-domain accuracis of CoOp and ProDA in seven out of eight situations with diverse type of distribution shifts. It fails in the cases with the source data for training (ImageNet for CoOp) or with a mild distribution shift (ImageNet-V2 for ProDA). In terms of the cases containing significant visual difference (ImageNet-Sketch), out-of-distribtuion shift (ImageNet-R) and natural adversarial noises (ImageNet-A), our EMPL consistently emerged victorious for the generalization across domains.

6.4. Cross-dataset Transfer

Task setup. We finally evaluate the baselines in the more challenging cross-dataset transfer setups, whose fundamentals are allowed to totally change across datasets (different tasks across different domains). In this case, the baseline prompt contexts are trained on ImageNet, then, required to access on the other target datasets with distinct knowledge.

Table 3. Comparison of single-prompt and multi-prompt learning baselines for cross-dataset transfer.(best viewed in color)

| | Source | Target | | | | | | | | | | |
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flower102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Avg
CoOp	**71.51**	93.7	89.14	64.51	68.71	85.35	18.47	64.15	41.92	46.39	66.55	63.88
CoOp(+EMPL)	70.89	94.16	90.21	65.29	71.52	86.21	23.16	67.13	46.93	**47.34**	68.07	66.49
CoCoOp	71.02	94.43	90.14	65.32	**71.88**	86.06	22.94	67.36	45.73	45.37	68.21	65.74
ProDA	71.41	**94.65**	90.22	64.81	70.69	85.57	22.23	**68.23**	43.33	45.78	67.86	65.89
ProDA(+EMPL)	71.17	94.63	**91.24**	**65.67**	71.76	**86.29**	**23.97**	67.98	**47.21**	46.87	**68.44**	**66.81**
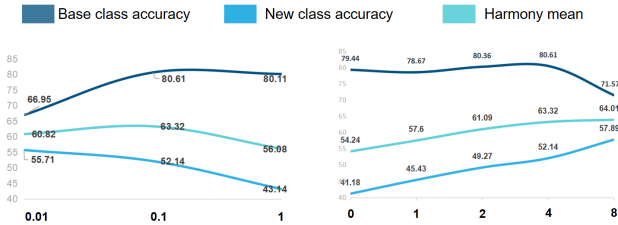
Figure 6. The trending curves when changing the value of λ and the number of new words for training (best viewed in color).

Figure 7. The statistic and trending curves when changing the number of prompt samples for training (best viewed in color).

Results. As demonstrated in Table.3, all baselines perform similarly in the source training set while behave differently across diverse target datasets, and in most cases, the multi-prompt learners outperform the single-prompt learners. In particular, EMPL variants outperform the other baselines in seven datasets and more importantly, they have significantly benefited their basic models with the accuracy increases in 18 of 20 transfer scenarios. More typically on the target datasets such as FGVCAircraft, SUN397, and DTD, EMPL raised their accuracies more than 3% and besides, it does not introudce any extra parameters to acheive this goal. On the other side, we also observe that EMPL results were not the state of the art (ImageNet, Caltech101, Flower102, and SUN397). The ImageNet case is similarly explained as what happened in Table.2-3, *i.e.*, the boost by EMPL is largely due to the new-class generalization. So testing EMPL on the source data may deemphasize this merit. As to Caltech101, Flower 102, and SUN397, it is observed that EMPL just slightly underperform the state-of-the-art models, *e.g.*, ProDA is 94.65 yet EMPL(+ProDA) is 94.63. As the number of promptgs was fixed to 8, their performances might be further improved by prompt augmentation.

6.5. Analysis

The number of prompts. The size of prompts used for each prompting inference sufficiently affects the final performance for arbitrary prompt distribution methods. So we provide the ablation for evaluating EMPL with the prompt number used for training. We evaluate EMPL variant model derived from CoOp based on DTD dataset, then, observe the change of the base-class, new-class accuracies, and their Harmonic means. As reported in Fig.7, the performance of
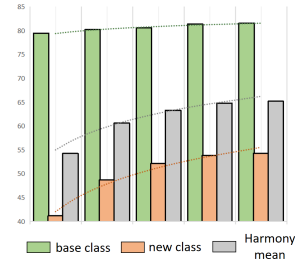
EMPL could be increased by drawing more prompt embeddings from the energy-based distribution.

Hyperparameters λ and $K-K'$. The open-vocabulary meta-learning objective (Eq.8) plays a key role of achieving the trade-off balance in base-to-new generalization. The implemenation is typically related with the hyperparameters λ and $K-K'$: the former determines the sensitivity to explore visual patterns with new-word prompts; the latter controls the ratio of how many new words would appear per training batch. We follow the ablation setup above and then change the value of λ in the range $\{0.01, 0.1, 1\}$, then, taking the same evaluation setup by changing the number of unseen words per training batch in the range $\{1,2,4,8\}$. As we have observed in Fig.6, λ typically trades off the in-domain and out-of-domain results where its larger value implies the objective with more attention to explore the uncertainty. It leads to the rise of new-class generalization, vice and versa. $K-K'$ is also related with the new-class generalization performance: increasing the number of new words leads to the improvement while it rapidly converges to the bottleneck.

Table 4. The trade-off between performance (the base-to-new generalization setup in DTD) and computation burden.

Transformer layers	Performance			Computation burden
	Base	New	H	Sec/iter
2	80.61	52.14	63.32	0.048
3	80.71	52.2	63.39	0.072
6	79.41	53.31	63.79	0.143
9	77.28	51.24	61.62	0.278

Positions for SGLD-based sampling. EMPL's training and prompting rely on SGLD that runs in the embedding

Table 5. Image-text retrieval results on MSCOCO and Flick30K.

	0%	0.5%			1%		
	CLIP	CoOp	CoCoOp	EMPL	CoOp	CoCoOp	EMPL
MSCOCO	53.35	53.10↓	54.50↑	**55.45**↑	53.58↑	56.40↑	**56.85**↑
Flickr30K	83.06	81.90↓	82.80↓	**83.94**↑	82.71↓	84.50↑	**85.63**↑

space. To justify the concern of backward sampling computation, we ablate EMPL (+CoOp) variants with the SGLD sampler applied to different positions in the text encoder and take sec / iter to measure how long it takes to generate a prompt embedding with a single RTX 3090 GPU. Table.4 shows that applying the SGLD sampling to the low-level space incurs huge computation overhead without obvious performance bonuses. It encourages us to take the two-layer backward pass to generate prompts across all experiments.

6.6. Vision-Language Information Retrievel

We finally provide the empirical study with respect to image-text retrieval tasks on MSCOCO and Flickr30K. We employed Karpathy split to separate MSCOCO into 113/5K/5K and Flickr30K into the amouts of 29,000 / 1,000 /1,000 for training / validation / test sets, respectively. We further construct the few-shot subsets for prompt tuning, with 0.5% and 1% instances drawn from their training sets, respectively. Given this, we train the prompt learners with these subsets, then evaluate the prompt learners' performance on their corresponding test sets using Recall at 1 (R@1) as our evaluation metric.We focus on the evaluation to CLIP, CoOp, CoCoOp and EMPL (+CoOp), where CLIP did not join prompt tuning and the other captions took as images' class labels. In Table.5, we observe that the CoOp-based models trained with 0.5% data in MSCOCO and Flick30K both suffer from the overfitting compared with CLIP. Encoding visual information by CoCoOp helps to alleviate, but failed to solve it in Flickr30K. In contrast, EMPL prevented CoOp from overfitting to the scarce training subsets to achieve the optimal results. With 1% training data, EMPL significantly improves CoOp, *e.g.*, **+3.37** for MSCOCO (1%) and **+2.92** for Flickr30K (1%), which outperformed the other baselines.

7. Conclusion, Limitation, and Future Work

In this paper, we have proposed a systematic overview to vision-language multi-prompt learning. In the discussion scope of CLIP, we revealed why multi-prompt learning strategies can improve cross-modal transferrability: (1) multi-prompt learning empirically reduces the modality gap with prompt augmentation and (2). single prompt learning provably suffers from non-identifiability issue while augmenting the prompt may alleviate. Given this obervations, we propose a new energy-based multi-prompt learning (EMPL) approach to improve the open-vocabulary generalization capability with regards to uncertainty modeling.

Our EMPL does not require any extra parameter introduced for CLIP, while its superiority has been theoretically and empirically supported by thorough experiments.

The drawback of EMPL mainly comes from its time cost for the prompt embedding inference . According to our device for training, we paid the triple time cost more than the CoOp original version and even more in terms of ProDA-based EMPL variants. For each inference for testing, we are encouraged to take double prompt embeddings compared with training phase to increase the performance, where we take the most certain class as our prediction results.

According to our discssion, it would be several promising treads in the future for multi-prompt learning. First, we only raise the first theoretic concern to multi-prompt learning since we have discovered the occurence of cross-modal non-identifiability behind single-prompt learner. Whereas why and how multi-prompt learners work out, still requiring for more sophisticated analytical studies with respect to learnability and optimization theories. Second, the prompt diversity is the key why multi-prompt learners outperform single-prompt learners. Therefore the research on the topic related with the prompt diversity is also inspiring. Finally, multi-prompt learning were always proposed as a time-consuimg approaches since they have to infer their prompts multiple times to predict each image. Reducing the inference cost would be a crucial issue in this field.

8. Ackonwledgemet

References

[1] Alvin Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002. 6

[2] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022. 2

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European*

Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, pages 446–461. Springer, 2014. 6

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. 1, 2

[5] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. *arXiv preprint arXiv:2210.01253*, 2022. 2, 6

[6] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pages 1542–1553. PMLR, 2020. 2, 3

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[9] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[11] 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. 1

[12] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 1, 2

[13] 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 *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[14] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 2

[15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 2

[16] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations*. 4

[17] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1

[19] 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. 1

[20] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, pages 56–73. Springer, 2022. 2

[21] 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

[22] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 6

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 6

[24] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[25] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[27] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 2, 4

[28] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022. 3

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[31] 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*, pages 5206–5215, 2022. 2, 3, 6

[32] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 1, 2

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

[34] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

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

[36] Bo Pang, Tian Han, Erik Nijkamp, Song-Chun Zhu, and Ying Nian Wu. Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33:21994–22008, 2020. 4

[37] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[38] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 2

[39] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2

[40] 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

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[42] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1, 2, 3

[43] 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. 6

[44] 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

[45] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[46] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 1

[47] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[48] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[49] Yezhen Wang, Bo Li, Tong Che, Kaiyang Zhou, Ziwei Liu, and Dongsheng Li. Energy-based open-world uncertainty modeling for confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9302–9311, 2021. 4, 5

[50] 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. 2, 6

[51] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591, 2017. 6

[52] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[53] Yuhui Zhang, Jeff Z HaoChen, Shih-Cheng Huang, Kuan-Chieh Wang, James Zou, and Serena Yeung. Diagnosing and rectifying vision models using language. In *The Eleventh International Conference on Learning Representations*. 2, 3

[54] 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, 3, 4, 6, 7

[55] 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, 3, 4, 6

[56] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1450–1459, 2021. 1