Debiased Fine-Tuning for Vision-Language Models by Prompt Regularization

Beier Zhu[1], Yulei Niu[2*], Saeil Lee[3], Minhoe Hur[4], Hanwang Zhang[1]

[1] Nanyang Technological University
[2] Columbia University
[3] HMGICS AIR Center
[4] AIRS Company, Hyundai Motor Group
beier002@e.ntu.edu.sg, yn.yuleiniu@gmail.com, saeil.lee@hmgics.com, minhoe.hur@hyundai.com,
hanwangzhang@ntu.edu.sg

Abstract

We present a new paradigm for fine-tuning large-scale vision-language pre-trained models on downstream task, dubbed Prompt Regularization (ProReg). Different from traditional fine-tuning which easily overfits to the downstream task data, ProReg uses the prediction by prompting the pretrained model to regularize the fine-tuning. The motivation is: by prompting the large model "a photo of a [CLASS]", the fill-in answer is only dependent on the pretraining encyclopedic knowledge while independent of the task data distribution, which is usually biased. Specifically, given a training sample prediction during fine-tuning, we first calculate its Kullback-Leibler loss of the prompt prediction and Cross-Entropy loss of the ground-truth label, and then combine them with a proposed sample-wise adaptive trade-off weight, which automatically adjusts the transfer between the pretrained and downstream domains. On various out-of-distribution benchmarks, we show the consistently strong performance of ProReg compared with conventional fine-tuning, zero-shot prompt, prompt tuning, and other state-of-the-art methods.

Introduction

Please think about it: when you want to train a vision model for a task, what is the first thing first? Most likely, you will download an off-the-shelf foundation model (Bommasani et al. 2021), *e.g.*, ResNet (He et al. 2016) or CLIP (Radford et al. 2021), pretrained on a large-scale dataset such as ImageNet (Deng et al. 2009) or image-text pairs collected from the Internet (Radford et al. 2021); then, remove its classifier head, plug in your own task-specific head to the penultimate layer, and finally fine-tune on your own task data. Such "pretrain, fine-tune" paradigm has become a nearly ubiquitous standard for CV community—from classification to generation (Viazovetskyi, Ivashkin, and Kashin 2020; Brock, Donahue, and Simonyan 2018), regions to pixels (He et al. 2017), and single modality to multi-modality (Anderson et al. 2018; Lu et al. 2019). The underlying empirical principle is that the pretrained model as initialization plays a role in regularization which reduces the variance of the fine-tuning model (Sutskever et al. 2013).

Despite the beneficial regularization, the pretraining knowledge has a negative impact, especially when the downstream task data is limited or biased (Wang et al. 2019; Yang et al. 2021): the early exposed encyclopedic or generic features from the pretrained model may mislead the fine-tuning to focus on the task-unrelated attributes, resulting in a biased fine-tuned model. Figure 1 shows three types of biases: Figure 1(a) contextual bias: images of training and test set contain distinct backgrounds, *e.g.*, if most of the training "bird" images are in "sky" background, the fine-tuning will misclassify "bird on ground" as "dog"; Figure 1(c) image style gap: test image style is unseen during training, *e.g.*, if training images are from art painting, cartoon and real domain, the fine-tuned model is able to classify the in-distribution art painting "dog" but is confused when testing on a sketch "dog"; Figure 1(e) language bias: for every question type, train and test sets have different prior distributions of answers, *e.g.*, if most VQA training "bananas" images are "yellow", it mistakes the answer "yellow" for question "what color are the bananas?", given an image of green bananas.

Recently, NLP community presents a tuning-free paradigm called "pretrained, prompt, predict" (Liu et al. 2021a) and it has been quickly migrated to CV tasks (Zhou et al. 2021; Tsimpoukelli et al. 2021; Yao et al. 2021) by using a pretrained multi-modal model (Kim, Son, and Kim 2021; Radford et al. 2021; Zhang et al. 2021). For example, image classification can be cast into the cloze prompt: "a photo of a [CLASS]", where the prediction can be found as the class word whose fill-in sentence is most similar to the query image. The similarity can be calculated directly from the pretrained model in a zero-shot learning fashion. As prompt is a rule-based query that has nothing to do with the downstream statistics, the prediction is expected to be independent to downstream domain and faithfully respects the pretrained knowledge.

Yet, relying too much on the general knowledge also hurts the domain-specific generalization. For example, as shown in Figure 1(b), although prompt can correctly focus on the foreground object, it is less discriminative to distinguish between "rat" and "cat" in domain-specific animal images; in Figure 1(d), prompt's prediction confidence is not as discriminative as that of fine-tuning; in Figure 1(f), prompt-based VQA is too general to perform the downstream task of counting "bananas". To this end, "prompt tuning" is pro-
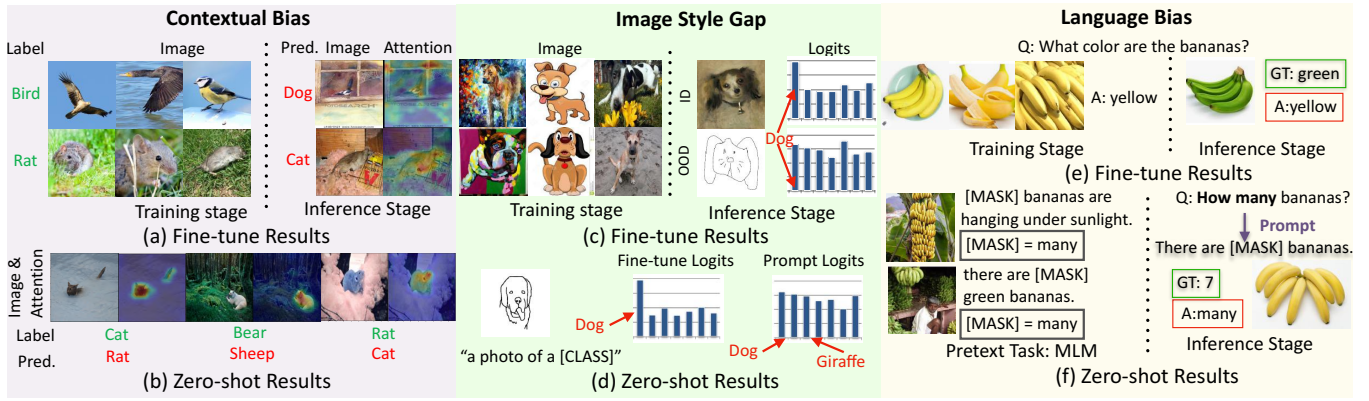
Figure 1: Examples of the contextual bias (He, Shen, and Cui 2020; Nam et al. 2020), image style bias (Li et al. 2017) and language bias (Goyal et al. 2017; Agrawal et al. 2018) caused by only relying on fine-tuning. (a&c&e): If the downstream data is biased, fine-tuning is biased. (b&d&f): Pretraining knowledge is too general to perform domain-specific downstream tasks.

posed to fine-tune the token embeddings in a prompt using the task data (Liu et al. 2021a,b; Han et al. 2021; Zhou et al. 2021). For example, the prefix "a photo of a" before the cloze blank [CLASS] can be replaced with a set of tunable parameters. Prompt tuning is essentially fine-tuning with fixed backbone and tunable head (*i.e.*, the prompt prefix). Therefore, it still inherits the above drawbacks of the biased fine-tuning.

In this paper, we present a new fine-tuning paradigm, called *Prompt Regularization* (ProReg), for a *just-right* knowledge transfer from pretrained model to fine-tuned model. As expected, ProReg can fine-tune the resultant model, neither biased towards the pretrained knowledge nor towards the downstream knowledge. We formulate the downstream knowledge as the ground-truth annotations in downstream tasks, and represent the pretrained knowledge with the "soft" label of the downstream data generated by prompting the pre-trained model. We then proposed the ProReg loss to enable learning from both of the knowledge. It is worth noting that different from traditional knowledge distillation that using a constant weight $\lambda \in (0, 1)$ to trade-off the contribution of the knowledge: $\mathcal{L}_{\text{kd}} = (1 - \lambda) \cdot \mathcal{L}_{\text{ce}} + \lambda \cdot \mathcal{L}_{\text{kl}}$, we propose a sample-wise adaptive weight to achieve a good trade-off between the two knowledge (Section). The proposed weight inspects whether the task-specific knowledge or the pre-trained general knowledge dominates the optimization process for each fine-tuning sample, which indeed requires a different ratio of the two knowledge types. We show that the estimation of the ratio evolves during the training process and can be automatically calculated on-the-fly.

We implement ProReg on top of two off-the-shelf large-scale pretrained models: CLIP (Radford et al. 2021) and ViLT (Kim, Son, and Kim 2021), which demonstrates that ProReg is applicable to different vision-language models that adopt masked language modeling or contrastive learning as pretraining tasks. We conduct extensive evaluations for ProReg on various out-of-distribution benchmarks, including BAR (Nam et al. 2020), NICO (He, Shen, and Cui 2020), PACS (Li et al. 2017) and DomainNet (Peng et al.

2019) for image classification tasks and VQA-CP (Agrawal et al. 2018) for visual question answering tasks. We demonstrate that: 1) ProReg consistently outperforms zero-shot prompt, conventional fine-tuning, and prompt tuning on all the datasets, 2) ProReg achieves compelling performance in both out-of-distribution and in-distribution settings. Thus, readers can feel free to use ProReg regardless of the training and testing distribution discrepancy.

Related Work

Vision-Language Models (VLM). Most existing VLMs use 1) Masked language modeling (Kim, Son, and Kim 2021; Li et al. 2019; Lu et al. 2019), 2) Image text pairing (Tan and Bansal 2019; Huang et al. 2020), and 3) Contrastive learning (Radford et al. 2021) as their pretraining objectives. Recently there is a line of adapting the existing VLMs to the downstream tasks. Conventional fine-tuning paradigm adds an additional classifier on top of the visual backbone (Linear Probe (Radford et al. 2021)) or additional feature adapter, (CLIP-Adapter (Gao et al. 2021)). Prompt-based learning that tunes the prompt to maximize the ground-truth token has gained its popularity, *e.g.*, CoOp (Zhou et al. 2021) and CoCoOp (Zhou et al. 2022). ProGrad (Zhu et al. 2022) bridges generalization gap by matching the gradient of prompt to the general knowledge. ProDA (Lu et al. 2022) introduces prompt distribution learning to adapts VLMs to downstream classification tasks. As discussed in Section , both of these two fine-tune paradigms may result in biased downstream models, in this work, we aims to fine-tune a de-biased VLM for downstream tasks.

OOD Generalization. In real world, test distribution may shift from the training distribution, such as domain generalization (Ben-David et al. 2007; Tzeng et al. 2017), long-tailed classification (Menon et al. 2020; Tang, Huang, and Zhang 2020), contextual bias (Nam et al. 2020; He, Shen, and Cui 2020), and language bias (Wu and Mooney 2019; Ramakrishnan, Agrawal, and Lee 2018; Cadene et al. 2019; Niu et al. 2021). WiSE (Wortsman et al. 2022) shares the same goal to improve OOD performance, which ensembles

the weights of zero-shot and fine-tuned models. However, it requires the fine-tuned model and the pre-trained model to have the same architecture. Differently, our ProReg is free of this requirement and allows the modification of architecture.

The ProReg Pipeline

Pre-trained Model

In this paper, we adopt two state-of-the-art VLM models as our backbones, *i.e.*, Constrastive Language-Image Pre-training model (CLIP (Radford et al. 2021) for image classification tasks and Vision-and-Language Transformer (ViLT (Kim, Son, and Kim 2021))) for VQA tasks.

CLIP collects large amounts of images with natural language description and is trained in a contrastive manner. The associated image and text pairs are regarded as positive pairs, while the mis-matched image and text pair are regarded as negative ones. The contrastive loss aims to maximize the cosine similarity of the positive pairs in the batch while minimize the cosine similarity of the negative pairs.

ViLT consists of stacked blocks that include a multi-headed self-attention and a multi-layer perceptron layer. It is pretrained with two pretext tasks: Image Text Matching (ITM) and Masked Language Modeling (MLM). For ITM, ViLT selects half of the sentences and replaces each of them with a mismatched sentence. Then, a binary classifier is added to predict whether the image and the sentence match with each other. MLM first randomly replace 15% of input text tokens with masked tokens, *e.g.*, [MASK], and then the model is optimized to predict the masked tokens, given other non-masked tokens and image feature.

For classification tasks, CLIP based models show superior performance over the ViLT one, thus we only report the CLIP results in the main paper while the image classification results of ViLT can be found in Appendix. For VQA tasks, the input text varies sample-wisely, it is difficult for CLIP model to infer and optimize. As a result, we only implement the ProReg on ViLT models. Note that our approach is applicable to broader vision-language models whose pre-trained pretext tasks include ITM and MLM objectives or contrastive objective.

Prompt-based Zero-shot Inference

Proper prompt can adapt the pre-trained model to downstream tasks without fine-tuning. In this section, we illustrate how CLIP performs zero-shot inference for image classification tasks and how ViLT can leverage carefully designed prompt to perform zero-shot prediction for VQA tasks.

Image Classification. We formulate image classification as an image-text matching problem where the text is designed by a simple static prompt template. Take the action recognition dataset BAR (Nam et al. 2020) in Figure 2(a) as an example. We first compute the text embedding $\mathbf{w}_i$ of "a person is [CLASS].", where the "[CLASS]" is replaced by i-th class name. Then the probability distribution of the image feature $\mathbf{x}$ over K classes is given by:

$$f(y = i|\mathbf{x}; \theta) = \frac{\exp(<\mathbf{x}, \mathbf{w}_i>/\tau)}{\sum_{j=1}^{K} \exp(<\mathbf{x}, \mathbf{w}_j>/\tau)} \quad (1)$$

Downs. Task	Pretext Task	Prompt Example
Image CLS	Contra.	• the person is [CLASS]. • a photo of a [CLASS].
VQA	MLM	• how many hats are there? $\xrightarrow{\text{Q2S}}$ there are [MASK] hats. • what color is the shirt? $\xrightarrow{\text{Q2S}}$ the color of the shirt is [MASK].
	ITM	• does this fruit grow on vines? $\xrightarrow{\text{Q2S}}$ this fruit grow on vines. • is the zebra sleeping? $\xrightarrow{\text{Q2S}}$ the zebra is sleeping.

Table 1: Prompt Examples. For VQA, the prompt is generated according to the question type. If the question is open-ended, we convert the question to a statement with masked token and use MLM head to predict the answer. If the question is closed-ended, we convert the question to a statement and use ITM head to predict "yes" or "no". For image classification, we adopt a simple static template and use cosine similarity to predict the class.

where τ is the temperature learned by CLIP.

Visual Question Answering (VQA). The design of prompts for VQA task is less straightforward because the text inputs in VQA (*i.e.*, questions) and pretraining tasks (*i.e.*, captions) have different forms. Notice that VQA dataset consists of two types of questions: open-ended questions (*e.g.*, "what color is the cake?") and closed-ended questions (*e.g.*, "any brown apples in the picture?"). We convert the question to a statement and name such design as **Question-to-Statement (Q2S)** prompt. Table 1 shows some prompt examples. Specifically, we use ITM for closed-ended questions and MLM for open-ended questions. For example, for a closed-ended question like "any brown apples in the picture?", the prompt statement is generated as "some brown apples in the picture.". The text embedding together with the visual feature are fed into ITM head to predict whether they match with each other. A high match score corresponds to the answer "yes" and a low score to "no". For an open-ended question like "what color is the cake?", the prompt is generated as "the color of the cake is [MASK].". Similar to Eq. (1), we search the candidate answer set to find the one has the highest probability.

Sample-wise Adaptive Trade-off Weight

In this section, we illustrate how to implement ProReg for CLIP models, the ProReg for ViLT models can be implemented analogously or refer to Appendix. Figure 2 (b) and (c) show the comparison of current fine-tuning works. Figure 2 (b) shows the pipeline of "pretrain, prompt, fine-tune" where the classification head is generated by the optimizing the continuous prompt. We name the "pretrain, fine-tune" paradigm in Figure 2(c) as *Conventional Fine-tuning*, which adds a classification head for prediction. The classification head can be randomly initialized or initialized by
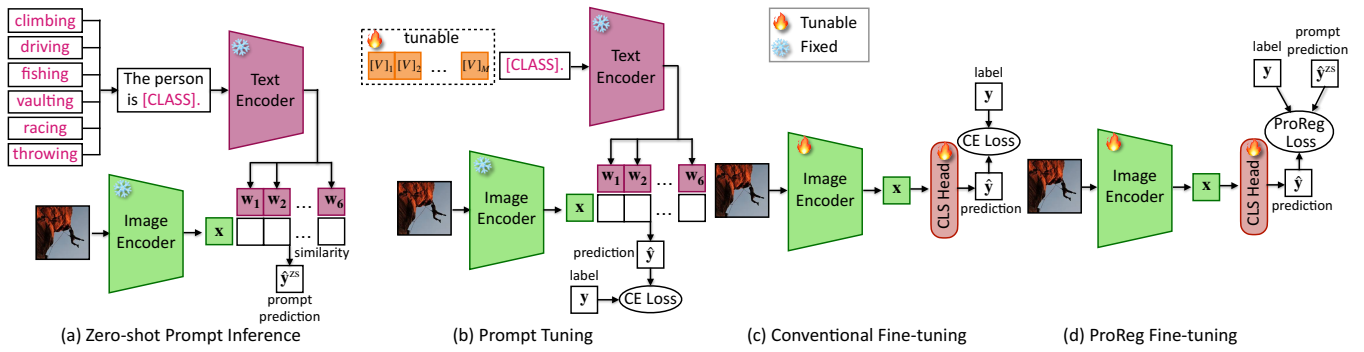
Figure 2: Comparisons of frameworks based on CLIP models for image classification. (a) Zero-shot Prompt Inference, *i.e.*, "pretrain, prompt", generates the classification weights by a text encoder thought prompting. (b) Prompt Tuning, *i.e.*, "pretrain, prompt, fine-tune", learns continuous prompt by minimizing the cross-entropy loss. (c) Conventional Fine-tuning, (*i.e.*, "pre-train, fine-tune"), adds classification head on top of the visual backbone and is optimized by cross-entropy loss. The classification head can be randomly initialized or initialized by the text encoder thought prompting . (d) ProReg fine-tuning. Different from conventional fine-tuning, the supervisions come from two sources: the ground-truth target from the downstream domain, and the zero-shot prompt prediction from the pretraining domain. The classification head of ProReg is initialized by prompting.

feeding the hand-craft prompt to the text encoder. The common characteristic is that both of the two paradigms only use the ground-truth labels as supervision. As discussed in Section , both of these two fine-tune paradigms may result in biased downstream models, especially when the task data is limited and biased (Yue et al. 2020; Wang et al. 2019; Yang et al. 2021).

We present a new fine-tuning paradigm, dubbed *Prompt Regularization* (ProReg), to transfer the knowledge from the pretraining domain to the downstream domain. ProReg considers two types of supervisions for optimization: task-specific downstream knowledge and task-agnostic general knowledge.

Task-specific downstream knowledge. The ground-truth labels serve as the task-specific knowledge and allows the model $f(;\theta)$ to be adapted to downstream task. The cross-entropy $\mathcal{L}_{ce}$ of an image $\mathbf{x}$ is obtained as:

$$\mathcal{L}_{ce}(\theta) = -\sum_i \mathbf{y}_i \log f_i(\mathbf{x};\theta), \qquad (2)$$

where $\mathbf{y}$ denotes the one-hot ground-truth vector, and the subscript i denotes the i-th label; $f(;\theta)$ is the classification model initialized by pretrained model.

Task-agnostic general knowledge. Compared to large-scale datasets for pre-training, the task-specific data for downstream task may be limited or even biased. As a result, only taking the ground-truth annotations as supervision may lead to biased fine-tuning. In order to achieve debiased fine-tuning, we also use task-agnostic pre-training knowledge as regularization. We take the zero-shot prompt prediction of the pre-trained model $\mathbf{y}^{zs}$ as the regularization knowledge, and use kullback-leibler divergence ($\mathcal{L}_{kl}$) between the fine-tuned model prediction and the zero-shot prompt prediction as the regularization term:

$$\mathcal{L}_{kl}(\theta) = -\sum_i \mathbf{y}_i^{zs} \log \frac{f_i(\mathbf{x};\theta)}{\mathbf{y}_i^{zs}}, \qquad (3)$$

To fine-tune the model neither biased towards the downstream domain nor the pre-trained knowledge domain, a straightforward intuition is knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), to combine the two terms with a constant weight $\lambda \in (0,1)$:

$$\mathcal{L}_{kd} = (1-\lambda) \cdot \mathcal{L}_{ce} + \lambda \cdot \mathcal{L}_{kl}, \qquad (4)$$

where we omit θ for simplicity. However, this simple solution overlooks that the trade-off of the two knowledge varies sample-wisely and evolves during training. In the next part, we will analyze the trade-off from the view of task-agnostic knowledge decomposition and propose a dynamic weight to balance the two loss terms for each training sample.

Decomposition of task-agnostic general knowledge. We decompose the regularization term $\mathcal{L}_{kl}$ in Eq. (3) as $\mathcal{L}_{kl} = \mathcal{L}_{ce} + (\mathcal{L}_{kl} - \mathcal{L}_{ce})$. The first term $\mathcal{L}_{ce}$ aims to learn task-specific knowledge. The second term $(\mathcal{L}_{kl} - \mathcal{L}_{ce})$ is the key component to learn supplementary knowledge that is provided by the task-agnostic general knowledge but not included in the task-specific downstream knowledge. To better understand the contribution of the two knowledge, we calculate their gradients *w.r.t* the logit $\mathbf{z}$ of the model f on the class t as:

$$\nabla_{\mathbf{z}_t} \mathcal{L}_{ce} = f_t - \mathbf{y}_t, \qquad (5)$$

$$\nabla_{\mathbf{z}_t}(\mathcal{L}_{kl} - \mathcal{L}_{ce}) = (f_t - \mathbf{y}^{zs}) - (f_t - \mathbf{y}_t) = \mathbf{y}_t - \mathbf{y}_t^{zs}. \qquad (6)$$

We have the following observations: First, the two gradients always have opposite directions, learning one knowledge well will inevitably lead to the deterioration of the other knowledge. If t is the true class, *i.e.*, $\mathbf{y}_t = 1$, then $\nabla_{\mathbf{z}_t} \mathcal{L}_{ce} = f_t - 1 < 0$ while $\nabla_{\mathbf{z}_t}(\mathcal{L}_{kl} - \mathcal{L}_{ce}) = 1 - \mathbf{y}_t^{zs} > 0$. If t is the false class, *i.e.*, $\mathbf{y}_t = 0$, $\nabla_{\mathbf{z}_t} \mathcal{L}_{ce} = f_t > 0$ while $\nabla_{\mathbf{z}_t}(\mathcal{L}_{kl} - \mathcal{L}_{ce}) = -\mathbf{y}_t^{zs} < 0$. The first observation reveals the conflict between task-specific knowledge learning and supplementary knowledge learning, which motivates us to balance the learning process of the two knowledge.

Second, the gradient $\mathbf{y}_t - \mathbf{y}_t^{zs}$ for learning supplementary knowledge is constant in different optimization steps, and

the magnitude varies sample-wisely. As a comparison, the gradient $f_t - \mathbf{y}_t$ for learning task-specific knowledge keeps updated during the training process (f_t changes after each optimization step). The above analysis motivates us to trade-off the two knowledge in a sample-wise and dynamic manner, dubbed ProReg:

$$\mathcal{L}_{\text{ProReg}} = \mathcal{L}_{\text{ce}} + w \cdot (\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}}) = (1-w) \cdot \mathcal{L}_{\text{ce}} + w \cdot \mathcal{L}_{\text{kl}}. \quad (7)$$

The sample-wise trade-off weight w aims to prevent the model from biasing to either task-specific knowledge and supplementary knowledge. There are two typical cases that affect the determination of w.

If $|\nabla_{\mathbf{z}_t} \mathcal{L}_{\text{ce}}|$ is much larger than $|\nabla_{\mathbf{z}_t}(\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}})|$, the task-specific knowledge will dominate the overall optimization direction of this sample. We assign a larger weight w to the term $(\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}})$ to emphasis the supplementary knowledge and prevent the model from biasing to the downstream knowledge. On the contrary, if $|\nabla_{\mathbf{z}_t}(\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}})|$ is much larger than $|\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{ce}}|$, the task-specific downstream knowledge might be neglected. We assign a smaller weight w to $(\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}})$ to guarantee the learning of downstream knowledge. From the above analysis, the trade-off weight w should be proportional to $|\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{ce}}|/|\nabla_{\mathbf{z}_t}(\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}})|$ to balance the two knowledge. From Eq. (5) and Eq. (6), we have $|\nabla_{\mathbf{z}_t}\mathcal{L}_{\text{ce}}|/|\nabla_{\mathbf{z}_t}(\mathcal{L}_{\text{kl}} - \mathcal{L}_{\text{ce}})| = \frac{\mathbf{y}_t - f_t}{\mathbf{y}_t - \mathbf{y}_t^{\text{zs}}}$. For simplifying analysis, we consider the binary classification case and assume the ground-truth label is t, *i.e.*, $\mathbf{y}_t = 1$. For true class t, we have $\frac{\mathbf{y}_t - f_t}{\mathbf{y}_t - \mathbf{y}_t^{\text{zs}}} = \frac{1 - f_t}{1 - \mathbf{y}_t^{\text{zs}}} \propto \frac{\mathbf{y}_t^{\text{zs}}}{f_t}$. For the false class $(1-t)$, we have $\frac{\mathbf{y}_{1-t} - f_{1-t}}{\mathbf{y}_{1-t} - \mathbf{y}_{1-t}^{\text{zs}}} = \frac{0 - (1 - f_t)}{0 - (1 - \mathbf{y}_t^{\text{zs}})} \propto \frac{\mathbf{y}_t^{\text{zs}}}{f_t}$. From Eq. (7), we also expect $w \in [0, 1]$ to guarantee the positive sign of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{kl}}$. Therefore, we design the trade-off weight as

$$w = \frac{f_t}{f_t + \mathbf{y}_t^{\text{zs}}} \propto \frac{\mathbf{y}_t^{\text{zs}}}{f_t}, \quad (8)$$

where t is the ground-truth label.

In our implementation, in addition to the sample-wisely trade-off weight w, we further introduce a hyper-parameter $\alpha > 0$ on $\mathcal{L}_{\text{kl}}$ in Eq. (7) to control the strength of trade-off globally. Our ProReg is formulated as:

$$\mathcal{L}_{\text{ProReg}} = (1-w) \cdot \mathcal{L}_{\text{ce}} + \alpha \cdot w \cdot \mathcal{L}_{\text{kl}} \quad (9)$$

Experiments

Datasets and Implementation Details

BAR (Nam et al. 2020) is a real-world image dataset for action classification, where the action is biased to the place. Examples of "climbing" class are shown in Figure 3(a), where most training background are rocks, while the one of testing images are snow.

NICO (He, Shen, and Cui 2020) dataset is designed for Out-of-Distribution (OOD) image classification, which has 2 subsets (animal and vehicle), 19 classes and 188 contexts. For each class, we select 3 different contexts for both training and test set, *e.g.*, the training contexts for "sheep" are "in water", "on road" and "on snow", while the test contexts are "at sunset", "on grass" and "in forset". Please kindly refer to Apprendix for more details of our setting.

(a) Examples of BAR Dataset

(b) Examples of PACS Dataset

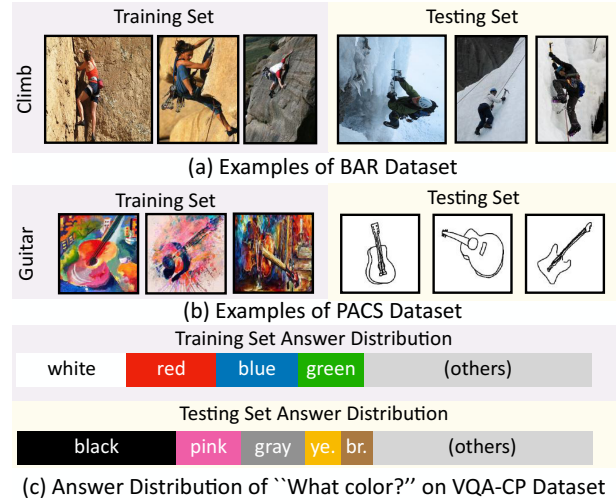(c) Answer Distribution of ``What color?'' on VQA-CP Dataset

Figure 3: Examples of OOD benchmarks.

PACS (Li et al. 2017) covers photo, sketch, cartoon and painting domains. Figure 3(b) shows some examples. The model is trained and validated on any three seen domains then tested on the rest unseen domain.

DomainNet (Peng et al. 2019) consists of images from 345 classes covering the "sktech", "real", "clipart" and "painting" domains. Here, we use the "sktech" domain as ID dataset, and use "clipart" and "painting" as the OOD datasets. Please kindly refer to Appendix for more details.

VQA-CP (Agrawal et al. 2018) is proposed to examine the generalizability of VQA models when the language prior varies significantly between training and testing splits. Figure 3(c) shows some answer distribution of the training and test set, *e.g.*, most of "what color " questions are answered as "white" in training set while "black " in the test set.

Experimental Details. For ViLT-based models, we followed the original fine-tuning settings in (Kim, Son, and Kim 2021), which adopt the ViLT-B/32 model with AdamW (Loshchilov and Hutter 2018) optimizer for 10 epochs for all datasets. For CLIP-based models, we used the ViT-B/32 backbone and adopted the ViLT fine-tuning settings, including the training epoch, optimizer, warmup schedule and image pre-processing, *etc*. α is set to 2 for all experiments. See Appendix for more details.

Fine-tuning Baselines. For CLIP-based models, we compared ProReg with 6 baselines. (1) zero-shot CLIP (Radford et al. 2021); (2) linear probe (Radford et al. 2021); (3) prompt tuning, *i.e.*, CoOp (Zhou et al. 2021); (4) an advance CLIP fine-tuning: CLIP-adapter (Gao et al. 2021); (5) conventional fine-tuning with randomly initialized classification head, which is denoted as FT; (6) conventional fine-tuning with classification head initialized from a text encoder thought prompting, which is denoted as FT++;

For ViLT-based models, we compared ViLT-ProReg with three baseline methods. (1) zero-shot ViLT. (2) conventional fine-tuning with randomly initialized classification head, which is denoted as FT. (3) conventional fine-tuning with classification head initialized from a text encoder thought

Method	Acc.
Zero-shot	86.7
Linear Probe	89.5
CoOp	90.2
CLIP-Adapter	90.4
FT	90.5
FT++	90.9
ProReg	**92.1**

Table 2: Accuracy (%) on BAR.

Method	ID	OOD	HM
Zero-shot	58.2	63.7	60.8
Linear Probe	70.4	52.9	60.4
CoOp	70.9	56.6	62.9
CLIP-Adapter	71.2	51.8	60.0
FT	**75.9**	55.8	64.3
FT++	75.8	60.6	67.4
ProReg	75.6	**64.4**	**69.2**

Table 3: Accuracy (%) on DomainNet dataset.

Method	OOD Domain				Avg
	A	C	P	S	
Zero-shot	89.8	96.3	99.4	85.8	92.8
Linear Probe	84.3	91.7	83.1	80.3	84.9
CoOp	92.6	96.1	98.1	85.0	93.0
CLIP-Adapter	93.6	96.6	97.7	84.4	93.1
FT	95.3	95.6	99.7	86.8	94.4
FT++	95.6	97.0	99.3	86.8	94.7
ProReg	**96.2**	**98.6**	**99.8**	**89.4**	**96.0**

Table 5: Accuracy (%) on PACS.

Method	Animal Subset			Vehicle Subset		
	ID	OOD	HM	ID	OOD	HM
Zero-shot	93.3	92.5	92.9	95.8	95.9	95.8
Linear Probe	99.0	87.0	92.6	99.4	87.5	93.1
CoOp	**99.3**	86.1	92.2	99.4	87.6	93.1
CLIP-Adapter	99.0	81.9	89.6	99.4	83.7	90.9
FT	98.7	85.3	91.5	**99.6**	88.9	93.9
FT++	98.9	91.7	95.2	99.4	90.8	94.9
ProReg	98.5	**94.4**	**96.4**	98.7	**94.1**	**96.3**

Table 4: Accuracy (%) on NICO dataset.

	Methods	**All**	Y/N	Num.	Other
ViLT	Q2S Zero-shot	43.62	**77.70**	12.02	33.93
	FT	45.68	44.16	14.44	55.12
	FT++	46.34	73.00	**16.29**	40.23
	ProReg	**54.89**	73.06	14.85	**55.58**

Table 6: Evaluations (Accurarcy%) on VQA-CP.

prompting, which is denoted as FT++. Please see Appendix for more details.

Evaluation Metrics. To evaluate the unbiasedness, we also report the in-domain (ID) accuracy, out-of-domain (OOD) accuracy and their harmonic mean for NICO and Domain-Net datasets. Specifically, ID test set has the same distribution with the training set while the distribution of OOD test set is different from the training one. A debiased model should have high performance on both ID and OOD testing, as a result, it should also have the highest harmonic mean.

Main Results

Image Classification. Results are shown in Table 2, Table 4, Table 5 and Table 3. Due to the powerful pretraining knowledge and hand-crafted prompt, the zero-shot CLIP model achieves strong performance. In particular, on NICO Vehicle subset (Table 4), the zero-shot CLIP exhibits harmonic mean of 95.8%. After training on the downstream data, we observed that ProReg demonstrates clear advantages over other fine-tuning methods on the OOD test sets, *e.g.*, on BAR test set, the ProReg outperforms other counterparts by at least 1.2%. Not surprisingly, on NICO and DomainNet datasets, we observed that other fine-tuning methods gain significant improvements on in-distribution accuracies but at a cost of performance loss on out-of-distribution accuracies compared to zero shot performance, *e.g.*, recently proposed CoOp (Zhou et al. 2021) increases the ID accuracies from 93.3% to 99.3% on NICO Animal subset (from 58.2% to 70.9% on DomainNet) but the OOD accuracies decreases from 92.5% to 86.1% (from 63.7% to 56.6% on Domain-Net). As a comparison, our ProReg obtains a good trade-off between on ID and OOD performance, thus achieving the best performance on harmonic means.

Table 5 reported the results on PACS. The zero-shot CLIP shows strong performance on Photo (P) and Cartoon (C) Domains, *e.g.*, zero-shot CLIP achieves 99.4% on photo domain. Besides the best performance in average accuracy, our ProReg significantly outperforms other fine-tuning methods on difficult unseen domains Sketch (S) and Cartoon (C), *e.g.*, ProReg gains 1.6% and 2.6% on sketch and cartoon domain compared to FT++ method. These results show the power of our ProReg to overcome diverse domain biases.

Visual Question Answering. Considering that the input text for VQA tasks varies sample-wisely, it is difficult for CLIP model to infer and optimize. We only implement the ProReg based on ViLT models and compared with zero-shot prompt, FT and FT++.

Although zero-shot prompt models have no access to the training data, thanks to the pretraining knowledge and our proposed question-to-statement (Q2S) prompt design, our zero-shot prompt achieves impressive results with 43.62% accuracy, which is a strong baseline. The results show that our ProReg framework shows strong performance under language bias, *e.g.*, ProReg achieves an overall accuracy of 54.89% with an impressive improvement $+8.55\%$ compared to conventional fine-tuning with classification initialized from prompt (FT++). More interestingly, conventional promptless fine-tune (FT) failed on "Yes/No" questions, while FT++ failed on "Other". As a comparison, ProReg performs relatively well on all the three questions types.

Ablation Studies

We further conduct ablation studies to answer the following questions.

Q1: What is the effect of the hyper-parameter α? We conducted the experiments on DomainNet dataset by varying α in Eq. (9), the results are shown in Figure 4. As expected, as α increases, the fine-tuned model is encouraged to learn more from pre-trained knowledge. As a result, the OOD performance will increase while the ID accuracy drops.
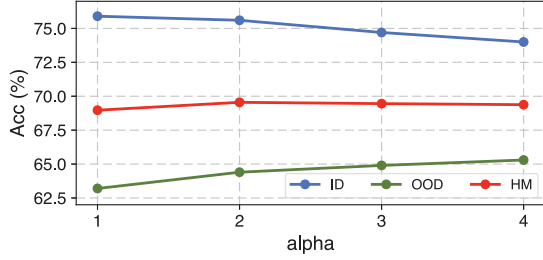
Figure 4: Effect of hyper-parameter α on DomainNet.
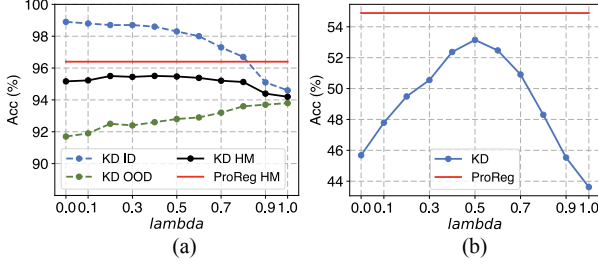

Figure 5: (a) KD results (ID, OOD and harmonic %) on NICO animal subset.(b) KD results (overall accuracy %) on VQA-CP with different weight λ.
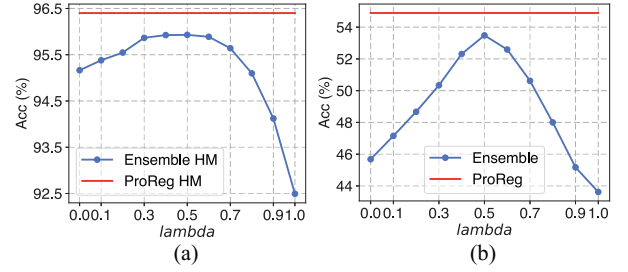

Figure 6: (a) Ensemble results (harmonic mean %) on NICO vehicle subset. (b) Ensemble results (overall accuracy %) on VQA-CP with different weight λ.
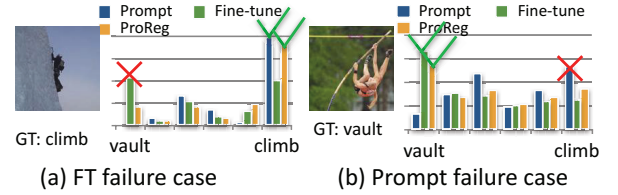

(a) FT failure case   (b) Prompt failure case

Figure 7: Qualitative examples of zero-shot prompt, conventional fine-tune and ProReg fine-tune on BAR dataset.

Q2: Can we blend the knowledge using conventional distillation that use constant trade-off weight? No, its performance is worse than ProReg fine-tune. We implemented traditional knowledge distillation strategy by varying the trade-off weight λ as described in Eq. (4). Figure 5 shows the results on NICO Vehicle and VQA-CP dataset. We observed that our ProReg model achieves superior performance than traditional knowledge distillation, *e.g.*, for VQA-CP datasets, even with the optimal λ, traditional KD (blue line) achieves the best performance of 53.15% with -1.74% gap to ProReg fine-tune (red line).

Q3: Can we directly ensemble the zero-shot model and traditional fine-tune model? No, it not only achieves worse performance than ProReg, but also doubles the inference time and the number of parameters. In Figure 6, we investigate whether combining the knowledge by ensembling the conventional fine-tuning and the zero-shot CLIP model can perform superior results. Specifically, given the prediction of fine-tuning model $\mathbf{y}^{ft}$ and the one from zero-shot model $\mathbf{y}^{zs}$, the ensembled prediction is formulated as $\mathbf{y}^{ens} = (1 - \lambda) \cdot \mathbf{y}^{ft} + \lambda \cdot \mathbf{y}^{zs}$, where $\lambda \in [0, 1]$. Empirical results on VQA-CP dataset and NICO vehicle dataset are plotted in Figure 6(a), where the highest harmonic mean wiht $\lambda = 0.5$ (blue line) still underperforms our ProReg (red line). For VQA-CP dataset, ensemble model reaches its optimal accuracy 53.48% with $\lambda = 0.5$ (blue line). The highest accuracy of ensemble is still surpassed by our ProReg result (54.89%, red line). Moreover, the ensemble model doubles the inference time and the number of parameters.

Qualitative results. We visualized a sports related question,

which is too domain-specific, the general knowledge learned by zero-shot prompt model is hard to answer such question. In the meanwhile, our ProReg inherits the knowledge from both domain and gives the right answer. Figure 7(a) and (b) show some failure cases of conventional fine-tune model and zero-shot ViLT model for BAR dataset. In Figure 7 (a), a "climb" image is mis-classified as "vault" by conventional fine-tune model, which can be attributed to the dataset bias, where the context of the "climb" training images are most rocks. In Figure 7 (b), a "vault" image is recognized as "dive" by the zero-shot ViLT model, we conjecture that the reason is that the knowledge of pole vaulting is not common in the pretrained domain. In both cases, ProReg predicts right answers by inheriting knowledge from the downstream data and the pretraining knowledge.

Conclusion

We presented an effective novel fine-tuning strategy: Prompt Regularization (ProReg), for deploying VLMs in downstream tasks. As its name implies, ProReg is motivated to make the best of the two worlds: pretrained general knowledge and task-specific knowledge, where the former is acquired by using prompt. The highlight of ProReg is the proposed sample-wise adaptive weight that trades off the training losses from the two worlds. Note that this weight is theoretically justified. Therefore, we believe that ProReg has a great potential for helping practitioners to fine-tune their own models efficiently. By extensive evaluations on three types of challenging OOD benchmarks, ProReg significantly outperformed zero-shot prompt, prompt tuning, conventional fine-tuning and other state-of-the-art methods.

Acknowledgments

References

Agrawal, A.; Batra, D.; Parikh, D.; and Kembhavi, A. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Ben-David, S.; Blitzer, J.; Crammer, K.; Pereira, F.; et al. 2007. Analysis of representations for domain adaptation. *NeurIPS*, 19: 137.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Cadene, R.; Dancette, C.; Cord, M.; Parikh, D.; et al. 2019. Rubi: Reducing unimodal biases for visual question answering. *NeurIPS*.

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.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2021. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *arXiv preprint arXiv:2110.04544*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Han, X.; Zhao, W.; Ding, N.; Liu, Z.; and Sun, M. 2021. PTR: Prompt Tuning with Rules for Text Classification. *arXiv preprint arXiv:2105.11259*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

He, Y.; Shen, Z.; and Cui, P. 2020. Towards Non-IID Image Classification: A Dataset and Baselines. *PR*, 107383.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *stat*, 1050: 9.

Huang, Z.; Zeng, Z.; Liu, B.; Fu, D.; and Fu, J. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2017. Deeper, broader and artier domain generalization. In *ICCV*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; and Neubig, G. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; and Tang, J. 2021b. GPT Understands, Too. *arXiv preprint arXiv:2103.10385*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *CVPR*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. In *ICLR*.

Nam, J.; Cha, H.; Ahn, S.-S.; Lee, J.; and Shin, J. 2020. Learning from Failure: De-biasing Classifier from Biased Classifier. In *NeurIPS*.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *ICCV*.

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 *ICML*.

Ramakrishnan, S.; Agrawal, A.; and Lee, S. 2018. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. In *NeurIPS*.

Sutskever, I.; Martens, J.; Dahl, G.; and Hinton, G. 2013. On the importance of initialization and momentum in deep learning. In *ICML*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*, 5100–5111.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. *NeurIPS*.

Tsimpoukelli, M.; Menick, J.; Cabi, S.; Eslami, S. A.; Vinyals, O.; and Hill, F. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *NeurIPS*.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *CVPR*.

Viazovetskyi, Y.; Ivashkin, V.; and Kashin, E. 2020. Style-gan2 distillation for feed-forward image manipulation. In *ECCV*. Springer.

Wang, Z.; Dai, Z.; Póczos, B.; and Carbonell, J. 2019. Characterizing and avoiding negative transfer. In *CVPR*.

Wortsman, M.; Ilharco, G.; Kim, J. W.; Li, M.; Kornblith, S.; Roelofs, R.; Lopes, R. G.; Hajishirzi, H.; Farhadi, A.; Namkoong, H.; et al. 2022. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7959–7971.

Wu, J.; and Mooney, R. 2019. Self-Critical Reasoning for Robust Visual Question Answering. *NeurIPS*.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *CVPR*.

Yao, Y.; Zhang, A.; Zhang, Z.; Liu, Z.; Chua, T.-S.; and Sun, M. 2021. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*.

Yue, Z.; Zhang, H.; Sun, Q.; and Hua, X.-S. 2020. Interventional Few-Shot Learning. *NeurIPS*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional Prompt Learning for Vision-Language Models. In *CVPR*.

Zhu, B.; Niu, Y.; Han, Y.; Wu, Y.; and Zhang, H. 2022. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*.