CPL: Counterfactual Prompt Learning for Vision and Language Models

Xuehai He[1] **Diji Yang**[1] **Weixi Feng**[2] **Tsu-Jui Fu**[2] **Arjun Akula**[3] **Varun Jampani**[3]
Pradyumna Narayana[3] **Sugato Basu**[3] **William Yang Wang**[2] **Xin Eric Wang**[1]
[1]UC Santa Cruz, [2]UC Santa Barbara, [3]Google
{xhe89,dyang39,xwang366}@ucsc.edu
{weixifeng,tsu-juifu,william}@ucsb.edu
{arjunakula,varunjampani,pradyn,sugato}@google.com

Abstract

Prompt tuning is a new few-shot transfer learning technique that only tunes the learnable prompt for pre-trained vision and language models such as CLIP. However, existing prompt tuning methods tend to learn spurious or entangled representations, which leads to poor generalization to unseen concepts. Towards non-spurious and efficient prompt learning from limited examples, this paper presents a novel **C**ounterfactual **P**rompt **L**earning (CPL) method for vision and language models, which simultaneously employs counterfactual generation and contrastive learning in a joint optimization framework. Particularly, CPL constructs counterfactual by identifying minimal non-spurious feature change between semantically-similar positive and negative samples that causes concept change and learns more generalizable prompt representation from both factual and counterfactual examples via contrastive learning. Extensive experiments demonstrate that CPL can obtain superior few-shot performance on different vision and language tasks than previous prompt tuning methods on CLIP. On image classification, we achieve a 3.55% average relative improvement on unseen classes across seven datasets; on image-text retrieval and visual question answering, we gain up to 4.09% and 25.08% relative improvements across three few-shot scenarios on unseen test sets respectively.[1]

1 Introduction

Pre-trained vision and language foundation models (Radford et al., 2021; Jia et al., 2021) have shown encouraging results toward open-domain visual-concept matching. Benefiting from prompt engineering (Song et al., 2022a; Liu et al., 2022), where free-form text prompts are designed for specific task goals, those foundation models can be easily transferred to a wide array of tasks under zero-shot and few-shot scenarios, including image classification (Deng et al., 2009), visual question answering (Shen et al., 2021), image-text retrieval (Jia et al., 2021), etc. But manually constructing prompts for vision and language models such as CLIP is a tedious, time-consuming process, which usually requires prior domain knowledge and leads to suboptimal solutions.

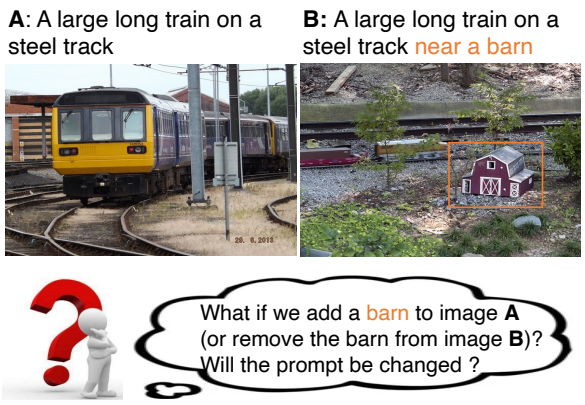

Figure 1: A conceptual overview of counterfactual prompt learning. CPL constructs counterfactuals by identifying non-spurious feature change that causally causes the prompt change. In this case, the "barn" feature is the essential cause between Prompt **A** and **B**.

Prompt tuning (Lester et al., 2021), on the other hand, liberates us from manual prompt engineering and automates this process. Prompt tuning methods (Ju et al., 2021; Lin et al., 2014; Zhou et al., 2022) are proposed to effectively transfer CLIP to image recognition tasks after tuning a learnable prompt with a few examples of the classes. However, those methods purely conduct empirical risk minimization (ERM) and optimize for predictive accuracy, which often produces spurious, inefficient, or entangled representations (Wang and Jordan, 2021). Therefore, the generalization ability of existing prompt tuning methods for vision and language models is limited, and they often fail to transfer well to unseen classes or concepts. For

[1]Our code is released at https://github.com/eric-ai-lab/CPL.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 3407 - 3418
December 7-11, 2022

example, the image classification performance of the SOTA method CoCoOp (Zhou et al., 2022) is similar or even degrades on unseen classes when compared with zero-shot CLIP.

Learning non-spurious representation for better generalization requires disentangling features that causally determine the prompts. One solution is counterfactual reasoning. Counterfactual ("counter to the facts") is a concept that describes the human capacity to learn from limited prior experiences by imagining the outcome of an alternative action that could have been taken. So we can do counterfactual intervention by asking "what if ..." questions in prompt learning. For example, as shown in Figure 1, a change in the visual feature of the barn would cause the label to change (if we view the two prompts as two labels).

Therefore, we introduce a new causality-based approach, **C**ounterfactual **P**rompt **L**earning (CPL), for non-spurious and efficient prompt learning. First, we introduce a text-based negative sampling strategy to discover the most semantically-similar negative sample based on text similarity. Then we generate a counterfactual example by identifying minimal non-spurious feature change between semantically-similar positive and negative samples that causally causes prompt change. Finally, we adopt contrastive learning in the joint optimization framework (with counterfactual construction) to tune the learnable prompts using both factual and counterfactual examples. The causally fine-tuned prompts will eventually guide vision-and-language foundation models to distinguish images from unseen concepts, thereby improving the generalization ability of prompt learning.

We extensively evaluate CPL using seven standard datasets for image classification, two for image-text-retrieval, and one for visual question answering (VQA). We show that CPL outperforms the baseline on all three tasks: on image classification, our method achieves 3.55% average relative improvement on unseen classes across the seven datasets in terms of accuracy; on image-text retrieval, our method improves the most (4.09% relative improvement in terms of Recall@1) when using 0.5% of total training instances on MSCOCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015); on VQA, we gain up to 25.08% relative improvement on the VQAv2 (Goyal et al., 2017a) dataset.

Our main contributions are summarized below:

- We introduce **C**ounterfactual **P**rompt **L**earning (CPL), a task-agnostic causality-based prompt learning method to effectively transfer CLIP to unseen concepts for different downstream tasks.
- We propose a text-based negative sampling strategy, where we compute BERTScore (Zhang et al., 2019) between text prompts, based on which we sample the most semantically-similar negative images.
- We introduce a optimization framework that simultaneously constructs counterfactuals by identifying minimal non-spurious feature change, and learns the generalized prompt representation from both factual and counterfactual examples.
- We conduct extensive experiments on image classification, image-text retrieval, and visual question answering, and validate the superiority of CPL to existing prompt tuning methods in transferring effectiveness on unseen concepts.

2 Related Work

Vision-and-Language Models. Vision-and-Language models pre-trained on large-scale image-text pairs have demonstrated great potential in multimodal representation learning (Jia et al., 2021; Yao et al., 2021; Yuan et al., 2021). Among them, the representative CLIP (Radford et al., 2021) benefits from 400M curated data and defines various prompt templates to carry out zero-shot image classification. However, those prompts still require hand-crafted designs. In this work, we automatically learn task-agnostic and task-relevant prompts without human priors. In addition, by considering the counterfactual examples, we can further improve various vision-and-language tasks, including visual question answering and image-text retrieval in a few-shot scenario.

Prompt Tuning. Many works focus on learning from discrete natural language prompts, e.g., AutoPrompt (Shin et al., 2020) elicits knowledge from language models with automatically generated discrete prompts. Lately, many other works (Zhou et al., 2021, 2022) directly tune prompts in continuous vector forms. Guo et al. (2021) introduces Q-Learning to optimize the soft prompt. P-Tuning v2 (Liu et al., 2021) shows that continuous

prompt tuning achieves the same performance as fine-tuning in various settings. Prompt tuning also receives great interest in the computer vision domain. For example, CoOp proposes a continuous prompt optimization strategy to avoid prompt design. CoCoOp (Zhou et al., 2022) extends CoOp by further learning an instance-conditional network to generate an input-conditional token for each image. However, these methods trained with empirical risk minimization (ERM) may learn to rely on correlations between class labels and spurious attributes by minimizing average training error (Zhang et al., 2022). They usually learn spurious, inefficient, and entangled representation, lacking generalization ability to unseen scenarios.

Counterfactual Reasoning. A number of recent works have investigated generating counterfactual images (Besserve et al., 2020), or counterfactual text in specific language domains (e.g., court view (Wu et al., 2020), dialogue generation (Zhu et al., 2020), Natural Language Inference (Kaushik et al., 2019; Gokhale et al., 2021), named entity recognition (Zeng et al., 2020)); On the vision end, Zhang et al. (2021) proposes to add intervention over the changed domain on images during the data-generation process and steer the generative model to produce counterfactual features to augment the training process. Agarwal et al. (2020) uses automated semantic image manipulations to generate synthetic data to make models more robust against spurious correlations; On the vision and language end, Chen et al. (2020) proposes to generate counterfactual VQA samples by masking critical objects in images or words in questions to augment the training data and gain a huge improvement on the VQAv2 dataset. Gokhale et al. (2020) proposes template-based counterfactual image augmentation methods. Fu et al. (2020) proposes a novel training strategy for visual language navigation that dynamically generates counterfactuals to account for unseen scenarios. To our best knowledge, CPL is the first to apply counterfactual generation to prompt-based few-shot learning for vision and language models.

Few-shot Learning. Recently, several few-shot efficient learners on vision (He et al., 2022) and language (Brown et al., 2020) tasks were proposed including CLIP. GPT (Brown et al., 2020), as a strong few-shot learner, is capable of performing a new language task by learning from only a few training instances. Frozen (Tsimpoukelli et al., 2021) is developed based on GPT and made into a multimodal few-shot learner by expanding the soft prompting to include a collection of images and text. Their method demonstrates strong few-shot capabilities on visual question answering and image classification tasks. Similarly, CoCa (Yu et al., 2022) is pre-trained from scratch and end-to-end using both web-scale data and annotated images by considering all labels as text, therefore unifying supervision for learning representations through natural language. It can achieve state-of-the-art performance with few-shot transfer or by minimal task-specific adaptation on a wide range of downstream vision-and-language tasks, including visual recognition, multimodal understanding, crossmodal retrieval, and image captioning. SimVLM (Wang et al., 2021b) is pre-trained with prefix language modeling on datasets with weak supervision. It exhibits its efficacy on few-shot captioning tasks. Even though all these models mentioned above can already achieve improvement on some few-shot tasks, how to exploit their few-shot reasoning ability using limited training examples still deserves the effort. In this work, we study this direction via the lens of prompt learning utilizing CLIP as a starting point.

3 Counterfactual Prompt Learning

3.1 Problem Formulation

Our goal is to learn generalizable prompt representation with limited data. The prompt in CLIP is divided into two parts: task-agnostic prompt $\boldsymbol{p}$ and task-relevant prompt $\boldsymbol{h}$. Task-agnostic prompt $\boldsymbol{p}$ is learned end-to-end automatically. The set of task-relevant prompts $\mathbb{H} = \{\boldsymbol{h}_0, \boldsymbol{h}_1, \ldots, \boldsymbol{h}_C\}$ is mapped from the label space $\mathbb{Y}$ with some predefined rules hinging on the task type, where C is the total number of classes. The final prompt $\boldsymbol{t}_c$ is the concatenation of the task-agnostic prompt and the task-relevant prompt fed into CLIP's text encoder: $\boldsymbol{t}_c = [\boldsymbol{p}, \boldsymbol{h}_c]$.

Existing works to this problem (Zhou et al., 2021, 2022) propose to first extract visual feature $\boldsymbol{v}$ of each input image by feeding it into CLIP's vision encoder F; and text embeddings are generated by feeding $\{\boldsymbol{t}_c\}_{c=1}^{C}$ into the CLIP's text encoder G. The probability of i-th class is computed as

$$p(\boldsymbol{t}_i \mid \boldsymbol{x}) = \frac{e^{\frac{<G(\boldsymbol{t}_i), \boldsymbol{v}>}{\tau}}}{\sum_{c=1}^{C} e^{\frac{<G(\boldsymbol{t}_c), \boldsymbol{v}>}{\tau}}}, \quad (1)$$

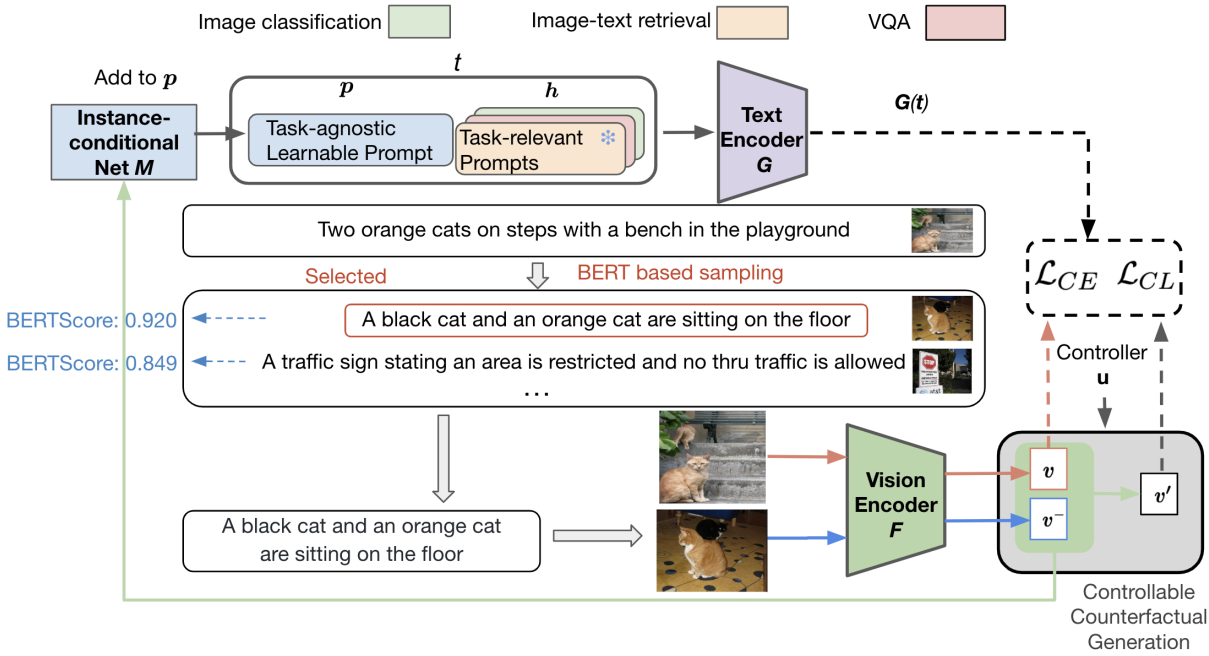

Figure 2: The counterfactual prompt learning framework. We freeze the vision encoder F and the text encoder G, and only optimize the task-agnostic prompts and the instance-conditioned net M (blue blocks). Please refer to Section 3.2 for the explanation.

where τ is the temperature parameter, $< \cdot >$ denotes the cosine similarity. Cross-entropy loss is then minimized and the gradients can be back-propagated via the text encoder G to update the learnable prompt representation $\boldsymbol{p}$. During training, the weights of CLIP always remain frozen. During inference, Eq. 1 is used to compute the probability for each class.

3.2 Method Overview

An overview of the Counterfactual Prompt Learning (CPL) framework is shown in Figure 2. For pre-processing, we construct task-relevant prompts for all training samples. The goal is to optimize the task-agnostic prompt $\boldsymbol{p}$.[2] During training, given a positive image-prompt pair, we first perform *text-based negative sampling* to find the most semantically-similar negative sample based on text similarity scores. Then we adopt a *controllable counterfactual generation* strategy to construct the counterfactual from the positive and negative samples in the visual feature space. Finally, we perform contrastive learning using both generated counterfactual image features and factual image features in a joint optimization framework to fine-tune the task-agnostic prompt $\boldsymbol{p}$, allowing the model to understand non-spurious semantic information and learn generalized prompt representations.

3.3 Controllable Counterfactual Generation

By viewing image feature $\boldsymbol{v}$ as a potential cause of the label, a non-spurious feature shall be a sufficient cause of the label. So we would like to generate counterfactuals by identifying minimal non-spurious feature change that causes the label change. The illustration of the counterfactual construction process is shown in Figure 3. Given positive image features $\boldsymbol{v}$ and negative image features $\boldsymbol{v}^-$, we can generate negative counterfactual image features $\boldsymbol{v}'$ as below:

$$\boldsymbol{v}' = (1 - \mathbf{u}) \circ \boldsymbol{v} + \mathbf{u} \circ \boldsymbol{v}^-, \tag{2}$$

where $\circ$ is the element-wise multiplication and $\mathbf{u}$ is the parameter controlling the amount of negative image feature that replaces the positive image feature. The negative image features are extracted from those images similar to the original image at the semantic level, which we will introduce in Section 3.4.

To capture the non-spuriousness, we would like to construct counterfactuals by replacing essential non-spurious features only. This can be achieved by minimizing the amount of feature change $\mathbf{u}^*$ to the original image that can causally incur label

[2]Together with the instance-conditional net $\boldsymbol{M}$ as introduced in Zhou et al. (2022). For simplicity, we will only use $\boldsymbol{p}$ hereafter as $\boldsymbol{p}$ and $\boldsymbol{M}$ are always optimized together.

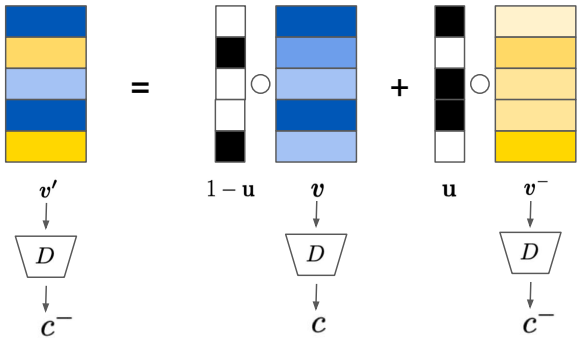

Figure 3: Counterfactual generation process. $\boldsymbol{v}$ and c are the positive image feature and label, while $\boldsymbol{v}^-$ and c^- are the negative image feature and label. $\circ$ is element-wise multiplication. By mixing $\boldsymbol{v}$ and $\boldsymbol{v}^-$, the counterfactual image feature $\boldsymbol{v'}$ is predicted as a negative label c^- by the discriminator D. $\mathbf{u}$ is minimized so a minimal change to the positive image feature $\mathbf{u}$ is captured here to causally change the label.

change:

$$\begin{aligned} \underset{\mathbf{u}^*}{\text{minimize}} \quad & \|\mathbf{u}^*\|_1 \\ \text{s.t.} \quad & \mathbf{u}^* = \arg\max_{\mathbf{u}} D_{c^-}(\boldsymbol{v'}). \end{aligned} \tag{3}$$

Given the factual and counterfactual features $\boldsymbol{v}$ and $\boldsymbol{v'}$, we aim to learn the prompt that can help CLIP better align visual features $\boldsymbol{v}$ and textual features $G(\boldsymbol{t})$ with same semantic meanings. This can be achieved by maximizing the mutual information (MI) between $\boldsymbol{v}$ and $G(\boldsymbol{t})$. Therefore, by minimizing the InfoNCE loss (Hjelm et al., 2018), we can maximize the lower bound on MI$(\boldsymbol{v}, G(\boldsymbol{t}))$. To this end, we define the contrastive objective function based on the InfoNCE estimator following Khosla et al. (2020):

$$\mathcal{L}_{CL}(\boldsymbol{p}, \mathbf{u}^*) = -log(\frac{e^{\frac{S(\boldsymbol{v}, G(\boldsymbol{t}))}{\tau}}}{e^{\frac{S(\boldsymbol{v}, G(\boldsymbol{t}))}{\tau}} + e^{\frac{S(\boldsymbol{v'}, G(\boldsymbol{t}))}{\tau}}}), \tag{4}$$

where $S(\cdot, \cdot)$ is normally the cosine similarity function and τ is the temperature value.

3.4 Text-based Negative Sampling

We then discuss how to perform negative sampling for constructing counterfactual features. As suggested in Robinson et al. (2020), good negative samples have different labels and are difficult to be distinguished from an anchor point, while their semantic representations are close (Suresh and Ong, 2021). Since not all negative samples can serve as useful negatives (Chuang et al., 2020), indiscriminate leverage of these data may harm model robustness and algorithm efficiency. Therefore, during training, in each batch, we only utilize the most semantically-similar one to generate counterfactual image features. Other image samples are filtered out.

Semantic concepts may be highly complex in the visual representations, and thus it is hard to directly measure semantic similarity in the visual space. While language is more expressive and naturally preserves semantic meanings. Therefore, we propose a text-based negative sampling method. We first measure the text similarity between prompts with BERTScore (Zhang et al., 2019), which computes pairwise cosine similarity between reference sentences and candidate sentences using BERT contextual embedding (Devlin et al., 2019). We compute a similarity matrix with the value of each element being:

$$\text{sim}(i, j) = \text{BERTScore}(\boldsymbol{h}_i, \boldsymbol{h}_j). \tag{5}$$

Denote $\mathcal{B}$ as the collection of sampled instances. During training, each prompt $\boldsymbol{h}_c \in \mathcal{B}$ $(1 \leq c \leq C$, where C is the size of sampled instances) can be treated as a query. Given a query prompt $\boldsymbol{h}_q$, its most semantically similar prompt (the one with the highest BERTScore) $\boldsymbol{h}_k$ is searched from $\mathcal{B}$. Then we use the CLIP vision encoder to obtain the features of the corresponding positive and negative images $\boldsymbol{v}$ and $\boldsymbol{v}^-$.

3.5 Joint Optimization

In addition to the contrastive learning loss as introduced in Eq. 4, we also adopt the standard cross-entropy loss for training:

$$\mathcal{L}_{\text{CE}}(\boldsymbol{p}) = -\sum_{c} \boldsymbol{y}_c \log p\left(\boldsymbol{t}_c \mid \boldsymbol{x}\right), \tag{6}$$

where $\boldsymbol{y}_c$ denotes the one-hot ground-truth annotation of the label. We treat all downstream tasks in this work as classification tasks, where the model predicts if the image and text prompt pair is matched or not.

Then the task-agnostic prompt $\boldsymbol{p}$ is learned by minimizing the weighted combination of contrastive learning loss and cross-entropy loss:

$$\mathcal{L}(\boldsymbol{p}) = \mathcal{L}_{CE}(\boldsymbol{p}) + \lambda \cdot \mathcal{L}_{CL}(\boldsymbol{p}, \mathbf{u}^*), \tag{7}$$

where λ determines the weight of $\mathcal{L}_{CL}$.

In fact, we can seek to put Eq. 3 and Eq. 7 in a single-stage optimization framework. The intuition is that we generate counterfactual image

Algorithm 1 Counterfactual Prompt Learning

$\mathbb{X}$: image space
$\mathbb{Y}$: label space
$\boldsymbol{h}_c$: task-relevant prompt for the c-th class
$\mathbb{H}$: the set of task-relevant prompts
$\boldsymbol{p}$: the task-agnostic prompt
$\boldsymbol{v}$: image features
$\boldsymbol{v}^-$: negative image features
$\mathbf{u}$: parameter controls the generation of counterfactual image features
function $\mathcal{CPL}(\mathbb{X}, \mathbb{Y})$
    $\mathbb{H} \leftarrow \mathbb{Y}$
    $\boldsymbol{t}_c \leftarrow [\boldsymbol{p}, \boldsymbol{h}_c]$
    for each i, j **do**
        $\text{sim}(i, j) = \text{BERTScore}(\boldsymbol{h}_i, \boldsymbol{h}_j)$ ▷ Eq. 5
    end for
    for q in the batch **do**
        $\boldsymbol{v} \leftarrow \boldsymbol{v_q}$
        Find the index k that maximize $\text{sim}(q, k)$ with the given index q
        $\boldsymbol{v}^- \leftarrow \boldsymbol{v_k}$
        Generate counterfactual image features ▷ Eq. 2
        $\mathcal{L}_{CE} \leftarrow$ cross-entropy loss ▷ Eq. 6
        $\mathcal{L}_{CL} \leftarrow$ contrastive loss ▷ Eq. 4
        Update $\boldsymbol{p}$ and $\mathbf{u}$ with the joint optimization loss ▷ Eq. 7
    end for
end function

features with minimal feature change that can maximize the negative prediction probability, and at the same time, utilize contrastive learning to learn the prompt that can guide CLIP to explicitly distinguish between factual images and counterfactual images. Putting all pieces together, we have:

$$\begin{aligned} \underset{\boldsymbol{p}, \mathbf{u}^*}{\text{minimize}} \quad & \mathcal{L}_{CE}(\boldsymbol{p}) + \lambda \cdot \mathcal{L}_{CL}(\boldsymbol{p}, \mathbf{u}^*) + \|\mathbf{u}^*\|_1 \\ \text{s.t.} \quad & \mathbf{u}^* = \arg\max_{\mathbf{u}} D_{c^-}(\boldsymbol{v'}) \\ \text{where} \quad \boldsymbol{v'} & = (1 - \mathbf{u}) \circ \boldsymbol{v} + \mathbf{u} \circ \boldsymbol{v}^-. \end{aligned} \tag{8}$$

In Eq. 8, the gradients can be back-propagated all the way through the text encoder G to the task-agnostic prompt, making use of the rich knowledge encoded in the pre-trained CLIP model to optimize the prompt.

Algorithm 1 presents the learning algorithm of CPL. In summary, given few input training samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, CPL consists of three main steps: (1) compute the similarity matrix between different text prompts within the sampled batch; (2) generate counterfactual image features; (3) optimize $\boldsymbol{p}$ and $\boldsymbol{u}$ with contrastive learning loss and cross-entropy loss.

3.6 Task-relevant Prompt Construction

We construct task-relevant prompts $\mathbb{H}$ for image classification, image-text retrieval, and visual question answering, respectively. For image classification, the prompts are class labels for each task; for image-text retrieval, captions for each image are adopted as prompts; for visual question answering, we first use a pre-trained generative T5 model (Raffel et al., 2019) to convert the question-answer pairs into declarative sentences referring to the VQA prompt generation method proposed in Song et al. (2022b). Then, motivated by Wei et al. (2022), we add additional category information into the prompt generated from templates based on the question type to help the model perform intermediate reasoning steps. Specifically, we add "The question is asking about others" for *Other* questions before the generated declarative sentence. In a similar vein, "The question is asking about yes or no" and "The question is asking about numbers" are added for *Yes/No* and *Number* questions.

4 Experiments

4.1 Tasks and Datasets

Image Classification. We employ seven publicly available image classification datasets used in CLIP: SUN397 (Xiao et al., 2010), Caltech101 (Griffin et al., 2007), ImageNet (Deng et al., 2009), OxfordPets (Parkhi et al., 2012), StandfordCars (Krause et al., 2013), Flowers102 (Nilsback and Zisserman, 2008), and Food101 (Bossard et al., 2014). These datasets constitute a comprehensive benchmark, which covers a diverse set of vision tasks including the classification of generic objects, fine-grained image recognition, action classification, etc. To evaluate the generalization ability of methods, we split those datasets into seen and unseen classes. Only images in the seen classes will be used for training. The setting follows the few-shot evaluation protocol in CLIP, where we use 16 shots for training and full test sets for testing.

Image-Text Retrieval. We consider two datasets for image-text retrieval: MSCOCO (Lin et al., 2014) and Flickr30K (Plummer et al., 2015). We adopt the widely used Karpathy split (Karpathy and Fei-Fei, 2015) for both the MSCOCO and Flickr30K datasets, where MSCOCO contains 113/5K/5K for train/validation/test. Flickr30K contains 29K/1K/1K images for train/validation/test. We construct few-shot setting subsets for both CoCoOp and CPL by taking 0.5%, 1%, and 3% of training instances. We train the model with the subsets and evaluate its performance on the complete

Classes	Method	SUN397	Caltech101	ImageNet	OxfordPets	StanfordCars	Flowers102	Food101	Average
Seen	CLIP	69.40	96.51	72.46	91.33	74.85	72.17	90.12	80.98
	CoCoOp	79.08 [+13.95]	97.66 [+1.19]	76.01 [+4.90]	95.18 [+4.22]	70.91 [-5.26]	**94.65** [+31.15]	90.67 [+0.61]	86.31 [+6.58]
	CPL (ours)	**81.05** [+16.79]	**97.70** [+1.23]	**78.81** [+8.76]	**96.69** [+5.87]	**75.51** [+0.88]	93.91 [+30.12]	**93.01** [+3.21]	**88.10** [+8.79]
Unseen	CLIP	75.40	94.10	68.09	97.04	74.95	**77.87**	91.30	82.68
	CoCoOp	76.83 [+1.90]	93.92 [-0.19]	70.44 [+3.45]	97.78 [+0.76]	73.09 [-2.48]	69.24 [-11.08]	91.53 [+0.25]	81.83 [-1.02]
	CPL (ours)	**80.19** [+6.35]	**94.94** [+0.89]	**73.17** [+7.46]	**98.81** [+1.82]	**78.90** [+5.27]	72.30 [-7.15]	**93.44** [+2.34]	**84.54** [+2.25]

Table 1: Result comparison between CPL and CoCoOp (Zhou et al., 2022) on seen and unseen classes across seven image classification datasets in terms of accuracy (%) under the few-shot setting. The relative difference (%) compared with CLIP is reported in color.

Training data used	Method	Flickr30k	MSCOCO	Average
0	CLIP	83.00	53.35	68.18
0.5%	CoCoOp	82.40 [-0.72]	55.55 [+4.12]	68.98 [+1.17]
	CPL (ours)	**85.64** [+3.18]	**57.91** [+8.55]	**71.78** [+5.28]
1%	CoCoOp	84.80 [+2.17]	56.62 [+6.13]	70.71 [+3.71]
	CPL (ours)	**86.91** [+4.71]	**58.43** [+9.52]	**72.67** [+6.59]
3%	CoCoOp	85.90 [+3.49]	58.08 [+8.87]	71.99 [+5.59]
	CPL (ours)	**87.74** [+5.71]	**59.96** [+12.39]	**73.85** [+8.32]

Table 2: Result comparison between CPL and CoCoOp on two image-text retrieval datasets, Flickr30k (Plummer et al., 2015) and MSCOCO (Lin et al., 2014), on the unseen test sets in terms of Recall@1 (%). The relative difference (%) over CLIP is reported in color.

Training data used	Method	VQAv2
0	CLIP	11.83
0.5%	CoCoOp	27.98 [+136.52]
	CPL w/o. Category Information	31.68 [+167.79]
	CPL	**33.39** [+182.25]
1%	CoCoOp	28.51 [+141.00]
	CPL w/o. Category Information	34.70 [+193.32]
	CPL	**35.66** [+201.44]
3%	CoCoOp	30.18 [+155.11]
	CPL w/o. Category Information	35.41 [+199.32]
	CPL	**36.32** [+207.02]

Table 3: Result comparison on the VQAv2 dataset (Goyal et al., 2017a) in terms of accuracy (%). The relative improvements over CLIP are reported in color. Incorporating category information into task-relevant prompts can further improve the performance.

test set. We use Recall at 1 (R@1) as the default evaluation metric.

Visual Question Answering. VQAv2 (Goyal et al., 2017b) is an extended dataset from the VQA (Antol et al., 2015) dataset. The questions are categorized into three types: *Number*, *Yes/No*, and *Other*. We set up the experiments following Anderson et al. (2018), which treats visual question answering as a classification problem: for each question, the model picks the corresponding answer from a given set of predefined most frequent candidate answers and matches it with the image. The questions are first converted into a masked template using the pre-trained T5 model and predefined rules. The infilled template along with the questions will be turned into prompts that naturally connect questions and answers. The model will predict whether the given prompt and image pairs are matched. We construct the few-shot setting by taking 0.5%, 1%, and 3% instances for training.

4.2 Implementation Details

Baselines. We mainly compare CPL with CoCoOp (Zhou et al., 2022), one of the earliest prompt tuning methods proposed for vision-and-language pre-trained models. CoCoOp considers each input image and injects the learnable instance-aware tokens into the context vectors as the final prompt. For a fair comparison, both CPL and CoCoOp adopt CLIP (Radford et al., 2021) as the pre-trained vision-and-language backbone and are compared with respect to their relative improvements over zero-shot CLIP.

Prompt Tuning. The task-agnostic prompt is randomly initialized from a zero-mean Gaussian distribution with the standard deviation 0.02, where we set length $L = 4$ by default. For vision and language tasks, in contrast to image classification, where an image is labeled by a category, the task-relevant prompts comprise more fine-grained details, usually a sentence. We here similarly tokenize the whole sentence using the CLIP word embedding (Radford et al., 2021), and feed the tokenized results to the text encoder with task-agnostic prompt vectors, to generate the language embedding for each prompt. In both the image-text retrieval and visual question answering, all data in the test set can be treated as belonging to unseen classes.

4.3 Main Results

Image Classification. The experimental results for image classification are shown in Table 1. With better prompts learned from counterfactual examples, our CPL method achieves clear advantages over CoCoOp for both seen and unseen classes across almost all datasets. Particularly on unseen classes, we gain an average relative improvement of 3.55%.

Meanwhile, CoCoOp shows its poor generalization ability. Specifically, we found that CoCoOp performs worse than CLIP on StandfordCars on both seen and unseen classes, and on Caltech101 and Flower102 on unseen classes, indicating that it tends to learn and leverage spurious relations and could not generalize well on unseen classes in some cases. We believe all these mentioned above can be sufficient evidence that the main idea of CPL, learning non-spurious prompt representation can aid CLIP adapting at test time, is practical.

Image-Text Retrieval. Table 2 reports results on image-text retrieval on the unseen test set. CPL can beat the zero-shot CLIP consistently across the three different settings, demonstrating that CPL can also learn better prompt representation and more effectively exploit the limited amount of data on image-text retrieval. Meanwhile, CoCoOp performs even worse than CLIP on Flickr30k using 0.5% training data, which suggests that a tiny quantity of training data for image-text retrieval can lead to spurious prompt representation if using naïve instance-conditional prompt tuning method.

Visual Question Answering. For visual question answering, the results are shown in Table 3. As can be seen, CPL surpasses the baseline CoCoOp with a relative improvement of up to 25.08% when using 1% instances for training. This proves the concept that CPL can be effective on more complicated vision-and-language tasks. In fact, visual question answering is more challenging for zero-shot CLIP which is pre-trained for image-text matching. During pre-training, CLIP sees most sentences similar to captions in image-text retrieval and those captions can be directly used as prompts; while for VQA, question-answer pairs have to be adapted into declarative prompts. Therefore, zero-shot CLIP has poor performance on VQA, but few-shot prompt tuning via CPL can help reduce the prompt domain gap significantly. Apart from the vanilla CPL method, we examined another variant

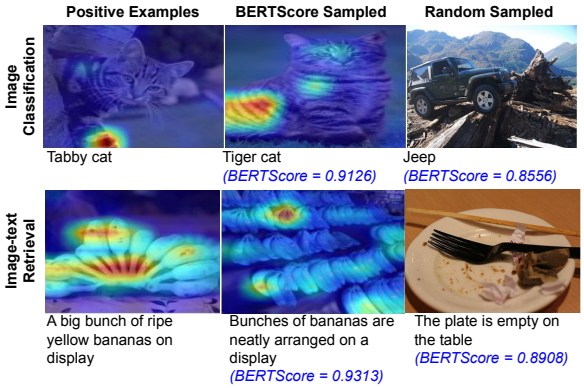

Figure 4: Visualization of the weights of the controller parameter $\mathbf{u}$ on images. The first column is the original positive examples; the second column is BERT-sampled negative examples; the third column is randomly-sampled negative examples for comparison. The BERTScore between the text prompts of positive examples and sampled examples are shown at the bottom.

of CPL where we do not add additional category information into the prompt (denoted as CPL w/o. Category Information), the results indicate that constructing task-relevant prompts by adding categorical information contributes to the improvement.

4.4 Ablation Analysis

Negative Sampling. We compare the random sampling vs. BERTScore sampling over ImageNet for image classification, MSCOCO for image-text retrieval, and VQAv2 for visual question answering in Table 4. With more challenging negative examples, BERTScore sampling leads to more effective prompt tuning and overbeats random sampling on all three tasks. The qualitative visualizations of the two sampling strategies are shown in Figure 4, from which it can be seen that BERTScore-sampled images are much more semantically similar to the original images.

Non-spurious Feature Visualization. We visualize the heatmap of the learned non-spurious feature weights in the image level in Figure 4. The weights are mainly centralized on the semantically meaningful regions that are aligned to the text prompts.

Number of Shots in Image Classification. We then study the effects of the number of shots on CPL for image classification. Following the few-shot evaluation protocol adopted in CLIP, we use 4, 8, and 16 shots for training on ImageNet. From Figure 5, increasing the number of shots keeps improving the performance of both two methods

Method	ImageNet	MSCOCO	VQAv2
Random sampling	75.28	57.78	33.01
BERTScore sampling	**76.02**	**58.43**	**35.66**

Table 4: Random sampling vs. BERTScore sampling for CPL over three tasks. On ImageNet, we measure the average accuracy across seen and unseen classes. On MSCOCO and VQAv2, we both use 1% instances for few-shot learning.

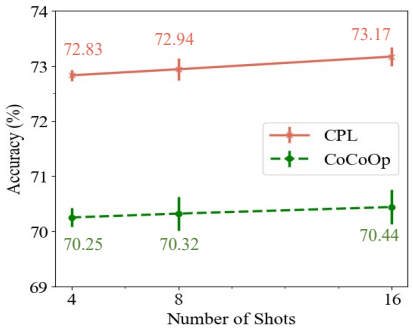

Figure 5: Accuracy comparison on ImageNet (Deng et al., 2009) unseen classes under three different shots. CPL performs better than CoCoOp consistently and has lower standard errors.

on unseen classes. Meanwhile, CPL outperforms CoCoOp under the three different settings and has lower standard errors.

Contribution of Contrastive Learning. In Section 3, we use the coefficient λ to weigh the contrastive learning loss and combine it with the cross-entropy loss. It is observed that the scale of contrastive learning loss is smaller, hence we try to use a larger λ to balance the two loss terms. Figure 6 shows the average accuracy result across seen and unseen classes on the SUN397 dataset under four different λ values. Note that when λ is zero, there is no contribution from the contrastive loss and the method actually learns the prompt using standard cross-entropy loss. From experimental results obtained on the SUN397 dataset, we can observe that using $\lambda = 1$ leads to the best performance.

5 Conclusion

In this paper, we propose a Counterfactual Prompt Learning (CPL) framework to avoid time-consuming prompt engineering and learn more generalizable prompt representation for vision and language models. We conduct abundant experiments on seven widely used image classification datasets, two image-text retrieval datasets, and one visual question answering dataset. Our proposed CPL

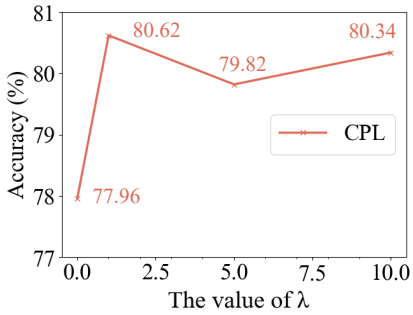

Figure 6: Ablation of four different λ values on the SUN397 dataset in terms of average accuracy (%). The performance of CPL peaks at $\lambda = 1$.

method outperforms the previous prompt tuning baseline and the zero-shot CLIP across the three tasks. In the future, we plan to develop more sophisticated methods based on CPL and extend CPL to other vision and language tasks.

Limitations

There are fairness issues in large pre-trained vision and language models such as CLIP. The proposed prompt learning method in this study automatically learns the prompt and does not address those issues in the pre-trained model. Considering the method is proposed for the few-shot setting, careful inspection and tuning are also needed when testing our method on other biased datasets. The methodologies proposed in Booth et al. (2021) and Wang et al. (2021a) may possibly be paired with CPL to potentially address the issues. Another limitation is the absence of explainability in CPL, which is a common problem with existing soft prompt tuning methods. Back-mapping tuned soft prompts representation to natural language is a way for interpretation; however, due to the limited size of vocabulary used by CLIP during the training, prior methods such as searching for the nearest words in the embedding space can not accurately match the vector to natural language. Expanding the dictionary size for CLIP embedding or developing more advanced back-mapping techniques can possibly address the limitation.

Acknowledgments

We would like to thank the support of the Google Ads Faculty Research Award. We also thank the anonymous reviewers for their thought-provoking comments. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the sponsor.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

M Besserve, A Mehrjou, R Sun, and B Schölkopf. 2020. Counterfactuals uncover the modular structure of deep generative models. In *Eighth International Conference on Learning Representations (ICLR 2020)*.

Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. 2021. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 268–277.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. 2020. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Tsu-Jui Fu, Xin Eric Wang, Matthew F Peterson, Scott T Grafton, Miguel P Eckstein, and William Yang Wang. 2020. Counterfactual vision-and-language navigation via adversarial path sampler. In *European Conference on Computer Vision*, pages 71–86. Springer.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020. Mutant: A training paradigm for out-of-distribution generalization in visual question answering. *arXiv preprint arXiv:2009.08566*.

Tejas Gokhale, Abhishek Chaudhary, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2021. Semantically distributed robust optimization for vision-and-language inference. *arXiv preprint arXiv:2110.07165*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017a. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017b. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Gregory Griffin, Alex Holub, and Pietro Perona. 2007. Caltech-256 object category dataset.

Han Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2021. Text generation with efficient (soft) q-learning. *arXiv preprint arXiv:2106.07704*.

Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. 2022. Parameter-efficient fine-tuning for vision transformers. *arXiv preprint arXiv:2203.16329*.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

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

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2021. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. 2022. Declaration-based prompt tuning for visual question answering. *arXiv preprint arXiv:2205.02456*.

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

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

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. 2021. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022a. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. 2022b. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*.

Varsha Suresh and Desmond C Ong. 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *arXiv preprint arXiv:2109.05427*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021a. Assessing multilingual fairness in pre-trained multimodal representations. *arXiv preprint arXiv:2106.06683*.

Yixin Wang and Michael I Jordan. 2021. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021b. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. 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.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiangji Zeng, Yunliang Li, Yuchen Zhai, and Yin Zhang. 2020. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. 2022. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiheng Zhang, Yongkang Wong, Xiaofei Wu, Juwei Lu, Mohan Kankanhalli, Xiangdong Li, and Weidong Geng. 2021. Learning causal representation for training cross-domain pose estimator via generative interventions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11280.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. *arXiv preprint arXiv:2203.05557*.

Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual off-policy training for neural dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3438–3448.