Supplementary Material for Prefix Conditioning

Abbreviation	Dataset	#Concepts	Train size	Test size	Source link
Food	Food-101	102	75,750	25,250	Tensorflow
CF10	CIFAR-10	10	50,000	10,000	Tensorflow
CF100	CIFAR-100	100	50,000	10,000	Tensorflow
VOC	VOC2007 classification	20	5,011	4,952	Tensorflow
DTD	Describable Textures	47	3,760	1,880	Tensorflow
Pets	Oxford-IIIT Pets	37	3,680	3,669	Tensorflow
Cal	Caltech-101	102	3,060	6084	Tensorflow
Flower	Oxford Flowers 102	102	1,020	6,149	Tensorflow
Patch	PatchCamelyon	2	294,912	32,768	Tensorflow
ESTAT	EuroSAT	10	N/A	27,000	Tensorflow
R45	Resisc45	45	N/A	31,500	Tensorflow

Table 1. Statistics of datasets used in zero-shot and linear probe.

A. Experimental Details

Dataset. Table 1 describes the statistics of dataset used for evaluation. We pick the test datasets based on UniCL [4] and availability in Tensorflow dataset. We use the test set to evaluate zero-shot recognition and linear probe while the train set is used to train a linear classifier. Note that since EuroSAT and Resisc45 utilize the training split for evaluation, we exclude the two datasets from linear probe evaluation. Also, since Oxford Flowers do not have many training samples (10 samples per class), we exclude the dataset from the evaluation too.

Data Augmentation. Following UniCL [4], only random cropping is applied to train all models for a fair comparison.

Computation. We use 32 Nvidia Tesla V100 GPUs to train all models. 4 nodes, where each node has 8 GPUs, are used to run experiments.

B. Additional Results

Attention Visualization. Fig. 1 visualizes attention weights for the class *forest area*, where a prompt template, *a tatto of*, is employed. The model focuses on a word, *forest* when prompt prefix is employed. In other two cases, the model also pays much attention to *tatoo* probably because the word should provide useful information to distinguish a sentence from others for image-caption contrastive learning. Fig. 2 represents attention for a real caption from CC3M. While the model conditioned with caption prefix and unconditional model attend to several words through many layers, the model conditioned with prompt prefix shows clear attention only in the first layer. Since the prompt-conditioned model has never seen the real caption during training, it fails in attending to discriminative words.

Class Name Shift. Test samples can be unseen with respect to image classification data in two ways (or combinations of two): 1) The image is similar to training distribution, but the class name used for testing is different from the image classification label. 2) Although the class label is the same, the image data comes from the different distributions. Datasets evaluated in the zeros-shot recognition include both two cases since class names and images are from different domains. 2) is analyzed in Subsection 4.3 of the main paper, *Robustness in image domain shift*. We analyze 1) by evaluating the recognition performance of ImageNet-1K by changing its class name from the one used during training. We find a synonym for each class with WordNet [1], where we exclude synonyms substantially similar to the original class name and obtain synonyms

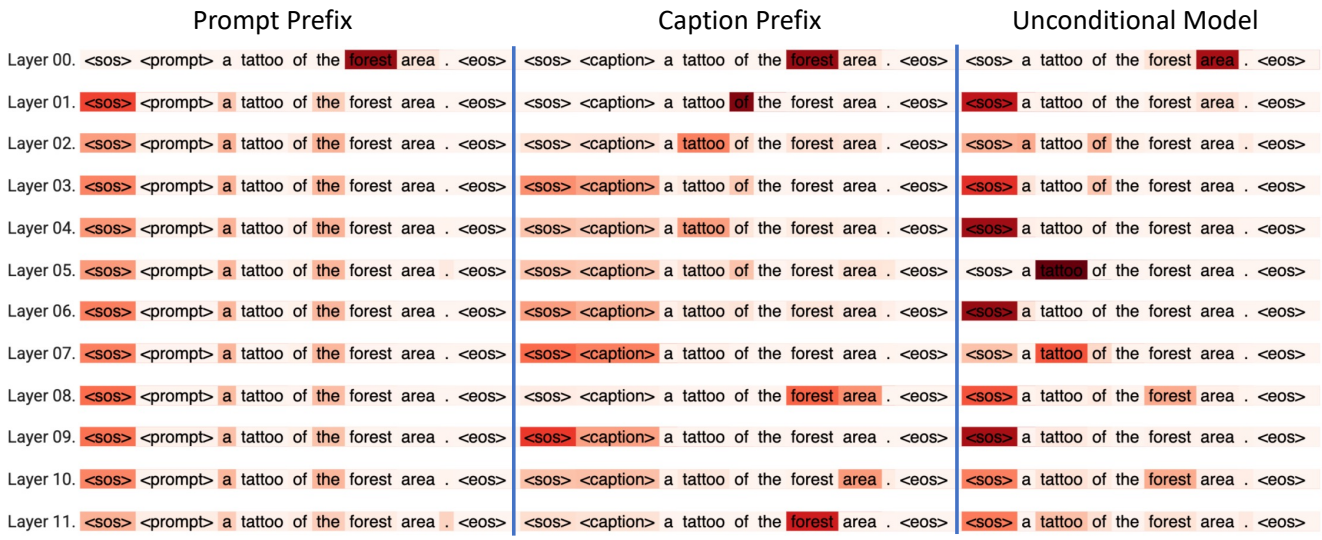


Figure 1. Attention visualization for a *class prompt*. Note that the attention weights are for and end token. Best viewed in color. The class name shown here is one of class prompts in the EUROSAT dataset. Different rows show the weights of different transformer layers. With a prompt prefix (leftmost), the model focuses on a class name (*forest area*) while caption prefix (middle) allows a model to pay attention to another noun, *tattoo*. By prefix conditioning, the attention of the model changes as intended.

Figure 2. Attention visualization for a *real caption*. Note that the attention weights are for and end token. Best viewed in color. The sentence shown here is from CC3M. Different rows show the weights of different transformer layers. Caption prefix conditioning helps to attend to many words while prompt conditioning fails to do that.

for 525 classes. Then, we use the synonym to classify images during evaluation. Since the input image distribution does not vary, we can evaluate the performance on the class name shift. If the model is robust to the change in the class name, the degrade in the performance should be small.

The first 6 rows of Table 2 describe the models trained with the original class names and evaluated on both original ones and synonyms, and the last two rows represent a model trained with synonyms, where the original class names are replaced with synonyms. Prompt prefix outperforms caption prefix with a large margin in testing with class names used in training time. Generally, caption prefix performs better when tested with the class names different from the ones used during training. Prompt prefix is tailored to handle class names employed during training time while caption prefix enables the language encoder to extract more general representations.

Interestingly, the choice of class names seems to significantly change the generalization as shown in the comparison between a model trained with synonyms and original class names. The original model decreases the accuracy more than 30% by changing the class name while the model trained with synonym decreases less than 20%.

Train Data	Train on Synonym	Prefix Training	Test-Time Prefix	Original		Synonym	
				top-1	top-5	top-1	top-5
IN1K + CC12M			N/A	69.3	89.3	31.2	49.5
IN1K + CC12M		✓	Prompt	**75.0**	**92.9**	**38.3**	54.8
IN1K + CC12M		✓	Caption	71.4	91.6	36.6	**56.7**
IN21K + CC12M			N/A	54.5	83.2	23.1	43.9
IN21K + CC12M		✓	Prompt	**69.9**	**92.4**	32.1	53.7
IN21K + CC12M		✓	Caption	65.3	90.6	**33.5**	**56.9**
IN21K + CC12M	✓	✓	Prompt	54.4	78.6	**70.8**	**92.8**
IN21K + CC12M	✓	✓	Caption	**54.5**	**82.6**	59.0	86.1

Table 2. Evaluation on the robustness to the class name shift using ImageNet-1K. *Original* refers to the subset of ImageNet-1K classes while *synonym* refers to their synonyms taken from Wordnet. The last two rows indicate the models trained with the synonyms, thus showing superior performance on *synonym* whereas degrading performance on *Original*.

Prefix Training	Test-time Prefix	CC3M				COCO			
		I2T@1	I2T@5	T2I@1	T2I@5	I2T@1	I2T@5	T2I@1	T2I@5
	N/A	21.8	47.4	21.0	45.7	23.9	49.5	18.7	43.2
✓	Prompt	13.1	31.3	8.1	21.8	17.2	38.1	16.8	37.7
✓	Caption	**22.6**	**47.5**	**21.6**	**46.1**	**24.7**	**49.7**	**19.7**	**43.9**

Table 3. Image-text retrieval results on CC3M and COCO. The performance is evaluated on the subset of CC3M and validation set of COCO. All models are trained on CC12M and ImageNet-21K. Caption conditioning (last row) slightly improves retrieval performance compared to the unconditional model (first row). Since prompt conditioning (middle) tailors a model for class-prompt, it fails to extract discriminative information from real captions.

Image-Caption Retrieval. In Table 3, we evaluate the performance of image-caption retrieval using the subset of CC3M (12288 pairs of image and caption) and COCO validation set (5000 pairs of image and caption), where all models are trained with CC12M and ImageNet-21K. First, our model (last row) slightly performs better than the model without conditioning (first row). Second, prompt prefix conditioning (second row) significantly performs worse than caption prefix conditioning (last row). Since the prompt prefix conditioning specializes a model for the class name prompts of ImageNet21K, the conditioning does not generalize well to real captions.

Larger Batch-size and Training Epochs. We examine the effect of increasing batch-size and training epochs in Table 4. In CLIP, increasing the batch-size and training epochs improves the performance of both ImageNet-1K and zero-shot recognition. On the other hand, the zero-shot performance of UniCL is not benefited from training with longer epochs (compare last and second to last row). UniCL attempts to ensure the invariance of images from the same classes by supervised contrastive loss while CLIP does not consider it. However, such invariance is not necessarily required in zero-shot recognition, which leads to the degraded performance.

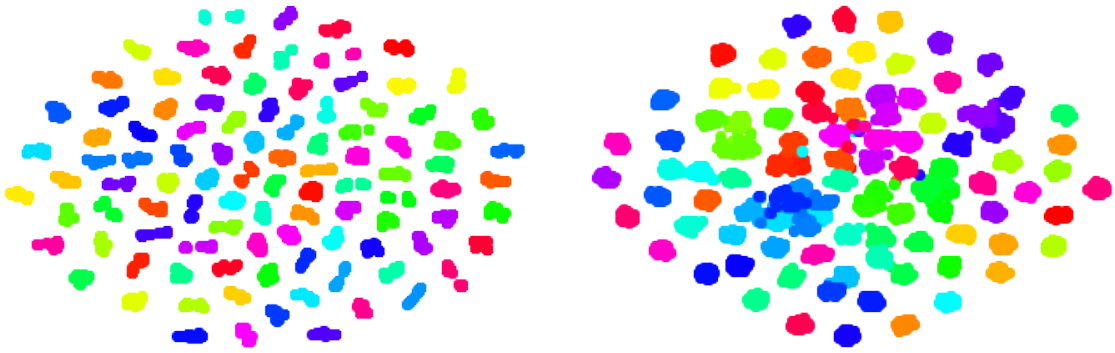
Comparison to Reported UniCL's Results. In the main paper, we provide our reproduced results of UniCL, which is based on our implementation, since the authors have not released the code and did not report the numerical accuracy of each zero-shot recognition. In this paragraph, we compare our approach and the reported performance of UniCL [4] and K-Lite [2] by aligning several hyper-parameters, e.g., batch-size and training epochs, using ImageNet-1K. When using ImageNet-22K and CC-15M for training, our method (batch-size:4096, training epochs: 30) shows 73.9 while UniCL (batch-size:4096, training epochs 32) reports 71.5. When using ImageNet-21K excluding ImageNet-1K and CC-15M, our method (batch-size:1024, training epochs 30) shows 49.7 whereas UniCL (batch-size: 4096, training epochs: 32) and K-Lite (batch-size: 4096, training epochs: 32) perform 46.6 and 48.7 respectively according to K-Lite results (See last two rows of Table 3 in [2]). These results suggest that our method performs better than the reported numbers of UniCL and K-Lite in ImageNet-1K. Also, the knowledge augmentation technique proposed by K-Lite can be complementary to our approach, thus combining two approaches is an interesting research direction.

T-SNE visualization for language features. Fig. 3 visualizes extracted language features (ImageNet-1K) conditioned with different prefixes. The prompt-prefix (left) has lower intra-class and higher inter-class variance, whereas caption-prefix (right) shows higher intra-class variance across prompts.

T-SNE visualization for image features. Fig. 4 visualizes image features from ImageNet-1K (blue) and CC3M (red).

Training Data		Objective	Batch-size	Epochs	Metric	
Classification	Caption				IN-1K	Zero-shot 11 datasets
ImageNet-21K	CC-12M	CLIP	1024	15	67.3	57.8
ImageNet-21K	CC-12M	CLIP	1024	30	**69.1**	**58.3**
ImageNet-22K	CC-15M	CLIP	1024	15	69.3	58.5
ImageNet-22K	CC-15M	CLIP	4096	15	71.1	59.5
ImageNet-22K	CC-15M	CLIP	4096	30	**72.2**	**59.8**
ImageNet-22K	CC-15M	UniCL	1024	15	69.7	58.5
ImageNet-22K	CC-15M	UniCL	4096	15	70.3	**60.4**
ImageNet-22K	CC-15M	UniCL	4096	30	**73.9**	58.9

Table 4. Performance comparison among different batch-size and training epochs. ImageNet-22K denotes the combination of ImageNet-21K and ImageNet-1K, CC-15M indicates that of CC-12M and CC-3M.



(a) Prompt conditioned (b) Caption conditioned

Figure 3. T-SNE [3] visualization of the class-prompt features of ImageNet-1K with different prefix conditions. Different colors indicate language embeddings of different classes. Prompt conditioning extracts more class discriminative representations than caption conditioning.

Since ImageNet-1K is object-centered while CC3M covers more diverse scenes, the distributions are separated. This is consistent across baseline (w/o conditioning) and our method (with conditioning).

Comparison between unconditioned and conditioned model by language features. Fig. 5 visualizes language features of ImageNet-1K class prompts (Blue) and CC3M captions (Red) for unconditioned (left) and conditioned (right) respectively. Note that the conditioned model utilizes prompt prefix for class prompts and caption prefix for real captions respectively. As seen from the visualization, unconditioned model cannot distinguish some prompts from captions of CC3M. This is probably because some captions are similar to class prompts of ImageNet. By contrast, the conditioned model differentiate class prompts from captions better than unconditioned model due to the prefix conditioning.

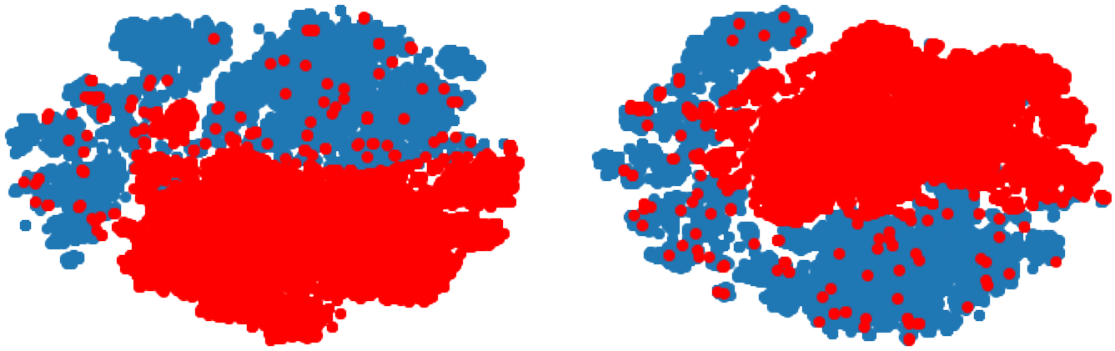
References

[1] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1

[2] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Anna Rohrbach, Zhe Gan, Lijuan Wang, Lu Yuan, et al. K-lite: Learning transferable visual models with external knowledge. *arXiv preprint arXiv:2204.09222*, 2022. 3

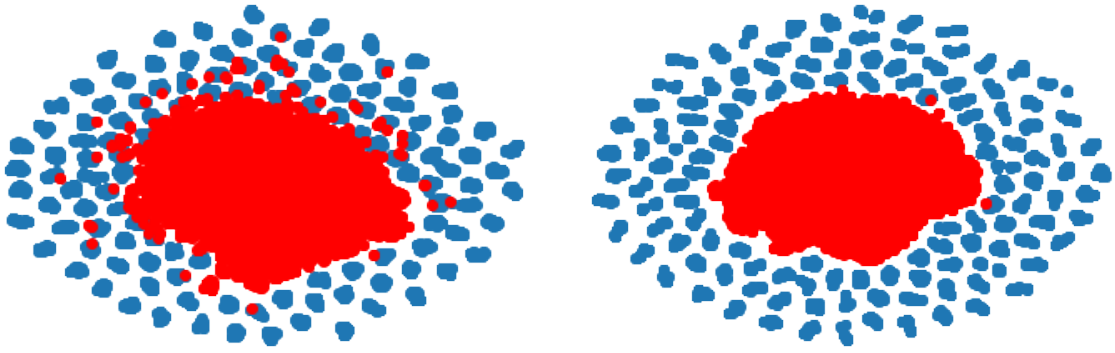
[3] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4, 5

[4] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. *arXiv preprint arXiv:2204.03610*, 2022. 1, 3



(a) W/O conditioning (b) With conditioning

Figure 4. T-SNE [3] visualization of the image features of ImageNet-1K (blue) and CC3M (red). Since ImageNet-1K is object-centered while CC3M covers more diverse scenes, the distributions are separated. This is consistent across baseline (w/o conditioning) and our method (with conditioning).



(a) Unconditioned model (b) Conditioned model

Figure 5. T-SNE [3] visualization of language features of ImageNet-1K class prompts (Blue) and CC3M captions (Red) for unconditioned (left) and conditioned (right) respectively. Our proposed condition better differentiates prompts from real captions.