Few-Shot Composition Learning for Image Retrieval with Prompt Tuning

Junda Wu*[1], Rui Wang*[2], Handong Zhao†[3], Ruiyi Zhang[3]
Chaochao Lu[4], Shuai Li[5], Ricardo Henao[2,6]
[1]New York University
[2]Duke University
[3]Adobe Research
[4]University of Cambridge
[5]Shanghai Jiao Tong University
[6]King Abdullah University of Science and Technology (KAUST)
jw6466@nyu.edu, {rui.wang16, ricardo.henao}@duke.edu,
{hazhao, ruizhang}@adobe.com, cl641@cam.ac.uk, shuaili8@sjtu.edu.cn

Abstract

We study the problem of composition learning for image retrieval, for which we learn to retrieve target images with search queries in the form of a *composition* of a reference image and a modification text that describes desired modifications of the image. Existing models of composition learning for image retrieval are generally built with large-scale datasets, demanding extensive training samples, *i.e.*, query-target pairs, as supervision, which restricts their application for the scenario of few-shot learning with only few query-target pairs available. Recently, prompt tuning with frozen pretrained language models has shown remarkable performance when the amount of training data is limited. Inspired by this, we propose a prompt tuning mechanism with the pretrained CLIP model for the task of few-shot composition learning for image retrieval. Specifically, we regard the representation of the reference image as a trainable visual prompt, prefixed to the embedding of the text sequence. One challenge is to efficiently train visual prompt with few-shot samples. To deal with this issue, we further propose a self-supervised auxiliary task via ensuring that the reference image can retrieve itself when no modification information is given from the text, which facilitates training for the visual prompt, while not requiring additional annotations for query-target pairs. Experiments on multiple benchmarks show that our proposed model can yield superior performance when trained with only few query-target pairs.

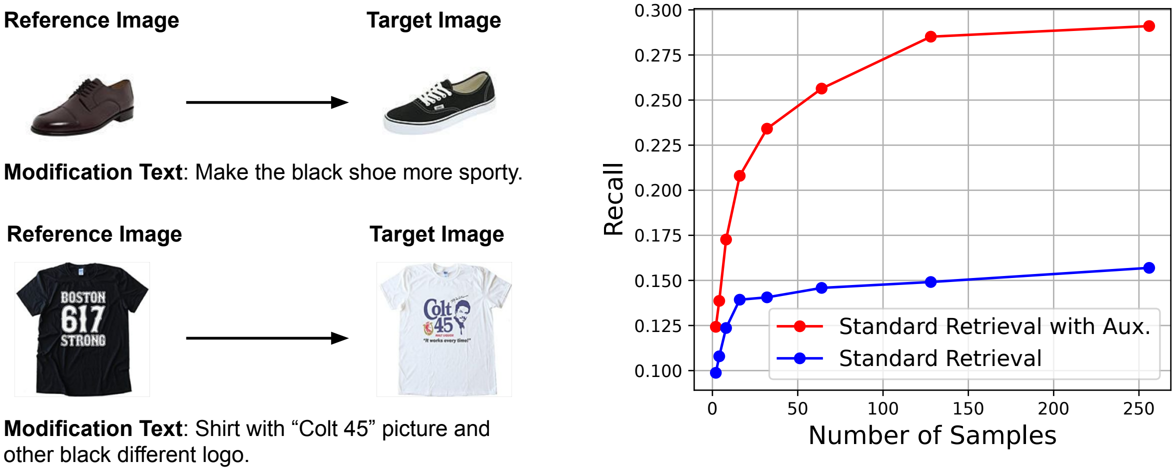
Introduction

The task of image retrieval generally involves searching for a *target* image given a user specified search *query* (Vo et al. 2019). The query can be formulated with different modalities, *e.g.*, images or text used as inputs in a recommendation system. We consider the case for which the query is composed of a reference image and modification text from the user, describing the expected modification between the reference and the target images. This scenario can be categorized as *composition learning* (Vo et al. 2019; Chen, Gong, and Bazzani 2020), since the query is constituted with multiple modalities (visual and language). For instance, in the first example of Figure 1a, we expect to modify the reference image (black shoe) with the modification text "make the black shoe more sporty.". The target image should be more sporty in style, while retaining the black color. Previous works generally consider high-resource training of retrieval models in a large scale, *i.e.*, with a substantial amount of user generated query-target pairs to train in a supervised manner (Vo et al. 2019; Chen, Gong, and Bazzani 2020). However, collecting data for image retrieval with coherent query-target pairs is usually expensive and time-consuming (Chen et al. 2021b), since the user data may not be available in abundance, especially for composition learning that demands both the image and text to constitute a valid query. Therefore, in this paper, we study the problem of few-shot composition learning for image retrieval. Specifically, we consider the scenario where only a few query-target pairs available for training.

Recently, prompt tuning for pretrained language models (Chen et al. 2021a,c; Liu et al. 2021b; Schick and Schütze 2020) has produced state of the art results for low-resource down stream tasks, for which the amount of training data is usually limited. Inspired by this, we propose a prompt tuning mechanism based on the pretrained CLIP model (Radford et al. 2021) for the task of few-shot composition learning for image retrieval. Specifically, we represent the reference image and modification text as a composed query sequence, by encoding the reference image as a trainable visual prompt, prefixed to the embedding of the text sequence. Then, retrieval is enabled via projecting the target image and the composed query sequence into a shared hidden space, with a target encoder and a query encoder, respectively. We implement the target encoder with the CLIP image encoder, and the query encoder based on the CLIP text encoder. With the visual prompt, the vision-language information in the query of our task can be conveniently encoded with the query encoder from the composed query sequence, as a normal text sequence. Additionally, this allow us to leverage the pretrained knowledge from the CLIP text encoder (during query encoding) to solve for the multi-modal interaction between

*These authors contributed equally.

†Corresponding Author



(a) Examples

(b) Retrieval Performance

Figure 1: (a) Examples of query-target pairs for the task of composition learning for image retrieval. (b) Retrieval performance with varying number of query-target pairs for training, measured by R@50 (higher the better, detailed in experiments) for dataset CelebA. *Standard Retrieval* is naively training with only (2). *with Aux.* means additionally training with the auxiliary loss (5). In experiment sections, we report results of training with $\{2, 4, 8, 16\}$ number of samples to simulate few-shot scenarios.

reference image and the modification text. To reduce the number of trainable parameters for few-shot learning while retaining the pretrained knowledge, we freeze the pretrained parameters of the CLIP encoders and train with the visual prompt embedding, which is in analogy with previous works of prompt tuning (Li and Liang 2021; Liu et al. 2022).

One challenge of the above prompt tuning mechanism is to efficiently train the visual prompt embeddings with only few samples. The vision-language information provided with the few-shot training data might not be sufficient in training a high quality projection from the reference image into the input semantic space of the query encoder (implemented with CLIP text encoder), so that the query encoder can recover the information of the reference image from the projected visual prompt and interact with the text for accurate retrieval. In our experiments, we empirically show that the visual prompt embeddings naively trained with the few-shot datasets, *e.g*, with the standard retrieval loss (2), can be spuriously correlated with non-informative tokens from the text, yielding inferior evaluation performance, as shown in Figure 1b. To deal with this problem, we propose to augment the training for the visual prompt embedding with a self-supervised auxiliary task, which encourages a reference image to retrieve itself when no modification information is given from the text. In this way, the visual prompt can be more sufficiently trained with the self-supervised vision-language information, *i.e.*, with the augmented queries. In Figure 1b, we show that our auxiliary can remarkably improve the retrieval performance. The contribution of this work can be summarized as follows:

- We investigate the task of few-shot composition learning for image retrieval, which so far has been under-explored by existing works.
- We propose a prompt tuning mechanism based on the CLIP model, along with a self-supervised auxiliary task that facilitates the training of the visual prompt embedding for the reference image, remarkably improving the performance of prompt tuning.
- Experimental results show that our method can produce state-of-the-art performance of image retrieval when only few training samples are available.

Related Work

Composition Learning for Image Retrieval

For the task of image retrieval (Ma et al. 2022; Jia et al. 2020), composition learning refers to the combined feature encoding for queries of different modalities, *e.g.*, images and text. Composition learning has become increasingly popular among recent works of image retrieval (Vo et al. 2019; Perez et al. 2018; Chen, Gong, and Bazzani 2020; Kim et al. 2021; Anwaar, Labintcev, and Kleinsteuber 2021; Han et al. 2017), owning to the fact that it enables a more comprehensive analysis of the user intent from the query data. However, the training of existing models of composition learning for image retrieval generally requires large-scale datasets. As mentioned above, such demands may not always be feasible. Therefore, we study few-shot composition learning for image retrieval, *i.e.*, the retrieval model is trained with only few query-target pairs.

We note that there are a few works regarding few-shot learning for image retrieval with single-modality query, *e.g.*, either images (Triantafillou, Zemel, and Urtasun 2017; Zhong, Chen, and Qian 2020; Wang, Gui, and Hebert 2017) or text (Chang et al. 2020). Among these, (Wang, Gui, and Hebert 2017) studies few-shot hash learning for hash-based image retrieval. This is out of the scope of our work, since we do not study hash-based image retrieval, where the focus is to learn hash functions for efficient feature matching during retrieval. (Zhong, Chen, and Qian 2020; Wang, Gui, and

Hebert 2017; Chang et al. 2020) study *metric-based* few-shot learning, *e.g.*, learning regularized feature space, for the task of image retrieval with single-modality query. In our work, we propose a new architecture with learnable prompt embeddings, *i.e.*, *architecture-based*, which is orthogonal to *metric-based* few-shot learning (Liu et al. 2021a). Additionally, we study image retrieval with composed query, which is different from those with single modality query. For instance, when the composed query contains a reference image and text regarding its expected modification, it would be inappropriate to learn with either the image or text query alone, since they contain complementary information that should be composed together during query encoding.

Prompt Tuning

Prompt tuning has become an increasingly popular approach for fine tuning with large pretrained models, due to its simplicity and efficiency when applied to resource-restricted tasks (Liu et al. 2021a). One branch of the works use hard prompts, *e.g.*, handcrafted template words for classification (Gao, Fisch, and Chen 2020; Hu et al. 2021), named entity recognition (Cui et al. 2021), *etc*. Recently, studies have shown that training with soft prompts (Hambardzumyan, Khachatrian, and May 2021; Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Liu et al. 2022), *i.e.*, trainable embedding vectors, can also yield state of the art performance, without the engineering required for template-word selection as with hard prompts. However, there are limited works exploring the use of prompt tuning, either hard or soft, for the image retrieval tasks with few-shot samples.

Composition Learning for Image Retrieval

Problem Formulation

The task of composition learning for image retrieval can be formulated as retrieving a target image I_t from a database B, given a query (I_r, S), with I_r and $S = [t_1, \cdots, t_n]$ denoting a reference image and the modification text sequence, respectively. I_t should be similar to I_r, while satisfying the modification specified in the text S. Figure 2 shows our general framework of training with composition learning for image retrieval. Our model is trained with both the retrieval loss L_{ret} and the auxiliary loss L_{aux}. We first explain the general architecture of our retrieval model with the retrieval loss L_{ret}. Then, we introduce the auxiliary loss L_{aux} for efficient training with few-shot datasets. The computation for the auxiliary loss L_{aux} is the same as L_{ret}, except with different input data.

Training for Image Retrieval

As mentioned in previous sections, for composition learning, we first construct a composed query sequence, denoted as S_q, via encoding the reference image I_r into a visual prompt embedding *ref* and prefix it to the text embedding sequence $[t_1, \cdots, t_n]$. Formally, S_q is defined as,

$$S_q = [ref; t_1, \cdots, t_n]. \tag{1}$$

For the convenience of notation, we do not differentiate the name of a token and its embedding. The compose query sequence query S_q and the target image I_t are encoded by query encoder E_q and the target encoder E_t, respectively, with their corresponding outputs, $f_q = E_q(S_q)$ and $f_t = E_t(I_t)$, lying in a shared hidden space. During training, we sample a negative dataset B_n, *s.t.*, $B_n = \{I_n | I_n \in B, I_n \neq I_t\}$, which is not shown in Figure 2 for simplicity. For each image I_n from B_n, we also encode I_n with the target encoder, $f_n = E_t(I_n)$. Let m be a thresholding value and $\mathcal{D}(\cdot, \cdot)$ denotes the cosine distance. The model of image retrieval can be trained with the following triplet loss (Vo et al. 2019; Chen, Gong, and Bazzani 2020; Yu et al. 2020; Anwaar, Labintcev, and Kleinsteuber 2021),

$$\begin{aligned} L_{ret} = L_{tpl}((I_r, S), I_t; B_n) = \frac{1}{|B_n|} \sum_{I_n \in B_n} \max(0, \\ \mathcal{D}(f_n, f_q) - \mathcal{D}(f_t, f_q) + m), \end{aligned} \tag{2}$$

such that encoded f_t and f_q are close to each other in the shared hidden space, while f_n and f_q should be apart from each other since $I_n \neq I_t$. L_{tpl} denoted the triplet loss. For inference, we can retrieve K images $\{I_b^k\}_{k=1}^K$ from the database B, who are the nearest neighbors of the encoded query representation f_q in the shared hidden space, *i.e.*, for each candidate image $I_b \in B$,

$$\mathcal{D}(E_t(I_b), f_q) \geq \mathcal{D}(E_t(I_b^k), f_q), \ k = 1, \cdots, K. \tag{3}$$

There can be different choices of E_q and E_t. In the proposed approach, we leverage the pretrained CLIP model, consisting of an image encoder C_I and a text encoder C_T, which are pretrained with a shared output space for the image and text. Specifically, we adopt the pretrained CLIP image encoder as our target encoder, *i.e.*, $E_t = C_I$. Our query encoder E_q is implement based on C_T, which will be explained later. $\{t_i\}_{i=1}^n$ are encoded with the pretrained CLIP embedding layer. To retain the knowledge from pretraining and reduce the number of trainable parameters for few-shot learning, *i.e.*, avoid overfitting, all the pretrained encoders and layers are fixed during training.

Auxiliary for Visual Prompt Tuning

The visual prompt *ref* is a projected embedding from the reference image I_r. In order to leverage the pretrained knowledge of CLIP for *ref* to be representative of I_r, as in Figure 2, we first encode I_r with the pretrained CLIP image encoder C_I. Then, we project with a linear layer l, *i.e.*,

$$ref = w \cdot f_r + b, \quad f_r = C_I(I_r). \tag{4}$$

$\{w \in \mathbb{R}^{d \times d}, b \in \mathbb{R}^d\}$ is the set of learnable parameters for the linear layer l (fully connected), which is purposed at projecting f_r from the output space of C_I into the input space of E_q (implemented with the CLIP text encoder), while keeping the semantics of I_r. However, as mentioned in the Introduction, along with the empirical results (Figure 1b), the information provided in the few-shot training data might not be sufficient to capture such a projection, resulting in biased correlation between *ref* and text tokens (Figure 3). In this paper, we propose an self-supervised auxiliary task to augment the training, so that the parameters of l,

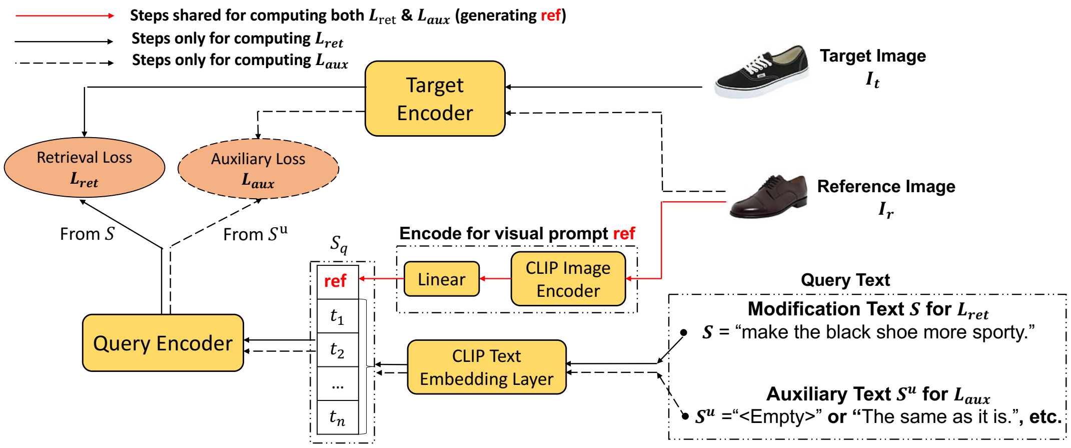

Steps shared for computing both L_{ret} & L_{aux} (generating ref)
Steps only for computing L_{ret}
Steps only for computing L_{aux}
Target Encoder
Target Image I_t
Retrieval Loss L_{ret}
Auxiliary Loss L_{aux}
Reference Image I_r
From S
From S^{u}
S_q
Encode for visual prompt ref
ref
Linear
CLIP Image Encoder
Query Text
t_1
t_2
...
t_n
Modification Text S for L_{ret}
S = “make the black shoe more sporty.”
Query Encoder
CLIP Text Embedding Layer
Auxiliary Text S^u for L_{aux}
S^u =“<Empty>” or “The same as it is.”, etc.

Figure 2: Our proposed framework of composition learning for image retrieval. ”<*Empty*>” means a text sequence of zero length. $\{t_i\}_{i=1}^n$ are token embeddings encoded from either S or S^u. *ref* is the visual prompt embedding, prefixed to $\{t_i\}_{i=1}^n$. The target encoder and query encoder can be designed with different choices. Here, we implement them with the pretrained CLIP image and text encoder, respectively. The computation for the loss L_{ret} and L_{aux} are the same except with different input data. Specifically, L_{ret} is computed with $\{(I_r, S), I_t\}$ and L_{aux} is computed with $\{(I_r, S^u), I_r\}$. The output from the query encoder is used to compute L_{ret} if we input with S, and compute L_{aux} if input with S^u. For simplicity, we omit the negative images I_n from the dataset B_n and positional encoding. All the CLIP encoders and layers remain fixed during training. We use the black arrows to indicate steps only for computing L_{ret} and dashed black arrows to indicate steps only for computing L_{aux}. The red arrows indicate steps that are shared for computing L_{ret} and L_{aux}, which are also steps for generating *ref*.

$\{w \in \mathbb{R}^{d\times d}, b \in \mathbb{R}^d\}$, can be learned more efficiently with the few-shot dataset.

Our intuition is that a valid model of composition learning for image retrieval should retrieval the original reference image when there is no information given from the text in the query, regarding modifications on the reference image. For example, when the text is given as *”Exactly what I want.”*, *”The same as it is.”*, or ”$<$ *Empty* $>$” (A zero length text sequence), the model should retrieval the reference image itself, since no modification is requested by the text. We denote such sentences/sequences that do not contain any modification information as the auxiliary text S^u. S^u can be handcrafted without referring to the reference image, as in the examples above. For a reference image I_r, we define our auxiliary loss as retrieving I_r, given S^u. Formally, we can define an augmented query-target pair $\{(I_r, S^u), I_r\}$, with which the auxiliary loss is,

$$L_{aux} = L_{tpl}((I_r, S^u), I_r; B_n^u) \tag{5}$$

B_n^u is a negative dataset for the auxiliary task, *s.t.*, $B_n^u \subset B$, $I_r \notin B_n^u$. Compared with the retrieval loss L_{ret} in (2), we replace the modification text S and the target image I_t with our auxiliary text S^u and the reference image I_r, respectively. Our auxiliary task is self-supervised with I_r, since we do not require either the target image I_t or the modification text S that should be coherent with I_r.

Remarks: How L_{aux} result in better *ref*? We can understand our proposed auxiliary task with (5) as a process of autoencoding, where we learn to encoded I_r (or f_r) into embedding *ref*, such that the query encoder E_q can decode from *ref* and recover the information of I_r via retrieving it from B. With this perspective, l has to project f_r (as *ref*) into the input space of E_q, *i.e.*, space of the pretrained text tokens, while retaining its semantics, otherwise E_q cannot correctly recover the information of I_r. Then, when the modification text S is given, E_q can more accurately capture the semantics of I_r from *ref*, so that the retrieved image can retain similar semantics as I_r except for the modification specified in S. In Figure 3, we also show the our auxiliary task can result in better alignment between *ref* and the text tokens.

Selection of S^u. Empirically, we found training with S^u=”$<$ *Empty* $>$” (zero-length sequence) have comparable performance as with S^u =*”Exactly what I want.”* or *”The same as it is.”*, etc. Therefore, for computational efficiency, we report results with S^u = ”$<$ *Empty* $>$” in the experiments. Note that tokenized text sequence of S^u = ”$<$ *Empty* $>$” with CLIP tokenizer is not of length zero, *i.e.*, still including stecial tokens of *[CLS]* and *[SEP]*. We also report results with S^u =*”Exactly what I want.”* or *”The same as it is.”*, etc., in the supplementary.

Query Encoding with E_q

The goal of composition encoding is to learn a fused vision-language representation from I_r and S, that is close to the embedding of the target image, $E_t(I_t)$, in the hidden space. We follow the framework of TIRG (Vo et al. 2019), which can be described as a residual process, whose results is computed from the outputs of a residual encoder and a gate encoder. We implement both the residual encoder and gate encoder using the pretrained CLIP text encoder C_T (frozen during training), with their inputs, S_{res} and S_{gate}, respectively, defined as,

$$S_{res} = [p_{res}; ref; t_1, \cdots, t_n] \tag{6}$$

$$S_{gate} = [p_g\, ; ref; t_1, \cdots, t_n], \tag{7}$$

$q_k \in \mathbb{R}^d$, $k = \{res, g\}$, is a single trainable plugin prompt to adapt C_T as the residual and gate encoder, respectively.

Since $q_k \in \mathbb{R}^d$, $k = \{res, g\}$ are small in size ($d = 512$), they will not overparameterize the model, which leads to overfitting with few-shot learning. In the Supplement, we exam the functionality of the $q_k \in \mathbb{R}^d$, $k = \{res, g\}$, showing that they are useful for few-shot learning. Since the focus of this paper is to effectively train the visual prompt embeddings in the few-shot scenario, please refer to the original TIRG paper (Vo et al. 2019) for details of the residual encoder and a gate encoder.

The Overall Objective for Training

For each query-target pair $\{(I_r, S), I_t\}$ from the few-shot dataset, we construct an auxiliary pair $\{(I_r, S^u), I_r\}$. Then, the model can be trained with both the retrieval loss (2) and the auxiliary loss (5). In this way, L_{aux} improves the samples efficiency of few-shot trainining, by taking full advantage of the few-shot dataset, allowing for more efficient training. The overall objective L for training is a combination of the original retrieval task in (2) with L_{ret} and our auxiliary task in (5) with L_{aux}, trained over samples of the few-shot datasets,

$$L = L_{ret} + \beta L^{u}_{tpl}, \tag{8}$$

where β is a balancing parameter. Since our proposed method is based on visual prompt tuning with the pretrained CLIP encoders, we denote our method as PromptCLIP.

Experiments

Datasets

We evaluate our proposed model on three datasets: FashionIQ (Guo et al. 2019), CelebA (Liu et al. 2015) and B2W (Forbes et al. 2019). FashionIQ consists of fashion images from 3 categories: Dresses, Shirts and Tops&Tees. CelebA has 202k face images from 10k identities. Each face image is annotated with 40 binary attributes. B2W consists of $3.5k$ images of birds with very long natural language description (with the average length of 35 words) about the query image and the target image.

Baselines and Ablations

We compare our method with several baselines: FiLM (Perez et al. 2018), VAL (Chen, Gong, and Bazzani 2020), Concat (Santoro et al. 2017), TIRG (Vo et al. 2019), ComposeAE (Anwaar, Labintcev, and Kleinsteuber 2021), DCNet (Kim et al. 2021) and Frozen (Tsimpoukelli et al. 2021). For a fair comparison, we use CLIP's (Radford et al. 2021) text encoder as the text feature extractor for all the baselines. Since FiLM and VAL are specifically designed on the CNN architecture, we do not change the model architectures for their visual encoders. For FiLM (Perez et al. 2018) and VAL (Chen, Gong, and Bazzani 2020), the CNNs are initialized by ResNet-50 (He et al. 2016) pretrained from ImageNet (Deng et al. 2009). For Concat (Santoro et al. 2017), TIRG (Vo et al. 2019), ComposeAE (Anwaar, Labintcev, and Kleinsteuber 2021), DCNet (Kim et al. 2021) and Frozen (Tsimpoukelli et al. 2021), we use the image encoder CLIP as the image feature extractor. All the pretrained encoders are fixed during training. Note that Frozen (Tsimpoukelli et al. 2021) is not designed for retrieval tasks. We adapt it for our task by 1) Encode the target image with C_T. 2) Encode the reference image and text with the multi-modal Frozen encoder as the query encoder 3) Train for image retrieval with (2).

In addition to the baselines above, we also compare PromptCLIP with the following ablation of our method:

- w/o L_{aux}: We disable L_{aux} by setting $\beta = 0$.
- w/o Linear: We delete the linear layer l that project f_r (output from C_I) to *ref* (input from E_q). This is to show that it is necessary to include the learnable l, since the output space of C_I and the input space of E_q (implemented with C_T) is different.

Implementation Details

Please note that we conduct hyperparameter sensitive analysis in the Appendix. We implement all the baselines and our method with PyTorch. The feature dimensions for both the text encoder and the image are 512. We set the margins m in Eq. 2 to 0.02. PromptCLIP's trainable prompts are initialized from a Gaussian distribution with zero-mean and the standard deviation of 0.02. The balancing parameter β in Eq. 8 is set to 0.5. The samples in the triplet sets are shuffled for each epoch. We use Adam (Kingma and Ba 2014) optimizer with the learning rate of 5×10^{-5} with a decay rate of 0.95. The negative dataset B_n and B^u_n for each training sample is the set of target or referenece images in the few-shot dataset, excluding the ground truth target or referenece image from the training sample itself, respectively, for (2) and (5).

We follow the few-shot setting in (Radford et al. 2021; Zhou et al. 2021) by sampling N-shot pairs of samples from each category of the dataset as the training data. We report on the numbers of shot $N = \{2, 4, 8, 16\}$. Each experiment is repeated 5 times with different random seeds and we report the mean and the standard error. Please refer to Supplement for more details.

Evaluation

Following previous works (Vo et al. 2019; Chen, Gong, and Bazzani 2020; Kim et al. 2021; Anwaar, Labintcev, and Kleinsteuber 2021), we use the recall at Rank K (R@K) as the evaluation metric, representing the likelihood that the ground truth image is included in the top K retrieved candidate images. For some datasets, the recall value may be near zeros when K is too low. Therefore, different datasets should be evaluated with different K values. Following (Anwaar, Labintcev, and Kleinsteuber 2021; Vo et al. 2019; Chen, Gong, and Bazzani 2020), we report results with different k of the R@k metrics for different datasets. Note that in the Tables 7, 8&9 of the supplement, we have also reported the results with k values that are not shown Tables 1, 2&3 in the main paper, respectively.

Visualization of Attentions

To investigate how our auxiliary loss L_{aux} affect the performance of image retrieval, we visualize how the query encoder E_q perceive the learnt *ref* embedding with and without L_{aux}, denoted as PromptCLIP and w/o L_{aux}. As mentioned above, the query encoder is composed of a residual encoder

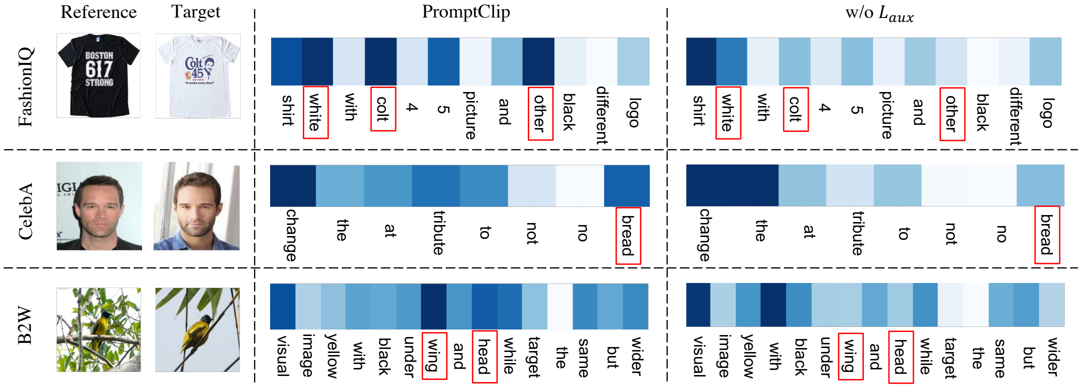


Figure 3: Attentions from the learned embedding of the reference image (*ref*) to text tokens, extracted from the last layer of the CLIP text encoder C_T. We input $[ref]$ appended by the text tokens into C_T. The shown attention values are *softmax* normalized attention scores, averaged over all the attention heads. Each row corresponds to the attentions of the same pair of the reference image and the text description. L_{aux} is the loss for our auxiliary task.

	R@5				R@10			
	2	4	8	16	2	4	8	16
FiLM	0.2±0.0	0.2±0.0	0.1±0.0	0.2±0.0	0.4±0.0	0.3±0.0	0.3±0.0	0.3±0.0
VAL	1.7±0.1	1.6±0.0	1.6±0.1	2.1±0.2	2.5±0.2	2.5±0.1	2.7±0.1	3.4±0.2
Concat	0.8±0.1	2.3±0.1	7.5±0.1	18.5±0.6	1.6±0.2	4.1±0.3	12.3±0.4	27.7±0.5
ComposeAE	5.7±0.5	9.5±0.5	16.4±0.6	26.4±0.7	9.6±0.6	15.6±0.8	25.7±0.6	37.5±0.7
DCNet	6.6±0.4	10.4±0.6	16.7±0.6	25.4±0.5	10.6±0.6	15.8±0.9	24.7±0.9	37.2±0.7
TIRG	6.2±0.3	9.8±0.4	18.5±0.2	25.6±1.1	9.7±0.5	15.3±0.6	27.3±0.6	37.4±1.1
Frozen	1.0±0.2	2.4±0.5	11.8±1.5	28.5±1.0	1.7±0.2	4.2±0.7	18.9±1.9	39.4±2.3
PromptCLIP	**13.2**±1.4	**23.6**±1.9	**37.2**±0.2	**43.5**±0.5	**20.2**±1.7	**35.7**±2.0	**52.3**±0.7	**59.9**±0.7
w/o L_{aux}	10.0±0.4	16.4±1.3	29.1±0.5	37.9±0.4	16.4±0.5	26.7±2.2	42.6±1.3	53.0±0.8
w/o Linear	2.4±0.5	4.5±0.4	13.3±0.7	23.9±0.6	4.4±1.3	7.5±1.3	20.1±1.5	33.9±2.0

Table 1. CelebA: Results of few-shot learning

and a gated encoder, which are both implemented with the CLIP text encoder C_T, with *ref* and the text tokens $\{t_i\}_{i=1}^n$ included in the inputs. Specifically, for each of the residual and gated encoder (both implemented with frozen C_T), we extract the attention scores from *ref* to the text tokens, using the last layer of C_T. The attention scores are averaged over attention heads and normalize with Softmax. We use the reference and modification text as inputs. In Figure 3, we draw the the resulting attention that is averaged between the residual and gated encoder, where the key words in the text are shown with red boxes. We find that the resulting attention with our auxiliary task (*PromptClip*) has more focus on the key words, *e.g.*, larger attention on *tail* and *wider* from the visual prompt *ref*. On the contrary, the text tokens and the *ref* are not as well correlated according to the attention from *w/o* L_{aux}, *i.e.*, without the auxiliary task ($\beta = 0$). For instance, in the first example, the *ref* has less focus on the key token *tail*, which contains important information regarding the modification on the reference image. Further, in the second and third examples, *ref* mistakenly attend to the no-informative word *the* and *with*, respectively. Figure 3 demonstrates that our auxiliary can enable the learnt *ref* to be better correlated it with the text tokens.

Results

We summarize results on the three considered datasets in Table 1, 2 and 3. We can observe that our proposed PromptCLIP consistently outperforms other baselines for all the three datasets, with a significant margin. It is also notable that most of the baselines, except Frozen, are not effective when being trained with only few samples on FashionIQ, *i.e.*, producing very low results. Here, we answer the following questions:

- *Why results with most of the baselines are very low on FashionIQ?* CelebA and B2W are datasets of human face and birds, respectively, *i.e.*, both are datasets of the same species. On the contrary, FashionIQ is a dataset of clothes, in which clothes of different categories, *e.g.*, for different parts of the body, may exhibit large variance in the shape or texture. With such a variance, it is more difficult to represent the data distribution of FashionIQ with only few samples. Thus, compared with CelebA and B2W, FashionIQ is more difficult for few-shot training.
- *Why Frozen is much higher than other baselines on FashionIQ?* This is caused by the following two factors: *a)* FashionIQ contains many absolute text descriptions

	R@10				R@50			
	2	4	8	16	2	4	8	16
FiLM	0.4±0.0	0.5±0.0	0.4±0.0	0.6±0.0	1.9±0.9	2.0±0.9	2.0±1.0	2.5±1.2
VAL	1.2±0.1	1.4±0.1	1.7±0.1	2.4±0.1	3.9±1.3	4.7±1.3	5.4±1.7	6.4±1.7
Concat	0.5±0.0	0.6±0.0	0.7±0.0	1.2±0.0	2.7±1.1	2.6±1.0	3.6±1.4	5.5±2.3
ComposeAE	2.8±0.0	3.1±0.2	3.5±0.2	4.0±0.1	9.2±3.6	10.4±4.1	10.9±4.0	12.2±3.9
DCNet	1.9±0.0	2.1±0.2	2.3±0.1	2.9±0.1	9.3±3.6	10.4±4.1	10.9±4.0	12.6±3.9
TIRG	1.9±0.0	2.0±0.2	2.3±0.1	2.9±0.0	9.3±3.7	10.4±4.2	10.9±4.0	12.8±4.2
Frozen	17.1±0.7	18.2±0.6	17.3±0.5	18.3±0.7	33.8±1.1	35.3±1.0	34.2±0.9	36.0±1.0
PromptCLIP	**18.9**±0.6	**19.8**±0.8	**19.9**±0.7	**21.2**±0.6	**36.9**±1.2	**38.5**±1.1	**38.4**±0.6	**40.4**±0.7
w/o L_{aux}	18.3±1.4	19.6±0.9	19.5±0.3	20.6±0.5	35.2±1.6	37.9±1.0	38.1±0.7	39.2±0.5
w/o Linear	16.2±0.3	17.0±0.3	16.3±0.2	17.3±0.1	31.7±0.2	32.6±0.6	32.3±0.6	32.6±0.6

Table 2. FashionIQ: Results of few-shot learning

	R@50				R@100			
	2	4	8	16	2	4	8	16
FiLM	2.7±0.3	2.4±0.4	3.3±0.3	3.3±0.3	6.2±0.5	5.8±0.7	7.3±1.0	7.5±0.7
VAL	4.2±0.3	5.8±0.4	7.5±0.3	8.4±0.4	10.6±0.8	11.9±0.5	13.8±0.2	15.3±0.5
Concat	3.2±0.2	3.5±0.2	4.9±0.2	5.9±0.5	7.8±0.3	8.6±0.5	11.1±0.4	12.3±0.6
ComposeAE	9.4±1.4	10.4±1.1	12.2±0.8	13.6±0.7	19.0±2.0	20.8±1.2	23.8±0.9	26.7±0.9
DCNet	9.4±1.5	10.1±1.1	12.9±0.8	14.4±0.7	19.3±1.9	20.8±1.3	23.7±1.2	25.6±0.8
TIRG	9.4±1.5	10.2±1.1	11.7±0.8	13.6±0.8	19.3±1.9	20.5±1.3	24.0±1.3	25.9±1.1
Frozen	6.4±0.5	7.1±0.9	7.7±0.7	9.8±1.2	16.0±0.8	15.5±0.8	15.4±0.8	18.5±1.5
PromptCLIP	11.7±1.3	**13.3**±1.5	**16.8**±0.7	**20.7**±1.2	**24.7**±1.4	**27.2**±1.3	**32.9**±1.2	**36.5**±0.7
w/o L_{aux}	9.1±1.2	8.5±0.9	9.7±0.4	11.9±0.6	19.9±1.3	18.9±0.8	19.7±0.6	23.5±0.9
w/o Linear	**12.9**±1.2	12.0±1.0	14.1±1.0	13.9±1.1	22.5±0.6	22.9±1.0	23.6±0.2	24.1±0.3

Table 3. B2W: Results of few-shot learning

about the target images, *i.e.*, we can infer the target image solely from the text in the query. For instance, in the first example of Figure 1 (a) of Supplementary, the query text has already contained the category information, ”polo”, and the color information, ”white”, with which we can retrieval the target image without referring to the reference image. *b)* There is no learnable parameters connecting the pretrained text features and the final output of the composition encoder (query encoder) of Frozen, while the pretrained image features are connected to the final output with linear layers. With these perspectives, when Frozen is trained with few samples in FashionIQ, the model can be learnt to disable the image information with those linear layers, so that the output from the final text encoder is only determined by the text input, which has been enough to get a decent retrieval performance.

On the other hand, from the ablation results, we find that the propose auxiliary task can make a remarkable improvement on the retrieval performance. We find that the improvement with L_{aux} is smaller on FashionIQ, compared with that on the other datasets. This may result from the fact that FashionIQ contains many absolute text (mentioned above), with which the model may have decent results without information of the reference image. For such case, improving the visual prompt with the auxiliary loss may not result in performance increase as significant as in the other two datasets. Additionally, the results from w/o Linear is usually significantly lower than PromptCLIP, indicating the importance of the linear layer l that projects from the outputs space of C_I into the input space of E_q.

Conclusions

In this paper, we study few-shot composition learning for image retrieval, a problem that is practical and currently under-explored by the existing literature. To alleviate the problem of overfitting as a result of few-shot learning, we propose a prompt tuning mechanism base on the pretrained CLIP image and text encoders. Specifically, we encode the reference image as a visual prompt, prefixed to the text embedding sequence. Additionally, we propose an auxiliary task that faciliate the training of the visual prompt for better the sample efficiency with few-shot learning. Experiments on different benchmarks show that our method can produce superior retrieval performance in the few-shot setup.

Acknowledgements

This work was supported by ONR N00014-18-1-2871-P00002-3. Students involved were also supported by Adobe Research. We would like to thank the anonymous reviewers for their insightful comments.

References

Anwaar, M. U.; Labintcev, E.; and Kleinsteuber, M. 2021. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1140–1149.

Chang, W.-C.; Yu, F. X.; Chang, Y.-W.; Yang, Y.; and Kumar, S. 2020. Pre-training tasks for embedding-based large-scale retrieval. *arXiv preprint arXiv:2002.03932*.

Chen, C.-Y.; Lin, H.-T.; Sugiyama, M.; and Niu, G. 2021a. On the Role of Pre-training for Meta Few-Shot Learning. In *Fifth Workshop on Meta-Learning at the Conference on Neural Information Processing Systems*.

Chen, W.; Liu, Y.; Wang, W.; Bakker, E.; Georgiou, T.; Fieguth, P.; Liu, L.; and Lew, M. S. 2021b. Deep image retrieval: A survey. *arXiv preprint arXiv:2101.11282*.

Chen, X.; Zhang, N.; Li, L.; Xie, X.; Deng, S.; Tan, C.; Huang, F.; Si, L.; and Chen, H. 2021c. Lightner: A lightweight generative framework with prompt-guided attention for low-resource ner. *arXiv preprint arXiv:2109.00720*.

Chen, Y.; Gong, S.; and Bazzani, L. 2020. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3001–3011.

Cui, L.; Wu, Y.; Liu, J.; Yang, S.; and Zhang, Y. 2021. Template-based named entity recognition using BART. *arXiv preprint arXiv:2106.01760*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Forbes, M.; Kaeser-Chen, C.; Sharma, P.; and Belongie, S. 2019. Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 708–717.

Gao, T.; Fisch, A.; and Chen, D. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Guo, X.; Wu, H.; Gao, Y.; Rennie, S.; and Feris, R. 2019. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 1(2): 7.

Hambardzumyan, K.; Khachatrian, H.; and May, J. 2021. WARP: Word-level Adversarial ReProgramming. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4921–4933. Online: Association for Computational Linguistics.

Han, X.; Wu, Z.; Huang, P. X.; Zhang, X.; Zhu, M.; Li, Y.; Zhao, Y.; and Davis, L. S. 2017. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, 1463–1471.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Li, J.; and Sun, M. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*.

Jia, X.; Zhao, H.; Lin, Z.; Kale, A.; and Kumar, V. 2020. Personalized Image Retrieval with Sparse Graph Representation Learning. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, 2735–2743. ACM.

Kim, J.; Yu, Y.; Kim, H.; and Kim, G. 2021. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence. AAAI*, 1–9.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

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.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68.

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

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Ma, H.; Zhao, H.; Lin, Z.; Kale, A.; Wang, Z.; Yu, T.; Gu, J.; Choudhary, S.; and Xie, X. 2022. EI-CLIP: Entity-aware Interventional Contrastive Learning for E-commerce Cross-modal Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 18030–18040.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Schick, T.; and Schütze, H. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Triantafillou, E.; Zemel, R.; and Urtasun, R. 2017. Few-shot learning through an information retrieval lens. *Advances in Neural Information Processing Systems*, 30.

Tsimpoukelli, M.; Menick, J.; Cabi, S.; Eslami, S.; Vinyals, O.; and Hill, F. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34.

Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.-J.; Fei-Fei, L.; and Hays, J. 2019. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6439–6448.

Wang, Y.-X.; Gui, L.; and Hebert, M. 2017. Few-shot hash learning for image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 1228–1237.

Yu, Y.; Lee, S.; Choi, Y.; and Kim, G. 2020. Curlingnet: Compositional learning between images and text for fashion iq data. *arXiv preprint arXiv:2003.12299*.

Zhong, Q.; Chen, L.; and Qian, Y. 2020. Few-shot learning for remote sensing image retrieval with maml. In *2020 IEEE International Conference on Image Processing (ICIP)*, 2446–2450. IEEE.

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