Prompting Vision-Language Fusion for Zero-Shot Composed Image Retrieval

Peng Wang FENGLIQIU@BUPT.EDU.CN
Zining Chen CHENZN@BUPT.EDU.CN
Zhicheng Zhao* ZHAOZC@BUPT.EDU.CN
Fei Su SUFEI@BUPT.EDU.CN
Beijing University of Posts and Telecommunications, Beijing, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Composed image retrieval (CIR) aims to retrieve target image given the combination of an image and a textual description as a query. Recently, benefiting from vision-language pretrained (VLP) models and large language models (LLM), the use of textual inversion or generating large-scale datasets has become a novel approach for zero-shot CIR task (ZS-CIR). However, the existing ZS-CIR models overlook one case where the textual description is often too brief or inherently inaccurate, making it challenging to effectively integrate the reference image into the query for retrieving the target image. To address this problem, we propose a simple yet effective method—prompting vision-language fusion (PVLF), which adapts representations in VLP models to dynamically fuse the vision and language (V&L) representation spaces. In addition, by injecting the context learnable prompt tokens in Transformer fusion encoder, the PVLF promotes the comprehensive coupling between V&L modalities, enriching the semantic representation of the query. We evaluate the effectiveness and robustness of our method on various VLP backbones, and the experimental results show that the proposed PVLF outperforms previous methods and achieves the state-of-the-art on two public ZS-CIR benchmarks (CIRR and FashionIQ).

Keywords: Composed image retrieval, zero-shot, prompting fusion, vision-language model

1. Introduction

Composed image retrieval (CIR) (Vo et al., 2019; Liu et al., 2021b; Wu et al., 2021), as one of the derived topics of cross-model retrieval, has a wide range of applications in various fields, including E-commerce platforms. In contrast to traditional uni-modal or cross-modal image retrieval systems, CIR employs a query that goes beyond simple unimodal queries such as images or textual descriptions. It is dedicated to combining a reference image and a relative caption as a query to retrieve the target image, as shown in Fig. 1 (c). Due to the high flexibility and customizability of this bi-modality query (*i.e.* users can express the concept through free-form textual descriptions), CIR has attracted rising attention.

Large scale VLP (Radford et al., 2021; Jia et al., 2021; Li et al., 2022, 2023) models have recently achieved tremendous success on various multi-modal downstream tasks. Owing to the exceptional performance of these VLP models on cross-modal alignment, some studies (Saito et al., 2023; Baldrati et al., 2023; Gu et al., 2023; Levy et al., 2023) have leveraged their powerful representation ability for ZS-CIR. Pic2Word (Saito et al., 2023)

* Corresponding author.

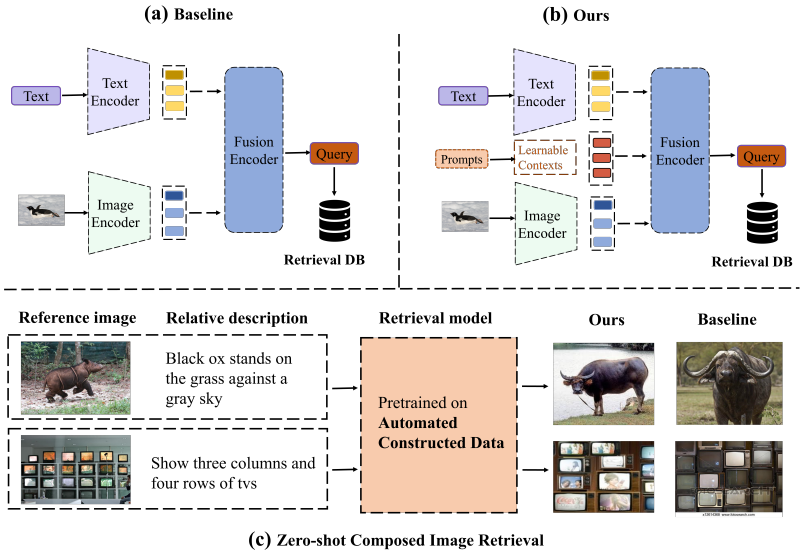

Figure 1: An illustration of our motivation. (a) TransAgg (Liu et al., 2023c). (b) Our proposed method PVLF. (c) Performance improvement compared to the baseline for the ZS-CIR.

proposes to leverage the linguistic capability of the language encoder in CLIP to generate embeddings that are close to the corresponding visual representations. SEARLE (Baldrati et al., 2023) transfers CIR to standard text-to-image retrieval by mapping the reference image into a learned pseudo-word token which is then concatenated with the relative caption. In addition, another line of recent researches (Gu et al., 2023; Liu et al., 2023c) have taken advantage of the powerful generation capabilities of the LLM to build scalable pipelines for ZS-CIR by generating high-quality triplets (I_r, T_r, I_t) composed of a reference image I_r, a relative caption T_r and a target image I_t, respectively. CompoDiff (Gu et al., 2023) proposes a dataset *SynthTriplets18M* and effectively trains the diffusion model. TransAgg (Liu et al., 2023c) constructs a high-quality *Laion-CIR* dataset through the designed templates and GPT-3 (Brown et al., 2020) to adaptively combine information from diverse modalities, as shown in Fig. 1 (a). However, the aforementioned methods overlook one crucial point, that is, the relative captions T_r within the current benchmark of the CIR field (CIRR (Liu et al., 2021b) and FashionIQ (Wu et al., 2021)) are extremely simple, or inherently inaccurate, making it challenging to learn the transformations from the reference image I_r to the target image I_t with a fine-grained perspective, and resulting in the significant deficiencies in both the effectiveness and the robustness.

We find that in the existing training data for supervised CIR models, the average length of T_r in triplets (I_r, T_r, I_t) is typically short, thereby the brief descriptions T_r are inadequate for accurately representing the transformations in CIR, which encompass diverse tasks such as domain conversion, scene or object composition or fashion-attribute manipulation etc. Such textual annotations struggle to fully leverage the wealth of knowledge encoded in pretrained language models.

Furthermore, prompt tuning (Liu et al., 2023a; Li and Liang, 2021; Jiang et al., 2020) has become a highly practical method in VLP models and LLMs, and has demonstrated the superior performance in both visual tasks (Jia et al., 2022; Wang et al., 2022; Zhou et al., 2022b,a) and multimodal domains (Khattak et al., 2023; Shen et al., 2024). Inspired by this, in this paper, we innovate the Prompting Vision-Language Fusion (PVLF) method for Zero-Shot Composed Image Retrieval. Different from the existing methods that utilize prompt tokens for either *uni-modal representations* shown in Fig. 2 (a) (b) or *multi-modal representations* shown in Fig. 2 (c) , we explore a new resolution for *multi-modal fusion*. Specifically, we propose to inject a small amount of task-specific context prompt tokens between the two modalities in fusion Transformer (Vaswani et al., 2017) encoder, as shown in Fig. 2 (d), to enable a more comprehensive and effective interaction. Firstly, we concatenate the textual tokens, context prompt tokens and visual tokens together and feed them into a fusion Transformer encoder, enabling three types of tokens to share a global attention integration. Secondly, we adaptively aggregate the thoroughly integrated features to obtain high-quality queries. In this way, PVLF improves the flexibility of textual descriptions, accurately integrates effective patch features, and effectively alleviates the imbalance modal representations, hence facilitating more comprehensive multimodal integration. Compared with the existing models, our retrieval examples demonstrate a noticeable improvement, as illustrated in Fig. 1 (c). Our contributions can be summarized as follows:

- To solve the insufficient description and semantic ambiguity in training triplets, context prompt learning is incorporated to obtain abundant and flexible textual features. To the best of our knowledge, it is the first time to introduce V&L prompt learning in ZS-CIR.
- By injecting a small number of learnable parameters, a simple yet effective Prompting Vision-Language Fusion (PVLF) method is proposed to enhance the modality fusion.
- Our method outperforms the SOTA models on the two public ZS-CIR benchmarks including FashionIQ and CIRR. Extensive experiments demonstrates the effectiveness and robustness of the proposed PVLF.

2. Related Work

2.1. Composed image retrieval

Composed image retrieval (CIR) (Liu et al., 2021b) is a variant task of retrieval within the context of multi-modal learning. Different from image-to-image or text-to-image retrievals, the goal of CIR is to generate joint-embedding features from both text and visual domains to retrieve the corresponding target image. Benefiting from the robust representation and generalization capabilities of VLP models (Radford et al., 2021; Li et al., 2022, 2023), ZS-CIR is firstly tacked by (Saito et al., 2023), which relies on a textual inversion network trained on the 3M unlabeled images using a cycle contrastive loss. Later, (Baldrati et al., 2023) train a new textual inversion network using fewer data and employs a weighted sum of distillation and regularization losses to achieve better results. In contrast, CompoDiff (Gu et al., 2023) proposes a novel dataset *SynthTriplets18M* and achieves impressive results by introducing diffusion model. CASE (Levy et al., 2023) addresses the ZS-CIR task by

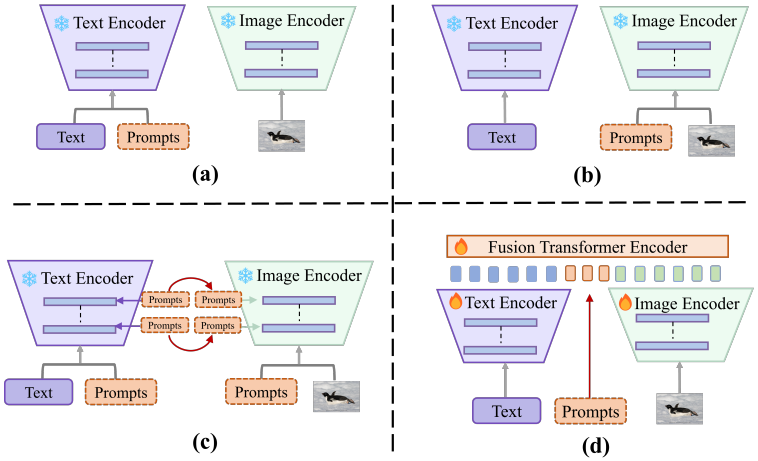

Figure 2: The architecture of (a) CoOp (textual prompt tuning), (b) VPT (visual prompt tuning), (c) MaPLe (multi-modal prompt learning), and (d) Ours.

employing an early fusion approach through the utilization of the BLIP (Li et al., 2022) and GPT-3 (Brown et al., 2020), further improving the results. Meanwhile, TransAgg (Liu et al., 2023c) proposes a retrieval-based pipeline for automatically constructing datasets for training, and achieves superior results by employing the Transformer-based feature fusion. However, existing methods rarely consider the relative descriptions used for supervising models may be too brief to accurately describe the transformations from the reference image to the target image, resulting in suboptimal query generation.

2.2. Prompt learning

Prompt learning is initially developed as a method for knowledge probing (Petroni et al., 2019), which aims to automate the process with the help of affordable-sized labeled data (Jiang et al., 2020). The motivation of prompt learning is to view pre-trained models (Devlin et al., 2018; Brown et al., 2020) as knowledge base to improve the practical applicability. Originating from NLP (Lester et al., 2021; Li and Liang, 2021), the model shows strong generalization to all kinds of downstream tasks. Due to the inherent sensitivity of hard-prompts, recent literatures (Lester et al., 2021; Li and Liang, 2021; Liu et al., 2023b, 2021a) propose to turn prompts into a set of continuous vectors and direct optimize them in an end-to-end manner as shown in Fig.2 (a). Meanwhile, the paradigm of prompt learning has also gradually gained popularity in computer vision (Jia et al., 2022; Wang et al., 2022; Rao et al., 2022) and multi-modal fields (Zhou et al., 2022b,a; Khattak et al., 2023; Shen et al., 2024), yielding promising results. VPT (Jia et al., 2022) introduces the visual prompt tuning method to pure vision backbones (Dosovitskiy et al., 2020), demonstrating the potential to serve as an alternative to fine-tuning visual backbones as shown in Fig. 2 (b). CoOp (Zhou et al., 2022b) brings continuous prompt learning to the vision domain for adaptation of VLP models. CoCoOp (Zhou et al., 2022a) solves CoOp's generalization issue by explicitly conditioning prompts on image instances. However, the prior works mainly

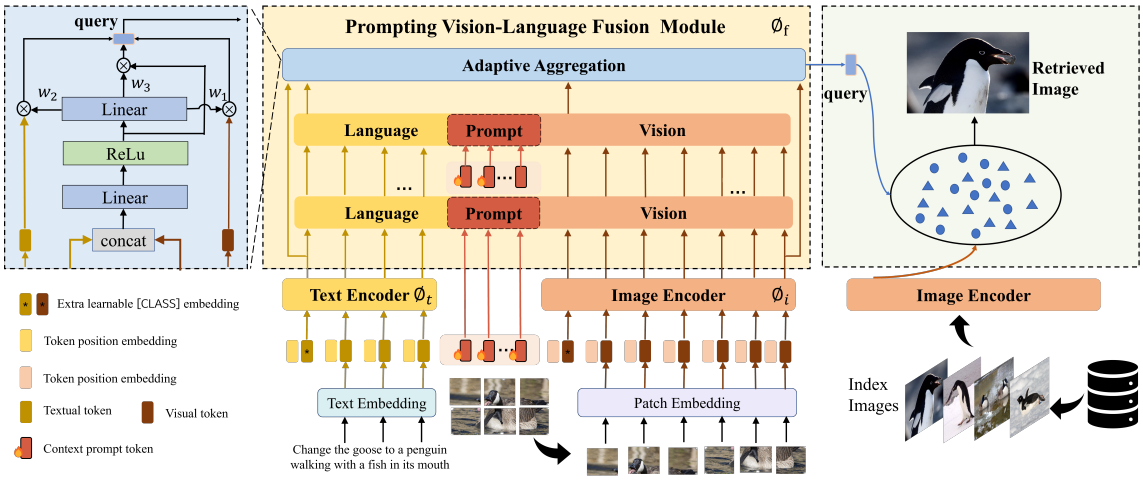

Figure 3: An overview of our proposed model architecture, which is primarily composed of three components: text encoder, image encoder and **P**rompting **V**ision-**L**anguage **F**usion (PVLF) module. The text encoder ϕ_t encodes the relative caption t, while the image encoder ϕ_i encodes reference image x and target image y. The pair (x, t) is transformed within the PVLF module ϕ_f, yielding a fusion query f_q that is utilized for retrieving the target image y in gallery images.

follow independent *uni-modal* representation learning. In *multi-modal* prompt learning, the work most closely to ours is MaPLe (Khattak et al., 2023) and MVLPT (Shen et al., 2024), whose architecture is shown in Fig.2 (c). Nevertheless, these methods primarily focus on *multi-modal representations*, our approach involves leveraging context prompt tokens for better *multi-modal fusion*.

3. Methods

In this section, we first introduce the specific details of the ZS-CIR task. Following that, we present the architecture of our method.

3.1. Problem Definition

Let $X = \{x_1, x_2, ...x_n\}$ denote as a set of query images, $T = \{t_1, t_2, ..., t_n\}$ as a set of query texts, and $Y = \{y_1, y_2, ..., y_n\}$ as a set of target images. The training set can be regarded as a collection of numerous triplets:

$$D = \{(x_i, t_i, y_i) \mid x_i \in \mathbb{R}^{H\times W\times 3}, y_i \in \mathbb{R}^{H\times W\times 3}\} \tag{1}$$

where $i = 1, 2, ...n$ and n is the number of samples. The CIR is defined as follows: given a pair of (x_i, t_i), we need to find the target image y_i from the gallery Y that best matches the semantic representation of (x_i, t_i). To achieve ZS-CIR, we leverage the VLP models (Radford et al., 2021; Li et al., 2022) to separately encode the images and texts. And our training objective is to ensure the fused query semantic representation $\phi_f(\phi_i(x_i), \phi_t(t_i))$,

which combines features from both modalities, remains as consistent as possible with the representation of the target image $\phi_i(y_i)$. Our objective can be represented as follows:

$$\phi_i(y_i) \approx \phi_f(\phi_i(x_i), \phi_t(t_i)) \tag{2}$$

where the image encoder, the text encoder, and the modality fusion module is defined as ϕ_i, ϕ_t, ϕ_f, respectively.

3.2. Network Architecture

As shown in Fig.3, our overall model architecture consists of three main components: text encoder, image encoder and the transformer-based **p**rompting **v**ision-**l**anguage **f**usion (PVLF) model. Unlike previous approachs (Li and Liang, 2021; Jia et al., 2022; Zhou et al., 2022b,a; Khattak et al., 2023; Shen et al., 2024) that employ prompts only for uni-modal or multi-modal representations, we propose a joint prompting approach where the context prompt tokens are injected between language modality and vision modality to establish interactions. And we build our approach on pre-trained vision-language models (Radford et al., 2021; Li et al., 2022).

3.2.1. Image Representation

We adopt the pretrained Vision Transformer(ViT) in VLP (Radford et al., 2021; Li et al., 2022) as our image encoder. For a plain ViT with N layers, an input image is divided into m fixed-sized patches. Each patch is then first embedded into d-dimensional latent space with positional encoding:

$$e_0^j = Embed(I_j) \qquad e_0^j \in \mathbb{R}^d, j = 1, 2, ...m \tag{3}$$

where we denote $E_k = \{e_k^j \in \mathbb{R}^d \mid j \in \mathbb{N}, 1 \leq j \leq m\}$ as the collection of image patch embeddings of the $(k+1)$-th Vision Transformer layer (L_{k+1}). Together with an extra learnable classfication token(`[CLS]`), the whole image encoder is formulated as:

$$[x_k, E_k] = L_k([x_{k-1}, E_{k-1}]) \quad k = 1, 2, ..., N \tag{4}$$

where $x_k \in \mathbb{R}^d$ denotes `[CLS]` token embedding at the L_{k+1} Transformer layer. N represents the number of layers. $[,.,]$ means the concatenation on the `[CLS]` token and visual embeddings. Finally, given a reference image x_i, we extract visual features from image encoder:

$$\mathcal{F}_x = \phi_i(x_i) \in \mathbb{R}^{|1+m| \times d} \tag{5}$$

3.2.2. Text Representation

Similar to the image representation, we adopt the pretrained BERT (Devlin et al., 2018) in VLP (Radford et al., 2021; Li et al., 2022) as our text encoder. For a reference text, we employ the lower-cased byte pair encoding (BPE) (Sennrich et al., 2015) to tokenize the textual description. After the textual description is bracketed with `[BOS]` and `[EOS]` tokens to indicate the start and end of sequence, the tokenized texts are fed into the Transformer layer.

$$[t_k, B_k] = L_k([t_{k-1}, B_{k-1}]) \quad k = 1, 2, ..., N \tag{6}$$

where we denote t_k as the `[CLS]` token embedding at the L_{k+1}, $B_k = \{b_k^i \in \mathbb{R}^d \mid i \in \mathbb{N}, 1 \leq i \leq n\}$ as the collection of textual embeddings of the L_{k+1}. Finally, given a reference text t_i, we extract textual features from text encoder:

$$\mathcal{F}_t = \phi_t(t_i) \in \mathbb{R}^{|1+n| \times d} \tag{7}$$

3.2.3. Prompting Vision-Language Fusion

In order to improve the flexibility of textual descriptions to accurately integrate patch features, we devise a simple yet effective method to guide vision-language fusion, facilitating a comprehensive interaction between vision and language modalities, as shown in figure 3. Specifically, after encoding the relative descriptions t_i and the reference image x_i, we consider introducing a set of learnable context prompt tokens after the textual tokens and before the visual tokens, as illustrated in PVLF module in figure 3. Subsequently, the textual tokens, context prompt tokens and visual tokens are concatenated together and jointly fed into the PVLF module to share the global attention computation.

It is worth noting that there exists prior works closely related to our approach on multi-modal prompt learning (Khattak et al., 2023; Shen et al., 2024). MaPLe (Khattak et al., 2023) is a dual-tower architecture that fine-tunes vision and language branches of CLIP together by sharing prompts across both modalities as shown in Fig. 2 (c). Although it employs shared prompts internally for modal connections, fundamentally, it still falls within the structure of class independent V&L prompting. In contrast, our model embodies a dual-tower representation supplemented by a single-tower fusion design. Our context prompt tokens are incorporated during the single-tower stage, where the self-attention is performed for all textual, prompt and visual tokens. MVLPT (Shen et al., 2024) incorporates cross-task knowledge into prompt tuning for V&L models which focuses on multitask learning, while our approach primarily aims to investigate the gains through prompt learning in the context of ZS-CIR. And we design two variants PVLF-Shallow, PVLF-Deep to facilitate the integration.

PVLF-Shallow In PVLF-Shallow, learnable context prompt tokens are injected into the first layer of Transformer encoder L_1 only. Let p denote the size of learnable context prompt tokens. Each prompt token is a learnable d-dimensional vector. During the forward, the textual tokens can be represented as $W_k = [t_k, B_k]$, the learnable context prompt tokens can be represented as $P = \{p^l \in \mathbb{R}^d \mid l \in \mathbb{N}, 1 \leq l \leq p\}$, the visual tokens can be represented as $V_k = [x_k, E_k]$, where $1 \leq k \leq N$. The PVLF-Shallow is as follows:

$$[W_1, Z_1, V_1] = L_1([W_0, P, V_0]) \tag{8}$$

$$[W_k, Z_k, V_k] = L_k([W_{k-1}, Z_{k-1}, V_{k-1}]) \tag{9}$$

where $Z_i \in \mathbb{R}^{p \times d}$ represents the context token embeddings computed by the i-th fusion Transformer encoder layer, k represents the depth of fusion encoder and $[W_k, Z_k, V_k] \in \mathbb{R}^{|1+n+p+1+m| \times d}$. If we denote $U_k = [W_k, Z_k, V_k]$, the computation for each layer of fusion Transformer encoder is illustrated as follows:

$$U' = MSA(U_k) + LN(U_k) \tag{10}$$

$$U_{k+1} = FFN(LN(U')) + LN(U') \tag{11}$$

where the multi-head self attention operator MSA, feed-forward network FFN and layer normalization (Ba et al., 2016) are applied. The MSA module enables comprehensive interactions among *textual* tokens, *context prompt* tokens and *visual* tokens at each layer. This stands as the key distinction compared to other multi-modal prompt methods (Khattak et al., 2023; Shen et al., 2024).

PVLF-Deep In PVLF-Deep, learnable context prompt tokens are injected into ***every*** fusion Transformer layer's input space. For $(i+1)$-th Layer L_{i+1}, we denote the collection of input learnable prompt tokens as $P_i = \{p_i^k \in \mathbb{R}^d \mid k \in \mathbb{N}, 1 \leq k \leq p\}$. The PVLF-Deep is as follows:

$$[W_k, \_, V_k] = L_k([W_{k-1}, P_{k-1}, V_{k-1}]) \tag{12}$$

Adaptive Aggregation Vision-language features are adaptively aggregated for a refined retrieval. We concatenate the global image feature x_N and the global text feature t_N together, and then feed them to a MLP to generate the fusion feature $f_u \in \mathbb{R}^d$. Subsequently, we map the fusion feature to three weight coefficients (w_1, w_2, w_3) and then employ them to individually weight and sum the (x_N, f_u, t_N) as follows:

$$f_q = (w_1, w_2, w_3) \cdot (x_N, f_u, t_N) \tag{13}$$

where $\langle \cdot \rangle$ is defined as the inner product between two matrixs. In the end, a (x_i, t_i) pair, propagated through the PVLF module, yields a fusion query for retrieving the target image.

3.3. Training Objective

Considering the class imbalance as well as the challenge of difficult samples, we adopt focal loss (Lin et al., 2017) as our training objective. Given a batch data of size B, the i-th query pair $(x^i,\ t^i)$ should be close to its positive target and far away from other negative instances:

$$p_t = \frac{exp[\mathcal{K}(f_q^i, \mathcal{F}_{x^i_{target}})/\tau]}{\sum_{j=1}^{B} exp[\mathcal{K}(f_q^i, \mathcal{F}_{x^i_{target}})/\tau]} \tag{14}$$

$$\mathcal{L} = -\alpha \cdot (1-p_t)^{\gamma} \cdot \log(p_t) \tag{15}$$

where $\tau = 0.01$ refers to the temperature parameter, and $\mathcal{K}(,)$ means the cosine similarity, f_q^i denotes the i-th query pair, $\mathcal{F}_{x^i_{target}}$ denotes the representation of i-th target image. We set the α as 1.0 and γ as 2.0 in experiments.

4. Experiments

In this section, we first describe our experimental setup including the datasets, and the implementation details. Then we comparatively assess our method against the most recent approaches to show the effectiveness of our method.

4.1. Experiment Setup

4.1.1. Training Datasets

We utilize the datasets constructed using Templates or LLM, as proposed in the TransAgg (Liu et al., 2023c), to tune our model. These datasets include Laion-CIR-Template and Laion-CIR-LLM, each contains around 16k triplets, as well as their combined counterpart Laion-CIR-combined, which comprises approximately 32k triplets.

4.1.2. Evaluation Datasets

We evaluate the performance on two standard benchmark datasets in ZS-CIR: FashionIQ (Wu et al., 2021) and CIRR (Liu et al., 2021b). FashionIQ contains 30,134 triplets from 77,684 images crawled from the web specifically designed for fashion retrieval, categorizing its contents into three categories: Dress, Toptee, and Shirt. CIRR comprises 21,552 real-life images taken from the natural language reasoning $NLVR^2$ (Suhr et al., 2018) dataset, which is built to overcome two common issues for CIR: non-complex images with too narrow domain and the high number of false-negatives. Noting that, similar to other work, we utilize both datasets for ZS-CIR evaluation and do not employ them to train the model.

4.1.3. Evaluation Metrics

Following the standard metrics in image retrieval, we report the average recall at rank-K (Recall@K) which is defined as the percentage of queries that correctly retrieve the ground-truth in the top-K results. Moreover, we also report the Recall_Subset@K which considers only the images in the subset of the query for CIRR.

4.1.4. Implementation Details

For the data pre-processing pipeline, we align with (Liu et al., 2023c) for a fair comparison. We choose AdamW (Loshchilov and Hutter, 2017) optimizer, initializing learning rate by 1e-4 with a cosine decay rate of 0.05 for fusion model. VLP models are applied to extract visual embeddings and textual embeddings seperately, while the learning rate of both image encoder and text encoder is set as 1e-6. We set the batch size to 64. And we use the PyTorch to conduct all the experiments on a single NVIDIA A100 80G GPU.

4.2. Quantitative Results

We compare our approach with several ZS-CIR methods for a fair comparison, including: 1) ***Text-only***: the similarity is computed using only the CLIP features of the relative caption; 2) ***Image-only***: retrieves the most similar images to the reference one via CLIP visual features; 3) ***Image+Text***: the summation of CLIP features of the reference image and the textual description; 4) ***PALAVRA*** (Cohen et al., 2022): a textual inversion-based two stage approach with a pre-trained mapping function; 5) ***Pic2Word*** (Saito et al., 2023): leverages the linguistic capability of the language encoder in CLIP to map the reference image into a pseudo-word token; 6) ***SEARLE*** (Baldrati et al., 2023): reduces ZS-CIR to standard text-to-image by mapping the reference image into a learned token which is then concatenated with the relative caption generated by GPT; 7) ***Context-I2W*** (Tang

et al., 2024): adaptively converts description-relevant image information into a pseudo-word token; 8) ***CompoDiff*** (Gu et al., 2023): pretrains the diffusion models on the proposed dataset *SynthTriplets18M*; 9) ***CASE*** (Levy et al., 2023): employs an early fusion approach through the utilization of the BLIP (Li et al., 2022) and GPT-3 on their dataset *LaSCo*. 10) ***TransAgg*** (Liu et al., 2023c): A Transformer-based adaptive aggregation model trained on the constructed*Laion-CIR* series datasets.

Following (Liu et al., 2023c), we train our model on the aforementioned constructed *Laion-CIR-Template*, *Laion-CIR-LLM* and *Laion-CIR-combined* datasets separately, and evaluate our models with the SOTA methods. We set the length of the context prompt token to 25, the depth of PVLF module to 2, the depth of the fusion Transformer to 2, and utilize the paradigm of PVLF-Deep as the default method.

Table 1: Zero-shot performance on FashionIQ (Wu et al., 2021) validation set. Best performance is in bold and second best is underlined.

Method	Backbone	Zero-shot	Triplets	Shirt		Dress		Toptee		Average	
				R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
Image-only	CLIP-B/32	✔	–	6.92	14.23	4.46	12.19	6.32	13.77	5.90	13.37
Text-only	CLIP-B/32	✔	–	19.87	34.99	15.42	35.05	20.81	40.49	18.7	36.84
Image+Text	CLIP-B/32	✔	–	13.44	26.25	13.83	30.88	17.08	31.67	14.78	29.60
Pic2Word (Saito et al., 2023) [CVPR'2023]	CLIP-L/14	✔	–	26.20	43.60	20.00	40.20	27.90	47.40	24.70	43.70
PALAVRA (Cohen et al., 2022) [ECCV'2022]	CLIP-B/32	✔	–	21.49	37.05	17.25	35.94	20.55	38.76	19.76	37.25
SEARLE-XL-OTI (Baldrati et al., 2023) [ICCV'2023]	CLIP-L/14	✔	–	30.37	47.49	21.57	44.47	30.90	51.76	27.61	47.90
CompoDiff (Gu et al., 2023) w/T5-XL [arXiv'2023]	ViT-L/14	✔	18m	**38.10**	52.48	**33.91**	47.85	**40.07**	52.22	**37.36**	50.85
Context-I2W (Tang et al., 2024) [AAAI'2024]	CLIP-L/14	✔	3m	29.70	48.60	23.10	45.30	30.60	52.90	27.80	48.90
TransAgg (Laion-CIR-Template) (Liu et al., 2023c) [BMVC'2023]	BLIP	✔	16k	32.83	52.31	27.67	49.38	35.70	58.08	32.07	53.26
TransAgg (Laion-CIR-LLM) (Liu et al., 2023c) [BMVC'2023]	BLIP	✔	16k	32.92	52.16	28.56	49.58	36.82	58.59	32.77	53.44
TransAgg (Laion-CIR-Combined) (Liu et al., 2023c) [BMVC'2023]	BLIP	✔	32k	34.45	53.97	30.24	51.91	38.40	59.51	34.36	55.13
PVLF (Laion-CIR-Template)	BLIP	✔	16k	33.90	53.09	28.16	50.62	37.02	59.87	33.03	54.53
PVLF (Laion-CIR-LLM)	BLIP	✔	16k	33.76	53.04	29.35	51.22	36.41	57.47	33.17	53.91
PVLF (Laion-CIR-Combined)	BLIP	✔	32k	36.61	**55.05**	31.58	**54.24**	38.85	**61.24**	35.68	**56.85**

Table 2: Zero-shot performance on CIRR test set. We report the results on Recall@1, Recall@5, Recall@10 and Recall@50. The Recall_subset@K metrics mean only considering the images within the subset of the query for the fine-grained retrieval evaluation.

Method	Backbone	Zero-shot	Triplets	Recall@K				Recall_subset@K		
				K=1	K=5	K=10	K=50	K=1	K=2	K=3
Image	CLIP-B/32	✔	–	6.89	22.99	33.68	59.23	21.04	41.04	60.31
Text	CLIP-B/32	✔	–	21.81	45.22	57.42	81.01	62.24	81.13	90.7
Image+Text	CLIP-B/32	✔	–	11.71	35.06	48.94	77.49	32.77	56.89	74.96
Pic2Word (Saito et al., 2023) [CVPR'2023]	CLIP-L/14	✔	–	23.90	51.70	65.30	87.80	–	–	–
PALAVRA (Cohen et al., 2022) [ECCV'2022]	CLIP-B/32	✔	–	16.62	43.49	58.51	83.95	41.61	65.3	80.94
SEARLE-XL-OTI (Baldrati et al., 2023) [ICCV'2023]	CLIP-L/14	✔	–	24.87	52.31	66.29	88.58	53.80	74.31	86.94
Context-I2W (Tang et al., 2024) [AAAI'2024]	CLIP-L/14	✔	3m	25.6	55.1	68.5	89.8	–	–	–
CompoDiff w/T5-XL (Gu et al., 2023) [arXiv'2023]	ViT-L/14	✔	18m	19.37	53.81	72.02	90.85	28.96	49.21	67.03
CASE Pre-LaSCo.Ca. (Levy et al., 2023) [AAAI'2024]	BLIP	✔	360k	35.40	65.78	78.53	94.63	64.29	82.66	61.91
TransAgg (Laion-CIR-Template) (Liu et al., 2023c) [BMVC'2023]	BLIP	✔	16k	38.1	68.42	79.08	93.51	70.34	86.42	94.28
TransAgg (Laion-CIR-LLM) (Liu et al., 2023c) [BMVC'2023]	BLIP	✔	16k	36.71	67.83	79.03	93.86	66.03	83.66	92.50
TransAgg (Laion-CIR-Combined) (Liu et al., 2023c) [BMVC'2023]	BLIP	✔	32k	37.87	68.88	79.6	93.86	69.79	86.09	93.93
PVLF (Laion-CIR-Template)	BLIP	✔	16k	39.78	70.87	81.18	94.28	**72.66**	**87.66**	94.48
PVLF (Laion-CIR-LLM)	BLIP	✔	16k	37.19	69.79	80.70	95.26	65.65	84.07	92.94
PVLF (Laion-CIR-Combined)	BLIP	✔	32k	**40.33**	**72.50**	**82.44**	**95.43**	72.64	87.37	**94.69**

FashionIQ Table 1 shows the results of our models on FashionIQ. In general, our approach outperforms the SOTA TransAgg (Liu et al., 2023c) for every category of the FashionIQ (shirt, dress, toptee and average), when trained on the same dataset. Additionally, our results remain consistent across all variants of the ***Laion-CIR*** series (including *Laion-CIR-Template, Laion-CIR-LLM* and *Laion-CIR-combined*), demonstrating the effectiveness of our approach. In specific, Our approach achieves 56.85 of the average recall@50, surpassing all previous ZS-CIR methods. Although our model exhibits a slightly lower performance on the average R@10 metric compared to CompoDiff w/T5-XL (Gu et al., 2023), it significantly outperforms (Gu et al., 2023) on the R@50 by a large margin. It is noteworthy that CompoDiff w/T5-XL utilizes a significantly larger dataset than our model.

CIRR As shown in Table 2, we report the results for CIRR test set. Overall, our model consistently achieves the best performance across various scenarios in all metrics, outperforming all previous methods. In addition, under the condition of fine-tuning with the same dataset (Liu et al., 2023c), our models also surpass baseline to varying degrees. Compared to CompoDiff which utilizes *18M* training triplets, we achieve the recall@1 score twice as high with a smaller training set. In contrast to baseline transagg (Liu et al., 2023c), our method demonstrates significant improvements across all scenarios. Notably, in more fine-grained *subset* scenarios, our approach outperforms all previous methods consistently by a more substantial margin as well, leading to new SOTA results. This improvement can be attributed to the greater flexibility and comprehensive integration of textual and visual tokens in our proposed PVLF, further improving the quality of the fusion query and demonstrating superior generalization performance.

4.3. Ablation Studies

In this section, we conduct a variety of ablation studies on two standard benchmark datasets to validate the robustness and the effectiveness of our proposed method.

Pretrained Backbone and Fine-tuning. To validate the effectiveness and the generalization of our proposed approach, we conduct comparisons with baseline (Liu et al., 2023c) on CIRR and FashionIQ under various fine-tuning strategies as shown in Table 3. Throughout this process, we remain other settings unchanged. As evident from the results, our method consistently outperforms baseline (Liu et al., 2023c) almost across all scenarios, demonstrating the effectiveness of our approach. In specific, our model exhibit a remarkable improvement in *average* recall on two datasets compared to the baseline. When tuning the full model on the two benchmarks CIRR and FashionIQ, the PVLF exceed the baseline by 4.29 and 1.08 respectively, which can be attributed to the PVLF's internal contextual prompt tokens, which exhibit more flexible and effective representations in linking textual and visual modalities.

Simultaneously, we also observe a slight degradation in performance when the CLIP backbone is frozen. We speculate that this may be attributed to a slight overfitting due to an abundance of context tokens when the parameters are frozen. It is necessary to highlight that our experiments in the Table 3 are conducted with a context token length of 25, without additionally specific fine-tuning for the CLIP backbone. Nevertheless, even in this scenario, the PVLF's average recall scores remain higher than the baseline.

Table 3: Generalization for different pretrained backbones and fine-tuning types on CIRR and FashionIQ. For CIRR, the average column denotes the mean of Recall@5 and Recall_subset@1. For FashionIQ, we report the average R@10 and R@50 of all three categories. We choose the TransAgg (Liu et al., 2023c) as our baseline for comparison, and report the results from (Liu et al., 2023c).

Backbone	Fine-tuning	Method	CIRR								FashionIQ		
			R@1	R@5	R@10	R@50	R_sub@1	R_sub@2	R_sub@3	**Average**	R@10	R@50	**Average**
CLIP-L/14	✗	**baseline**	25.04	53.98	67.59	88.94	55.33	76.82	88.94	53.66	28.57	48.29	38.43
		PVLF(ours)	24.06	54.7	67.9	89.59	53.36	75.34	87.37	54.03	28.75	48.74	38.75
		improvement	-0.98	+0.72	+0.31	+0.65	-1.97	-1.48	-1.57	+0.37	+0.18	+0.45	+0.32
	only text encoder	**baseline**	27.9	58.27	71.01	91.30	60.48	80.31	90.75	59.38	30.61	50.38	40.50
		PVLF(ours)	30.23	61.37	74.45	92.13	62.04	81.12	91.00	61.71	32.29	52.92	42.61
		improvement	+2.33	+3.1	+3.44	+0.83	+1.56	+0.81	+0.25	+2.33	+1.68	+2.54	+2.11
	both	**baseline**	33.04	64.39	76.27	93.45	63.37	82.27	92.22	63.89	32.63	53.65	43.14
		PVLF(ours)	33.84	66.92	78.37	93.97	62.98	82.87	92.05	65.00	32.85	53.61	43.23
		improvement	+0.80	+2.53	+2.10	+0.52	-0.39	+0.60	-0.17	+1.11	+0.22	-0.04	+0.09
BLIP	✗	**baseline**	34.89	64.75	76.24	92.22	66.34	83.76	92.92	65.55	26.95	46.1	36.53
		PVLF(ours)	35.16	66.32	77.42	93.16	67.35	84.45	92.90	66.84	29	48.97	38.98
		improvement	+0.27	+1.57	+1.18	+0.94	+1.01	+0.69	-0.02	+1.29	+2.05	+2.87	+2.45
	only text encoder	**baseline**	38.1	68.42	79.08	93.51	70.34	86.42	94.28	69.38	32.07	53.26	42.67
		PVLF(ours)	40.54	71.99	81.75	95.05	71.95	87.40	94.12	71.97	34.17	55.46	44.81
		improvement	+2.44	+3.57	+2.67	+1.54	+1.61	+0.98	-0.16	+2.59	+2.10	+2.20	+2.14
	both	**baseline**	37.18	67.21	77.92	93.43	69.34	85.68	93.62	68.28	34.64	55.72	45.18
		PVLF(ours)	40.33	72.50	82.44	95.43	72.64	87.37	94.69	72.57	35.68	56.85	46.26
		improvement	+3.15	+5.29	+4.52	+2.00	+3.30	+1.69	+1.07	+4.29	+1.04	+1.13	+1.08

Table 4: Ablation studies on the CIRR dataset to investigate the effect on different context prompt token length.

Recall@k	**Context prompt length-l**						
	l =5	l =10	l =15	l =20	l =25	l =30	l =35
$k=1$	39.49	39.66	39.63	38.22	**40.32**	39.20	39.15
$k=5$	72.23	71.85	72.11	70.41	**72.49**	71.58	71.65
$k=10$	82.06	82.11	82.30	81.44	**82.43**	82.11	82.13
$k=50$	95.24	95.53	95.21	95.07	**95.43**	95.36	95.37

Context prompt length. How many context prompt tokens should be used? We employ BLIP (Li et al., 2022) as the pre-trained backbone and conduct ablation studies on the CIRR (Liu et al., 2021b) dataset to investigate the impact of context prompt token length. We use random initialization for all context tokens. As shown in Table. 4, the trends in Recall@1, Recall@5, Recall@10 are nearly aligned, with each achieving its highest recall nearly at the length of 25. However, the optimal prompt token length for Recall@50 is 10, which suggests that as the number of recalled samples increases, the stability of the recall rate trends becomes more erratic. Meanwhile, we observe that an excessive length of prompt tokens tends to adversely affect the performance of the model. This phenomenon is also observed in other studies (Jia et al., 2022; Khattak et al., 2023; Shen et al., 2024), which we believe to be normal to some extent. The optimal context prompt length is likely to vary across different tasks.

Table 5: Ablation studies on the CIRR and FashionIQ datasets to investigate the impact of the prompt depth on model performance. The J means the prompt depth.

Method	CIRR				FashionIQ		
	R@1	R@5	R@10	R@50	R@10	R@50	Avg
TransAgg	37.18	67.21	77.92	93.43	34.64	55.72	45.18
PVLF-$J = 1$	38.53	71.05	81.51	95.21	35.28	56.77	46.03
improvement	+1.35	+3.84	+3.59	+1.78	+0.64	+1.05	+0.85
PVLF-$J = 2$	**40.33**	**72.50**	**82.44**	**95.43**	**35.68**	**56.85**	**46.26**
improvement	+3.15	+5.29	+4.52	+2.00	+1.04	+1.13	+1.08
PVLF-$J = 3$	39.27	70.92	82.09	95.41	34.43	56.14	45.29
improvement	+2.09	+3.71	+4.17	+1.98	-0.21	+0.42	+0.09

Prompt depth. We conduct ablation studies on the depth of the prompt depth J as shown in Table 5. In general, the performance improves as context prompt depth increases. We observe that the performance sensitivity increases when context prompt tokens are inserted in deep layers, which is also reported in (Jia et al., 2022; Khattak et al., 2023). We conjecture that in the context of ZS-CIR, it is highly likely that the model is overfitting. Overall, the deeper context prompts may yield better results to some extent, which can be attributed to the model's incorporation of additional context parameters at each layer to enhance the robustness and thoroughness of the fusion between different modalities. And it is worth noting that J is also the number of layers in fusion Transformer. By default, we use $J = 2$ to finetune the pretrained models to achieve the state-of-the-art performance.

5. Conclusion

To address ZS-CIR, we present Prompting Vision-Language Fusion, which explores the potential of introducing V&L prompt learning methods into the ZS-CIR. We firstly identify the issue of semantic ambiguity in training triplets, which may result in suboptimal performance. Subsequently, we propose to inject a certain number of context prompt tokens in the input space, facilitating a more comprehensive modality fusion to construct higher-quality queries. Extensive experiments demonstrate our method achieves SOTA performance. Additionally, it showcases remarkable robustness across different datasets and various fine-tuning strategies.

6. Acknowledgment

This work is supported by The Key R&D Program of Yunnan Province (202102AE09001902-2).

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. *arXiv preprint arXiv:2303.15247*, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer, 2022.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. Compodiff: Versatile composed image retrieval with latent diffusion. *arXiv preprint arXiv:2303.11916*, 2023.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

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

Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Data roaming and early fusion for composed image retrieval. *arXiv preprint arXiv:2303.09429*, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023a.

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

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *AI Open*, 2023b.

Yikun Liu, Jiangchao Yao, Ya Zhang, Yanfeng Wang, and Weidi Xie. Zero-shot composed text-image retrieval. *arXiv preprint arXiv:2306.07272*, 2023c.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, October 2021b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314, 2023.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5656–5667, 2024.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

Yuanmin Tang, Jing Yu, Keke Gai, Jiamin Zhuang, Gang Xiong, Yue Hu, and Qi Wu. Context-i2w: Mapping images to context-dependent words for accurate zero-shot composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5180–5188, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317, 2021.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.