2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CAP: Contextual Enhancement and Adaptive Prompting Network for Zero-Shot Composed Image Retrieval

Dian Chen[1†], Bo Li[2†], Ying Qin[1(✉)], Qingwen Li[2], Hong Li[2], and Shikui Wei[1]

[1] Beijing Jiaotong University, Beijing, China
`{chendian,yingqin,shkwei}@bjtu.edu.cn`
[2] Information Technology Research Institute of China Tower, China
`{libo88872,liqw3,lihong}@chinatowercom.cn`

Abstract. This paper focuses on the zero-shot Composed Image Retrieval (ZS-CIR) task, which only requires unlabeled images or image-title pairs for model training. Previous work has utilized textual inversion networks to form queries by combining "a photo of" fixed templates with pseudo-words projected from reference image features into the text embedding space. However, fixed prompt templates offer limited performance improvement for the model and can affect the learning of instance-specific contextual information in open-domain tasks. To address these issues, we propose a zero-shot composed image retrieval framework based on contextual enhancement and adaptive prompting (CAP), which consists of a Contextual Enhancement Module (CEM) and an Adaptive Prompting module (APM). CEM introduces bi-directional LSTM re-parameterized learnable prompts, and APM decouples the retrieval instances and maps the different features to the corresponding prompt parameters. These two modules cooperate to construct the optimal prompts adapted to the retrieval instances. Extensive qualitative and quantitative experiments on three datasets show that our model has a good generalization and better performance compared to state-of-the art methods.

Keywords: Composed image retrieval · Zero-shot · Contrastive learning.

1 Introduction

Composed Image Retrieval (CIR) refers to the task of retrieving a target image given a reference image and modified text as a query [1]. Existing approaches to Composed Image Retrieval, which fall into the supervised learning paradigm, train models on labeled triplets of the <reference image, modified text, target image> designed specifically for this task. However, constructing such a triple dataset requires the collection of pairs of images (i.e., distinct but similar images), followed by the manual construction of descriptions that reflect the differences between the two images [2], thus requiring high cost. If the CIR model is applied to a new domain, the triplets for

† These authors contributed equally to this work.

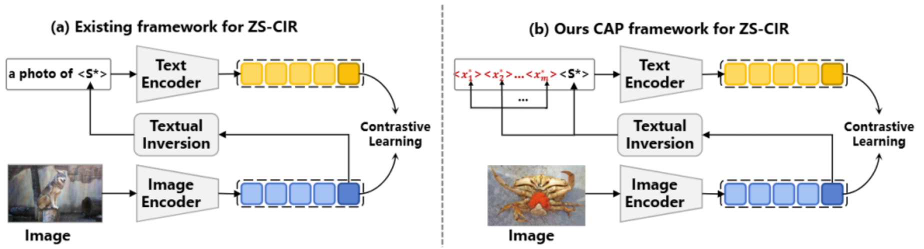


Fig. 1. An illustration of our improvement. (a) Existing framework for ZS-CIR. (b) Ours CAP framework for ZS-CIR

that domain should be constructed, and the model should be retrained. These factors limit the scalability of supervised methods into broader open-domains. To address these constraints and enhance the generalization capabilities of CIR models, researchers have proposed the Zero Shot Composed Image Retrieval (ZS-CIR) [3] task, which requires only unlabeled images or image-title pairs to train the model. ZS-CIR can utilize publicly available datasets covering a wide range of domains and semantic categories. As shown in **Fig. 1**(a), the main idea of the ZS-CIR is to train a textual inversion network, which serves as a function that transforms image features into the text embedding space to generate pseudo-words. The manually designed prompt template "a photo of" combined with the pseudo-word mapped into the text embedding vector constitutes the text, which is then fed into a text encoder (e.g.,CLIP [4]) to obtain text features. The weights of the textual inversion network are updated through contrastive learning of the text features with the image features. This process aims to achieve close alignment of text and image semantics, even in cases where only image data is available. The manually crafted prompts in existing methods are designed to provide the text encoder with comprehensive contextual information about the reference image. However, these manually designed prompt templates are not optimal, and a single pseudo-word fails to adequately decouple the attributes and other information within the retrieved image instance, which hinders the provision of adaptive prompts for retrieval instances.

To address the above two issues, we propose a novel zero-shot composed image retrieval framework named CAP, which is based on contextual enhancement and adaptive prompting. Compared to conventional ZS-CIR methods and inspired by prompt learning, our work focuses on learning adaptive prompts instead of using fixed prompts, as depicted in **Fig. 1**(b). We propose the Contextual Enhancement Module (CEM) to learn the most effective templates for the training dataset. It introduces learnable parameters in place of fixed manual prompts and employs the Bidirectional long short-term memory network (BiLSTM) to reparameterize the learnable parameters. The BiLSTM is capable of handling both forward and reverse data inputs, enabling the prompt embeddings to consider contextual information and form interdependencies. To prevent the model from forgetting the features learned during the CLIP encoding phase when updating its weights, we further employ residual connections for the BiLSTM. The CEM can learn effective prompts for the training dataset. However, once training is completed, during the inference phase, the learned prompts are loaded from the saved model for each query instance. When extending retrieval instances to an open-domain unseen by the model, these prompts exhibit certain limitations. To address this, we propose the Adaptive Prompting Module (APM), which generates unique prompts for each retrieval instance by decoupling and mapping different aspects of image feature information and generating prompt offsets for each prompt embedding. By combining these prompt offsets with the learnable parameters of the CEM, we can generate specific prompts

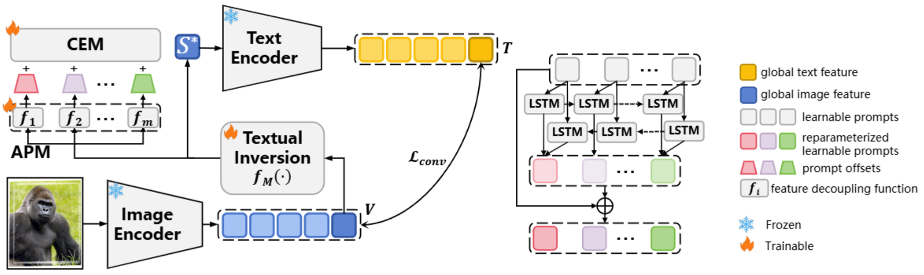
for retrieval instances, thereby enhancing the generalization and transferability of the CIR model across unseen open-domains. The major contributions can be summarized as follows:

- We propose a novel zero-shot network CAP, a framework that focuses on providing text encoders with more complete contextual information about the reference image by learning context-dependent prompt.
- To overcome the challenge that fixed templates trained on a specific dataset cannot be adapted to retrieved instances, we propose the APM to facilitate the mapping of different features of an image to distinct prompts.
- Extensive experimental results demonstrate that our proposed method outperforms state-of-the-art methods across three different datasets.

2 Related Work

2.1 Zero-shot Composed Image Retrieval

The rapid expansion of vision-language pre-trained (VLP) models has significantly improved the performance of supervised CIR tasks, attributed to their advanced cross-modal alignment and feature extraction capabilities. Nonetheless, the conventional supervised learning framework necessitates the need for costly annotated image-text triples for model training. To address these issues, researchers have proposed ZS-CIR. The goal of ZS-CIR is to train models using unlabeled images or image-text pairs. During the inference phase, ZS-CIR leverages a reference image and an adapted text description to automatically retrieve the target image. For instance, Pic2Word [3] utilizes a text inversion method to transform image features into individual pseudo-words in the CLIP text embedding space. iSEARLE [5] with the same text inversion method uses GPT-driven regularization loss and distillation loss for pre-training to process image features in a more fine-grained way. LinCIR [6] trains exclusively on text datasets by projecting latent embeddings into token space and replacing keywords to create new texts. All of these models utilize text inversion and fixed manual templates as prompts, overlooking the role that prompts play in pre-trained models. By fine-tuning the prompt, we can craft the most effective prompts to provide contextual information that is most relevant to the image, thereby enabling the constructed text to more accurately express the content of the image.



(a) Overall Arichitecture (b) CEM

Fig. 2. (a) The Architecture of our proposed CAP. m feature decoupling functions (f_i) are combined to form the APM. (b) The specific structure of CEM.

2.2 Prompt Learning

Prompt learning has been widely used in VLP models and Large Language Models (LLMs). It originated in the field of Natuarl Language Processing(NLP), where it served to improve practical utility by using pre-trained language models as a knowledge base [7]. Recent research has transformed the approach of prompt into using a set of continuous vectors for direct end-to-end optimization [8], and has introduced prompting learning to adapt VLP models in the visual domain. CoOp [9] introduces sequential prompt learning into the visual domain, facilitating the adaptation of VLP models, Furthermore, CoCoOp [10] addresses the generalization issue of CoOp by employing image instances as prompts. PRE [11] leverages a prompt encoder to re-parameterize the prompt embeddings, thereby enhancing the exploration of domain-specific knowledge. Inspired by the concept of prompt learning, in this work, we harness the idea of prompt learning to enhance the generalization capability of ZS-CIR.

3 Proposed Method

The overall architecture of our proposed CAP is shown in **Fig. 2**(a). In the training phase, image features are mapped to the text embedding space to form pseudo-words using the textual inversion network, while prompts are generated collaboratively using CEM and APM. The prompt and pseudo-words form a pseudo-phrase describing the contextual information of the image and are fed into the text encoder. Finally, contrastive learning is employed to minimize the distance between image and text features, thus updating the model parameters. In the inference phase, the learned prompts, pseudo-words mapped from image features, and modified text are combined to generate composite query features, which are compared with the image features in the database for effective retrieval.

3.1 Contextual Enhancement Module

Inspired by prompt learning, CEM is designed to construct a prompt that encapsulate contextual information, with the aim of finding a more appropriate prompt than the manual template “a photo of ”. To construct the prompt, initially$m$learnable parameters are introduced into the text embedding space, which can be denoted as,

$$\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m] \in \mathbb{R}^{m \times d_t}, \tag{1}$$

where the dimension of d_t corresponds to the dimension of the textual embedding of CLIP, being 768. Since the prompt embeddings are supposed to be interdependent, we employ a BiLSTM network to handle both forward and reverse sequences of data inputs. This approach allows the network to capture contextual information effectively. To ensure that the prompt retains the features learned during the CLIP encoding phase, we integrate residual connectivity into the BiLSTM. This integration helps in preserving the learned features while enabling the model to adapt and learn from new data. The process can be mathematically represented as follows:

$$\boldsymbol{c}_i = BiLSTM(\boldsymbol{x}_i) + \boldsymbol{x}_i, i = 1, \ldots, m. \tag{2}$$

$$\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_m] \in \mathbb{R}^{m \times d_t}, \tag{3}$$

where $BiLSTM(\cdot)$ denotes BiLSTM network. $\boldsymbol{C}$ is the prompt embeddings after reparameterization. In addition, the module introduces a pseudo-word $\boldsymbol{S}^*$ into the text branch, which contains the main information in the image as mapped by the textual reverse network. $\boldsymbol{S}^*$ is obtained in accordance with the baseline model Pic2Word [3] through a textual inversion network f_M,

$$\boldsymbol{S}^* = f_M(\boldsymbol{V}), \tag{4}$$

where Vdenotes the image feature extracted by the image encoder of CLIP. The module has two additional parameters, $\boldsymbol{t}_s$ and $\boldsymbol{t}_e$, which denote the embeddings of the start and end of the sentence in the CLIP text branch, respectively. The purpose of $\boldsymbol{t}_s$ and $\boldsymbol{t}_e$ is to enable the model to recognize the length of a given text segment. During the training process, these two parameters remain fixed along with the gradient updates. The input to the text encoder can be represented as:

$$\boldsymbol{T}_c = [\boldsymbol{t}_s, \boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_m, \boldsymbol{S}^*, \boldsymbol{t}_e]. \tag{5}$$

3.2 Adaptive Prompting Module

CEM is able to learn optimal prompt but cannot adapt to open-domain images. When different query image instances are input, we further decouple the information in $\boldsymbol{S}^*$, which contains the image information. We introduce m feature decoupling functions $\{f_j\}_{j=1}^{m}$ to map the different features in $\boldsymbol{S}^*$to prompt offsets $\boldsymbol{B} = [\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_m] \in \mathbb{R}^{m \times d_t}$,

$$\boldsymbol{b}_j = f_j(\boldsymbol{i}), j = 1, \ldots, m, \tag{6}$$

where $f(\cdot)$ is a multilayer perceptron (MLP) with four linear layers. The m feature decouple functions have non-shared parameters. The decoupled prompt offset $\boldsymbol{B}$ is then combined with the prompts $\boldsymbol{T}_c$ learned by the CEM to obtain the final prompts. The final text feature $\boldsymbol{T}$ can be represented as,

$$\boldsymbol{p}_k = \boldsymbol{c}_k + \boldsymbol{b}_k, k = 1, \dots, m, \tag{7}$$

$$\boldsymbol{T} = [\boldsymbol{t}_s, \boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_m, \boldsymbol{S}^*, \boldsymbol{t}_e]. \tag{8}$$

Finally, we align the constructed text features $\boldsymbol{T}$ with the image features $\boldsymbol{V}$.

3.3 Loss Function

We normalize the obtained text features $\boldsymbol{T}$ and image features $\boldsymbol{V}$, respectively,

$$\overline{\boldsymbol{T}} = \frac{\boldsymbol{T}}{\parallel \boldsymbol{T} \parallel}, \tag{9}$$

$$\overline{\boldsymbol{V}} = \frac{\boldsymbol{V}}{\parallel \boldsymbol{V} \parallel}. \tag{10}$$

We aim to minimize the distance between an image feature and its corresponding text feature, while maximizing the distances to other text features, and vice versa. To achieve this, we apply a contrastive function as follows:

$$\mathcal{L}_{t2v}(\bar{\boldsymbol{t}}, \bar{\boldsymbol{v}}) = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\exp\left(\kappa(\bar{\boldsymbol{t}}_j, \bar{\boldsymbol{v}}_j)/\tau\right)}{\sum_{k=1}^{B} \exp\left(\kappa(\bar{\boldsymbol{t}}_j, \bar{\boldsymbol{v}}_k)/\tau\right)}, \tag{11}$$

$$\mathcal{L}_{v2t}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{t}}) = -\frac{1}{B}\sum_{j=1}^{B} \log \frac{\exp\left(\kappa(\bar{\boldsymbol{v}}_j, \bar{\boldsymbol{t}}_j)/\tau\right)}{\sum_{k=1}^{B} \exp\left(\kappa(\bar{\boldsymbol{v}}_j, \bar{\boldsymbol{t}}_k)/\tau\right)}, \tag{12}$$

where κ is a function computing the cosine distance using the vector dot product. B represents the batch size, and τ is a temperature coefficient, which controls the strength of the penalty for negative samples. The final loss function is the sum of the losses of the two comparisons, as follows,

$$\mathscr{L}_{conv} = \mathscr{L}_{t2v}(\bar{\boldsymbol{t}}, \bar{\boldsymbol{v}}) + \mathscr{L}_{v2t}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{t}}) \tag{13}$$

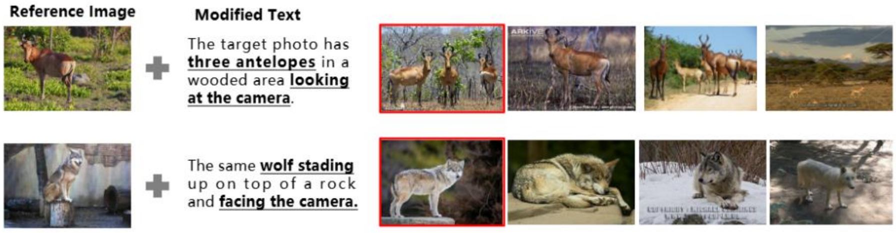

Reference Image
Modified Text
The target photo has three antelopes in a wooded area looking at the camera.
The same wolf stading up on top of a rock and facing the camera.

Fig. 3. Qualitative results from our approach CAP on CIRR dataset. The target images are highlighted by red boxes.

4 Experiment

4.1 Datasets

We employ CC3M [12] as a training dataset with the aim of evaluating the performance of CAP in downstream CIR tasks. To evaluate the model's retrieval capabilities and generalization abilities in open-domains, we have adopted the CIRR [13] as the standard CIR benchmarking dataset. Additionally, we conduct domain conversion experiments utilizing ImageNet [14] and object combination experiments with the COCO [16].

CIRR [13] is a realistic image retrieval dataset that includes a diverse range of common daily life scenes paired with artificially generated, modified texts. These scenes are sourced from the Natural Language Visual Reasoning dataset NLVR^2 [15], forming 36,554 triplets from 21,552 images.

ImageNet [14], consisting of 200 diverse categories, is used to evaluate model performance via domain conversion tasks. Considering the noise in the annotaions, we choose the domains of *cartoon*, *origami*, *toy*, and *sculpture* to evaluate the model's performance.

COCO [16] contains images of 91 object categories, reflecting common daily life scenarios, and is used for our object combination experiments. In these experiments, the reference image is presented without its background while the modified text lists additional instances in English. The target image is the original, unprocessed version of the reference image.

Table 1. Experiment results on CIRR dataset. "*Image only*" retrieves the target image using image features alone; "*Text only*" retrieves the target image using text features alone; "*Image+Text*" retrieves the target image using the average of image and text features.

Methods	**R@1**	**R@5**	**R@10**	**R@50**	**R-s@1**	**R-s@2**	**R-s@3**	**Average**
Image only	7.06	25.02	35.73	59.53	21.05	41.74	61.78	23.04
Text only	21.31	46.45	57.57	78.93	**63.43**	**81.51**	**90.65**	54.94
Image+Text	12.84	36.95	50.37	78.23	33.92	58.96	76.56	35.47
Pic2Word*	23.18	51.52	64.00	86.30	54.08	75.48	86.80	32.80
Pic2Word [3]	23.90	51.70	65.30	87.80	-	-	-	-
SEARLE-XL [17]	24.24	52.48	66.29	**88.84**	53.76	75.01	88.19	53.12
LinCIR [6]	25.04	53.25	66.68	-	57.11	77.37	88.89	55.18
CIReVL [18]	24.55	52.31	64.92	86.34	59.54	79.88	89.69	55.93
CAP	**25.86**	**55.80**	**68.95**	88.47	57.90	78.83	89.43	**56.85**

4.2 Implementation Details

In this work, we utilize CLIP (ViT-L/14 version) as the feature extraction backbone for images and a Transformer as the text extraction backbone. Both image and text features are projected into a common 768-dimensional embedding space. The number of learnable parameters introduced in CEM is 3. During training, we apply the AdamW optimizer [19] with a batch size of 1024, a learning rate of 1e^{-4}, a learning rate decay weight of 0.1, and a learning rate warm-up iteration of 1×10^{4}, training for a total of 40 epochs. All experiments are conducted using the PyTorch neural network framework on a single NVIDIA A100 GPU. We measure retrieval performance using Recall@K (R@K) [20], which calculates the percentage of evaluation queries that retrieve the target image within the top-K results.

Table 2. Experiment results on ImageNet.

Methods	**Cartoon**		**Origami**		**Toy**		**Sculpture**		**Average**
	R@10	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	**R@10**	**R@50**	
Image only	0.3	4.6	0.2	1.7	0.6	6.0	0.4	4.3	2.3
Text only	0.2	0.9	0.6	3.2	0.5	1.8	0.2	2.1	1.2
Image+Text	2.2	13.1	2.5	12.7	1.6	11.2	1.4	10.7	6.9
Pic2Word	8.0	21.9	13.5	5.6	8.7	21.6	10.0	23.8	16.7
CAP	**9.7**	**24.9**	**15.4**	**27.0**	**10.1**	**25.2**	**10.5**	**26.7**	**18.7**

Table 3. Composition experiment results on COCO.

Methods	**R@1**	**R@5**	**R@10**	**Average**
Image only	8.2	14.9	19.0	14.0
Text only	6.0	16.3	23.6	15.3
Image+Text	10.4	20.4	26.7	19.2
Pic2Word	11.5	24.8	33.4	29.1
CAP	**12.30**	**26.51**	**35.07**	**30.79**

4.3 Comparison with State-of-the-art Methods

We compared our method with the baseline. **Table 1** shows the results of our method on CIRR dataset. It can be observed that our method's CAP performance significantly outperforms the state-of-the-art methods on CIRR benchmark dataset. Notably, the average recall metric, calculated as (R@5+R-s@1)/2, shows an improvement of 4.05% over the Pic2Word on CIRR dataset, demonstrating the generality and effectiveness of our model. Additionally, our method performs well in composition experiments on the COCO dataset and domain conversion experiments with ImageNet, as illustrated in **Table 2** and **Table 3**. Compared with Pic2word, an average recall improvement of 2.0% and 1.69%, respectively, indicating the model's domain generalization capability.

These results substantiate the superiority of our approach, which we attribute to the enhanced and adaptive prompts that better guide the context.

4.4 Ablation Study

Table 4. Ablation study on three datasets. The "*" indicates that the decoupling function parameters of APM are not shared, meaning that image information cannot be decoupled.

Methods	CEM	APM	ImageNet	COCO	CIRR
Baseline	-	-	16.65	29.10	52.80
CEM	√	-	17.61	28.71	53.84
APM	-	√	16.82	28.89	53.38
CAP*	√	√*	18.35	28.82	51.54
CAP	√	√	**18.68**	**30.79**	**56.85**

Table 5. Ablation study of prompt numbers on ImageNet, COCO, CIRR

Num	ImageNet	COCO	CIRR	Avg
1	16.25	28.60	53.41	32.75
2	16.28	29.60	55.32	33.73
3	**18.68**	**30.79**	**56.85**	**35.44**
4	17.16	27.15	53.80	32.70
5	12.98	29.83	52.06	31.62

Table 4 presents the ablation results of CAP on three datasets. It can be observed that CEM and APM have a positive effect on the model's performance. The * indicates parameter sharing in the information decoupling function of the APM module, which means that image information cannot be decoupled. Experimental results reveal that this configuration yields worse performance compared to CAP with decoupled information. It demonstrates that decouple image information has a beneficial effect on the model. In addition, the collaboration between CEM and APM shows effectiveness in the ablation experiments, since it can promote the prompt to obtain more contextual information, facilitating the alignment of textual semantics with image semantics.

Table 5 presents the ablation results of prompt numbers on three datasets. When the number of prompts is set to 3, the model achieves optimal performance across three datasets. It can be observed that with a smaller number of prompts, the model's ability to represent image information through the context learned by prompts is limited. Conversely, when the number of prompts is large, overfitting may occur, leading to a decline in the model's generalization ability. Therefore, selecting an appropriate amount of parameters can maximize the model's performance.

4.5 Qualitative Results

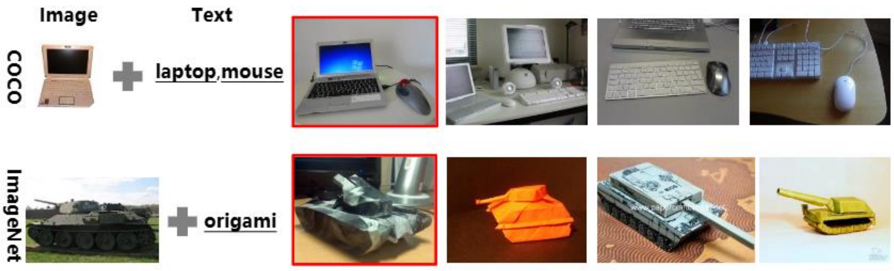


Fig. 4. Qualitative results from our approach CAP on COCO and ImageNet datasets. The target images are highlighted by red boxes.

To further evaluate the effectiveness of our method, we visualized the retrieval results of our proposed method on CIRR, COCO, and ImageNet, as shown in **Fig. 3** and **Fig. 4**, respectively. Each row reports the reference image, the modified text, and the top-4 search results. The retrieval examples demonstrate that our method not only effectively guides the text branch to obtain contextual information but also enhances domain generalization capabilities. For instance, in the CIRR dataset, the search results not only capture the main subject of the reference image, which is antelopes, but also meet the keywords “three” and “looking at the camera”. In COCO, the main subject of the reference image, the laptop, is retained, and the keyword “mouse” is also satisfied. In the ImageNet domain conversion experiment, our model demonstrated its capability by successfully retrieving images and accurately identifying tank images within the origami domain. It can be clearly seen that our approach excels at converting the domain of input image features.

5 Conclusion

We propose the CAP framework to handle the challenge of zero-shot CIR task, which consists of two models, CEM and APM. The CEM aims to introduce reparameterized learnable prompts, while the APM decouples image information by mapping distinct information elements to separate prompt parameters. Together, they synergize to address the issue of poor generalization of manually designed prompts in open-domains by dynamically capturing richer contextual information, thereby enhancing the model’s retrieval performance. Through extensive experiments, the superiority of our method is demonstrated on three multi-modal datasets.

Acknowledgments This work is supported by the National Key R&D Program of China National Key R&D（2024YFC3014900）

References

1. Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J.: Composing text and image for image retrieval - an empirical odyssey. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6432–6441 (2019).
2. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2105–2114 (2021).
3. Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T.: Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19305–19314 (2023).
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021).
5. Baldrati, A., Agnolucci, L., Bertini, M., Bimbo, A.: Zero-shot composed image retrieval with textual inversion. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 15292–15301 (2023)
6. Gu, G., Chun, S., Kim, W., Kang, Y., Yun, S.: Language-only efficient training of zero-shot composed image retrieval. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13225–13234 (2024).
7. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
8. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. AI Open 5, 208–215 (2024).
9. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision 130, 2337 – 2348 (2021).
10. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16795–16804 (2022).
11. Minh, A.P.T.: Pre: Vision-language prompt learning with reparameterization encoder. ArXiv abs/2309.07760 (2023).
12. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Annual Meeting of the Association for Computational Linguistics (2018).
13. Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on reallife images with pre-trained vision-and-language models. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2105–2114 (2021).
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009).
15. Suhr, A., Zhou, S., Zhang, I., Bai, H., Artzi, Y.: A corpus for reasoning about natural language grounded in photographs. ArXiv abs/1811.00491 (2018).

16. Lin, T.Y., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision (2014).
17. Agnolucci, L., Baldrati, A., Bertini, M., Bimbo, A.: isearle: Improving textual inversion for zero-shot composed image retrieval. ArXiv abs/2405.02951 (2024).
18. Karthik, S., Roth, K., Mancini, M., Akata, Z.: Vision-by-language for training-free compositional image retrieval. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024).
19. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017).
20. Patel, Y., Tolias, G., Matas, J.: Recall@k surrogate loss with large batches and similarity mixup. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7492–7501 (2022).